Non-Gaussian Graphical Models: Estimation with Score Matching and Causal Discovery under Zero-Inflation

Shiqing Yu

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2020

Reading Committee:

Mathias Drton, Chair

Ali Shojaie, Chair

Yen-Chi Chen

Program Authorized to Offer Degree:
Statistics

University of Washington

Abstract

Non-Gaussian Graphical Models: Estimation with Score Matching and Causal Discovery under Zero-Inflation

Shiqing Yu

Co-Chairs of the Supervisory Committee:
Professor Mathias Drton
Department of Mathematics, Technische Universität München

Associate Professor Ali Shojaie
Department of Biostatistics

Graphical models specify conditional independence relations between variables. These include undirected graphical models and directed graphical models, the latter of which also capture causal relationships. This dissertation considers challenges but also opportunities brought about by non-Gaussianity in undirected and directed graphical models.

A common challenge in estimating parameters of probability density functions is the intractability of the normalizing constant, which hinders the use of maximum likelihood estimation, especially for non-Gaussian graphical models or when the density is supported on a subspace of $\mathbb{R}^m$. This dissertation presents a generalized form of score matching for distributions supported on general domains. The proposed approach generalizes the original forms proposed in Hyvärinen (2005) for $\mathbb{R}^m$ and Hyvärinen (2007) for $\mathbb{R}^m_+$, while avoiding direct calculation of the normalizing constant, and yielding closed-form estimates for exponential families of continuous distributions. We generalize the regularized score matching method of Lin et al. (2016), and apply it to a general class of pairwise interaction graphical models, establishing strong theoretical guarantees.

Motivated by modern RNA sequencing technologies that provide gene expression measurements from single cells, this dissertation also addresses the challenge of causal discovery

from zero-inflated expression patterns in single cell data. In particular, we propose directed graphical models based on Hurdle conditional distributions to explore cause-effect relationships among the genes. We show that, under a natural and weak assumption, the exact directed acyclic graph for our model can be identified. We propose methods for graph recovery and show simulated experiments that validate the identifiability and graph estimation methods in practice.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY

AUC: The area under the ROC curve.

BIC: The Bayesian information criterion.

DAGS: Directed acyclic graphs; directed graphs with no directed cycles.

EBIC: The extended Bayesian information criterion.

FDR: The false discovery rate.

FPR: The false positive rate.

GGMS: Gaussian graphical models.

HURDLE CONDITIONALS: See Definition 10.

HURDLE POLYNOMIALS: See Definition 12.

ROC CURVE: Receiver operating characteristic curve.

TPR: The true positive rate.

ACKNOWLEDGMENTS

First of all, I would like to thank my PhD advisors, Mathias Drton and Ali Shojaie, whose support and guidance have made this dissertation possible. I would also like to thank Yen-Chi Chen, Steve Mooney and Dmitriy Drusvyatskiy for their inputs and for serving on my supervisory committee.

Next, I would like to thank the faculty, the department staff, and my fellow PhDs in the Department of Statistics at the University of Washington for their great help and assistance during my time here. I would also like to thank the Department of Mathematical Sciences at the University of Copenhagen for their hospitality during my visit.

Furthermore, I would like to thank Divakar Viswanath at the University of Michigan for his help and guidance during my graduate school application process. I would also like to thank Xiang Yu for guiding me through my undergraduate research project at the University of Michigan.

In addition, I would like to thank Prof. Zhang, who had been treating me like his grandson since I was born, and encouraged me to pursue a PhD when I was in high school. May his soul rest in peace.

Last but not the least, I would like to thank my parents, my grandfather and my deceased grandmother for bringing me up and giving me support over the last 25 years. Many thanks to Carmen for her emotional support and patience during my PhD. I would also like to thank my friends and other family relatives for their help and encouragement. Finally, I would like to express my gratitude to all my deceased family and friends who unfortunately are no longer able to see me reaching this stage of my life, for leaving me with good memories in my life, and may they rest in peace.

DEDICATION

to my family

Chapter 1

INTRODUCTION AND BACKGROUND

Graphical models specify conditional independence relations among variables. Non-Gaussian models provide the flexibility to express more complicated interactions among the variables, especially for distributions that are naturally supported on some proper subset of the m-dimensional real space $\mathbb{R}^m$, or those that are not absolutely continuous with respect to the Lebesgue measure. This comes with both opportunities and significant challenges. A challenge for undirected graphical models is the fact that non-Gaussian models as well as Gaussian models that are restricted to a proper subset of $\mathbb{R}^m$ generally have intractable normalizing constants, making maximum likelihood estimation computationally difficult. For directed graphical models, which capture causal relationships among the variables, non-Gaussian modeling is again more involved, but here it also offers an opportunity for causal discovery. Indeed, for many Gaussian settings, they may only be identifiable up to an equivalence class and thus preventing causal discovery. In this dissertation, we address these opportunities and challenges in detail.

1.1 Score Matching

Probability density functions, especially those for multivariate graphical models, are often defined only up to a normalizing constant. In high-dimensional settings, calculating the normalizing constant is often computationally intensive and intractable except in some special cases such as Gaussian densities (Gaussian graphical models) on the m-dimensional real space $\mathbb{R}^m$. The situation becomes worse when the distributions are defined only on a proper subset of $\mathbb{R}^m$. For example, even the Gaussian densities on $\mathbb{R}^m_+$ have intractable normalizing constants except for special cases e.g. with diagonal covariance matrices. This inability to calculate normalizing constants makes density estimation challenging and maximum likelihood estimation becomes less ideal.

Fortunately, Hyvärinen (2005) introduced the method of *score matching* that comes to our rescue. Given a continuous distribution P_0 supported on $\mathbb{R}^m$ with Lebesgue density p_0, let $\mathcal{P}$ be a family of distributions with twice continuously differentiable densities. The score matching estimator of p_0 using $\mathcal{P}$ as a model is the minimizer of the expected ℓ_2

distance between the gradients of the true log density $\log p_0$ on $\mathbb{R}^m$ and a proposed log density $\log p$ from $\mathcal{P}$, namely $\int_{\mathbb{R}^m} p_0(\boldsymbol{x})\|\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}) - \nabla_{\boldsymbol{x}} \log p_0(\boldsymbol{x})\|_2^2 \,\mathrm{d}\boldsymbol{x}$. The loss depends on p_0, but integration by parts can be used to rewrite it in a form that can be approximated by averaging over the sample without knowing p_0. A key feature of score matching is that normalizing constants cancel in gradients of log-densities, allowing for simple treatment of models with intractable normalizing constants. For exponential families, the loss is quadratic in the canonical parameter, making optimization straightforward.

If the considered distributions are supported on a proper subset of $\mathbb{R}^m$, then the integration by parts arguments underlying the score matching estimator may fail due to discontinuities at the boundary of the support. For data supported on the non-negative orthant $\mathcal{D} \equiv \mathbb{R}^m_+$, Hyvärinen (2007) addresses this problem by modifying the loss to $\int_{\mathbb{R}^m} p_0(\boldsymbol{x})\|\nabla \log p(\boldsymbol{x}) \odot \boldsymbol{x} - \nabla \log p_0(\boldsymbol{x}) \odot \boldsymbol{x}\|_2^2 \,\mathrm{d}\boldsymbol{x}$, where $\odot$ denotes entrywise multiplication. In this loss, boundary effects are dampened by multiplying gradients element-wise with the identity functions x_j.

In Chapter 2, we propose *generalized score matching* methods for densities supported on $\mathcal{D} \equiv \mathbb{R}^m_+$ based on element-wise multiplication with slowly growing/bounded functions $\boldsymbol{h}^{1/2}(\boldsymbol{x})$ in place of $\boldsymbol{x}$. We show that this modification drastically improves estimation accuracy, both theoretically and empirically. This work is published in Yu et al. (2018, 2019b).

In Chapter 3, we further extend the generalized score matching methods to densities supported on more general domains $\mathcal{D} \subseteq \mathbb{R}^m$ that may be either bounded or unbounded. This includes densities supported on simplex domains, for which density estimation has been difficult due to their sum-to-one constraint, with important applications such as compositional microbiome data. Our generalization completes the framework for estimation of continuous distributions supported on various types of domains with intractable normalizing constants.

Liu and Kanamori (2019), on the other hand, extended our results in Chapter 2 to bounded open subspaces $\mathcal{D}$ of $\mathbb{R}^m$ that have piecewise smooth boundaries by minimizing $\sup_{g\in\mathcal{G}} \int_{\mathcal{D}} p_0(\boldsymbol{x})g(\boldsymbol{x})\|\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}) - \nabla_{\boldsymbol{x}} \log p_0(\boldsymbol{x})\|_2^2 \,\mathrm{d}\boldsymbol{x}$, where $\mathcal{G} \equiv \{g | g(\boldsymbol{x}) = 0 \,\forall \boldsymbol{x} \in \partial\mathcal{D}$ and g is 1-Lipschitz continuous$\}$, $\partial\mathcal{D}$ being the boundary of $\mathcal{D}$. The $\sup_{g\in\mathcal{G}}$ in the loss is obtained at $g_0(\boldsymbol{x}) \equiv \min_{\boldsymbol{x}'\in\partial\mathcal{D}} \|\boldsymbol{x} - \boldsymbol{x}'\|$, the distance of $\boldsymbol{x}$ to $\partial\mathcal{D}$. In Chapter 3, we extend

their slightly different approach to unbounded subsets of $\mathbb{R}^m$, but we empirically show that their approach and its extension seem inferior to our generalized score matching estimators in many settings.

1.2 Undirected Graphical Models

Graphical models specify conditional independence relations among variables in a random vector $\boldsymbol{Y}$ indexed by the nodes $\mathcal{V}$ of a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with edge set $\mathcal{E}$ (Maathuis et al., 2019). Models based on undirected graphs may be used to explore conditional independence between any two variables Y_V and Y_U given all others $(Y_W)_{W \neq U,V}$, as represented by the absence of an edge between V and U in $\mathcal{E}$.

Among undirected graphical models, largely due to their tractability, Gaussian graphical models (GGMs) have gained great popularity. The conditional independence graph of a multivariate normal vector $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is determined by the *inverse covariance matrix* $\mathbf{K} \equiv \boldsymbol{\Sigma}^{-1}$, also termed *concentration* or *precision matrix*. Specifically, X_i and X_j are conditionally independent given all other variables if and only if the (i, j)-th and the (j, i)-th entries of $\mathbf{K}$ are both zero. This simple relation underlies a rich literature including Drton and Perlman (2004), Meinshausen and Bühlmann (2006), Yuan and Lin (2007) and Friedman et al. (2008), among others.

More recent work has provided tractable procedures also for non-Gaussian undirected graphical models. This includes Gaussian copula models (Liu et al., 2009; Dobra and Lenkoski, 2011; Liu et al., 2012), Ising models (Ravikumar et al., 2010), other exponential family models (Chen et al., 2015; Yang et al., 2015), as well as semi- or non-parametric estimation techniques (Fellinghauer et al., 2013; Voorman et al., 2014). In this dissertation, we apply our generalized score matching method to a class of pairwise interaction power models, introduced in Section 1.4, that generalizes Gaussian and non-negative Gaussian random variables, recently considered by Lin et al. (2016) and Yu et al. (2016), as well as square root graphical models proposed by Inouye et al. (2016) when the sufficient statistic function is a pure power.

1.3 Directed Graphical Models and Causal Discovery

Models based on directed acyclic graphs (DAGs), for which $\mathcal{E}$ is comprised of directed edges, capture conditional independence structure that naturally arises from cause-effect relationships between the variables. In biology and genetics, graphical models have been applied to infer the structure of gene regulatory networks based on measurements of gene expression (Maathuis et al., 2019, Sections 20-21). Traditional technologies produce expression levels aggregated over hundreds or thousands of individual cells, and these bulk measurements are frequently modeled using the assumption of Gaussianity. In directed GGMs the exact structure of the underlying DAG cannot be identified from purely observational data, and the target of inference becomes an equivalence class of DAGs. For instance, one cannot differentiate between $V \to U$ and $U \to V$ when the variables are assumed bivariate normal. In the Gaussian case, directed graphical models posit linear functional relationships between the variables coupled with additive Gaussian noise. A more recent line of work emphasizes that directed graphical models that alter this assumption to nonlinear functional relationships and additive noise (Peters et al., 2014), or linear relations and non-Gaussian noise (Shimizu et al., 2006; Wang and Drton, 2020), or linear relations with homoscedastic Gaussian noise (Peters and Bühlmann, 2013; Chen et al., 2019) are amenable to causal discovery in the sense that different DAGs are no longer equivalent.

More recent technology obtains sequencing measurements of mRNA present in single cells. This new technology, as well as the larger sample sizes it provides, promise to give more information than bulk measurements, but at the same time bring in a unique new challenge. At the single cell level, genes appear as "on" with positive single cell gene expression levels, or as "off" with the recorded measurements zero or negligible (McDavid et al., 2019). A novel undirected graphical model that deals with this zero-inflation was introduced by McDavid et al. (2019). Since more information can be inferred from single-cell sequencing data, one would hope that the data can also be analyzed using more informative directed graphical models, and that we can infer which variables (genes) are the causes of change

in other variables (expression levels of other genes). In Chapter 4, we formulate directed graphical models for zero-inflated data, and prove that under a weak assumption one can recover the exact DAG from the joint distribution. We use simulation studies to support our identifiability theory and justify the use of two methods for estimating the DAG in practice.

1.4 Score Matching Applied to Pairwise Interaction Graphical Models

In this section we introduce the *pairwise interaction power models*, also called *a-b models*, that we examine in Chapters 2 and 3. In particular, we assume the models have (Lebesgue) probability density functions proportional to

$$\exp\left\{-\frac{1}{2a}\boldsymbol{x}^a\mathbf{K}\boldsymbol{x}^a+\frac{1}{b}\boldsymbol{\eta}^\top\boldsymbol{x}^b\right\} \tag{1.1}$$

supported on some general domain $\mathcal{D}\subseteq\mathbb{R}^m$. Here $a\geq 0$ and $b\geq 0$ are known constants, and $\mathbf{K}\in\mathbb{R}^{m\times m}$ and $\boldsymbol{\eta}\in\mathbb{R}^m$ are unknown parameters of interest. For $a=0$ we define ${\boldsymbol{x}^a}^\top\mathbf{K}\boldsymbol{x}^a/a\equiv(\log\boldsymbol{x})^\top\mathbf{K}(\log\boldsymbol{x})$ and for $b=0$ we define $\boldsymbol{\eta}^\top\boldsymbol{x}^b/b\equiv\boldsymbol{\eta}^\top(\log\boldsymbol{x})$. This class of models is motivated by the form of important univariate distributions, including gamma and truncated normal distributions.

This model class provides a framework for pairwise interaction that is concrete yet rich enough to capture key differences in how densities may behave; in particular, if $\mathcal{D}$ is a product set of the form $\mathcal{D}_1\times\cdots\times\mathcal{D}_m$ with one-dimensional sets $\mathcal{D}_1,\ldots,\mathcal{D}_m$, X_i and X_j are conditionally independent given all others if and only if $\kappa_{ij}=\kappa_{ji}=0$ in the interaction matrix $\mathbf{K}$, just as for Gaussian graphical models. We will develop estimators of $(\boldsymbol{\eta},\mathbf{K})$ and the associated conditional independence graph using the proposed *generalized score matching*.

A special case of (1.1) are (truncated) Gaussian graphical models, with $a=b=1$. Let $\boldsymbol{\mu}\in\mathbb{R}^m$, and let $\mathbf{K}$ be a positive definite matrix. Then a random vector $\boldsymbol{X}\in\mathcal{D}$ follows a (truncated) normal distribution for mean parameter $\boldsymbol{\mu}$ and inverse covariance parameter $\mathbf{K}$, in symbols $\boldsymbol{X}\sim\mathrm{TN}_{\mathcal{D}}(\boldsymbol{\mu},\mathbf{K})$, if it has density proportional to

$$\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\mathbf{K}(\boldsymbol{x}-\boldsymbol{\mu})\right\} \tag{1.2}$$

on domain $\mathcal{D} \subseteq \mathbb{R}^m$. We refer to $\mathbf{\Sigma} = \mathbf{K}^{-1}$ as the covariance parameter of the distribution, and note that the $\boldsymbol{\eta}$ parameter in (1.1) is $\mathbf{K}\boldsymbol{\mu}$. Another special case of (1.1) is the exponential square root graphical models in Inouye et al. (2016), where $a = b = 1/2$ and $\mathcal{D} \equiv \mathbb{R}^m_+$.

Lin et al. (2016) proposed estimating the truncated GGMs supported on $\mathcal{D} \equiv \mathbb{R}^m_+$ using the score matching loss from Hyvärinen (2007) plus an ℓ_1 loss on the off-diagonal elements of $\mathbf{K}$. We propose a similar regularized loss using our generalized score matching loss for estimating the a-b models above. The loss in Lin et al. (2016) may be unbounded from below, and thus infinitely many minimizers exist; in Chapter 2 we fix this problem by introducing *diagonal multipliers*, effectively imposing an elastic net-type penalty (Zou and Hastie, 2005) on $\mathbf{K}$. In addition, Lin et al. (2016) requires a sample size $n = \Omega\left((\log m)^8\right)$ for GGMs on $\mathbb{R}^m_+$, while for our generalized score matching estimator the sample complexity can be reduced to $n = \Omega(\log m)$ for any domain $\mathcal{D}$ that is a finite disjoint union of convex sets. In Chapter 3, we present similar theoretical guarantees for general a-b models on a general domain $\mathcal{D}$.

1.5 Notation

Constant scalars, vectors, and functions are written in lower-case (e.g., a, $\boldsymbol{a}$), random scalars and vectors in upper-case (e.g., X, $\boldsymbol{X}$). Regular font is used for scalars (e.g. a, X), and boldface for vectors (e.g. $\boldsymbol{a}$, $\boldsymbol{X}$). Graph nodes are written in unbolded upper-case (e.g. U, V), and domains, sets of nodes, vectors of functions in calligraphic font (e.g. $\mathcal{D}$, $\mathcal{V}$, $\mathcal{H}$). Matrices are in upright bold, with constant matrices in upper-case ($\mathbf{K}$, $\mathbf{M}$) and random matrices holding observations in lower-case ($\mathbf{x}$, $\mathbf{y}$).

Subscripts refer to entries in vectors and columns in matrices. When used as a subscript of a vector, a set of nodes/indices selects the corresponding entries from the vector, e.g. $\boldsymbol{y}_{\mathcal{V}} = (y_V)_{V \in \mathcal{V}}$. Superscripts refer to rows in matrices. So X_j is the j-th component of a random vector $\boldsymbol{X}$. For a data matrix $\mathbf{x} \in \mathbb{R}^{n \times m}$, each row comprising one observation of m variables/features, $X_j^{(i)}$ is the j-th feature for the i-th observation. For a matrix $\mathbf{K} = [\kappa_{ij}]_{i,j} \in \mathbb{R}^{q \times q}$, $\operatorname{diag}(\mathbf{K}) = (\kappa_{11}, \ldots, \kappa_{qq})^\top$ denotes its diagonal, and for a vector $\boldsymbol{v} \in \mathbb{R}^q$, $\operatorname{diag}(\boldsymbol{v})$ is the $q \times q$ diagonal matrix with diagonal entries $v_1, \ldots, v_q$. Stacking the columns

of $\mathbf{K}$ gives its vectorization $\mathrm{vec}(\mathbf{K}) = (\kappa_{11}, \ldots, \kappa_{q1}, \kappa_{12}, \ldots, \kappa_{q2}, \ldots, \kappa_{1r}, \ldots, \kappa_{qr})^\top$.

For $a \geq 1$, the ℓ_a-norm of a vector $\boldsymbol{v} \in \mathbb{R}^q$ is denoted

$$\|\boldsymbol{v}\|_a = \left(\sum_{j=1}^{q} |v_j|^a \right)^{1/a},$$

with $\|\boldsymbol{v}\|_\infty = \max_{j=1,\ldots,q} |v_j|$. A matrix $\mathbf{K} = [\kappa_{ij}]_{i,j} \in \mathbb{R}^{q \times r}$ has Frobenius norm

$$|||\mathbf{K}|||_F \equiv \|\mathrm{vec}(\mathbf{K})\|_2 \equiv \sqrt{\sum_{i=1}^{q} \sum_{j=1}^{r} \kappa_{ij}^2},$$

and max norm $\|\mathbf{K}\|_\infty \equiv \|\mathrm{vec}(\mathbf{K})\|_\infty \equiv \max_{i,j} |\kappa_{ij}|$. Its ℓ_a-ℓ_b operator norm is

$$|||\mathbf{K}|||_{a,b} \equiv \max_{\boldsymbol{x} \neq \mathbf{0}} \frac{\|\mathbf{K}\boldsymbol{x}\|_b}{\|\boldsymbol{x}\|_a}$$

with shorthand notation $|||\mathbf{K}|||_a \equiv |||\mathbf{K}|||_{a,a}$; for instance, $|||\mathbf{K}|||_\infty \equiv \max_{i=1\ldots,q} \sum_{j=1}^{r} |\kappa_{ij}|$.

For a function $f : \mathbb{R}^m \to \mathbb{R}$, we define $\partial_j f(\boldsymbol{x})$ as the partial derivative with respect to x_j, and $\partial_{jj} f(\boldsymbol{x}) = \partial_j \partial_j f(\boldsymbol{x})$. For vector-valued $\boldsymbol{f} : \mathbb{R}^m \to \mathbb{R}^m$, $\boldsymbol{x} \mapsto (f_1(\boldsymbol{x}), \ldots, f_m(\boldsymbol{x}))$, we let $\boldsymbol{f}'(\boldsymbol{x}) = (\partial_1 f_1(\boldsymbol{x}), \ldots, \partial_m f_m(\boldsymbol{x}))^\top$ be the vector of partial derivatives. Likewise $\boldsymbol{f}''(x)$ is used for second partial derivatives. For $a \in \mathbb{R}$, let $\boldsymbol{f}^a(\boldsymbol{x}) \equiv (f_1^a(\boldsymbol{x}), \ldots, f_m^a(\boldsymbol{x}))^\top$. For two compatible functions f and g, $f \circ g$ denotes their function composition. For $\boldsymbol{a}$, $\boldsymbol{b} \in \mathbb{R}^m$, $\boldsymbol{a} \odot \boldsymbol{b} \equiv (a_1 b_1, \ldots, a_m b_m)^\top$.

Finally, the symbol $\mathbb{1}_A(\cdot)$ denotes the indicator function of a set A, while $\mathbf{1}_n \in \mathbb{R}^n$ is the vector of all 1's. When it is clear from the context, $\mathbb{E}_0$ denotes the expectation under a true distribution P_0.

Chapter 2

GENERALIZED SCORE MATCHING FOR NON-NEGATIVE DATA

2.1 Structure of Chapter

In this chapter we discuss our *generalized score matching* estimator for continuous distributions supported on $\mathbb{R}^m_+$, with applications of *regularized general score matching* to the pairwise interaction power models (see (1.1)). The chapter is organized as follows. Section 2.2 introduces score matching and our proposed *generalized score matching* for distributions on $\mathbb{R}^m_+$. In Section 2.3, we apply generalized score matching to exponential families, with univariate truncated normal distributions as an example. *Regularized generalized score matching* for graphical models and the notion of *diagonal multipliers* are formulated in Section 2.4. The estimators for pairwise interaction power models are shown in Section 2.5, while theoretical consistency results are presented in Section 2.6, where we treat the probabilistically most tractable case of truncated GGMs. Simulation results and applications to RNAseq data are given in Section 2.7. Proofs for theorems in Sections 2.2–2.6 are presented in Appendices A.1 and A.2. Additional experimental results are presented in Appendix A.3. The work in this chapter is published in Yu et al. (2018) and Yu et al. (2019b).

2.2 Score Matching

In this section, we review the original score matching and develop our generalized score matching estimators.

2.2.1 Original Score Matching

Let $\boldsymbol{X}$ be a random vector taking values in $\mathbb{R}^m$ with distribution P_0 and density p_0. Let $\mathcal{P}$ be a family of distributions of interest with twice continuously differentiable densities supported on $\mathbb{R}^m$. Suppose $P_0 \in \mathcal{P}$. The *score matching loss* for $P \in \mathcal{P}$, with density p, is given by

$$J(P) = \int_{\mathbb{R}^m} p_0(\boldsymbol{x}) \|\nabla \log p(\boldsymbol{x}) - \nabla \log p_0(\boldsymbol{x})\|_2^2 \, \mathrm{d}\boldsymbol{x}. \tag{2.1}$$

The gradients in (2.1) can be thought of as gradients with respect to a hypothetical location parameter, evaluated at the origin (Hyvärinen, 2005). The loss $J(P)$ is minimized if and only

if $P = P_0$, which forms the basis for estimation of P_0. Importantly, since the loss depends on p only through its log-gradient, it suffices to know p up to a normalizing constant. Under mild conditions, (2.1) can be rewritten as

$$J(P) = \int_{\mathbb{R}^m} p_0(\boldsymbol{x}) \sum_{j=1}^{m} \left[\partial_{jj} \log p(\boldsymbol{x}) + \frac{(\partial_j \log p(\boldsymbol{x}))^2}{2} \right] \mathrm{d}\boldsymbol{x}, \tag{2.2}$$

plus a constant independent of p. The integral in (2.2) can be approximated by a sample average; this alleviates the need for knowing the true density p_0, and provides a way to estimate p_0.

2.2.2 Generalized Score Matching for Non-Negative Data

When the true density p_0 is supported on a proper subset of $\mathbb{R}^m$, the integration by parts underlying the equivalence of (2.1) and (2.2) may fail due to discontinuity at the boundary. For distributions supported on the non-negative orthant, $\mathbb{R}^m_+$, Hyvärinen (2007) addressed this issue by instead minimizing the *non-negative score matching loss*

$$J_+(P) = \int_{\mathbb{R}^m_+} p_0(\boldsymbol{x}) \|\nabla \log p(\boldsymbol{x}) \odot \boldsymbol{x} - \nabla \log p_0(\boldsymbol{x}) \odot \boldsymbol{x}\|_2^2 \, \mathrm{d}\boldsymbol{x}. \tag{2.3}$$

This loss can be motivated via gradients with respect to a hypothetical scale parameter (Hyvärinen, 2007). Under mild conditions, $J_+(P)$ can again be rewritten in terms of an expectation of a function independent of p_0, thus allowing one to form a sample loss.

In this work, we consider generalizing the non-negative score matching loss as follows.

Definition 1. *Let $\mathcal{P}_+$ be the family of distributions of interest, and assume every $P \in \mathcal{P}_+$ has a twice continuously differentiable density supported on $\mathbb{R}^m_+$. Suppose the m-variate random vector $\boldsymbol{X}$ has true distribution $P_0 \in \mathcal{P}_+$, and let p_0 be its twice continuously differentiable density. Let $h_1, \ldots, h_m : \mathbb{R}_+ \to \mathbb{R}_+$ be a.s. positive functions that are absolutely continuous in every bounded sub-interval of $\mathbb{R}_+$, and set $\boldsymbol{h}(\boldsymbol{x}) = (h_1(x_1), \ldots, h_m(x_m))^\top$. For $P \in \mathcal{P}_+$ with density p, the* generalized $\boldsymbol{h}$-score matching loss *is*

$$J_{\boldsymbol{h}}(P) = \int_{\mathbb{R}^m_+} \frac{1}{2} p_0(\boldsymbol{x}) \|\nabla \log p(\boldsymbol{x}) \odot \boldsymbol{h}(\boldsymbol{x})^{1/2} - \nabla \log p_0(\boldsymbol{x}) \odot \boldsymbol{h}(\boldsymbol{x})^{1/2}\|_2^2 \, \mathrm{d}\boldsymbol{x}, \tag{2.4}$$

where $\boldsymbol{h}^{1/2}(\boldsymbol{x}) \equiv (h_1^{1/2}(x_1), \ldots, h_m^{1/2}(x_m))^\top$.

Proposition 1. *The distribution* P_0 *is the unique minimizer of* $J_{\boldsymbol{h}}(P)$ *for* $P \in \mathcal{P}_+$.

Proof. First, observe that $J_{\boldsymbol{h}}(P) \geq 0$ and $J_{\boldsymbol{h}}(P_0) = 0$. For uniqueness, suppose $J_{\boldsymbol{h}}(P_1) = 0$ for some $P_1 \in \mathcal{P}_+$. Let p_0 and p_1 be the respective densities. By assumption $p_0(\boldsymbol{x}) > 0$ a.s. and $h_j^{1/2}(\boldsymbol{x}) > 0$ a.s. for all $j = 1, \ldots, m$. Therefore, we must have $\nabla \log p_1(\boldsymbol{x}) = \nabla \log p_0(\boldsymbol{x})$ a.s., or equivalently, $p_1(\boldsymbol{x}) = \text{const} \times p_0(\boldsymbol{x})$ almost surely in $\mathbb{R}_+^m$. Since p_1 and p_0 are continuous densities supported on $\mathbb{R}_+^m$, it follows that $p_1(\boldsymbol{x}) = p_0(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}_+^m$. □

Choosing all $h_j(x) = x^2$ recovers the loss from (2.3). In our generalization, we will focus on using functions h_j that are increasing but are bounded or grow rather slowly. This will alleviate the need to estimate higher moments, leading to better practical performance and improved theoretical guarantees.

We will consider the following assumptions:

(A1) $p_0(\boldsymbol{x}) h_j(x_j) \partial_j \log p(\boldsymbol{x}) \Big|_{x_j \searrow 0^+}^{x_j \nearrow +\infty} = 0\,, \quad \forall \boldsymbol{x}_{-j} \in \mathbb{R}_+^{m-1}, \quad \forall p \in \mathcal{P}_+;$

(A2) $\mathbb{E}_{p_0} \|\nabla \log p(\boldsymbol{X}) \odot \boldsymbol{h}^{1/2}(\boldsymbol{X})\|_2^2 < +\infty, \;\; \mathbb{E}_{p_0} \|(\nabla \log p(\boldsymbol{X}) \odot \boldsymbol{h}(\boldsymbol{X}))'\|_1 < +\infty, \;\; \forall p \in \mathcal{P}_+,$

where $\partial_j \log p(\boldsymbol{x}) \equiv \frac{\partial \log p(\boldsymbol{y})}{\partial y_j}\Big|_{\boldsymbol{y}=\boldsymbol{x}}$, $f(\boldsymbol{x}) \Big|_{x_j \searrow 0^+}^{x_j \nearrow +\infty} \equiv \lim_{x_j \nearrow +\infty} f(\boldsymbol{x}) - \lim_{x_j \searrow 0} f(\boldsymbol{x})$, "$\forall p \in \mathcal{P}_+$" is a shorthand for "for all p being the density of some $P \in \mathcal{P}_+$", and the prime symbol denotes component-wise differentiation. While the second half of (A2) was not made explicit in Hyvärinen (2005, 2007), (A1)-(A2) were both required for integration by parts and Fubini-Tonelli to apply.

Once the forms of p_0 and p are given, sufficient conditions for $\boldsymbol{h}$ for Assumptions (A1)-(A2) to hold are easy to find. In particular, (A1) and (A2) are easily satisfied and verified for exponential families.

Integration by parts yields the following theorem which shows that $J_{\boldsymbol{h}}$ from (2.4) is an expectation (under P_0) of a function that does not depend on p_0, similar to (2.2). The proof is given in Appendix A.1.1.

Theorem 2. *Under (A1) and (A2), the loss from (2.4) equals*

$$J_{\boldsymbol{h}}(P) = \int_{\mathbb{R}_+^m} p_0(\boldsymbol{x}) \sum_{j=1}^{m} \Big[h_j'(x_j)\partial_j(\log p(\boldsymbol{x})) + h_j(x_j)\partial_{jj}(\log p(\boldsymbol{x})) + \frac{1}{2} h_j(x_j) \left(\partial_j(\log p(\boldsymbol{x}))\right)^2 \Big] \, \mathrm{d}\boldsymbol{x} \quad (2.5)$$

plus a constant independent of p.

Given a data matrix $\mathbf{x} \in \mathbb{R}^{n \times m}$ with rows $\boldsymbol{X}^{(i)}$, we define the sample version of (2.5) as

$$\hat{J}_{\boldsymbol{h}}(P) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} \Big\{ h_j'(X_j^{(i)})\partial_j(\log p(\boldsymbol{X}^{(i)})) + h_j(X_j^{(i)}) \left[\partial_{jj}(\log p(\boldsymbol{X}^{(i)})) + \frac{1}{2} \left(\partial_j(\log p(\boldsymbol{X}^{(i)}))\right)^2 \right] \Big\}. \quad (2.6)$$

Subsequently, for a distribution P with density p, we let $J_{\boldsymbol{h}}(p) \equiv J_{\boldsymbol{h}}(P)$. Similarly, when a distribution $P_{\boldsymbol{\theta}}$ with density $p_{\boldsymbol{\theta}}$ is associated to a parameter vector $\boldsymbol{\theta}$, we write $J_{\boldsymbol{h}}(\boldsymbol{\theta}) \equiv J_{\boldsymbol{h}}(p_{\boldsymbol{\theta}}) \equiv J_{\boldsymbol{h}}(P_{\boldsymbol{\theta}})$. We apply similar conventions to the sample version $\hat{J}_{\boldsymbol{h}}(P)$. We note that this type of loss is also treated in slightly different settings in Parry (2016) and Almeida and Gidas (1993).

Remark 1. In the one-dimensional case, using the notation in Parry et al. (2012), $J_{\boldsymbol{h}}(P)$ and $\hat{J}_{\boldsymbol{h}}(P)$ correspond to $d(P_0, P)$ and $S(x, P)$, respectively, and can be generated by $\phi(x, p, p_1) \equiv -h(x)p_1^2/(2p)$ (c.f. Equations (39), (51), (53) and Section 10.1 therein). Thus Theorem 2 follows from this correspondence. While (A1) is equivalent to the condition implied by the boundary divergence $d_b = 0$ in that paper, (A2), which we assume for invoking Fubini-Tonelli due to multi-dimensionality, is not present. On the other hand, while Parry (2016) treats the multivariate case, it does not cover the connection between our $J_{\boldsymbol{h}}$ and $\hat{J}_{\boldsymbol{h}}$. Since ϕ is concave but not strictly concave in (p, p_1), the results in Parry (2016) only imply that P_0 is *a* minimizer, a weaker conclusion than Proposition 1.

2.3 Exponential Families

In this section, we study the case where $\mathcal{P}_+ \equiv \{p_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ is an exponential family comprising continuous distributions with support $\mathbb{R}_+^m$. More specifically, we consider densities that are indexed by the canonical parameter $\boldsymbol{\theta} \in \mathbb{R}^r$ and have the form

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{\theta}^\top \boldsymbol{t}(\boldsymbol{x}) - \psi(\boldsymbol{\theta}) + b(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathbb{R}_+^m, \tag{2.7}$$

where $\boldsymbol{t}(\boldsymbol{x}) \in \mathbb{R}_+^r$ comprises the sufficient statistics, $\psi(\boldsymbol{\theta})$ is a normalizing constant depending on $\boldsymbol{\theta}$ only, and $b(\boldsymbol{x})$ is the base measure, with $\boldsymbol{t}$ and b a.s. differentiable with respect to each component. Define $\boldsymbol{t}_{\boldsymbol{j}}'(\boldsymbol{x}) \equiv (\partial_j t_1(\boldsymbol{x}), \ldots, \partial_j t_r(\boldsymbol{x}))^\top$ and $b_j'(\boldsymbol{x}) \equiv \partial_j b(\boldsymbol{x})$.

Theorem 3. *Under Assumptions (A1)-(A2) from Section 2.2.2, the empirical generalized $\boldsymbol{h}$-score matching loss (2.6) can be rewritten as a quadratic function in $\boldsymbol{\theta} \in \mathbb{R}^r$:*

$$\hat{J}_{\boldsymbol{h}}(p_{\boldsymbol{\theta}}) = \frac{1}{2}\boldsymbol{\theta}^\top \boldsymbol{\Gamma}(\mathbf{x})\boldsymbol{\theta} - \boldsymbol{g}(\mathbf{x})^\top \boldsymbol{\theta} + \text{const}, \quad \textit{where} \tag{2.8}$$

$$\boldsymbol{\Gamma}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} h_j(X_j^{(i)})\boldsymbol{t}_j'(\boldsymbol{X}^{(i)})\boldsymbol{t}_j'(\boldsymbol{X}^{(i)})^\top \quad \textit{and} \tag{2.9}$$

$$\boldsymbol{g}(\mathbf{x}) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\left[h_j(X_j^{(i)})b_j'(\boldsymbol{X}^{(i)})\boldsymbol{t}_j'(\boldsymbol{X}^{(i)}) + h_j(X_j^{(i)})\boldsymbol{t}_j''(\boldsymbol{X}^{(i)}) + h_j'(X_j^{(i)})\boldsymbol{t}_j'(\boldsymbol{X}_i)\right] \tag{2.10}$$

are sample averages of functions of the data matrix $\mathbf{x}$ only.

Define $\boldsymbol{\Gamma}_0 \equiv \mathbb{E}_{p_0}\boldsymbol{\Gamma}(\mathbf{x})$, $\boldsymbol{g}_0 \equiv \mathbb{E}_{p_0}\boldsymbol{g}(\mathbf{x})$, and $\boldsymbol{\Sigma}_0 \equiv \mathbb{E}_{p_0}[(\boldsymbol{\Gamma}(\mathbf{x})\boldsymbol{\theta}_0 - g(\mathbf{x}))(\boldsymbol{\Gamma}(\mathbf{x})\boldsymbol{\theta}_0 - g(\mathbf{x}))^\top]$.

Theorem 4. *Suppose that*

(C1) $\boldsymbol{\Gamma}$ is a.s. invertible, and

(C2) $\boldsymbol{\Gamma}_0$, $\boldsymbol{\Gamma}_0^{-1}$, $\boldsymbol{g}_0$ and $\boldsymbol{\Sigma}_0$ exist and are entry-wise finite.

Then the minimizer of (2.8) is a.s. unique with closed-form solution $\hat{\boldsymbol{\theta}} \equiv \boldsymbol{\Gamma}(\mathbf{x})^{-1}\boldsymbol{g}(\mathbf{x})$. Moreover,

$$\hat{\boldsymbol{\theta}} \to_{a.s.} \boldsymbol{\theta}_0 \quad \textit{and} \quad \sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \to_d \mathcal{N}_r\left(\mathbf{0}, \boldsymbol{\Gamma}_0^{-1}\boldsymbol{\Sigma}_0\boldsymbol{\Gamma}_0^{-1}\right) \quad \textit{as } n \to \infty.$$

Theorems 3 and 4 are proved in Appendix A.1.2. Theorem 3 clarifies the quadratic nature of the loss, and Theorem 4 provides a basis for asymptotically valid tests and confidence intervals for the parameter $\boldsymbol{\theta}$. Note that Condition (C1) holds if and only if $h_j(X_j) > 0$ a.s. and $[\boldsymbol{t}_j'(\boldsymbol{X}^{(1)}), \ldots, \boldsymbol{t}_j'(\boldsymbol{X}^{(n)})] \in \mathbb{R}^{r \times n}$ has rank r a.s. for some $j = 1, \ldots, m$.

The conclusion in Theorem 4 indicates that, similar to the estimator in Hyvärinen (2007) with $h_j(x) = x^2$, the closed-form solution for our generalized $\hat{\boldsymbol{\theta}}$ allows one to consistently estimate the canonical parameter in an exponential family distribution without needing to calculate the often complicated normalizing constant $\psi(\boldsymbol{\theta})$ or resort to numerical methods. Computational details are explicated in Section 2.5.3.

Below we illustrate the estimator $\hat{\boldsymbol{\theta}}$ in the case of univariate truncated normal distributions. We assume (A1)-(A2) and (C1)-(C2) throughout.

Example 3.1. *Univariate ($m = r = 1$) truncated normal distributions for mean parameter μ and variance parameter σ^2 have density*

$$p_{\mu,\sigma^2}(x) \propto \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}, \quad x \in \mathbb{R}_+. \tag{2.11}$$

If σ^2 is known but μ unknown, then writing the density in canonical form as in (2.7) yields

$$p_\theta(x) \propto \exp\left\{\theta t(x) + b(x)\right\}, \quad \theta \equiv \frac{\mu}{\sigma^2}, \quad t(x) \equiv x, \quad b(x) = -\frac{x^2}{2\sigma^2}.$$

Given an i.i.d. sample $X_1, \ldots, X_n \sim p_{\mu_0,\sigma^2}$, the generalized h-score matching estimator of μ is

$$\hat{\mu}_h \equiv \frac{\sum_{i=1}^n h(X_i)X_i - \sigma^2 h'(X_i)}{\sum_{i=1}^n h(X_i)}.$$

If $\lim_{x \searrow 0^+} h(x) = 0$, $\lim_{x \nearrow +\infty} h^2(x)(x-\mu_0)p_{\mu_0,\sigma^2}(x) = 0$ and the expectations are finite (for example, when $h(x) = o(\exp(Mx^2))$ for $M < \frac{1}{4\sigma^2}$), then

$$\sqrt{n}(\hat{\mu}_h - \mu_0) \to_d \mathcal{N}\left(0, \frac{\mathbb{E}_0[\sigma^2 h^2(X) + \sigma^4 h'^2(X)]}{\mathbb{E}_0^2[h(X)]}\right).$$

We recall that the Cramér-Rao lower bound *(i.e. the lower bound on the variance of any unbiased estimator) for estimating μ is*

$$\frac{\sigma^4}{\mathrm{var}(X - \mu_0)}.$$

Example 3.2. *Consider the univariate truncated normal distributions from (2.11) in the setting where the mean parameter μ is known but the variance parameter $\sigma^2 > 0$ is unknown. In canonical form as in (2.7), we write*

$$p_\theta(x) \propto \exp\{\theta t(x) + b(x)\}, \quad \theta \equiv \frac{1}{\sigma^2}, \quad t(x) \equiv -(x-\mu)^2/2, \quad b(x) = 0.$$

Given an i.i.d. sample $X_1, \ldots, X_n \sim p_{\mu,\sigma_0^2}$, the generalized h-score matching estimator of σ^2 is

$$\hat{\sigma}_h^2 \equiv \frac{\sum_{i=1}^n h(X_i)(X_i-\mu)^2}{\sum_{i=1}^n h(X_i) + h'(X_i)(X_i-\mu)}.$$

If, in addition to the assumptions in Example 3.1, $\lim_{x \nearrow +\infty} h^2(x)(x-\mu)^3 p_{\mu,\sigma_0^2}(x) = 0$, then

$$\sqrt{n}(\hat{\sigma}_h^2 - \sigma_0^2) \to_d \mathcal{N}\left(0, \frac{2\sigma_0^6 \mathbb{E}_0[h^2(X)(X-\mu)^2] + \sigma_0^8 \mathbb{E}_0[h'^2(X)(X-\mu)^2]}{\mathbb{E}_0^2[h(X)(X-\mu)^2]}\right).$$

Moreover, the Cramér-Rao lower bound *for estimating σ^2 is*

$$\frac{4\sigma_0^8}{\mathrm{var}(X-\mu)^2}.$$

Remark 2. In Example 3.2, if $\mu_0 = 0$, then $h(x) \equiv 1$ also satisfies (A1)-(A2) and (C1)-(C2) and one recovers the sample variance $\frac{1}{n}\sum_i X_i^2$, which obtains the Cramér-Rao lower bound.

In these examples, there is a benefit in using a bounded function h, which can be explained as follows. When $\mu \gg \sigma$, there is effectively no truncation to the Gaussian distribution, and our method adapts to using low moments in (2.4), since a bounded and increasing $h(x)$ becomes almost constant as it reaches its asymptote for x large. Hence, we effectively revert to the original score matching (recall Section 2.2.1). In the other cases, the truncation effect is significant and our estimator uses higher moments accordingly.

Figure 2.1 plots the asymptotic variance of $\hat{\mu}_h$ from Example 3.1, with $\sigma = 1$ known. Efficiency as measured by the Cramér-Rao lower bound divided by the asymptotic variance is also shown. We see that two truncated versions of $\log(1+x)$ have asymptotic variance close to the Cramér-Rao bound. This asymptotic variance is also reflective of the variance for smaller finite samples.

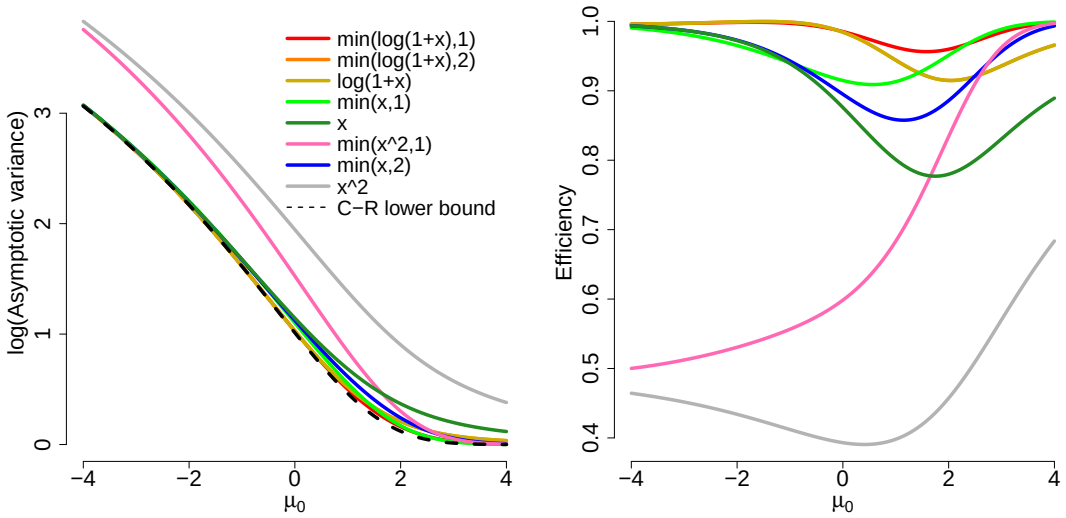


Figure 2.1: Log of asymptotic variance and efficiency with respect to the Cramér-Rao bound for $\hat{\mu}_h$ ($\sigma^2 = 1$ known).

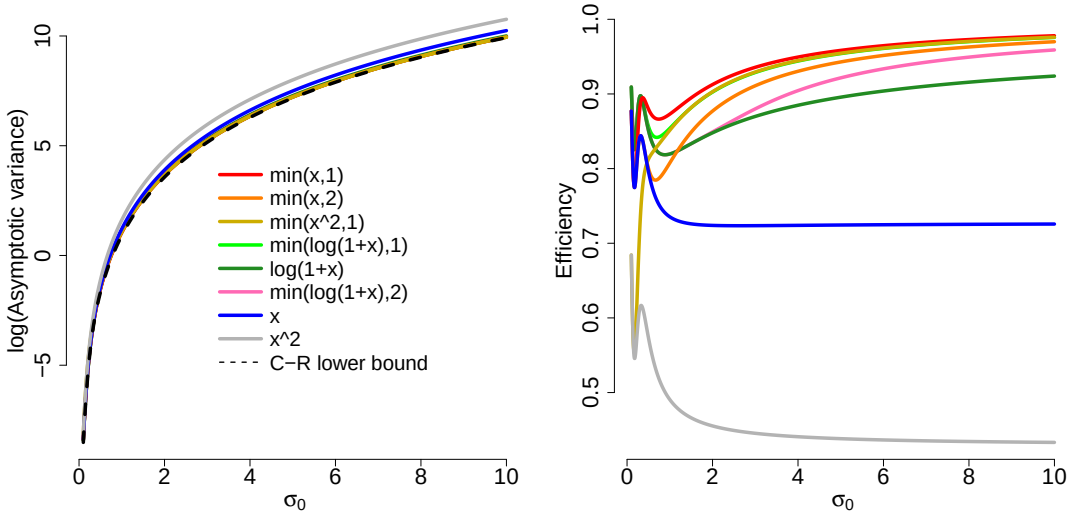


Figure 2.2: Log of asymptotic variance and efficiency with respect to the Cramér-Rao bound for $\hat{\sigma}_h^2$ ($\mu = 0.5$ known).

Figure 2.2 is the analog of Figure 2.1 for $\hat{\sigma}_h^2$ from Example 3.2 with $\mu = 0.5$ known. While the specifics are a bit different the benefits of using bounded or slowly growing h are again clear. We note that when σ is small, the effect of truncation to the positive part of the real line is small.

In both plots we order/color the curves based on their overall efficiency, so they have different colors in one from the other, although the same functions are presented. For all functions presented here (A1)–(A2) and (C1)–(C2) are satisfied.

2.4 Regularized Generalized Score Matching

In high-dimensional settings, when the number r of parameters to estimate may be larger than the sample size n, it is hard, if not impossible, to estimate the parameters consistently without turning to some form of regularization. More specifically, for exponential families, condition (C1) in Section 2.3 fails when $r > n$. A popular approach is then the use of ℓ_1 regularization to exploit possible sparsity.

Let the data matrix $\mathbf{x} \in \mathbb{R}^{n\times m}$ comprise n i.i.d. samples from distribution P_0. Assume P_0 has density p_0 belonging to an exponential family $\mathcal{P}_+ \equiv \{p_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$, where $\boldsymbol{\Theta} \subseteq \mathbb{R}^r$. Adding an ℓ_1 penalty to (2.8), we obtain the regularized generalized score matching loss

$$\frac{1}{2}\boldsymbol{\theta}^\top \boldsymbol{\Gamma}(\mathbf{x})\boldsymbol{\theta} - \boldsymbol{g}(\mathbf{x})^\top \boldsymbol{\theta} + \lambda\|\boldsymbol{\theta}\|_1 \tag{2.12}$$

as in Lin et al. (2016). The loss in (2.12) involves a quadratic smooth part as in the familiar lasso loss for linear regression. However, although the matrix $\boldsymbol{\Gamma}$ is positive semidefinite, the regularized loss in (2.12) is not guaranteed to be bounded unless the tuning parameter λ is sufficiently large—a problem that does not occur in lasso. We note that here, and throughout, we suppress the dependence on the data $\mathbf{x}$ for $\boldsymbol{\Gamma}(\mathbf{x})$, $\boldsymbol{g}(\mathbf{x})$ and derived quantities.

For a more detailed explanation, note that by (2.9), $\boldsymbol{\Gamma} = \mathbf{H}^\top\mathbf{H}$ for some $\mathbf{H} \in \mathbb{R}^{nm\times r}$. In the high-dimensional case, the rank of $\boldsymbol{\Gamma}$, or equivalently that of $\mathbf{H}$, is at most $nm < r$. Hence, $\boldsymbol{\Gamma}$ is not invertible and $\boldsymbol{g}$ does not necessarily lie in the column span of $\boldsymbol{\Gamma}$. Let $\mathrm{Ker}(\boldsymbol{\Gamma})$

be the kernel of $\boldsymbol{\Gamma}$. Then there may exist $\boldsymbol{\nu} \in \mathrm{Ker}(\boldsymbol{\Gamma})$ with $\boldsymbol{g}^\top \boldsymbol{\nu} \neq 0$. In this case, if

$$0 \leq \lambda < \sup_{\boldsymbol{\nu} \in \mathrm{Ker}(\boldsymbol{\Gamma})} |\boldsymbol{g}^\top \boldsymbol{\nu}| / \|\boldsymbol{\nu}\|_1,$$

there exists $\boldsymbol{\nu} \in \mathrm{Ker}(\boldsymbol{\Gamma})$ with $\frac{1}{2}\boldsymbol{\nu}^\top \boldsymbol{\Gamma} \boldsymbol{\nu} = 0$ and $-\boldsymbol{g}^\top \boldsymbol{\nu} + \lambda \|\boldsymbol{\nu}\|_1 < 0$. Evaluating at $\boldsymbol{\theta}(a) = a \cdot \boldsymbol{\nu}$ for scalar $a > 0$, the loss becomes $a\left(-\boldsymbol{g}^\top \boldsymbol{\nu} + \lambda \|\boldsymbol{\nu}\|_1\right)$, which is negative and linear in a, and thus unbounded below. In this case no minimizer of (2.12) exists for small values of λ. This issue also exists for the estimators from Zhang and Zou (2014) and Liu and Luo (2015), which correspond to score matching for GGMs. We note that in the context of estimating the interaction matrix in pairwise models, $r = m^2$; thus, the condition $nm < r$ reduces to $n < m$, or $n < m + 1$ when both $\mathbf{K}$ and $\boldsymbol{\eta}$ in (2.14) below are estimated.

To circumvent the unboundedness problem, we add small values $\gamma_\ell > 0$ to the diagonal entries of $\boldsymbol{\Gamma}$, which become $\boldsymbol{\Gamma}_{\ell,\ell} + \gamma_\ell$, $\ell = 1, \ldots, r$. This is in the spirit of work such as Ledoit and Wolf (2004) and corresponds to an elastic net-type penalty (Zou and Hastie, 2005) with weighted ℓ_2 penalty $\sum_{\ell=1}^r \gamma_\ell \theta_\ell^2$. After this modification, $\boldsymbol{\Gamma}$ is positive definite, our regularized loss is strongly convex in $\boldsymbol{\theta}$, and a unique minimizer exists for all $\lambda \geq 0$. For the special case of truncated GGMs, we will show that a result on consistent estimation holds if we choose $\gamma_\ell = \delta_0 \boldsymbol{\Gamma}_{\ell,\ell}$ for a suitably small constant $\delta_0 > 0$, for which we propose a particular choice to avoid tuning. This choice of γ_ℓ depends on the data through $\boldsymbol{\Gamma}_{\ell,\ell}$.

Definition 2. *For* $\boldsymbol{\gamma} \in \mathbb{R}_+^r \backslash \{\mathbf{0}\}$, *let* $\boldsymbol{\Gamma}_{\boldsymbol{\gamma}} \equiv \boldsymbol{\Gamma} + \mathrm{diag}(\boldsymbol{\gamma})$. *The* regularized generalized $\boldsymbol{h}$-score matching estimator *with tuning parameter* $\lambda \geq 0$ *and* amplifier $\boldsymbol{\gamma}$ *is the estimator*

$$\hat{\boldsymbol{\theta}} \in \operatorname*{argmin}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \hat{J}_{\boldsymbol{h},\lambda,\boldsymbol{\gamma}}(\boldsymbol{\theta}) \equiv \operatorname*{argmin}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \frac{1}{2}\boldsymbol{\theta}^\top \boldsymbol{\Gamma}_{\boldsymbol{\gamma}}(\mathbf{x})\boldsymbol{\theta} - \boldsymbol{g}(\mathbf{x})^\top \boldsymbol{\theta} + \lambda \|\boldsymbol{\theta}\|_1. \tag{2.13}$$

In the case where $\boldsymbol{\gamma} = (\delta - 1)\mathrm{diag}(\boldsymbol{\Gamma})$ for some $\delta > 1$, we also call δ the *multiplier*. We note that $\hat{\boldsymbol{\theta}}$ from (2.13) is a *piecewise linear* function of λ (Lin et al., 2016).

2.5 Score Matching for Graphical Models for Non-negative Data

In this section we apply our generalized score matching estimator to a general class of graphical models for non-negative data.

2.5.1 A General Framework of Pairwise Interaction Models

We consider the class of pairwise interaction power models with density introduced in (1.1). We recall the form of the density:

$$p_{\boldsymbol{\eta},\mathbf{K}}(\boldsymbol{x}) \propto \exp\left(-\frac{1}{2a}\boldsymbol{x}^{a\top}\mathbf{K}\boldsymbol{x}^{a} + \boldsymbol{\eta}^{\top}\frac{\boldsymbol{x}^{b}-\mathbf{1}_m}{b}\right)\mathbb{1}_{\mathbb{R}^m_+}(\boldsymbol{x}), \tag{2.14}$$

where a and b are known constants, and the interaction matrix $\mathbf{K}$ and the vector $\boldsymbol{\eta}$ are parameters. When $b = 0$, we use the convention that $\frac{x^0-1}{0} \equiv \log x$ and apply the logarithm element-wise. Our focus will be on the interaction matrix $\mathbf{K}$ that determines the conditional independence graph through its support $S(\mathbf{K}) \equiv \{(i,j) : \kappa_{ij} \neq 0\}$. However, unless $\boldsymbol{\eta}$ is known or assumed to be zero, we also need to estimate $\boldsymbol{\eta}$ as a nuisance parameter. In the case where we assume $\boldsymbol{\eta} \equiv \mathbf{0}$ is known (i.e. the linear part $(\boldsymbol{x}^b - \mathbf{1}_m)/b$ is not present), we call the distribution (and the corresponding estimator) a *centered* distribution (estimator), in contrast to the general case termed *non-centered* when we assume $\boldsymbol{\eta} \neq \mathbf{0}$ or unknown.

We first give a set of sufficient conditions for the density to be valid, i.e., the right-hand side of (2.14) to be integrable. The proof is given in Appendix A.1.3.

Theorem 5. *Define conditions*

(CC1) $\mathbf{K}$ *is* strictly co-positive*, i.e.,* $\boldsymbol{v}^{\top}\mathbf{K}\boldsymbol{v} > 0$ *for all* $\boldsymbol{v} \in \mathbb{R}^m_+\backslash\{\mathbf{0}\}$*;*

(CC2) $2a > b > 0$*;*

(CC3) $a > 0$*,* $b = 0$*, and* $\eta_j > -1$ *for* $j = 1, \ldots, m$ *(*$\boldsymbol{\eta} \succ -\mathbf{1}_m$*).*

In the non-centered case, if (CC1) and one of (CC2) and (CC3) holds, then the function on the right-hand side of (2.14) is integrable over $\mathbb{R}^m_+$*. In the centered case, (CC1) and* $a > 0$ *are sufficient.*

We emphasize that (CC1) is a weaker condition than positive definiteness. Criteria for strict co-positivity are discussed in Väliaho (1986).

2.5.2 Implementation for Different Models

In this section we give some implementation details for the regularized generalized $\boldsymbol{h}$-score matching estimator defined in (2.13) applied to the pairwise interaction models from (2.14). We let $\boldsymbol{\Psi} \equiv \left(\mathbf{K}^\top, \boldsymbol{\eta}\right)^\top \in \mathbb{R}^{(m+1)\times m}$. The unregularized loss is then

$$\hat{J}_{\boldsymbol{h}}(P) = \frac{1}{2}\mathrm{vec}(\boldsymbol{\Psi})^\top \boldsymbol{\Gamma}(\boldsymbol{x})\mathrm{vec}(\boldsymbol{\Psi}) - \boldsymbol{g}(\boldsymbol{x})^\top \mathrm{vec}(\boldsymbol{\Psi}).$$

The general form of the matrix $\boldsymbol{\Gamma}$ and the vector $\boldsymbol{g}$ in the loss were given in equations (2.8)–(2.10). Here $\boldsymbol{\Gamma} \in \mathbb{R}^{(m+1)m\times(m+1)m}$ is block-diagonal, with the j-th $\mathbb{R}^{(m+1)\times(m+1)}$ block

$$\begin{aligned}
\boldsymbol{\Gamma}_j(\mathbf{x}) &\equiv \begin{bmatrix} \boldsymbol{\Gamma}_{11,j} & \boldsymbol{\Gamma}_{12,j} \\ \boldsymbol{\Gamma}_{12,j}^\top & \Gamma_{22,j} \end{bmatrix} \qquad (2.15)\\
&\equiv \frac{1}{n}\sum_{i=1}^n \begin{bmatrix} h_j\left(X_j^{(i)}\right) {X_j^{(i)}}^{2a-2} \boldsymbol{X}^{(i)a} \boldsymbol{X}^{(i)a\top} & -h_j\left(X_j^{(i)}\right) {X_j^{(i)}}^{a+b-2} \boldsymbol{X}^{(i)a} \\ -h_j\left(X_j^{(i)}\right) {X_j^{(i)}}^{a+b-2} \boldsymbol{X}^{(i)a\top} & h_j\left(X_j^{(i)}\right) {X_j^{(i)}}^{2b-2} \end{bmatrix} \\
&= \frac{1}{n}\mathbf{y}^\top\mathbf{y}, \quad \mathbf{y} \equiv \left[-\left(\sqrt{\boldsymbol{h}_j(\boldsymbol{X}_j)} \odot \boldsymbol{X}_j^{a-1}\right) \odot \mathbf{x}^a \quad \sqrt{\boldsymbol{h}_j(\boldsymbol{X}_j)} \odot \boldsymbol{X}_j^{b-1}\right] \in \mathbb{R}^{n,m+1}, \quad (2.16)
\end{aligned}$$

where the $\odot$ product between a vector and a matrix means an elementwise multiplication of the vector with each *column* of the matrix, and $\boldsymbol{h}_j(\boldsymbol{X}_j) \equiv \left[h_j(X_j^{(1)}), \ldots, h_j(X_j^{(n)})^\top\right] \in \mathbb{R}^m$. Furthermore, $\boldsymbol{g} \equiv \begin{bmatrix} \mathrm{vec}(\mathbf{g}_1) \\ \boldsymbol{g}_2 \end{bmatrix} \in \mathbb{R}^{(m+1)m}$, where $\mathbf{g}_1$ and $\boldsymbol{g}_2$ correspond to each entry of $\mathbf{K}$ and $\boldsymbol{\eta}$, respectively. The j-th column of $\mathbf{g}_1 \in \mathbb{R}^{m\times m}$, written as $\boldsymbol{g}_{1,j}(\mathbf{x})$, is

$$\frac{1}{n}\sum_{i=1}^n \left(h_j'\left(X_j^{(i)}\right) {X_j^{(i)}}^{a-1} + (a-1)h_j\left(X_j^{(i)}\right) {X_j^{(i)}}^{a-2}\right)\boldsymbol{X}^{(i)a} + a h_j\left(X_j^{(i)}\right) {X_j^{(i)}}^{2a-2} \boldsymbol{e}_{j,m},$$

where $\boldsymbol{e}_{j,m}$ is the m-vector with 1 at the j-th position and 0 elsewhere, and the j-th entry of $\boldsymbol{g}_2 \in \mathbb{R}^m$ is

$$g_{2,j} = \frac{1}{n}\sum_{i=1}^n -h_j'\left(X_j^{(i)}\right) {X_j^{(i)}}^{b-1} - (b-1)h_j\left(X_j^{(i)}\right) {X_j^{(i)}}^{b-2}.$$

These formulae also hold for $b = 0$ since $\boldsymbol{\Gamma}$ and $\boldsymbol{g}$ only depend on the gradient of the log density, and $\frac{\mathrm{d}(x^b-1)/b}{\mathrm{d}x} = x^{b-1}$ also holds for $b = 0$. In the centered case where we know

$\boldsymbol{\eta}_0 \equiv \mathbf{0}$, we only estimate $\mathbf{K} \in \mathbb{R}^{m\times m}$, and $\boldsymbol{\Gamma} \in \mathbb{R}^{m^2\times m^2}$ is still block-diagonal, with the j-th block being the $\boldsymbol{\Gamma}_{11,j}$ submatrix in (2.15), while $\boldsymbol{g}$ is just $\mathrm{vec}(\mathbf{g}_1)$. Since b only appears in the $\boldsymbol{\eta}$ part of the density, the formulae only depend on a in the centered case.

We emphasize that it is indeed necessary to introduce amplifiers $\boldsymbol{\gamma} \succ \mathbf{0}$ or a multiplier $\delta > 1$ in addition to the ℓ_1 penalty. It is clear from (2.16) that $\mathrm{rank}(\boldsymbol{\Gamma}_j) \le \min\{n, m+1\}$ (or $\min\{n, m\}$ if centered). Thus, $\boldsymbol{\Gamma}$ is non-invertible when $n \le m$ (or $n < m$ if centered) and $\boldsymbol{g}$ need not lie in its column span.

We claim that including amplifiers/multipliers for the submatrices $\boldsymbol{\Gamma}_{11,j}$ only is sufficient for unique existence of a solution for all penalty parameters $\lambda \ge 0$. To see this, consider any nonzero vector $\boldsymbol{\nu} \in \mathbb{R}^{m+1}$. Partition it as $\boldsymbol{\nu} \equiv (\boldsymbol{\nu}_1, \boldsymbol{\nu}_2)$ with $\boldsymbol{\nu}_1 \in \mathbb{R}^m$. Let $\boldsymbol{\Gamma}_{j,\boldsymbol{\gamma}}$ be our amplified version of the matrix $\boldsymbol{\Gamma}_j$ from (2.19), so

$$\boldsymbol{\Gamma}_{j,\boldsymbol{\gamma}} \;=\; \begin{pmatrix} \boldsymbol{\Gamma}_{11,j} + \mathrm{diag}(\gamma_1, \ldots, \gamma_m) & \boldsymbol{\Gamma}_{12,j} \\ \boldsymbol{\Gamma}_{12,j}^\top & \Gamma_{22,j} \end{pmatrix}.$$

As $\boldsymbol{\Gamma}_j$ itself is positive semidefinite, we find that if at least one of the first m entries of $\boldsymbol{\nu}$ is nonzero then

$$\boldsymbol{\nu}^\top \boldsymbol{\Gamma}_{j,\boldsymbol{\gamma}} \boldsymbol{\nu} \;\ge\; \boldsymbol{\nu}^\top \boldsymbol{\Gamma}_j \boldsymbol{\nu} + \sum_{k=1}^{m} \nu_k^2 \boldsymbol{\gamma}_k \;\ge\; \sum_{k=1}^{m} \nu_k^2 \boldsymbol{\gamma}_k \;>\; 0.$$

If only the last entry of $\boldsymbol{\nu}$ is nonzero then

$$\boldsymbol{\nu}^\top \Gamma_{j,\boldsymbol{\gamma}} \boldsymbol{\nu} \;=\; \nu_{m+1}^2 \Gamma_{22,j} \;>\; 0$$

almost surely; recall that $\Gamma_{22,j} = \frac{1}{n}\sum_{i=1}^n h_j\left(X_j^{(i)}\right) X_j^{2b-2}$. We conclude that $\boldsymbol{\Gamma}_{j,\boldsymbol{\gamma}}$ (and thus the entire amplified $\boldsymbol{\Gamma}$) is a.s. positive definite, which ensures unique existence of the loss minimizer.

Given the formulae for $\boldsymbol{\Gamma}$ and $\boldsymbol{g}$, one adds the ℓ_1 penalty on $\boldsymbol{\Psi}$ to get the regularized loss (2.22). Our methodology readily accommodates two different choices of the penalty parameter λ for $\mathbf{K}$ and $\boldsymbol{\eta}$. This is also theoretically supported for truncated GGMs, since if the ratio of the respective values $\lambda_{\mathbf{K}}$ and $\lambda_{\boldsymbol{\eta}}$ is fixed, the proof of the theorems in Section 2.6 can be easily modified by replacing $\boldsymbol{\eta}$ by $(\lambda_{\boldsymbol{\eta}}/\lambda_{\mathbf{K}})\boldsymbol{\eta}$. To avoid picking two tuning parameters,

one may also choose to remove the penalty on $\boldsymbol{\eta}$ altogether by profiling out $\boldsymbol{\eta}$ and solve for $\hat{\boldsymbol{\eta}} \equiv \boldsymbol{\Gamma}_{22}^{-1}\left(\boldsymbol{g}_2 - \boldsymbol{\Gamma}_{12}^\top \mathrm{vec}(\hat{\mathbf{K}})\right)$, with $\hat{\mathbf{K}}$ the minimizer of the profiled loss

$$\hat{J}_{\boldsymbol{h},\lambda,\boldsymbol{\gamma},\mathrm{profile}}(\mathbf{K}) \equiv \frac{1}{2}\mathrm{vec}(\mathbf{K})^\top \boldsymbol{\Gamma}_{\boldsymbol{\gamma},11.2}\mathrm{vec}(\mathbf{K}) - (\boldsymbol{g}_1 - \boldsymbol{\Gamma}_{12}\boldsymbol{\Gamma}_{22}^{-1}\boldsymbol{g}_2)^\top \mathrm{vec}(\mathbf{K}) + \lambda\|\mathbf{K}\|_1, \tag{2.17}$$

where the Schur complement $\boldsymbol{\Gamma}_{\boldsymbol{\gamma},11.2} \equiv \boldsymbol{\Gamma}_{\boldsymbol{\gamma},11} - \boldsymbol{\Gamma}_{12}\boldsymbol{\Gamma}_{22}^{-1}\boldsymbol{\Gamma}_{12}^\top$ is a.s. positive definite such that the profiled estimator exists a.s. for all $\lambda \geq 0$. This profiled approach corresponds to choosing $\lambda_{\boldsymbol{\eta}}/\lambda_{\mathbf{K}} = 0$. A detailed theoretical analysis of the profiled estimator is beyond the scope of this chapter, however. We note that in the other extreme, with $\lambda_{\boldsymbol{\eta}}/\lambda_{\mathbf{K}} = +\infty$, the non-centered estimator reduces to the estimator from the centered case.

Example 5.3. *The truncated normal model comprises the density*

$$p_{\boldsymbol{\mu},\mathbf{K}}(\boldsymbol{x}) \;\propto\; \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \mathbf{K}(\boldsymbol{x}-\boldsymbol{\mu})\right\} \mathbb{1}_{[0,\infty)^m}(\boldsymbol{x}). \tag{2.18}$$

This corresponds to (2.14) with $a = b = 1$, and $\boldsymbol{\eta} = \mathbf{K}\boldsymbol{\mu}$. The j-th $(m+1)\times(m+1)$ block of $\boldsymbol{\Gamma}(\mathbf{x})$ is

$$\frac{1}{n}\begin{bmatrix} \mathbf{x}^\top \mathrm{diag}(\boldsymbol{h}_j(\boldsymbol{X}_j))\mathbf{x} & -\mathbf{x}^\top \boldsymbol{h}_j(\boldsymbol{X}_j) \\ -\boldsymbol{h}_j(\boldsymbol{X}_j)^\top \mathbf{x} & \boldsymbol{h}_j(\boldsymbol{X}_j)^\top \mathbf{1}_n \end{bmatrix}. \tag{2.19}$$

Partitioning the vector $\boldsymbol{g}(\mathbf{x})$ into m subvectors $\boldsymbol{g}_j(\mathbf{x}) \in \mathbb{R}^{m+1}$, where the entries of $\boldsymbol{g}_j(\mathbf{x})$ correspond to column $\boldsymbol{\Psi}_j$, the k-th entry of $\boldsymbol{g}_j(\mathbf{x})$ is

$$g_{jk}(\mathbf{x}) \equiv \begin{cases} \frac{1}{n}\sum_{i=1}^n h_j'\left(X_j^{(i)}\right) X_k^{(i)} & \text{if } k \leq m,\ k \neq j, \\ \frac{1}{n}\sum_{i=1}^n h_j'\left(X_j^{(i)}\right) X_k^{(i)} + h_j\left(X_j^{(i)}\right) & \text{if } k = j, \\ -\frac{1}{n}\sum_{i=1}^n h_j'\left(X_j^{(i)}\right) & \text{if } k = m+1. \end{cases} \tag{2.20}$$

Example 5.4. *The exponential square-root graphical model in Inouye et al. (2016) has*

$$p_{\boldsymbol{\eta},\mathbf{K}}(\boldsymbol{x}) \propto \exp\left(-\sqrt{\boldsymbol{x}}^\top \mathbf{K}\sqrt{\boldsymbol{x}} + 2\boldsymbol{\eta}^\top\sqrt{\boldsymbol{x}}\right)\mathbb{1}_{[0,\infty)^m}(\boldsymbol{x}),$$

which corresponds to (2.14) with $a = b = 1/2$. We refer to this as the exponential *model. In this case, the j-th $\mathbb{R}^{(m+1)\times(m+1)}$ block of $\boldsymbol{\Gamma}$ is*

$$\boldsymbol{\Gamma}_j(\mathbf{x}) \equiv \frac{1}{n}\sum_{i=1}^n \frac{h_j\left(X_j^{(i)}\right)}{X_j^{(i)}} \begin{pmatrix} -\sqrt{\boldsymbol{X}^{(i)}} \\ 1 \end{pmatrix}\left(-\sqrt{\boldsymbol{X}^{(i)}}^\top, 1\right)$$

and $\boldsymbol{g} = \mathrm{vec}(\mathbf{g}_0)$*, where the* j*-th column of* $\mathbf{g}_0 \in \mathbb{R}^{(m+1)\times m}$ *is*

$$\boldsymbol{g}_j(\mathbf{x}) \equiv \frac{1}{n}\sum_{i=1}^{n} \frac{2h_j'\left(X_j^{(i)}\right)X_j^{(i)} - h_j\left(X_j^{(i)}\right)}{2{X_j^{(i)}}^{3/2}} \begin{pmatrix} \sqrt{\boldsymbol{X}^{(i)}} \\ -1 \end{pmatrix} + \frac{h_j\left(X_j^{(i)}\right)}{2X_j^{(i)}} \boldsymbol{e}_{j,m+1}.$$

Example 5.5. *If* $a = 1/2$ *and* $b = 0$*, then (2.14) becomes*

$$p_{\boldsymbol{\eta},\mathbf{K}}(\boldsymbol{x}) \propto \exp\left(-\sqrt{\boldsymbol{x}}^{\top}\mathbf{K}\sqrt{\boldsymbol{x}} + \boldsymbol{\eta}^{\top}\log(\boldsymbol{x})\right)\mathbb{1}_{(0,\infty)^m}(\boldsymbol{x}). \tag{2.21}$$

If $\mathbf{K}$ *is diagonal in this case, then* $\boldsymbol{X} \sim p_{\boldsymbol{\eta},\mathbf{K}}$ *has independent entries with* X_j *following the gamma distribution with rate* κ_{jj} *and shape* $\eta_j + 1$*, which gives an intuition for condition (CC3)* $\eta_j > -1$ *in Theorem 5. We can thus view (2.21) as a multivariate gamma distribution with pairwise interactions among the covariates, and call this the* gamma *model. For this model, the* j*-th block of* $\boldsymbol{\Gamma}$ *is*

$$\boldsymbol{\Gamma}_j(\mathbf{x}) \equiv \frac{1}{n}\sum_{i=1}^{n} \frac{h_j\left(X_j^{(i)}\right)}{{X_j^{(i)}}^2} \begin{pmatrix} -\sqrt{X_j^{(i)}\boldsymbol{X}^{(i)}} \\ 1 \end{pmatrix} \left(-\sqrt{X_j^{(i)}\boldsymbol{X}^{(i)}}^{\top}, 1\right)$$

and the part of $\boldsymbol{g}$ *corresponding to* $\mathbf{K}_j$ *is*

$$\boldsymbol{g}_{1,j}(\mathbf{x}) \equiv \frac{1}{n}\sum_{i=1}^{n} \frac{2h_j'\left(X_j^{(i)}\right)X_j^{(i)} - h_j\left(X_j^{(i)}\right)}{2{X_j^{(i)}}^{3/2}} \sqrt{\boldsymbol{X}^{(i)}} + \frac{h_j\left(X_j^{(i)}\right)}{2X_j^{(i)}} \boldsymbol{e}_{j,m},$$

while the part for η_j *is*

$$g_{2.j}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \frac{h_j\left(X_j^{(i)}\right)}{{X_j^{(i)}}^2} - \frac{h_j'\left(X_j^{(i)}\right)}{X_j^{(i)}}.$$

We note that the $\boldsymbol{\Gamma}_{11,j}$ sub-matrix of $\boldsymbol{\Gamma}_j$ and the $\boldsymbol{g}_{1,j}$ sub-vector of $\boldsymbol{g}_j$ for the gamma model are the same as those for the exponential model, since $a = 1/2$ in both cases and the parts involving $\mathbf{K}$ in the densities are the same.

2.5.3 Computational Details

In the most general exponential family setting, as in Eq. (2.8)–(2.10) in Theorem 3, the time complexity for forming $\boldsymbol{\Gamma} \in \mathbb{R}^{r\times r}$ and $\boldsymbol{g} \in \mathbb{R}^r$ is $\mathcal{O}\left(nm(f_{b'}(m) + r^2 + r(f_{t'}(m) + f_{t''}(m)))\right)$.

Here $f_{b'}(m)$ is the average time complexity for calculating $\partial_j b(\boldsymbol{x})$ over $j = 1, \ldots, m$, and similarly $f_{t'}(m)$ for $\partial_j t_\ell(\boldsymbol{x})$ and $f_{t''}(m)$ for $\partial_{jj} t_\ell(\boldsymbol{x})$ over $j = 1, \ldots, m$ and $\ell = 1 \ldots, r$. In many applications, however, these three functions would be constant in m, thus giving an $\mathcal{O}(nmr^2)$ computational complexity, with the dominating term coming from the operations for $\boldsymbol{t}'_j \boldsymbol{t}'^{\top}_j$ in $\boldsymbol{\Gamma}$ since $\boldsymbol{\Gamma}$ is of dimension $r \times r$.

For pairwise interaction power models, $r = m^2$ and the formula above becomes $\mathcal{O}(nm^5)$. However, since $\boldsymbol{\Gamma}$ is block-diagonal with only m^3 nonzero entries and by the special form of $\boldsymbol{t}(\boldsymbol{x}) = \boldsymbol{x}^a \boldsymbol{x}^{a\top}$, the true complexity is in fact $\mathcal{O}(nm^3)$.

While the introduction of the ℓ_1 penalty inevitably precludes the estimator from having a closed-form solution and introduces non-differentiability, state-of-art numerical optimization algorithms, such as coordinate-descent (Friedman et al., 2007), can be applied for fast estimation. To speed up estimation, one can usually use warm starts using the solution from the previous λ's, as well as lasso-type strong screening rules (Tibshirani et al., 2012) to eliminate components of $\hat{\boldsymbol{\theta}}$ that are known a priori to have zero estimates.

In our implementation for pairwise interaction models of Section 2.5.1 (available in the `genscore` R package), we optimize our loss functions with respect to a symmetric matrix $\hat{\mathbf{K}}$; in the non-centered case the vector $\hat{\boldsymbol{\eta}}$ is also included. We use a coordinate-descent method analogous to Algorithm 2 in Lin et al. (2016), where in each step we update each element of $\hat{\mathbf{K}}$ and $\hat{\boldsymbol{\eta}}$ based on the other entries from the previous steps, while maintaining symmetry. In our simulations in Section 2.7 we always scale the data matrix by column ℓ_2 norms before proceeding to estimation. Note that estimation of $\hat{\mathbf{K}}$ without symmetry can be parallelized as the loss can be decomposed into a sum over the columns.

2.5.4 Choice of the Function $\boldsymbol{h}$

In this subsection we discuss the requirements on the function $\boldsymbol{h}$ as well as some reasonable choices of $\boldsymbol{h}$.

Requirements on $\boldsymbol{h}$

In Section 2.2.2, we presented two assumptions (A1) and (A2) under which the generalized score-matching loss is valid, i.e., the integration by parts is justified and Theorem 2 holds. In this section, we present some sufficient (and nearly necessary) requirements on $\boldsymbol{h}$ such that (A1) and (A2) are satisfied.

Definition 3. *Suppose* $\boldsymbol{h} : \mathbb{R}_+^m \to \mathbb{R}_+^m$ *with* $\boldsymbol{h}(\boldsymbol{x}) = (h_1(x_1), \ldots, h_m(x_m))^\top$. *We write that* $\boldsymbol{h} \in \mathcal{H}_{a,b}$ *(for simplicity we omit the dependency on* m*) if for all* $j = 1, \ldots, m$*:*

i) h_j *is absolutely continuous in every bounded sub-interval of* $\mathbb{R}_+$*, and thus has derivative* h_j' *a.s.;*

ii) $h_j(x) > 0$ *a.s. on* $\mathbb{R}_+$*;*

iii) h_j *and* h_j' *are both bounded by some piecewise powers of* x *a.s. on* $\mathbb{R}_+$*;*

iv) $\lim_{x \searrow 0+} h_j(x)/x_j^q = 0$*, where* $q \equiv \begin{cases} \max\{1-a, 1-b\} & \text{if } b > 0, \\ 1 - \eta_{0,j} & \text{if } b = 0. \end{cases}$

Theorem 6. *Assume every* P *in the family of distribution* $\mathcal{P}_+$ *satisfies (CC1)–(CC3) and thus has finite normalizing constants. If* $\boldsymbol{h} \in \mathcal{H}_{a,b}$*, then (A1) and (A2) are satisfied.*

In centered models, where $\boldsymbol{\eta} \equiv \mathbf{0}$, we can assume $b = 2a$ and iv) in the definition of $\mathcal{H}_{a,2a}$ has $q = 1 - a$. For truncated GGMs, $a = b = 1$, so iv) in Definition 3 is simply $\lim_{x_j \searrow 0^+} h_j(x_j) = 0$.

In the case of $b = 0$, $\boldsymbol{\eta}$ is an unknown parameter, and (CC3) requires each of its component to be greater than -1. If one has prior information on $\boldsymbol{\eta}$ or restricts the parameter space for $\boldsymbol{\eta}$, the requirement reduces to $h_j(x_j) = o(x_j^{1-\eta_{0,j}})$ as $x_j \searrow 0^+$. Otherwise, it suffices to require $h_j(x_j) = o(x_j^2)$. Note that this is only a condition for $x_j \searrow 0^+$, and the globally quadratic behavior of $h_j(x_j) = x_j^2$ from the original score matching is not needed on the entire $\mathbb{R}_+$, leaving opportunities for improvements.

Reasonable Choices of $\boldsymbol{h}$

Assume a common univariate h for all components in $\boldsymbol{h}$. Inspired by Theorem 6, we consider h that behaves like a power of x both as $x \nearrow +\infty$ and as $x \searrow 0^+$. Since the requirements on the two tails are separate, we can choose h to be a piecewise defined function that joins two powers with possibly different degrees. In other words, $h(x) = \min(x^{p_1}, cx^{p_2})$ for some powers $p_1 \geq p_2 \geq 0$ and constant $c > 0$. Only one constant c is required since generalized score matching is invariant to scaling of h. In determining the exact power of p_1 we have the following considerations:

a) In the centered case:

   (i) (A1) and (A2): Theorem 6 requires that $p_1 \geq 1 - a$.

   (ii) "Controlled $\boldsymbol{\Gamma}$ and $\boldsymbol{g}$ for $\boldsymbol{x}^a$": We propose avoiding poles at the origin for the entries of $\boldsymbol{\Gamma}$ and $\boldsymbol{g}$. The formula for $\boldsymbol{\Gamma}_{11}$ in (2.16) shows that to this end $\sqrt{h(x)}x^{a-1}$ needs to have a non-negative degree. This requires $p_1 \geq 2 - 2a$. The formula for $\boldsymbol{g}_1$ similarly shows that $h'(x)x^{a-1}$, $h(x)x^{a-2}$ and $h(x)x^{2a-2}$ all need to have a non-negative degree for small x. This requires $p_1 \geq 2 - a$.

b) In the non-centered case, in addition to (i) and (ii),

   (iii) (A1) and (A2): Theorem 6 requires $p_1 \geq \max\{1-a, 1-b\}$ for $b > 0$, or $1-\min_j \eta_{0,j}$ for $b = 0$.

   (iv) "Controlled $\boldsymbol{\Gamma}$ and $\boldsymbol{g}$ for $\boldsymbol{x}^b$": From the definition of $\boldsymbol{\Gamma}_{22}$ and $\boldsymbol{g}_2$ and by the same reasoning as above, $\sqrt{h(x)}x^{b-1}$, $h'(x)x^{b-1}$ and $h(x)x^{b-2}$ need to be non-negative powers of x, thus requiring $p_1 \geq \max\{2 - b, 2 - 2b\} = 2 - b$.

The choice of p_2, is only relevant for large data points. Our main consideration is then merely how well $\boldsymbol{\Gamma}$ and $\boldsymbol{g}$ concentrate on their true population values (Theorem 7). From this perspective, our intuition is that p_2 should be chosen small so that the tails of the

distributions of the entries of $\mathbf{\Gamma}$ and $\boldsymbol{g}$ are well-behaved. Thus, we can choose $p_2 = 0$, in which case $h(x) = \min(x^{p_1}, c)$ is a truncated power.

2.5.5 Tuning Parameter Selection

By treating the unpenalized loss (i.e., $\lambda = 0$, $\boldsymbol{\gamma} = 0$) as a negative log-likelihood, we may use the extended Bayesian Information Criterion (eBIC) to choose the tuning parameter (Chen and Chen, 2008; Foygel and Drton, 2010). Consider the centered case as an example. Let $\hat{S}^\lambda \equiv \{(i,j) : \hat{\kappa}_{ij}^\lambda \neq 0, i < j\}$, where $\hat{\mathbf{K}}^\lambda$ be the estimate associated with tuning parameter λ. The eBIC is then

$$\text{eBIC}(\lambda) = -n\text{vec}(\hat{\mathbf{K}})^\top \mathbf{\Gamma}(\mathbf{x})\text{vec}(\hat{\mathbf{K}}) + 2n\boldsymbol{g}(\mathbf{x})^\top \text{vec}(\hat{\mathbf{K}}) + |\hat{S}^\lambda| \log n + 2\log \binom{p(p-1)/2}{|\hat{S}^\lambda|},$$

where $\hat{\mathbf{K}}$ can be either the original estimate associated with λ, or a refitted solution obtained by restricting the support to $\hat{S}^\lambda$.

We use the eBIC instead of the ordinary BIC (Bayesian Information Criterion) since the BIC tends to choose an overly complex model when the model space is large, as encountered in the high-dimensional setting. The extension in eBIC comes from the last term in the above display which can be motivated by a prior distribution under which the number of edges in the conditional independence graph is uniformly distributed; see also Żak-Szatkowska and Bogdan (2011) and Barber and Drton (2015).

2.6 Theory for Graphical Models

In our regularized generalized score matching framework, we introduced the amplifiers and the multipliers to address the inexistence problem. We also proposed using a general function $\boldsymbol{h}$ in place of $\boldsymbol{x}^2$ as a means to improve estimation accuracy. This section provides a theoretical analysis of these two aspects.

In Section 2.6.1, we present the theory for our regularized generalized score matching estimators for general pairwise interaction models before going into the details for the special

cases of (truncated) GGMs. Next, we show that a specific choice of amplifiers/multipliers yields consistent estimation without the need for tuning. This point is important even in the case of Gaussian models on all of $\mathbb{R}^m$. Therefore, in Section 2.6.2 we digress from non-negative data and consider the original score matching of Hyvärinen (2005) for centered Gaussian distributions. Finally, in Section 2.6.3, we derive probabilistic results for $\hat{\boldsymbol{\Psi}}$ based on Theorem 7, justifying the benefits of using a general bounded $\boldsymbol{h}$ over $\boldsymbol{x}^2$ in the non-negative setting. As the most important models from the class of pairwise interaction power models over $\mathbb{R}^m_+$, we only treat truncated GGMs since they have the most tractable concentration bounds; this case also provides a comparison to Corollary 2 in Lin et al. (2016), which uses $\boldsymbol{x}^2$.

2.6.1 Theory for Pairwise Interaction Models

The graphical models we treat are parametrized by the interaction matrix $\mathbf{K}$ and the coefficients $\boldsymbol{\eta}$ on $(\boldsymbol{x}^b - \mathbf{1}_m)/b$. It is convenient to accommodate this setting with a matrix-valued parameter $\boldsymbol{\Psi} \in \mathbb{R}^{r_1 \times r_2}$ (in place of $\boldsymbol{\theta}$) and specify our regularized $\boldsymbol{h}$-score matching loss as

$$\hat{J}_{\boldsymbol{h},\lambda,\boldsymbol{\gamma}}(\boldsymbol{\Psi}) \equiv \underset{\boldsymbol{\Psi} \in \mathbb{R}^{r_1 \times r_2}}{\operatorname{argmin}} \frac{1}{2} \operatorname{vec}(\boldsymbol{\Psi})^\top \boldsymbol{\Gamma}_{\boldsymbol{\gamma}}(\mathbf{x}) \operatorname{vec}(\boldsymbol{\Psi}) - \boldsymbol{g}(\mathbf{x})^\top \operatorname{vec}(\boldsymbol{\Psi}) + \lambda \|\boldsymbol{\Psi}\|_1. \tag{2.22}$$

In the non-centered case we thus take $\boldsymbol{\Psi} = [\mathbf{K}, \boldsymbol{\eta}]^\top \in \mathbb{R}^{m(m+1) \times m}$. In the centered case, $\boldsymbol{\Psi}$ is simply the $m \times m$ interaction matrix $\mathbf{K}$. Following related prior work such as Lin et al. (2016), for ease of proof we allow the matrix $\mathbf{K}$ to be nonsymmetric, which allows us to decouple optimization over the different columns of $\mathbf{K}$ or $\boldsymbol{\Psi}$, while in our implementations we ensure that $\mathbf{K}$ is symmetric.

Definition 4. *Let $\boldsymbol{\Gamma}_0 \equiv \mathbb{E}_0 \boldsymbol{\Gamma}(\mathbf{x})$ and $\boldsymbol{g}_0 \equiv \mathbb{E}_0 \boldsymbol{g}(\mathbf{x})$ be the population versions of $\boldsymbol{\Gamma}(\mathbf{x})$ and $\boldsymbol{g}(\mathbf{x})$ under the distribution given by a true parameter matrix $\boldsymbol{\Psi}_0$. The support of a matrix $\boldsymbol{\Psi}$ is $S(\boldsymbol{\Psi}) \equiv \{(i,j) : \psi_{ij} \neq 0\}$, and we let $S_0 = S(\boldsymbol{\Psi}_0)$. For a matrix $\boldsymbol{\Psi}_0$, we define $d_{\boldsymbol{\Psi}_0}$ to be the maximum number of non-zero entries in any column, and $c_{\boldsymbol{\Psi}_0} \equiv |||\boldsymbol{\Psi}_0|||_{\infty,\infty}$. Writing $\boldsymbol{\Gamma}_{0,AB}$ for the $A \times B$ submatrix of $\boldsymbol{\Gamma}_0$, we define*

$$c_{\boldsymbol{\Gamma}_0} \equiv |||(\boldsymbol{\Gamma}_{0,S_0S_0})^{-1}|||_{\infty,\infty}. \tag{2.23}$$

Finally, $\mathbf{\Gamma}_0$ *satisfies* the irrepresentability condition with incoherence parameter $\alpha \in (0, 1]$ and edge set S_0 *if*

$$|||\mathbf{\Gamma}_{0,S_0^c S_0}(\mathbf{\Gamma}_{0,S_0 S_0})^{-1}|||_{\infty,\infty} \leq (1-\alpha). \tag{2.24}$$

Our analysis of the regularized generalized $\boldsymbol{h}$-score matching estimator builds on the following theorem taken from Lin et al. (2016, Theorem 1).

Theorem 7. *Suppose* $\mathbf{\Gamma}_0$ *has* $\mathbf{\Gamma}_{0,S_0 S_0}$ *invertible and satisfies the irrepresentability condition (2.24) with incoherence parameter* $\alpha \in (0, 1]$*. Assume that*

$$\|\mathbf{\Gamma}_{\boldsymbol{\gamma}}(\mathbf{x}) - \mathbf{\Gamma}_0\|_\infty < \epsilon_1, \qquad \|\boldsymbol{g}(\mathbf{x}) - \boldsymbol{g}_0\|_\infty < \epsilon_2, \tag{2.25}$$

with $d_{\mathbf{\Psi}_0}\epsilon_1 \leq \alpha/(6c_{\mathbf{\Gamma}_0})$*. If*

$$\lambda > \frac{3(2-\alpha)}{\alpha} \max\{c_{\mathbf{\Psi}_0}\epsilon_1, \epsilon_2\},$$

then the following holds:

(a) The regularized generalized $\boldsymbol{h}$*-score matching estimator* $\hat{\mathbf{\Psi}}$ *minimizing (2.22) is unique, with support* $\hat{S} \equiv S(\hat{\mathbf{\Psi}}) \subseteq S_0$*, and satisfies*

$$\|\hat{\mathbf{\Psi}} - \mathbf{\Psi}_0\|_\infty \leq \frac{c_{\mathbf{\Gamma}_0}}{2-\alpha}\lambda.$$

(b) If

$$\min_{1 \leq j < k \leq m} |\mathbf{\Psi}_{0,jk}| > \frac{c_{\mathbf{\Gamma}_0}}{2-\alpha}\lambda,$$

then $\hat{S} = S_0$ *and* $\operatorname{sign}(\hat{\mathbf{\Psi}}_{jk}) = \operatorname{sign}(\mathbf{\Psi}_{0.jk})$ *for all* $(j, k) \in S_0$*.*

This result is deterministic, and the improvement of our generalized estimator over the one in Lin et al. (2016) is in its probabilistic guarantees, as shown for truncated GGMs in Theorems 10 and 11 in Section 2.6.3. Before going into these examples, we state a general corollary.

Corollary 8. *Under the assumptions of Theorem 7, the matrix* $\hat{\mathbf{\Psi}}$ *minimizing (2.22) satisfies*

$$\begin{aligned} \|\hat{\mathbf{\Psi}} - \mathbf{\Psi}_0\|_F &\leq \frac{c_{\mathbf{\Gamma}_0}}{2-\alpha}\lambda\sqrt{|S_0|} \leq \frac{c_{\mathbf{\Gamma}_0}}{2-\alpha}\lambda\sqrt{d_{\mathbf{\Psi}_0} m}, \\ \|\hat{\mathbf{\Psi}} - \mathbf{\Psi}_0\|_2 &\leq \frac{c_{\mathbf{\Gamma}_0}}{2-\alpha}\lambda \min(\sqrt{|S_0|}, d_{\mathbf{\Psi}_0}). \end{aligned}$$

2.6.2 Revisiting Gaussian Score Matching

In this section we consider estimating the inverse covariance matrix $\mathbf{K}$ of a centered Gaussian distribution $\mathrm{N}(\mathbf{0}, \mathbf{K})$, which has density proportional to (1.2) on all of $\mathbb{R}^m$. As shown, e.g., in Example 1 of Lin et al. (2016), the ℓ_1-regularized score matching loss then takes the form

$$\frac{1}{2}\mathrm{tr}(\mathbf{K}\mathbf{K}\mathbf{x}\mathbf{x}^\top) - \mathrm{tr}(\mathbf{K}) + \lambda\|\mathbf{K}\|_1, \tag{2.26}$$

which can be written as (2.12) with $\boldsymbol{\theta} = \mathrm{vec}(\mathbf{K})$, $\boldsymbol{\Gamma} = \mathrm{diag}(\mathbf{x}\mathbf{x}^\top, \ldots, \mathbf{x}\mathbf{x}^\top)$ and $\boldsymbol{g} = \mathrm{vec}(\mathbf{I}_m)$. Thus, in general, the kernel of $\boldsymbol{\Gamma}$ need not be orthogonal to $\boldsymbol{g}$, and for λ small the loss can be unbounded from below as discussed above. Hence, an amplifier/multiplier on the diagonals of $\boldsymbol{\Gamma}$ is needed. We have the following theorem on the estimator using the amplification.

Theorem 9. *Suppose the data matrix* $\mathbf{x}$ *holds* n *i.i.d. copies of* $\boldsymbol{X} \sim \mathrm{N}(\mathbf{0}, \mathbf{K}_0)$. *Adopt the amplifying in Section 2.4 and redefine the loss in (2.26) as*

$$\frac{1}{2}\mathrm{tr}(\mathbf{K}\mathbf{K}\mathbf{G}) - \mathrm{tr}(\mathbf{K}) + \lambda\|\mathbf{K}\|_1, \quad \mathbf{G}_{jk} = (\mathbf{x}\mathbf{x}^\top)_{jk}\left(\mathbb{1}_{\{j\neq k\}} + (\delta-1)\mathbb{1}_{\{j=k\}}\right), \tag{2.27}$$

where $1 < \delta < 2 - \left(1 + 80\sqrt{\log m/n}\right)^{-1}$. *Let* $\hat{\mathbf{K}}$ *be the resulting estimator. Let* $c^* \equiv 12800\left(\max_j \boldsymbol{\Sigma}_{0,jj}\right)^2$ *and* $c_1 = 4c_{\boldsymbol{\Gamma}_0}/\alpha$. *If for some* $\tau > 2$, *the regularization parameter and the sample size satisfy*

$$\lambda > 2c_{\mathbf{K}_0}(2-\alpha)\sqrt{c^*(\tau\log m + \log 4)/n}/\alpha,$$
$$n > \max(c^* c_1^2 d_{\mathbf{K}_0}^2, 2)(\tau\log m + \log 4),$$

then $\|\hat{\mathbf{K}} - \mathbf{K}_0\|_\infty \leq \frac{c_{\boldsymbol{\Gamma}_0}}{2-\alpha}\lambda$ *with probability* $1 - m^{2-\tau}$.

In Corollary 1 of Lin et al. (2016) the same results were shown with $c^* \equiv 3200\left(\max_j \boldsymbol{\Sigma}_{0,jj}\right)^2$ when a unique minimizer exists, but the existence was not guaranteed.

2.6.3 Generalized Score Matching for Truncated GGMs

Next, we provide theory for the regularized generalized $\boldsymbol{h}$-score matching estimator $\hat{\boldsymbol{\Psi}}$ in the special case of truncated GGMs. Again, assume a common h for all components in $\boldsymbol{h}$.

Theorem 10. *Suppose the data matrix* $\mathbf{x}$ *holds* n *i.i.d. copies of* $\boldsymbol{X} \sim \mathrm{TN}(\mathbf{0}, \mathbf{K}_0)$*, where the mean parameter is known to be zero. Assume that* $\boldsymbol{h} \in \mathcal{H}_{1,1}$ *and that* $0 \leq h \leq M$*,* $0 \leq h' \leq M'$ *a.s. for constants* M, M'*, and choose* $\boldsymbol{\gamma} = (\delta - 1)\mathrm{diag}(\boldsymbol{\Gamma})$ *with*

$$1 < \delta < C(n, m) \equiv 2 - \left(1 + 4e \max\{6 \log m/n, \sqrt{6 \log m/n}\}\right)^{-1}.$$

Suppose that the $\boldsymbol{\Gamma}_{0,S_0S_0}$ *block of* $\boldsymbol{\Gamma}_0$ *is invertible and* $\boldsymbol{\Gamma}_0$ *satisfies the irrepresentability condition (2.24) with* $\alpha \in (0, 1]$ *and true edge set* S_0*. Define* $c_{\boldsymbol{X}} \equiv 2 \max_j \left(2\sqrt{(\mathbf{K}_0^{-1})_{jj}} + \sqrt{e}\, \mathbb{E}_0 X_j\right)$*. If for* $\tau > 3$ *the sample size and the regularization parameter satisfy*

$$n > \mathcal{O}\left(\tau \log m \max\left\{\frac{M^2 c_{\boldsymbol{\Gamma}_0}^2 c_{\boldsymbol{X}}^4 d_{\mathbf{K}_0}^2}{\alpha^2}, \frac{M c_{\boldsymbol{\Gamma}_0} c_{\boldsymbol{X}}^2 d_{\mathbf{K}_0}}{\alpha}\right\}\right), \tag{2.28}$$

$$\lambda > \mathcal{O}\left[(M c_{\mathbf{K}_0} c_{\boldsymbol{X}}^2 + M' c_{\boldsymbol{X}} + M)\left(\sqrt{\frac{\tau \log m}{n}} + \frac{\tau \log m}{n}\right)\right], \tag{2.29}$$

then the following statements hold with probability $1 - m^{3-\tau}$*:*

(a) The regularized generalized $\boldsymbol{h}$*-score matching estimator* $\hat{\mathbf{K}}$ *that minimizes (2.22) is unique, has its support included in the true support,* $\hat{S} \equiv S(\hat{\mathbf{K}}) \subseteq S_0$*, and satisfies*

$$\begin{aligned}
\|\hat{\mathbf{K}} - \mathbf{K}_0\|_\infty &\leq \frac{c_{\boldsymbol{\Gamma}_0}}{2 - \alpha}\lambda, \\
|||\hat{\mathbf{K}} - \mathbf{K}_0|||_F &\leq \frac{c_{\boldsymbol{\Gamma}_0}}{2 - \alpha}\lambda\sqrt{|S_0|}, \\
|||\hat{\mathbf{K}} - \mathbf{K}_0|||_2 &\leq \frac{c_{\boldsymbol{\Gamma}_0}}{2 - \alpha}\lambda \min(\sqrt{|S_0|}, d_{\mathbf{K}_0}),
\end{aligned}$$

where $c_{\boldsymbol{\Gamma}_0}$ *is defined in (2.23).*

(b) Moreover, if

$$\min_{j,k:(j,k)\in S_0} |\kappa_{0,jk}| > \frac{c_{\boldsymbol{\Gamma}_0}}{2 - \alpha}\lambda,$$

then $\hat{S} = S_0$ *and* $\mathrm{sign}(\hat{\kappa}_{jk}) = \mathrm{sign}(\kappa_{0,jk})$ *for all* $(j, k) \in S_0$*.*

The theorem is proved in Appendix A.1.4, where details on the dependencies on constants are provided. A key ingredient of the proof is a tail bound on $\|\boldsymbol{\Gamma}_{\boldsymbol{\gamma}} - \boldsymbol{\Gamma}_0\|_\infty$, which

features products of the $X_j^{(i)}$'s. In Lin et al. (2016), the products are up to fourth order. Using bounded $\boldsymbol{h}$, our products automatically calibrates to a quadratic polynomial when the observed values are large, and resort to higher moments only when they are small. This leads to improved bounds and convergence rates, underscored in the new requirement on the sample size n, which should be compared to $n \geq \mathcal{O}(d_{\mathbf{K}_0}^2(\log m^\tau)^8)$ in Lin et al. (2016).

For the non-centered case, by definition, $c_{\boldsymbol{\Psi}_0} \equiv |||\boldsymbol{\Psi}_{\mathbf{0}}^\top|||_{\infty,\infty} \leq c_{\mathbf{K}_0} + \|\boldsymbol{\eta}_0\|_\infty$, $d_{\boldsymbol{\Psi}_0} \leq d_{\mathbf{K}_0} + 1$. The proof given for Theorem 10 goes through again here, and we have the following consistency results.

Theorem 11. *Suppose the data matrix holds n i.i.d. copies of $\boldsymbol{X} \sim \mathrm{TN}(\boldsymbol{\mu}_0, \mathbf{K}_0)$. Assume that $\boldsymbol{h} \in \mathcal{H}_{1,1}$ and that $0 \leq h \leq M$, $0 \leq h' \leq M'$ a.s. for constants M, M'. Let $\boldsymbol{\gamma}$ be a vector of amplifiers that are non-zero only for the diagonal entries of the matrices $\boldsymbol{\Gamma}_{11,j}$, amplifying those by $(\delta - 1)\mathrm{diag}(\boldsymbol{\Gamma}_{11,j})$ with*

$$1 < \delta < C(n,m) \equiv 2 - \left(1 + 4e \max\{6 \log m/n, \sqrt{6 \log m/n}\}\right)^{-1}.$$

Suppose further that $\boldsymbol{\Gamma}_{0,S_0S_0}$ is invertible and satisfies the irrepresentability condition (2.24) with $\alpha \in (0,1]$. Define $c_{\boldsymbol{X}} \equiv 2\max_j \left(2\sqrt{(\mathbf{K}_0^{-1})_{jj}} + \sqrt{e}\,\mathbb{E}_0 X_j\right)$. Suppose for $\tau > 3$ the sample size and the regularization parameter satisfy

$$n > \mathcal{O}\left(\tau \log m \max\left\{\frac{M^2 c_{\boldsymbol{\Gamma}_0,\boldsymbol{\Psi}_0}^2 c_{\boldsymbol{X}}^4 d_{\boldsymbol{\Psi}_0}^2}{\alpha^2}, \frac{M c_{\boldsymbol{\Gamma}_0,\boldsymbol{\Psi}_0} c_{\boldsymbol{X}}^2 d_{\boldsymbol{\Psi}_0}}{\alpha}\right\}\right), \tag{2.30}$$

$$\lambda > \mathcal{O}\left[(M c_{\boldsymbol{\Psi}_0} c_{\boldsymbol{X}}^2 + M' c_{\boldsymbol{X}} + M)\left(\sqrt{\frac{\tau \log m}{n}} + \frac{\tau \log m}{n}\right)\right], \tag{2.31}$$

where $c_{\boldsymbol{\Gamma}_0,\boldsymbol{\Psi}_0}$ is $c_{\boldsymbol{\Gamma}_0}$ as in (2.23) but with notation $\boldsymbol{\Psi}_0$ to differentiate it from the centered case. Then the following statements hold with probability $1 - m^{3-\tau}$:

(a) The regularized generalized $\boldsymbol{h}$-score matching estimator $\hat{\boldsymbol{\Psi}}$ that minimizes (2.22) is unique, has its support included in the true support, $\hat{S} \equiv S(\hat{\boldsymbol{\Psi}}) \subseteq S_0$, and satisfies

$$\|\hat{\mathbf{K}} - \mathbf{K}_0\|_\infty \leq \frac{c_{\boldsymbol{\Gamma}_0,\boldsymbol{\Psi}_0}}{2-\alpha}\lambda, \qquad \|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0\|_\infty \leq \frac{c_{\boldsymbol{\Gamma}_0,\boldsymbol{\Psi}_0}}{2-\alpha}\lambda,$$

$$|||\hat{\mathbf{K}} - \mathbf{K}_0|||_F \leq \frac{c_{\mathbf{\Gamma}_0,\mathbf{\Psi}_0}}{2-\alpha}\lambda\sqrt{|S_0|}, \qquad |||\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0|||_F \leq \frac{c_{\mathbf{\Gamma}_0,\mathbf{\Psi}_0}}{2-\alpha}\lambda\sqrt{|S_0|},$$
$$|||\hat{\mathbf{K}} - \mathbf{K}_0|||_2 \leq \frac{c_{\mathbf{\Gamma}_0,\mathbf{\Psi}_0}}{2-\alpha}\lambda \min\left(\sqrt{|S_0|}, d_{\mathbf{\Psi}_0}\right), \quad |||\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0|||_2 \leq \frac{c_{\mathbf{\Gamma}_0,\mathbf{\Psi}_0}}{2-\alpha}\lambda \min\left(\sqrt{|S_0|}, d_{\mathbf{\Psi}_0}\right).$$

(b) Moreover, if

$$\min_{j,k:(j,k)\in S_0} |\kappa_{0,jk}| > \frac{c_{\mathbf{\Gamma}_0}}{2-\alpha}\lambda \qquad \textit{and} \qquad \min_{j:(m+1,j)\in S_0} |\eta_{0,j}| > \frac{c_{\mathbf{\Gamma}_0}}{2-\alpha}\lambda,$$

then $\hat{S} = S_0$ *and* $\operatorname{sign}(\hat{\kappa}_{jk}) = \operatorname{sign}(\kappa_{0,jk})$ *for all* $(j,k) \in S_0$ *and* $\operatorname{sign}(\hat{\eta}_j) = \operatorname{sign}(\eta_{0j})$ *for* $(m+1,j) \in S_0$.

Remark 3. The quantity $c_{\boldsymbol{X}}$ in Theorem 11 depends on $\mathbb{E}_0 X_j$, which in turn depends on the structure of both $\boldsymbol{\mu}_0$ and $\mathbf{K}_0$. If $\mu_{0,j}$ is large compared to $(\mathbf{K}_0)^{-1}_{jj}$, then $c_{\boldsymbol{X}}$ seems to scale with $\boldsymbol{\mu}_0$, which negatively impacts the guarantees stated in Theorem 11. However, as in the one-dimensional case for estimation of μ_0 (Example 3.1), our estimator should automatically adapt to the large mean parameter. This suggests that it might be possible to improve our analysis involving $c_{\boldsymbol{X}}$.

2.7 Numerical Experiments

In this section, we compare the performance of our estimator with different choices of $\boldsymbol{h}$ to the existing approaches for pairwise interaction power models. In our simulation experiments, we consider $m = 100$ variables and $n = 80$ and $n = 1000$ samples, corresponding to high- and low-dimensional settings. We also tried intermediate sample sizes between these two extremes, but found no interesting result worth reporting. For $n = 80$, amplification is necessary. Except in Section 2.7.2, the amplifier is set based on Theorem 10 to $\delta = C(n,m) = 1.8647$ for truncated GGMs. The same amplifier is also used for settings with other a and b. For $n = 1000$, we consider $\delta = 1$, i.e., no amplification, and $\delta = C(n,m) = 1.6438$ (again, based on Theorem 10). Throughout, we assume a common univariate h for all components in $\boldsymbol{h}$.

2.7.1 Structure of $\mathbf{K}$

The underlying interaction matrices are selected as follows: Proceeding as in Section 4.2 of Lin et al. (2016), the graph is chosen to have 10 disconnected subgraphs, each containing $m/10$ nodes. Thus, $\mathbf{K}_0$ is block-diagonal. In each block, each lower-triangular element is set to 0 with probability $1-\pi$ for some $\pi \in (0,1)$, and is otherwise drawn from Uniform$[0.5, 1]$. The upper triangular elements are determined by symmetry. The diagonal elements of $\mathbf{K}_0$ are chosen as a common positive value such that the minimum eigenvalue of $\mathbf{K}_0$ is 0.1.

We generate 5 different true precision matrices $\mathbf{K}_0$, and run 10 trials with each of these precision matrices. For $n = 1000$, we choose $\pi = 0.8$, which is in accordance with Lin et al. (2016). For $n = 80$, we set $\pi = 0.2$. This way $n/(d^2_{\mathbf{K}_0} \log m)$ is roughly constant; recall Theorems 10 and 11 for truncated GGMs.

In Appendix A.3, we report results on *Erdös-Rényi graphs*, which lead to similar conclusions.

2.7.2 Truncated GGMs

Given our focus on truncated GGMs and their relevance in graphical modeling applications, we start with experiments for these models.

Choice of $\boldsymbol{h}$

Our estimator requires choosing a function $\boldsymbol{h} : \mathbb{R}^m_+ \to \mathbb{R}^m_+$. For simplicity, we will always specify $\boldsymbol{h}(\boldsymbol{x}) = (h(x_1), \ldots, h(x_m))$ for a single non-decreasing univariate function $h : \mathbb{R}_+ \to \mathbb{R}_+$, i.e. all coordinates share the same h function.

As previously explained, $\boldsymbol{h} \in \mathcal{H}_{a,b}$ is a sufficient condition for assumptions (A1)-(A2), as well as (C1)-(C2) in the case of unregularized estimators. Only in the proofs of our theoretical guarantees in Section 2.6 for truncated GGMs, did we require h to be bounded and to have bounded derivatives. As motivated by the discussion in Section 2.5.4, we consider truncated and untruncated powers, $\min(x, c)$ and x (since $2 - a = 2 - b = 1$); we evaluate this choice

by contrasting them with powers $x^{1.5}$ and x^2. We also explore functions like $\log(1+x)$ that seem natural and are linear near 0. In particular, we make a further comparison to functions linear near 0 with a finite asymptote as $x \nearrow +\infty$ but differentiable everywhere: MCP- (Fan and Li, 2001) and SCAD-like (Zhang, 2010) functions defined below. The results we report are based on selections of best performing choices of h.

$$\mathrm{SCAD}(x;\lambda,\gamma) \equiv \begin{cases} \lambda x & \text{if } 0 \le x \le \lambda, \\ \frac{2\gamma\lambda x - x^2 - \lambda^2}{2(\gamma-1)} & \text{if } \lambda < x < \gamma\lambda, \\ \frac{\lambda^2(\gamma+1)}{2} & \text{if } x \ge \gamma\lambda; \end{cases} \quad \mathrm{MCP}(x;\lambda,\gamma) \equiv \begin{cases} \lambda x - \frac{x^2}{2\gamma} & \text{if } 0 \le x \le \gamma\lambda, \\ \frac{1}{2}\gamma\lambda^2 & \text{if } x > \gamma\lambda. \end{cases}$$

We do not observe any clear relationship between features such as convexity, differentiability or the slope of h at 0, and performance of the estimator. Nonetheless, for many choices of rather simple functions h, our estimator provides a significant improvement over existing methods. In particular, most h functions that behave linearly for small x, namely $\log(1+x)$ and x and their truncations, and additionally MCP and SCAD, always perform better than $x^{1.5}$ and x^2. This agrees with our discussion in Section 2.5.4, where $2 - a = 1$ is a reasonable choice of the power for small x; also see Section 2.7.3. However, we conclude that there is no real gain from making the function smoother by using MCP or SCAD.

Truncated Centered GGMs: For data from a truncated centered Gaussian distribution, we compare our generalized score matching estimator with various choices of h, to *SpaCE JAM* (SJ, Voorman et al., 2014), which estimates graphs using additive models for conditional means, a pseudo-likelihood method *SPACE* (Peng et al., 2009) in the reformulation of Khare et al. (2015), *graphical lasso* (GLASSO, Yuan and Lin, 2007; Friedman et al., 2008), the *neighborhood selection* estimator (NS) of Meinshausen and Bühlmann (2006), and *nonparanormal SKEPTIC* (Liu et al., 2012) with Kendall's τ. Recall that the choice of $h(x) = x^2$ corresponds to the estimator from Lin et al. (2016).

Centered, $n = 80$, multiplier 1.8647					
$\min(\log(1+x), c)$			$\min(x, c)$		
c	Mean	sd	c	Mean	sd
∞	0.694	0.033	∞	0.702	0.031
2	0.694	0.033	3	0.702	0.031
1	0.692	0.033	2	0.698	0.033
0.5	0.664	0.038	1	0.686	0.030
MCP$(1, c)$			SCAD$(1, c)$		
c	Mean	sd	c	Mean	sd
10	0.701	0.032	10	0.702	0.031
5	0.700	0.032	5	0.701	0.032
1	0.672	0.036	2	0.696	0.033
$x^{1.5}$: (0.683, 0.030)			x^2: (0.630, 0.029)		
GLASSO (0.600,0.032)			SPACE: (0.587, 0.031)		
NS: (0.587,0.031)			SJ: (0.540,0.036)		

Centered, $n = 1000$, multiplier 1					
$\min(\log(1+x), c)$			$\min(x, c)$		
c	Mean	sd	c	Mean	sd
2	0.826	0.015	2	0.820	0.014
∞	0.826	0.015	3	0.820	0.015
1	0.824	0.014	∞	0.819	0.015
0.5	0.804	0.015	1	0.817	0.014
MCP$(1, c)$			SCAD$(1, c)$		
c	Mean	sd	c	Mean	sd
5	0.824	0.015	2	0.823	0.014
10	0.822	0.015	5	0.822	0.015
1	0.810	0.015	10	0.821	0.015
$x^{1.5}$: (0.782,0.014)			x^2: (0.732,0.015)		
SPACE: (0.780,0.015)			NS: (0.779,0.015)		
GLASSO (0.764,0.014)			SJ: (0.703,0.015)		

Centered, $n = 1000$, multiplier 1.6438					
$\min(\log(1+x), c)$			$\min(x, c)$		
c	Mean	sd	c	Mean	sd
∞	0.857	0.011	3	0.855	0.011
2	0.857	0.011	∞	0.855	0.011
1	0.855	0.011	2	0.854	0.011
0.5	0.833	0.012	1	0.847	0.011
MCP$(1, c)$			SCAD$(1, c)$		
c	Mean	sd	c	Mean	sd
5	0.857	0.011	5	0.856	0.011
10	0.856	0.011	10	0.855	0.011
1	0.840	0.012	2	0.855	0.011
$x^{1.5}$: (0.812,0.011)			x^2: (0.736,0.011)		
SPACE: (0.780,0.015)			NS: (0.779,0.015)		
GLASSO (0.764,0.014)			SJ: (0.703,0.015)		

Table 2.1: Mean and standard deviation of areas under the ROC curves (AUC) using different estimators in the centered setting, with $n = 80$ and multiplier 1.8647, or $n = 1000$ and multiplier 1 and 1.6438. Methods include our estimator with different choices of h, GLASSO, SPACE, neighborhood selection (NS), and Space JAM (SJ).

The ROC (*receiver operating characteristic*) curves for different estimators are shown in

Figure 2.3 on Page 39. Each plotted curve corresponds to the average of 50 ROC curves, where the averaging is based on the vertical averaging from Algorithm 3 in Fawcett (2006), and is mean AUC-preserving. The x and y axes of each ROC curve represent the false positive and true positive rates at varying levels of penalty parameter λ, defined as

$$\text{FPR} \equiv \frac{|\hat{S}_{\text{off}} \backslash S_{0,\text{off}}|}{m(m-1) - |S_{0,\text{off}}|} \qquad \text{and} \qquad \text{TPR} \equiv \frac{|\hat{S}_{\text{off}} \cap S_{0,\text{off}}|}{|S_{0,\text{off}}|},$$

where $S_{0,\text{off}} \equiv \{(i,j) : i \neq j \wedge \kappa_{0,ij} \neq 0\}$, and $\hat{S}_{\text{off}} \equiv \{(i,j) : i \neq j \wedge \hat{\kappa}_{ij} \neq 0\}$.

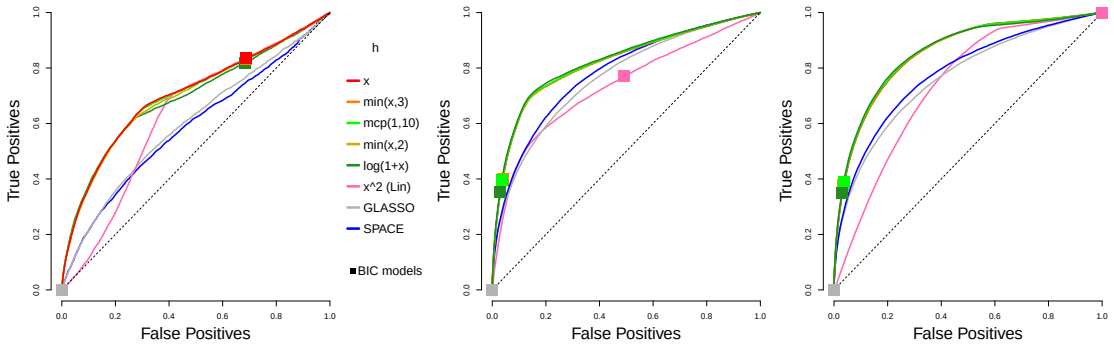
To reduce clutter, we only report the results for the top performing competing methods. In particular, results for nonparanormal SKEPTIC are omitted, as the method always performs the worst in our experiments. The corresponding means and standard deviations of AUCs (*areas under the curves*) over 50 curves are given in Table 2.1.

Looking at the mean AUCs, with the standard deviations in mind, all choices of h considered here perform better than $h(x) = x^2$ from Hyvärinen (2007) and Lin et al. (2016) and the competing methods. The results for $n = 1000$ in Table 2.1 also show that the multiplier does help improve the AUCs, a matter to be discussed in Section 2.7.2.

Truncated Non-Centered GGMs: We generate data from a truncated non-centered Gaussian distribution with both parameters $\boldsymbol{\mu}$ and $\mathbf{K}$ unknown. In each trial, we form the true $\mathbf{K}_0$ as in Section 2.7.1, and generate each component of $\boldsymbol{\mu}_0$ independently from the normal distribution with mean 0 and standard deviation 0.5.

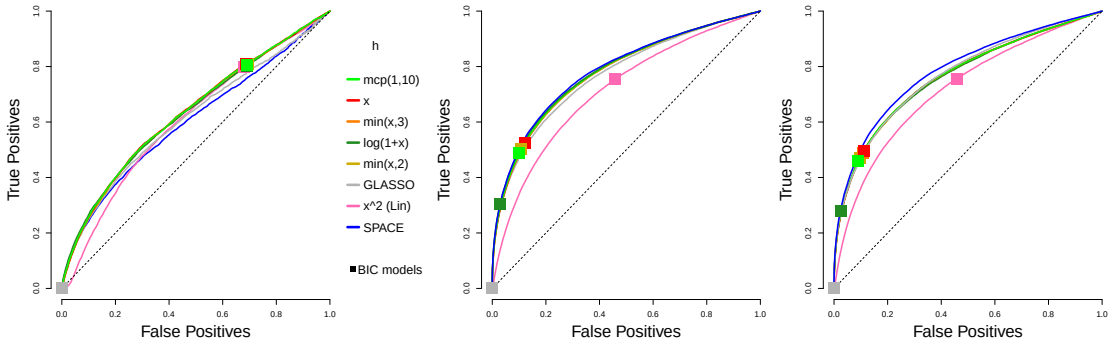
We compare the performance of our *profiled* estimator based on (2.17), with different h functions, but with no penalty on $\boldsymbol{\eta} \equiv \mathbf{K}\boldsymbol{\mu}$, to SPACE, SpaCE JAM (SJ), GLASSO, and neighborhood selection (NS). As before, we consider 50 trials. Representative ROC curves are plotted in Figure 2.4, and the corresponding AUCs are summarized in Table 2.2.

Even without tuning the extra penalty parameter on $\boldsymbol{\eta} \equiv \mathbf{K}\boldsymbol{\mu}$, our profiled estimator beats the competing methods by a large margin when $n = 80$. With multipliers 1 and $n = 1000$, our estimators still do better than Space JAM and GLASSO, and have performance comparable to other competing methods. It might appear that the performance of our



(a) $n = 80$, mult $= 1.8647$ (b) $n = 1000$, mult $= 1$ (c) $n = 1000$, mult $= 1.6438$

Figure 2.3: Average ROC curves of our centered estimator with various choices of h, compared to SPACE and GLASSO, for the truncated centered GGM case; $m = 100$ variables and $n = 80$ or 1000 samples are considered. Squares indicate average true positive rate (TPR) and false positive rate (FPR) of models picked by eBIC with refitting for the estimator in the same color.



(a) $n = 80$, mult $= 1.8647$ (b) $n = 1000$, mult $= 1$ (c) $n = 1000$, mult $= 1.6438$

Figure 2.4: Average ROC curves of our non-centered profiled estimator with various choices of h, compared to SPACE and GLASSO, for the truncated non-centered GGM case; $m = 100$ variables and $n = 80$ or 1000 samples are considered.

Non-centered profiled, $n = 80$, multiplier 1.8647					
$\min(\log(1+x), c)$			$\min(x, c)$		
c	Mean	sd	c	Mean	sd
∞	0.632	0.032	∞	0.634	0.032
2	0.632	0.032	3	0.634	0.032
1	0.631	0.032	2	0.632	0.032
0.5	0.619	0.033	1	0.628	0.032
MCP$(1, c)$			SCAD$(1, c)$		
c	Mean	sd	c	Mean	sd
10	0.634	0.032	5	0.634	0.032
5	0.634	0.032	10	0.634	0.032
1	0.622	0.032	2	0.634	0.032
$x^{1.5}$: (0.623,0.031)			x^2: (0.607,0.030)		
GLASSO: (0.614,0.029)			NS: (0.604,0.028)		
SPACE: (0.602,0.029)			SJ: (0.561,0.036)		

Non-centered profiled, $n = 1000$, multiplier 1					
$\min(\log(1+x), c)$			$\min(x, c)$		
c	Mean	sd	c	Mean	sd
∞	0.783	0.020	2	0.779	0.020
2	0.783	0.020	∞	0.779	0.020
1	0.782	0.020	3	0.779	0.020
0.5	0.767	0.021	0.5	0.758	0.020
MCP$(1, c)$			SCAD$(1, c)$		
c	Mean	sd	c	Mean	sd
5	0.782	0.020	2	0.780	0.020
10	0.780	0.020	5	0.780	0.020
1	0.771	0.021	10	0.779	0.020
$x^{1.5}$: (0.751,0.019)			x^2: (0.713,0.018)		
SPACE: (0.786,0.020)			NS: (0.785,0.02)		
GLASSO (0.770,0.019)			SJ: (0.720,0.019)		

Non-centered profiled, $n = 1000$, multiplier 1.6438					
$\min(\log(1+x), c)$			$\min(x, c)$		
c	Mean	sd	c	Mean	sd
∞	0.764	0.018	∞	0.766	0.019
2	0.764	0.018	3	0.765	0.019
1	0.762	0.018	2	0.764	0.018
0.5	0.738	0.018	1	0.753	0.018
MCP$(1, c)$			SCAD$(1, c)$		
c	Mean	sd	c	Mean	sd
10	0.766	0.019	10	0.766	0.019
5	0.766	0.019	5	0.766	0.019
1	0.745	0.018	2	0.763	0.018
$x^{1.5}$: (0.748,0.018)			x^2: (0.718,0.017)		
SPACE: (0.786,0.020)			NS: (0.785,0.020)		
GLASSO (0.770,0.019)			SJ: (0.720,0.019)		

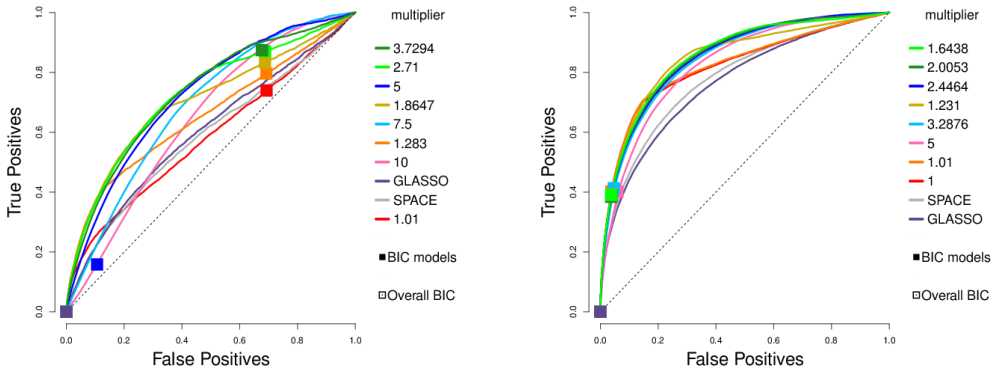
Table 2.2: Mean and standard deviation of AUC using different profiled estimators in the non-centered setting, with $n = 80$ and multiplier 1.8647, or $n = 1000$ and multipliers 1 and 1.6438. Methods include our estimator with different choices of h, GLASSO, SPACE, neighborhood selection (NS), and Space JAM (SJ).

estimators deteriorate with a multiplier larger than 1; however, as we will see, there can be significant improvement in AUCs if we tune an additional parameter for the multiplier. As in the centered case, the leading h functions in each category perform similarly, and the exact choice is not crucial. Subsequently, we will simply use $h(x) = \min(x, 3)$.

Choice of multiplier

Truncated Centered GGMs: In Figure 2.5, the ROC curves for GLASSO, SPACE, and our estimator with $h(x) = \min(x, 3)$, but with different levels of amplification, via different choices of multipliers δ, are compared for the centered case of Section 2.7.2.

While Theorem 10 guarantees consistency only for $\delta < C(n, m)$, we observe that there can be a gain from going beyond the *upper-bound multiplier* $C(n, m)$, which is 1.8647 for $n = 80$ and 1.6438 for $n = 1000$ (when $n = 1000$, $C(n, m)$ turns out to be the best-performing multiplier). However, the effect deteriorates fast as the multiplier grows larger. The figure suggests that while some additional gains are possible by tuning over the choice of multiplier, the *upper-bound multiplier* is a good default.



(a) $n = 80$ (b) $n = 1000$

Figure 2.5: Performance of $\min(x, 3)$ for truncated centered GGMs using different multipliers, compared to GLASSO and SPACE, in the centered setting, $n = 80$ or 1000.

Truncated Non-Centered GGMs : In Figure 2.6, we consider the non-centered case of Section 2.7.2, and use the non-profiled estimator; that is, the non-centered estimator with ℓ_1 penalty on both $\mathbf{K}$ and $\boldsymbol{\eta} \equiv \mathbf{K}\boldsymbol{\mu}$. The ROC curves are compared to competing methods GLASSO and SPACE. For the choice of amplification in our estimator, we consider the upper-bound multiplier $C(n, m)$ from Theorem 11 as the default. We refer to this as *high* amplification. We also consider lower amplification, with $\delta = 2-(1+24e\log m/n)^{-1}$, referred to as *medium*. For $n = 1000$, we also consider a *low* multiplier 1, which corresponds to no amplification. We compare these possible defaults to a finer grid of multipliers of which we show some representatives in the plots.

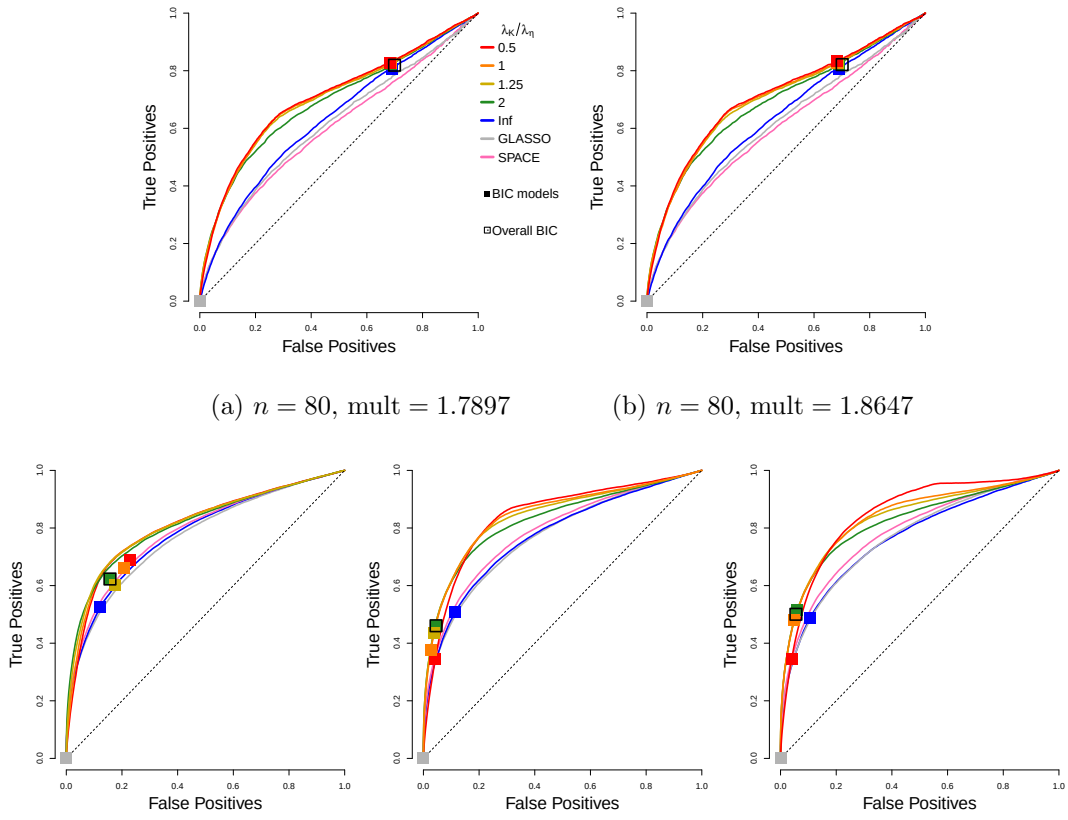
We see that among our defaults, the upper-bound choice $C(n, m)$ performs best. Some additional gains are possible by tuning the multiplier over a grid of values containing this choice. Moreover, we see that it can be beneficial to tune over both $\lambda_{\mathbf{K}}$ and $\lambda_{\mathbf{K}}/\lambda_{\boldsymbol{\eta}}$.

We remark that while for each run, the best model picked by BIC falls on the ROC curve, a few squares are off the curve in Figure 2.6 (c). This is because these squares correspond to the average of the true and false positive rates of the chosen BIC models over 50 runs, potentially due to multimodality of the distribution of the models. Nonetheless, in all cases, the average of the models picked by BIC tuned over both $\lambda_{\mathbf{K}}$ and $\lambda_{\mathbf{K}}/\lambda_{\boldsymbol{\eta}}$ looks reasonable.

2.7.3 Other a/b Models

We now turn to the non-Gaussian ($a \neq 1$ or $b \neq 1$) setting. Based on the observations in Section 2.5.4, we focus on functions of type $\min(x^p, c)$ for some power $p > 0$ and truncation point $c > 0$. For simplicity, for the non-centered models we use the profiled estimator (2.17) (i.e., $\lambda_{\boldsymbol{\eta}} = 0$) and use the multiplier $C(n, m)$ in Theorem 10 for truncated GGMs as a guidance. We note that tuning over the $\lambda_{\boldsymbol{\eta}}$ parameter and the multiplier can potentially give a significant improvement as seen in Section 2.7.2.

These simulations suggest that among the class of functions of the form $\min(x^p, c)$, x^{2-a} or $\min(x^{2-a}, c)$ with a moderately large c can be used as the default choice of $h(x)$. This agrees with our findings in Section 2.7.2. We note that bounded h functions were only



(a) $n = 80$, mult $= 1.7897$ (b) $n = 80$, mult $= 1.8647$

(c) $n = 1000$, mult $= 1$ (d) $n = 1000$, mult $= 1.2310$ (e) $n = 1000$, mult $= 1.6438$

Figure 2.6: Performance of the non-centered estimator with $h(x) = \min(x, 3)$. Each curve corresponds to a different choice of $\lambda_{\mathbf{K}}/\lambda_{\boldsymbol{\eta}}$. Squares indicate models picked by eBIC with refit. The square with black outline has the highest eBIC among all models (combinations of $\lambda_{\mathbf{K}}$, $\lambda_{\boldsymbol{\eta}}$). Multipliers correspond to medium or high for $n = 80$, and low, medium or high for $n = 1000$, respectively.

used in the proof for truncated GGMs, and picking a moderately large truncation point can correspond to having an untruncated power.

Exponential Setting

For the exponential models, $a = b = 1/2$. Since $a = b$, for both centered and non-centered settings, based on the principle in Section 2.5.4, choosing $h(x) = \min(x^{3/2}, c)$ satisfies (A1) and (A2) and also ensures that entries in $\mathbf{\Gamma}$ and $\boldsymbol{g}$ are bounded (for small x), while choosing $h(x) = \min(\sqrt{x}, c)$ only guarantees (A1) and (A2).

In Figure 2.7, we present the AUCs for the ROC curves of edge recovery with different choices of $h(x) = \min(x^{\text{pow}}, c)$. As before, we set $n = 80$ or 1000 and $m = 100$, but we use an $\boldsymbol{\eta}_0$ with each component uniformly equal to -0.5, 0 or 0.5; for $\boldsymbol{\eta}_0 \equiv \mathbf{0}$, we assume this information is known and use the centered estimator. The results suggest that pow $= 3/2 = 2 - a$ is the best choice of power. For this optimal choice, the performance improves with larger c, so x^{2-a} gives the best results. For sub-optimal powers, including truncation gives better results.

Gamma Setting

The centered gamma models reduce to the centered exponential models. Thus, in this section, we only consider the non-centered settings, with $a = 1/2$, $b = 0$. From Section 2.5.4, we have the following choices:

- $\min(x^2, c)$ both satisfies (A1)–(A2) and ensures $\mathbf{\Gamma}$ and $\boldsymbol{g}$ are bounded;
- $\min(x^{\max\{3/2, 1-\min_j \eta_{0,j}\}}, c)$ ensures (A1)–(A2) and bounds $\mathbf{\Gamma}_{11}$ and $\boldsymbol{g}_1$; by default without prior information on $\boldsymbol{\eta}_0$ this is $\min(x^2, c)$;
- $\min(x^{3/2}, c)$ satisfies both conditions on the interaction part only ($\boldsymbol{x}^a$), but does not guarantee (A1)–(A2);

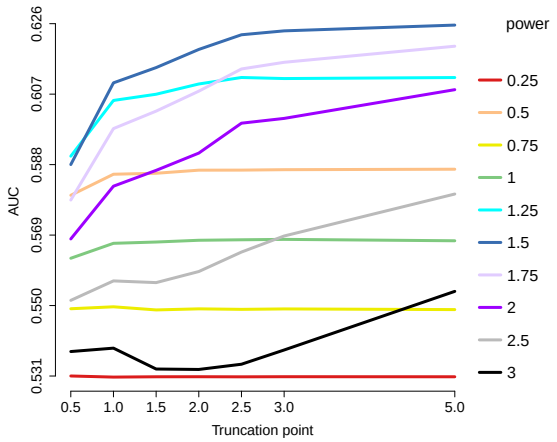
- $\min(x^{1/2}, c)$ satisfies the sufficient conditions for (A1)–(A2) on the interaction only.

The results are shown in Figure 2.8, where we consider $n = 80$, 1000, and $\boldsymbol{\eta} = \pm 0.5\mathbf{1}_{100}$. They suggest that pow $= 2-a = 1.5$ works consistently well, although slightly outperformed by 1 and 1.25 in one case. As in the exponential case, with the optimal power it is beneficial to choose a large truncation point, or work with an untruncated power $x^{1.5}$. We conclude that the performance is likely only dependent on the $(2-a)$ power requirement for the $\boldsymbol{x}^{a\top}\mathbf{K}\boldsymbol{x}^a$ part or $2 - \min_j \eta_{0,j}$; simulations in the next section rule out the possibility of the latter.

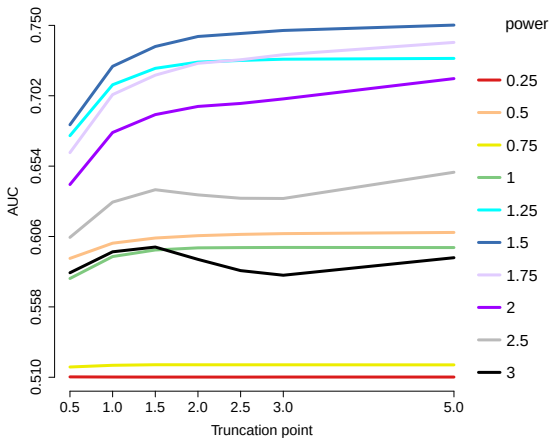
Other Choices of a and b

In this section, we consider other choices of a and b. Specifically, $a = 3/2$ and $b = 1/2$ or 0. These combinations are chosen to confirm, in a more extreme setting, that the performance is mainly determined by requirements on the power based on a, which correspond to choosing a power of $1-a$ or $2-a$, but not those on b (or on $\boldsymbol{\eta}$ when $b = 0$) that correspond to $1-b$ and $2-b$. The relationship between these two settings is analogous to that between the exponential and gamma models (same a, b nonzero/zero).

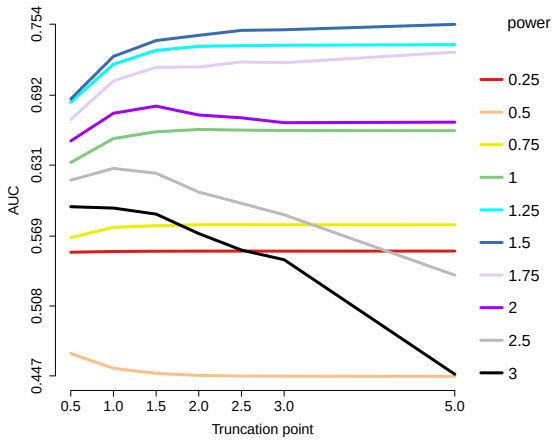
The results are shown in Figures 2.9 and 2.10, and indeed confirm that $x^{2-a} = x^{0.5}$ consistently gives the optimal results, even though $\boldsymbol{\eta}^\top \boldsymbol{x}^b$ is in favor of $x^{2-b} = x^{1.5}$ for $b = 0.5$, and $\boldsymbol{\eta}^\top \log(\boldsymbol{x})$ is in favor of x^2 or at least $x^{1-\min_j \eta_{0,j}}$ when $b = 0$. There are two possible explanations for the optimality of $2-a$ over $\max\{2-a, 2-b\}$ or $\max\{2-a, 1-\min_j \eta_{0,j}\}$: (1) The AUC metric is measured only on our interest, edge recovery for the interaction matrix, which only depends on $\boldsymbol{x}^a$; (2) using the profiled estimator weakens the effect of b.



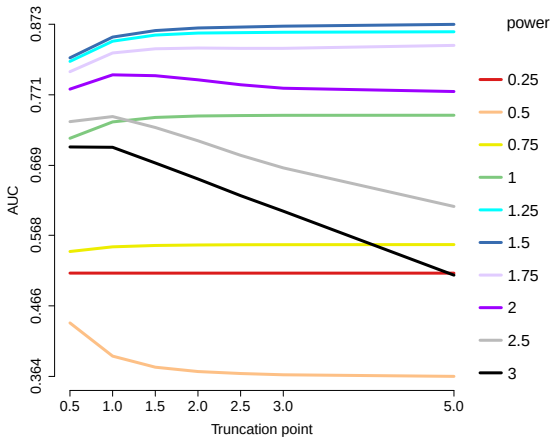
(a) $n = 80$, $\boldsymbol{\eta} = -0.5\mathbf{1}_{100}$, profiled estimator



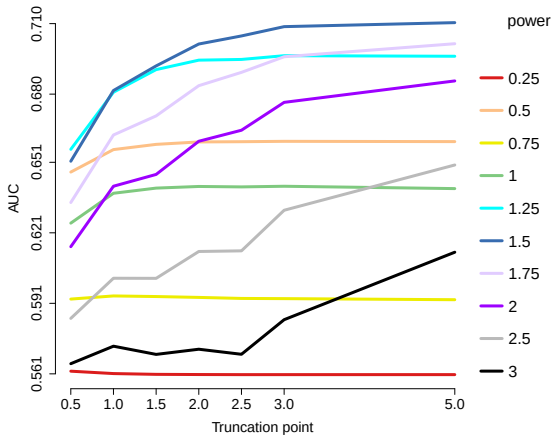
(b) $n = 1000$, $\boldsymbol{\eta} = -0.5\mathbf{1}_{100}$, profiled estimator



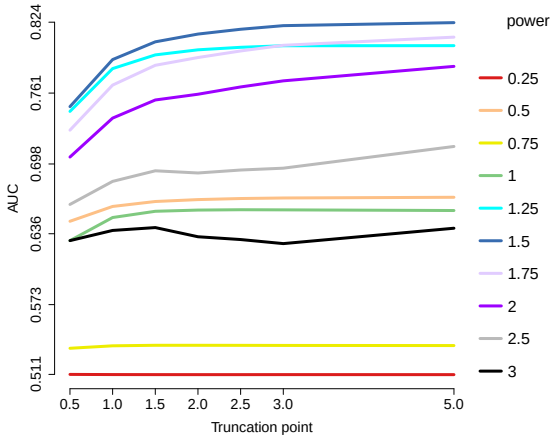
(c) $n = 80$, $\boldsymbol{\eta} = \mathbf{0}$, centered estimator



(d) $n = 1000$, $\boldsymbol{\eta} = \mathbf{0}$, centered estimator

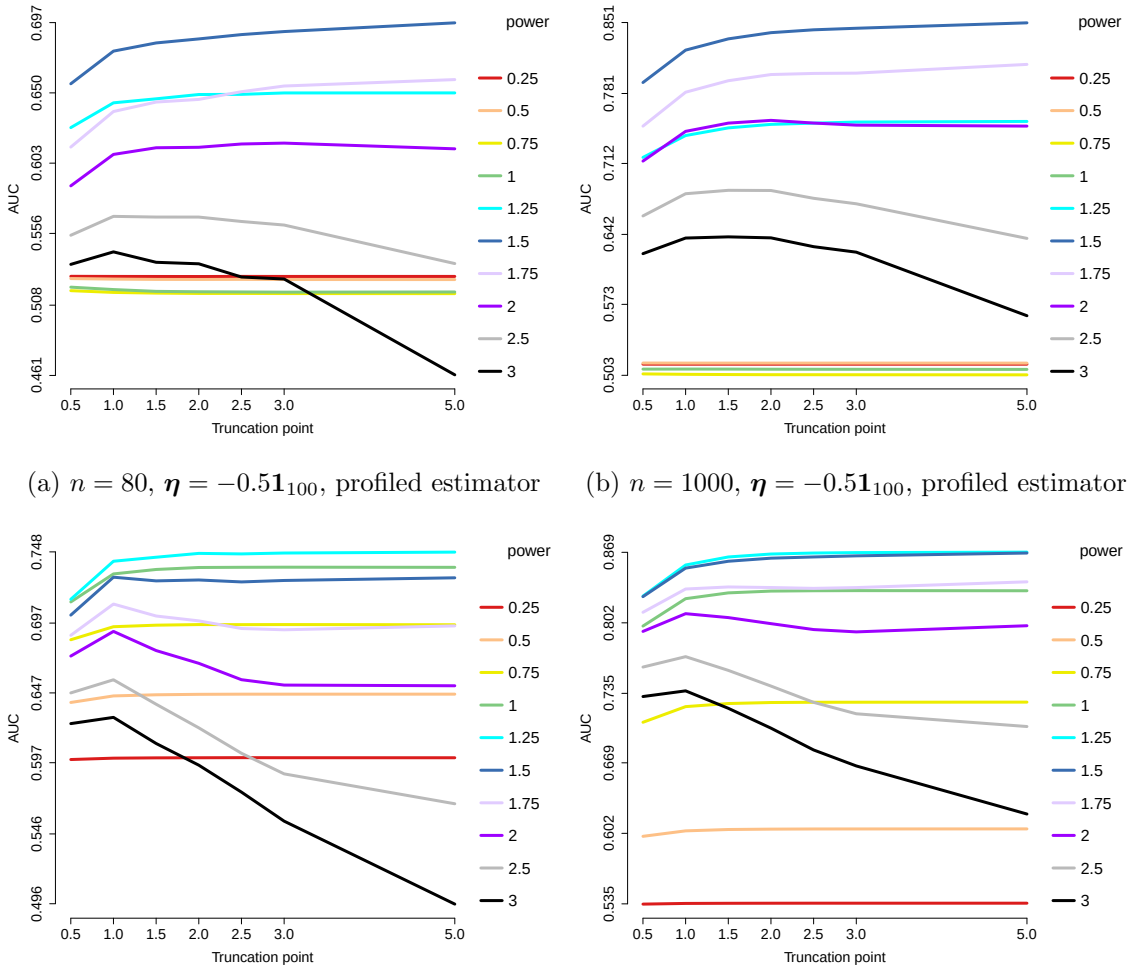


(e) $n = 80$, $\boldsymbol{\eta} = 0.5\mathbf{1}_{100}$, profiled estimator



(f) $n = 1000$, $\boldsymbol{\eta} = 0.5\mathbf{1}_{100}$, profiled estimator

Figure 2.7: AUCs for edge recovery using generalized score matching for the exponential models. Each curve represents a different choice of power p in $h(x) = \min(x^p, c)$, and the x axis marks the truncation point c. Colors are sorted by p.



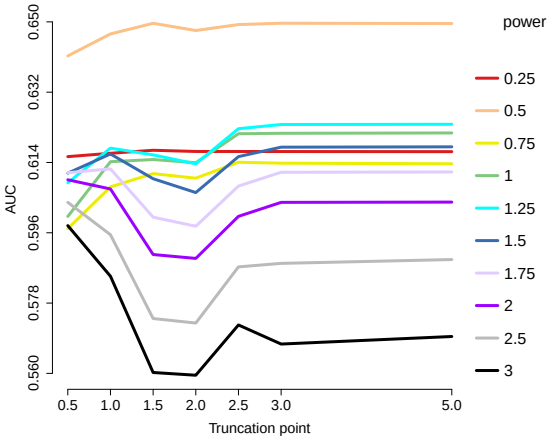
(a) $n = 80$, $\boldsymbol{\eta} = -0.5\mathbf{1}_{100}$, profiled estimator

(b) $n = 1000$, $\boldsymbol{\eta} = -0.5\mathbf{1}_{100}$, profiled estimator

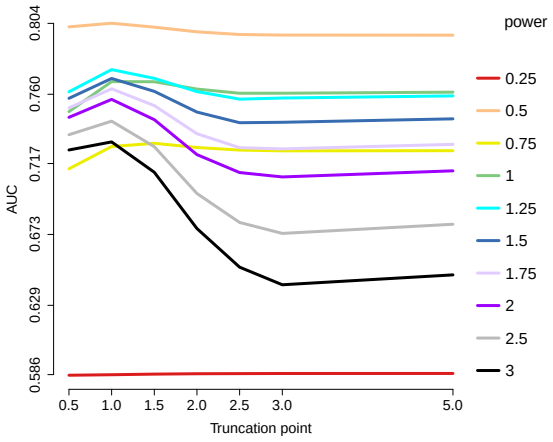
(c) $n = 80$, $\boldsymbol{\eta} = 0.5\mathbf{1}_{100}$, profiled estimator

(d) $n = 1000$, $\boldsymbol{\eta} = 0.5\mathbf{1}_{100}$, profiled estimator

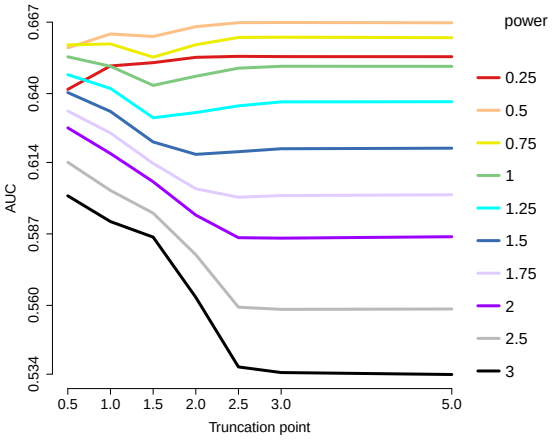
Figure 2.8: AUCs for edge recovery using generalized score matching for the gamma models.



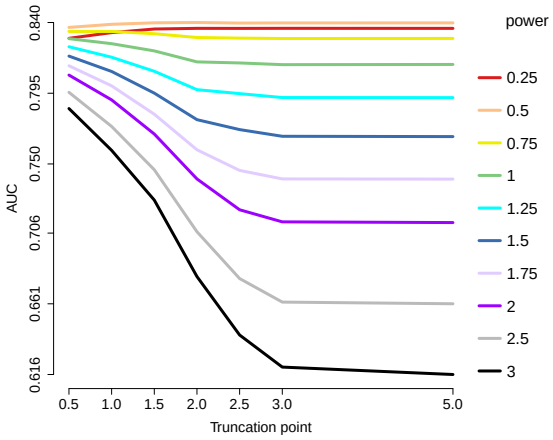
(a) $n = 80$, $\boldsymbol{\eta} = -0.5\mathbf{1}_{100}$, profiled estimator



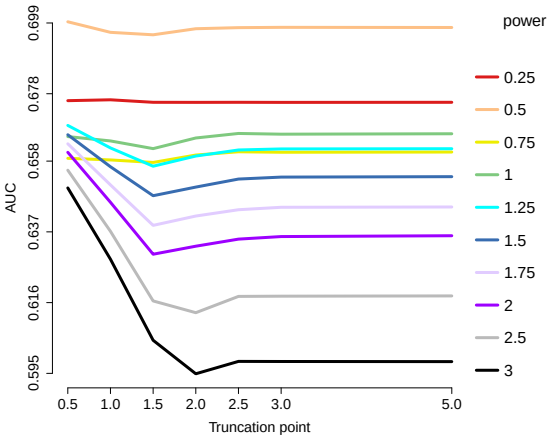
(b) $n = 1000$, $\boldsymbol{\eta} = -0.5\mathbf{1}_{100}$, profiled estimator



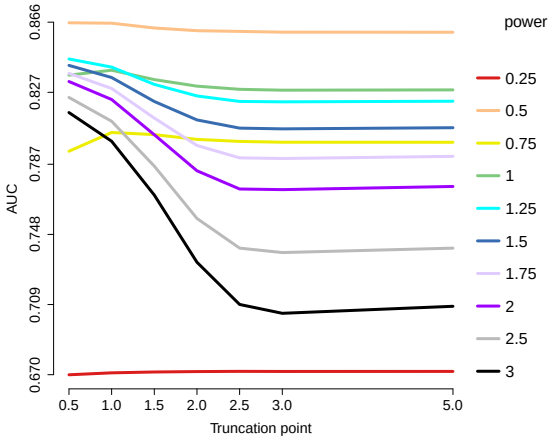
(c) $n = 80$, $\boldsymbol{\eta} = \mathbf{0}$, centered estimator



(d) $n = 1000$, $\boldsymbol{\eta} = \mathbf{0}$, centered estimator

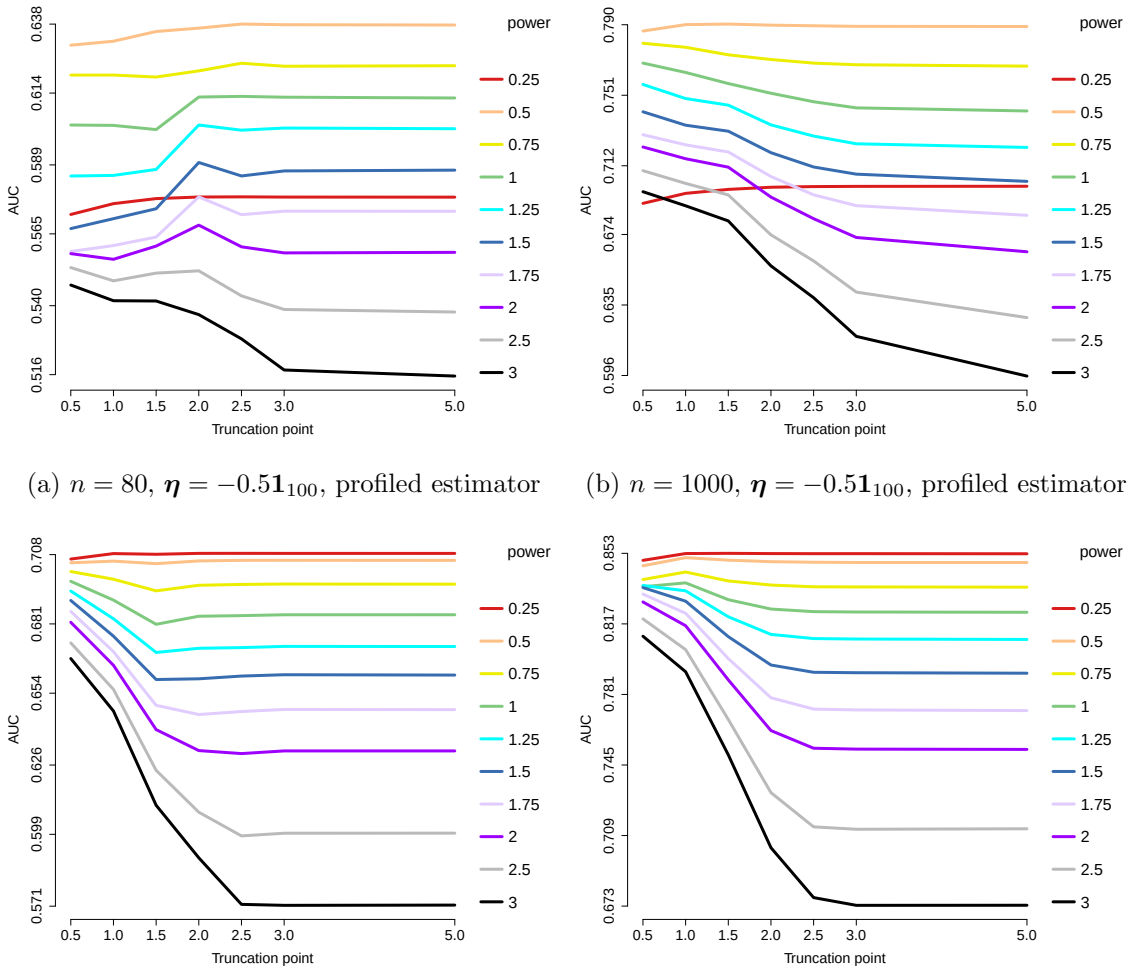


(e) $n = 80$, $\boldsymbol{\eta} = 0.5\mathbf{1}_{100}$, profiled estimator



(f) $n = 1000$, $\boldsymbol{\eta} = 0.5\mathbf{1}_{100}$, profiled estimator

Figure 2.9: AUCs for edge recovery using generalized score matching for $a = 3/2$, $b = 1/2$.



(a) $n = 80$, $\boldsymbol{\eta} = -0.5\mathbf{1}_{100}$, profiled estimator

(b) $n = 1000$, $\boldsymbol{\eta} = -0.5\mathbf{1}_{100}$, profiled estimator

(c) $n = 80$, $\boldsymbol{\eta} = 0.5\mathbf{1}_{100}$, profiled estimator

(d) $n = 1000$, $\boldsymbol{\eta} = 0.5\mathbf{1}_{100}$, profiled estimator

Figure 2.10: AUCs for edge recovery using generalized score matching for $a = 3/2$, $b = 0$.

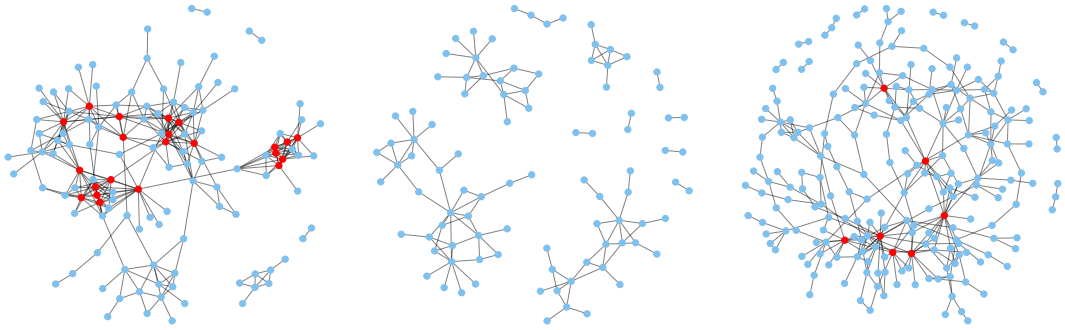
2.7.4 RNAseq Data

In this section we apply our regularized generalized $\boldsymbol{h}$-score matching estimator for truncated non-centered GGMs to RNAseq data also studied in Lin et al. (2016), since the same model is considered therein. The data consists of $n = 487$ prostate adenocarcinoma samples from The Cancer Genome Atlas (TCGA) data set. Following Lin et al. (2016), we focus on $m = 333$ genes that belong to the known cancer pathways in the Kyoto Encyclopedia of Genes and Genomes (KEGG) and that have no more than 10% missing values. Missing values are set to 0. We choose $h(x) = \min(x, 3)$ and use the upper-bound multiplier (*high*), as discussed in Section 2.7.2. For simplicity, we use the profiled estimator, and choose the regularization parameter λ so that the estimated graph has exactly $m = 333$ edges, all these choices being as in Lin et al. (2016).

We compare our graph to the one in Lin et al. (2016), which corresponds to $h(x) = x^2$ with no multiplier. Shown in Figure 2.11 are the estimated graphs, with their intersection in the middle. To improve visualization, isolated nodes are removed and the layouts are optimized for each plot. Red-colored points are the "hub nodes", namely nodes with degree at least 10. In Figure 2.12, we plot the same graphs in a fixed layout optimized for the graph corresponding to $h = \min(x, 3)$, and include the isolated nodes.

Out of 333 edges, the two estimated graphs share 117 edges in common. Assuming that edges are placed at random between nodes and the two graphs are independent, the distribution of the number R of common edges follow a hypergeometric distribution, so $P(R = r) = \frac{\binom{m}{r}\binom{m(m-1)/2-m}{m-r}}{\binom{m(m-1)/2}{m}}$. For $m = 333$ the probability of at least 117 common edges is essentially zero. The large number of shared edges between the two methods can be explained by the fact that they both minimize the same underlying score-matching loss.

The graph using $h(x) = \min(x, 3)$ has much more isolated nodes (204) than the other (108), and has a slightly smaller max degree (16 versus 19). Table 2.3 provides another way of comparing between the two graphs by listing the genes with the highest node degrees.

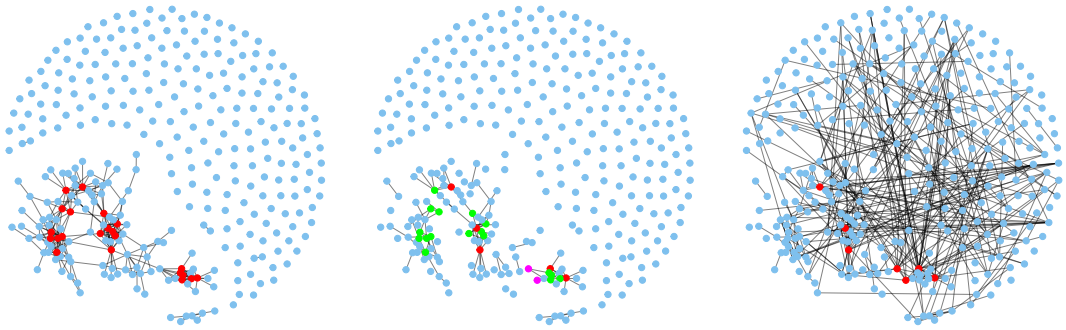


(a) $\min(x, 3)$ (b) Common edges (c) Lin et al.

Figure 2.11: Graphs estimated by regularized generalized score matching estimator with $h(x) = \min(x, 3)$ with upper-bound multiplier (left) and $h(x) = x^2$ with no multiplier (Lin et al., 2016, right), and their intersection graph (middle). Isolated nodes with no edges are removed, and the layout is optimized for each plot. In (a) and (c), red points indicate nodes with degree at least 10 ("hub nodes").

$\min(x, 3)$ with multiplier 1.63	Lin et al.
LAMB3 (16)	CCNE2 (19)
PIK3CG (16)	**PIK3CG (16)**
MMP2 (15)	BRCA2 (13)
GLI2 (13)	**BIRC5 (12)**
LAMA4 (13)	**LAMB3 (10)**
PDGFRB (13)	**PIK3CD (10)**
PIK3CD (13)	SKP2 (10)
RASSF5 (13)	HRAS (9)
BIRC5 (12)	STAT5B (9)
FLT3 (12)	**GSTP1(8)**
GSTP1 (12)	**PDGFRB (8)**
LAMA2 (12)	
RAC2 (12)	

Table 2.3: List of genes with the highest node degrees in each estimated graph.



(a) $\min(x, 3)$ (b) Common edges (c) Lin et al.

Figure 2.12: Graphs estimated by regularized generalized score matching estimator with $h(x) = \min(x, 3)$ with upper-bound multiplier (left) and $h(x) = x^2$ with no multiplier (Lin et al., 2016, right), and their intersection graph (middle). Isolated nodes are included and the layout is fixed across plots and optimized for graph (a). In (b) the red nodes are hub nodes shared by both graphs, the green ones are hub nodes in graph (a) only, and the magenta ones are hub nodes in graph (c) only.

In Table 2.3 we list the top ten genes in terms of node degree for both estimated graphs. Due to ties, 13 genes are listed for $h(x) = \min(x, 3)$ and 11 for Lin et al. (2016). As noted in Lin et al. (2016), genes with high node-degrees are known to be important in biological networks (Carter et al., 2004; Jeong et al., 2001; Han et al., 2004). Among these top genes, six are common in both graphs, and are discussed in Lin et al. (2016). We next elaborate on the evidence supporting the first four of the newly discovered genes.

- MMP2 (Matrix metalloproteinase 2): According to Trudel et al. (2003), increased MMP-2 expression is an independent predictor of decreased prostate cancer disease-free survival. Morgia et al. (2005) state that activity of MMP-2 can be useful in diagnosis, therapy, and assessment of malignant progression in prostate cancer.

- GLI2 (GLI family zinc finger 2): GLI2 is a primary mediator of the hedgehog signaling pathway, which has been reported in prostate cancer, and plays a critical role in the malignant phenotype of prostate cancer cells (Thiyagarajan et al., 2007). Its increased level of expression is also related to AI prostate cancer, and may be a therapeutic target in castrate-resistant prostate cancer (Narita et al., 2008).

- LAMA4 (Laminin subunit alpha 4): LAMA4 is consistently upregulated in benign prostatic hyperplasia when compared to normal prostate tissues (Luo et al., 2002).

- RASSF5 (RAS association domain family member 5): The combination of RASSF5 along with four other DNA methylation markers can effectively differentiate between benign prostate biopsy cores from non-cancer patients and cancer cores, and can be used to identify patients at risk without repeat biopsies (Brikun et al., 2014).

We note that the two methods indeed use different estimators (different h functions and multipliers), and it is thus not surprising to see that some of the top genes by one method are not among those for the other. In particular, CCNE2, BRCA2, SKP2 and STAT5B, while previously reported as newly discovered in Lin et al. (2016), are dropped by our new

analysis. Testing and inference (potentially using bootstrapping) is an important problem but is beyond the scope of this chapter.

2.8 Discussion

In this chapter, we proposed a generalization of the score matching estimator of Hyvärinen (2007), based on scaling the log-gradients to be matched with a suitably chosen function $\boldsymbol{h}$. The generalization retains the advantages of Hyvärinen's method: Estimates can be computed without knowledge of normalizing constants, and for canonical parameters of exponential families, the estimation loss is a quadratic function.

For high-dimensional exponential family graphical models, following Lin et al. (2016), we add an ℓ_1 penalty to regularize the generalized score matching loss. One practical issue that is overlooked in Lin et al. (2016) is the fact that the score matching loss can be unbounded from below for a small tuning parameter, when the dimension m exceeds the sample size n. We fix this issue by amplifying the diagonal entries in the quadratic matrix in the definition of the generalized score matching loss by a factor/multiplier, and we give an upper bound on that multiplier that guarantees consistency.

As examples we consider *pairwise interaction power models* on the non-negative orthant $\mathbb{R}^m_+$. Specifically, the considered models are exponential families in which the log density is the sum of pairwise interactions between entries of powers $\boldsymbol{x}^a$ plus linearly weighted effects $\boldsymbol{x}^b$, or $\log(\boldsymbol{x})$ when $b = 0$. Our main interest is in the matrix of interaction parameters whose support determines the distributions' conditional independence graph. The considered framework covers truncated normal distributions ($a = b = 1$), exponential square root graphical models ($a = b = 1/2$) from Inouye et al. (2016), as well as a class of multivariate gamma distributions ($a = 1/2$, $b = 0$).

In the case of multivariate truncated normal distributions, where the conditional independence graph is given by the underlying Gaussian inverse covariance matrix, the sample size required for the consistency of our method using bounded $\boldsymbol{h}$ is $\Omega(d^2 \log m)$, where d is the degree of the graph. This matches the rates for Gaussian graphical models in Ravikumar

et al. (2011) and Lin et al. (2016). In contrast, the sample complexity for truncated Gaussian models given in Lin et al. (2016) is $\Omega(d^2 \log^8 m)$.

For the considered class of pairwise interaction models, we recommend using the function $\boldsymbol{h}$ with coordinates $h_j(x) = \min\left(x^{2-a}, c\right)$ for some moderately large c, or simply $h_j(x) = x^{2-a}$. While this choice is effective, it would be an interesting problem for future work to develop a method that adaptively chooses an optimized function $\boldsymbol{h}$ from data.

Chapter 3

GENERALIZED SCORE MATCHING FOR GENERAL DOMAIN

3.1 Introduction

In this chapter, we aim to further extend the generalized score matching loss to complicated and possibly unbounded domains $\mathcal{D}$ not limited to $\mathbb{R}^m$ and $\mathbb{R}^m_+$, and apply the method to estimation of graphical models supported on $\mathcal{D}$. In particular, we allow $\mathcal{D}$ to be any *componentwise countable* domain defined in Definition 5, which requires that the induced domain of one component x_j fixing all other $\boldsymbol{x}_{-j}$ is a union of at most countably many intervals in $\mathbb{R}$; this should cover most but pathological cases in practice. We propose composing the $\boldsymbol{h}$ function in the loss defined in Chapter 2 (Yu et al., 2019b) with a distance function $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_m) : \mathcal{D} \to \mathbb{R}^m_+$, with $\psi_j(\boldsymbol{x})$ the distance of x_j to its induced boundary while fixing the other components $\boldsymbol{x}_{-j}$, truncated from above by some user-selected constant C_j. We then use $(\boldsymbol{h} \circ \boldsymbol{\psi})$ in place of $\boldsymbol{h}$ and show that the generalized score matching loss can again be approximated by an empirical loss, which is quadratic in the canonical parameters for exponential family distributions.

Since the density corresponding to the minimizer of the expected (true) loss is equal to the true density a.e. $\boldsymbol{x} \in \mathcal{D}$, the method above only makes sense for $\mathcal{D}$ with positive Lebesgue measure in $\mathbb{R}^m$. As an important example of domains with zero Lebesgue measure in $\mathbb{R}^m$, we also present results for the m-simplex domain $\left\{\boldsymbol{x} \in \mathbb{R}^m_+ \middle| \ \boldsymbol{x} \succ \mathbf{0},\ \mathbf{1}_m^\top \boldsymbol{x} = 1\right\}$ by profiling out $x_m \equiv 1 - \mathbf{1}_{m-1}^\top \boldsymbol{x}_{-m}$. Notably, we revisit the A^{m-1} models for simplex domains introduced in Aitchison (1985), which corresponds to $a = b = 0$ with $\mathbf{K}\mathbf{1}_m = \mathbf{K}^\top \mathbf{1}_m = \mathbf{0}_m$ in the pairwise interaction power models introduced in Section 1.4. We show that estimation for models on the simplex can be done with little additional effort compared to general a-b models on $\mathcal{D}$ with positive Lebesgue measure.

In Chapter 2 we proved theoretical consistency results for high-dimensional settings requiring sample size $n = \Omega(\log m)$ for Gaussian graphical models (GGMs) restricted to $\mathbb{R}^m_+$. In this chapter we show that this sample complexity also holds for GGMs restricted to any domain $\mathcal{D}$ that is a finite disjoint union of convex sets, as well as general a-b models on bounded subsets of $\mathbb{R}^m_+$ with positive measure for $a \geq 0$ and on simplex domains for $a > 0$.

On unbounded domains for $a > 0$ and on simplex domains for $a = 0$ (notably, the A^{m-1} models) we require the sample size n to be $\Omega(\log m)$ times an unknown constant that may weakly depend on m.

The structure of the chapter is as follows. In Section 3.2 we provide an introduction to score matching and in Section 3.3 we detail our new methodology and theory necessary for the method to be usable in practice. In Section 3.4 we revisit the pairwise interaction power *a-b* models and the regularized generalized score matching estimator for exponential families, while in Sections 3.5 and 3.6 we focus on application of our method to graphical models on domains with positive Lebesgue measure and simplex domains, respectively. Theoretical results and numerical experiments are demonstrated in Sections 3.7 and 3.8, respectively. We apply our method to a DNA methylation dataset in Section 3.9. Longer proofs are included in Appendix B. An implementation that incorporates various types of domain $\mathcal{D}$ is available in the `genscore` R package.

3.2 Preliminaries

Suppose $\boldsymbol{X} \in \mathbb{R}^m$ is a random vector with distribution function P_0 supported on domain $\mathcal{D} \subseteq \mathbb{R}^m$ and a twice continuously differentiable probability density function p_0 with respect to the Lebesgue measure restricted to $\mathcal{D}$. Let $\mathcal{P}(\mathcal{D})$ be a family of distributions of interest with twice continuously differentiable densities on $\mathcal{D}$. The goal is to estimate p_0 by picking the distribution P from $\mathcal{P}(\mathcal{D})$ with density p that minimizes some empirical loss that measures the distance between p and p_0.

3.2.1 Original Score Matching on $\mathbb{R}^m$ (Hyvärinen, 2005)

The original *score matching* loss proposed by Hyvärinen (2005) for $\mathcal{D} \equiv \mathbb{R}^m$ is given by

$$J_{\mathbb{R}^m}(P) \equiv \frac{1}{2} \int_{\mathbb{R}^m} p_0(\boldsymbol{x}) \|\nabla \log p(\boldsymbol{x}) - \nabla \log p_0(\boldsymbol{x})\|_2^2 \, \mathrm{d}\boldsymbol{x}, \tag{3.1}$$

in which the gradients can be thought of as gradients with respect to a hypothetical location parameter and evaluated at the origin (Hyvärinen, 2005). The log densities enable estimation

without calculating the normalizing constants of p and p_0. Under mild conditions, using integration by parts the loss can be rewritten as

$$J_{\mathbb{R}^m}(P) \equiv \int_{\mathbb{R}^m} p_0(\boldsymbol{x}) \sum_{j=1}^{m} \left[\partial_{jj} \log p(\boldsymbol{x}) + \frac{1}{2} \left(\partial_j \log p(\boldsymbol{x}) \right)^2 \right] \mathrm{d}\boldsymbol{x}$$

plus a constant independent of p. One can thus use a sample average to approximate the loss without knowing the true density p_0.

3.2.2 Score Matching on $\mathbb{R}^m_+$ (Hyvärinen, 2007; Yu et al., 2018, 2019b)

Consider $\mathcal{D} \equiv \mathbb{R}^m_+$. Let $\boldsymbol{h} : \mathbb{R}^m_+ \to \mathbb{R}^m_+$, $\boldsymbol{x} \mapsto (h_1(x_1), \ldots, h_m(x_m))^\top$, where $h_1, \ldots, h_m : \mathbb{R}_+ \to \mathbb{R}_+$ are almost surely positive functions that are absolutely continuous in every bounded sub-interval of $\mathbb{R}_+$. The *generalized $\boldsymbol{h}$-score matching loss* proposed in Chapter 2 (Yu et al., 2018, 2019b) is

$$J_{\boldsymbol{h},\mathbb{R}^m_+}(P) \equiv \frac{1}{2} \int_{\mathbb{R}^m_+} p_0(\boldsymbol{x}) \left\| \nabla \log p(\boldsymbol{x}) \odot \boldsymbol{h}^{1/2}(\boldsymbol{x}) - \nabla \log p_0(\boldsymbol{x}) \odot \boldsymbol{h}^{1/2}(\boldsymbol{x}) \right\|_2^2 \mathrm{d}\boldsymbol{x}. \tag{3.2}$$

The score matching loss for $\mathbb{R}^m_+$ originally proposed by Hyvärinen (2007) is a special case of (3.2) with $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{x}^2$. In Chapter 2 (Yu et al., 2018, 2019b) we proved that by choosing $h_1, \ldots, h_m$ that grow slowly and preferably bounded the estimation efficiency can be significantly improved. Under assumptions that for all $P \in \mathcal{P}(\mathbb{R}^m_+)$ with density p,

(A0.1) $p_0(x_j; \boldsymbol{x}_{-j}) h_j(x_j) \partial_j \log p(x_j; \boldsymbol{x}_{-j}) \Big|_{x_j \nearrow +\infty}^{x_j \searrow 0^+} = 0, \quad \forall \boldsymbol{x}_{-j} \in \mathbb{R}^{m-1}_+ \ \forall j;$

(A0.2) $\mathbb{E}_{p_0} \left\| \nabla \log p(\boldsymbol{X}) \odot \boldsymbol{h}^{1/2}(\boldsymbol{X}) \right\|_2^2 < +\infty, \quad \mathbb{E}_{p_0} \left\| (\nabla \log p(\boldsymbol{X}) \odot \boldsymbol{h}(\boldsymbol{X}))' \right\|_1 < +\infty,$

where $f(\boldsymbol{x}) \Big|_{x_j \nearrow +\infty}^{x_j \searrow 0^+} \equiv \lim_{x_j \nearrow +\infty} f(\boldsymbol{x}) - \lim_{x_j \searrow 0^+} f(\boldsymbol{x})$, the loss (3.2) can be rewritten as

$$J_{\boldsymbol{h},\mathbb{R}^m_+}(P) \equiv \int_{\mathbb{R}^m_+} p_0(\boldsymbol{x}) \sum_{j=1}^{m} \Bigg[h_j'(x_j) \partial_j (\log p(\boldsymbol{x})) + h_j(x_j) \partial_{jj} (\log p(\boldsymbol{x})) + \frac{1}{2} h_j(x_j) [\partial_j (\log p(\boldsymbol{x}))]^2 \Bigg] \mathrm{d}\boldsymbol{x} \tag{3.3}$$

plus a constant independent of p. One can thus use the minimizer of the empirical loss (3.3) as an estimator of p_0.

3.2.3 Truncated Score Matching on Bounded Open Subsets of $\mathbb{R}^m$ (Liu and Kanamori, 2019)

Liu and Kanamori (2019) approach the problem of density estimation on a bounded open subset $\mathcal{D} \subsetneq \mathbb{R}^m$ with a piecewise smooth boundary $\partial\mathcal{D}$ by minimizing the following "maximally weighted score matching" loss

$$J_{g_0,\mathcal{D}}(P) \equiv \sup_{g\in\mathcal{G}} \frac{1}{2} \int_{\mathbb{R}^m_+} g(\boldsymbol{x}) p_0(\boldsymbol{x}) \left\| \nabla \log p(\boldsymbol{x}) - \nabla \log p_0(\boldsymbol{x}) \right\|_2^2 \mathrm{d}\boldsymbol{x}. \tag{3.4}$$

with $\mathcal{G} \equiv \{g | g(\boldsymbol{x}) = 0, \forall \boldsymbol{x} \in \partial\mathcal{D}$ and g is L-Lipschitz continuous$\}$ for some constant $L > 0$. They prove that the maximum is obtained with $g_0(\boldsymbol{x}) \equiv L \inf_{\boldsymbol{x}'\in\partial\mathcal{D}} \|\boldsymbol{x} - \boldsymbol{x}'\|_2$, i.e. the ℓ_2 distance of $\boldsymbol{x}$ to the boundary of $\mathcal{D}$, and with integration by parts similar to the previous methods, (3.4) can be estimated using the empirical loss which can be calculated with a closed form.

3.3 Generalized Score Matching for General Domains

3.3.1 Assumption on the Domain

For any index $j = 1, \ldots, m$, write $\mathcal{C}_{j,\mathcal{D}}(\boldsymbol{x}_{-j}) \equiv \{y \in \mathbb{R} : (y; \boldsymbol{x}_{-j}) \in \mathcal{D}\}$ as a set-valued function of values $\boldsymbol{x}_{-j}$ of the other components. This is the j-th coordinate of the intersection between $\mathcal{D}$ and the line $\{(y; \boldsymbol{x}_{-j}) : y \in \mathbb{R}\}$. Also write $\mathcal{S}_{-j,\mathcal{D}} \equiv \{\boldsymbol{x}_{-j} : \mathcal{C}_{j,\mathcal{D}}(\boldsymbol{x}_{-j}) \neq \varnothing\} \subset \mathbb{R}^{m-1}$, a projection of $\mathcal{D}$ onto $\mathbb{R}^{m-1}$. For notational simplicity we may drop the dependency of $\mathcal{C}_j$ and $\mathcal{S}_{-j}$ on $\mathcal{D}$.

Definition 5. *We call a domain $\mathcal{D} \subseteq \mathbb{R}^m$* componentwise countable *if for any index $j = 1, \ldots, m$ and any $\boldsymbol{x}_{-j} \in \mathcal{S}_{-j,\mathcal{D}}$, we can write*

$$\mathcal{C}_{j,\mathcal{D}}(\boldsymbol{x}_{-j}) \equiv \bigcup_{k=1}^{K_j(\boldsymbol{x}_{-j})} \left[a_{k,j}(\boldsymbol{x}_{-j}), b_{k,j}(\boldsymbol{x}_{-j})\right], \tag{3.5}$$

where $1 \leq K_j \leq +\infty$, $-\infty \leq a_{k,j} \leq b_{k,j} \leq +\infty$ for all $1 \leq k \leq K_j$, and further if the boundary set

$$\partial\mathcal{D} \equiv \left\{ \boldsymbol{x} \in \mathcal{D} : \exists j = 1, \ldots, m,\ x_j \in \bigcup_{k=1}^{K_j(\boldsymbol{x}_{-j})} \{a_{k,j}(\boldsymbol{x}_{-j}), b_{k,j}(\boldsymbol{x}_{-j})\} \backslash \{\pm\infty\} \right\} \tag{3.6}$$

is a Lebesgue-null set in $\mathbb{R}^m$; here each bounded interval in (3.5) can be open/closed/half-open half-closed and must be non-overlapping.

That is, for any j and any $\boldsymbol{x}_{-j}$ such that there exists some x_j with $(x_j; \boldsymbol{x}_{-j}) \in \mathcal{D}$, the intersection between $\mathcal{D}$ and the line $\{(y_j; \boldsymbol{x}_{-j}) : \boldsymbol{y} \in \mathbb{R}\}$, when projected to the j-th dimension, is a union of at most countably many disjoint (bounded or unbounded) intervals in $\mathbb{R}$, and its "boundary points" must form a Lebesgue-null set. Note that $K_j(\boldsymbol{x}_{-j})$ cannot be 0 since $\mathcal{C}_{j,\mathcal{D}}(\boldsymbol{x}_{-j}) \neq \varnothing$ by definition of $\mathcal{S}_{-j,\mathcal{D}}$.

3.3.2 Generalized Score Matching Loss for General Domains

We first define the notion of the *truncated componentwise distance*.

Definition 6. *Fix positive constants $\boldsymbol{C} \succ \boldsymbol{0}_m$. Given a non-empty* componentwise countable *domain $\mathcal{D} \subseteq \mathbb{R}^m$ as in Definition 5 and a vector $\boldsymbol{x} \in \mathcal{D}$, define the* truncated componentwise distance of $\boldsymbol{x}$ to the boundary of $\mathcal{D}$ *as*

$$\boldsymbol{\varphi}_{\boldsymbol{C},\mathcal{D}}(\boldsymbol{x}) \equiv (\varphi_{C_1,\mathcal{D},1}(\boldsymbol{x}), \ldots, \varphi_{C_m,\mathcal{D},m}(\boldsymbol{x})) \in \mathbb{R}^m_+, \tag{3.7}$$

$$\varphi_{C_j,\mathcal{D},j}(\boldsymbol{x}) \equiv \begin{cases} C_j, & \text{if } a_{k,j} = -\infty \text{ and } b_{k,j} = +\infty, \\ \min(C_j, b_{k,j} - x_j), & \text{if } a_{k,j} = -\infty \text{ and } x_j \leq b_{k,j} < +\infty, \\ \min(C_j, x_j - a_{k,j}, b_{k,j} - x_j) & \text{if } -\infty < a_{k,j} \leq x_j \leq b_{k,j} < +\infty, \\ \min(C_j, x_j - a_{k,j}) & \text{if } -\infty < a_{k,j} \leq x_j \text{ and } b_{k,j} = +\infty, \end{cases} \tag{3.8}$$

where k is such that $a_{k,j} \leq x_j \leq b_{k,j}$ as in (3.5).

We thus propose the *generalized $(\boldsymbol{h}, \boldsymbol{C}, \mathcal{D})$-score matching loss* for a general domain $\mathcal{D}$ similar to (3.2) with $\boldsymbol{h}$ applied to $\boldsymbol{\varphi}_{\boldsymbol{C},\mathcal{D}}(\boldsymbol{x})$ instead of to $\boldsymbol{x}$:

Definition 7. *Suppose the true distribution P_0 has a twice continuously differentiable density p_0 supported on $\mathcal{D} \subseteq \mathbb{R}^m$, a non-empty componentwise countable domain (Definition 5). Given positive constants $\boldsymbol{C} \succ \boldsymbol{0}$, and $\boldsymbol{h} : \mathbb{R}^m_+ \to \mathbb{R}^m_+$, $\boldsymbol{y} \mapsto (h_1(y_1), \ldots, h_m(y_m))$ with*

$h_1, \ldots, h_m : \mathbb{R}_+ \to \mathbb{R}_+$, *the* generalized $(\boldsymbol{h}, \boldsymbol{C}, \mathcal{D})$-score matching loss *in* $P \in \mathcal{P}(\mathcal{D})$ *with density* p *is defined as*

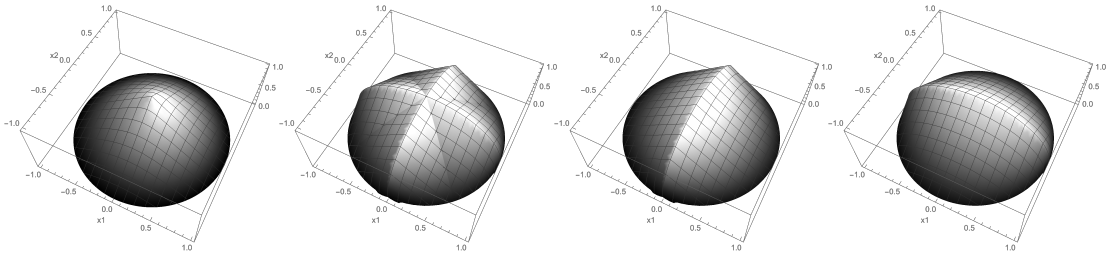
$$J_{\boldsymbol{h},\boldsymbol{C},\mathcal{D}}(P) \equiv \frac{1}{2} \int_{\mathcal{D}} p_0(\boldsymbol{x}) \Big\| \nabla \log p(\boldsymbol{x}) \odot \left(\boldsymbol{h} \circ \boldsymbol{\varphi}_{\boldsymbol{C},\mathcal{D}}\right)^{1/2}(\boldsymbol{x}) - \nabla \log p_0(\boldsymbol{x}) \odot \left(\boldsymbol{h} \circ \boldsymbol{\varphi}_{\boldsymbol{C},\mathcal{D}}\right)^{1/2}(\boldsymbol{x}) \Big\|_2^2 \mathrm{d}\boldsymbol{x}. \quad (3.9)$$

3.3.3 Comparison to Previous Work

The intuition of (3.9) is applying the loss from (3.2) to $(\boldsymbol{h} \circ \boldsymbol{\varphi}_{\boldsymbol{C},\mathcal{D}})$ in place of $\boldsymbol{h}$. Thus, the generalized score matching loss (3.2) in Chapter 2 becomes a special case with $\mathcal{D} = \mathbb{R}^m_+$ and $\boldsymbol{C} = +\infty^m$, since $\boldsymbol{\varphi}_{+\infty^m,\mathbb{R}^m_+}(\boldsymbol{x}) = \boldsymbol{x}$. In Chapter 2 we suggested using an $\boldsymbol{h}$ with each component a bounded function, which is now incorporated into the definition of $\boldsymbol{\varphi}$ with finite truncation points $\boldsymbol{C}$. When $\boldsymbol{h}(\boldsymbol{x}) = \mathbf{1}_m$, $(\boldsymbol{h} \circ \boldsymbol{\varphi}_{\boldsymbol{C},\mathcal{D}})^{1/2}(\boldsymbol{x}) \equiv \mathbf{1}_m$ and it corresponds to the original score-matching for $\mathbb{R}^m$ in Hyvärinen (2005). If $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{x}^2$, $\boldsymbol{C} = +\infty^m$ and $\mathcal{D} \equiv \mathbb{R}^m_+$, $(\boldsymbol{h} \circ \boldsymbol{\varphi}_{\boldsymbol{C},\mathcal{D}})^{1/2}(\boldsymbol{x}) \equiv x^2$ corresponds to the estimator in Hyvärinen (2007) and Lin et al. (2016).

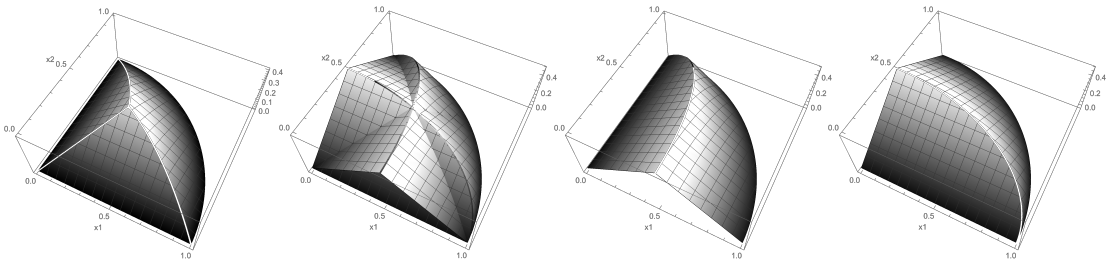
The $\boldsymbol{\varphi}_{\boldsymbol{C},\mathcal{D}}$ function transforms an $\boldsymbol{x} \in \mathcal{D}$ into the component-wise distance vector taking values in $\mathbb{R}^m_+$, and the rest remains the same as in Chapter 2. It is thus a natural extension to the work and requires little extra work to be done as $\boldsymbol{\varphi}_{\boldsymbol{C},\mathcal{D}}$ usually has closed-form solution and can be computationally trivial.

We note that for a bounded $\mathcal{D}$, our approach is different than using a uniform $g_0(\boldsymbol{x})$ the ℓ_2 distance of $\boldsymbol{x}$ to the boundary of $\mathcal{D}$ as in Liu and Kanamori (2019), in that we decompose the distance for each component and apply an extra $\boldsymbol{h}$ function. In Figure 3.1 we illustrate the two components of our $\boldsymbol{\varphi}$ function on the 2-d unit disk $x_1^2 + x_2^2 \leq 1$, considered below in Example 3.8, compared to the g_0 function in Liu and Kanamori (2019). While $g_0(\boldsymbol{x}) = 1 - \sqrt{x_1^2 + x_2^2}$, choosing $\boldsymbol{C} \succeq \mathbf{1}_2$, we have $\varphi_1(\boldsymbol{x}) = \sqrt{1 - x_2^2} - |x_1|$ and $\varphi_2(\boldsymbol{x}) = \sqrt{1 - x_1^2} - |x_2|$. Similarly, in Figure 3.2 we plot these functions for the 2-d unit disk restricted to $\mathbb{R}^2_+$ considered in Example 3.9, where $\varphi_1(\boldsymbol{x}) = \min\left\{x_1, \sqrt{1 - x_2^2} - x_1\right\}$, $\varphi_2(\boldsymbol{x}) = \min\left\{x_2, \sqrt{1 - x_1^2} - x_2\right\}$, $g_0(\boldsymbol{x}) = \min\left\{x_1, x_2, 1 - \sqrt{x_1^2 + x_2^2}\right\}$.



(a) g_0 (b) $\varphi_{\mathbf{1}_2,\mathcal{D},1}$ & $\varphi_{\mathbf{1}_2,\mathcal{D},2}$ (c) $\varphi_{\mathbf{1}_2,\mathcal{D},1}$ (d) $\varphi_{\mathbf{1}_2,\mathcal{D},2}$

Figure 3.1: Comparison of g_0, $\varphi_{\mathbf{1}_2,\mathcal{D},1}$ and $\varphi_{\mathbf{1}_2,\mathcal{D},2}$ on $\mathcal{D} \equiv \{\boldsymbol{x} \in \mathbb{R}^2 : \|\boldsymbol{x}\|_2 < 1\}$.



(a) g_0 (b) $\varphi_{\mathbf{1}_2,\mathcal{D},1}$ & $\varphi_{\mathbf{1}_2,\mathcal{D},2}$ (c) $\varphi_{\mathbf{1}_2,\mathcal{D},1}$ (d) $\varphi_{\mathbf{1}_2,\mathcal{D},2}$

Figure 3.2: Comparison of g_0, $\varphi_{\mathbf{1}_2,\mathcal{D},1}$ and $\varphi_{\mathbf{1}_2,\mathcal{D},2}$ on $\mathcal{D} \equiv \{\boldsymbol{x} \in \mathbb{R}^2_+ : \|\boldsymbol{x}\|_2 < 1\}$.

3.3.4 Form of the Truncated Componentwise Distance $\boldsymbol{\varphi}$

Example 3.6 ($\mathbb{R}^m$ and $\mathbb{R}^m_+$)**.** *$\boldsymbol{\varphi}_{\boldsymbol{C},\mathbb{R}^m}(\boldsymbol{x}) = \boldsymbol{C}$ and $\boldsymbol{\varphi}_{\boldsymbol{C},\mathbb{R}^m_+}(\boldsymbol{x}) = (\min(C_1, x_1), \ldots, \min(C_m, x_m))$ by definition.*

Example 3.7 (Unit hypercube)**.** *Consider the unit hypercube $\mathcal{D}$ as the convex hull of 2^m points $\{\boldsymbol{x} \in \mathbb{R}^m : |x_1| = \cdots = |x_m| = 1/2\}$. The domain of x_j is $[-1/2, 1/2]$ independent of the other components and so $\varphi_{C_j,\mathcal{D},j}(\boldsymbol{x}) = \min\{C_j, 1/2 - |x_j|\}$. Since the hypercube is bounded by nature, it is natural to drop the truncation by C_j and simply use $\boldsymbol{\varphi}_{\mathcal{D}}(\boldsymbol{x}) = \mathbf{1}_m/2 - |\boldsymbol{x}|$.*

Example 3.8 ($\mathcal{L}^q$ balls)**.** *Consider the $\mathcal{L}^q$ ball with radius $r > 0$ and $q \geq 1$ defined by $\mathcal{D} \equiv \{\boldsymbol{x} \in \mathbb{R}^m : \|\boldsymbol{x}\|_q \leq r\}$. Given a point $\boldsymbol{x} \in \mathcal{D}$ and for $j = 1, \ldots, m$, the domain for x_j while fixing $\boldsymbol{x}_{-j}$ is simply $[-(r^q - \mathbf{1}^\top |\boldsymbol{x}_{-j}|^q)^{1/q}, (r^q - \mathbf{1}^\top |\boldsymbol{x}_{-j}|^q)^{1/q}]$, and so $\varphi_{C_j,\mathcal{D},j}(\boldsymbol{x}) = \min\left\{C_j, (r^q - \mathbf{1}^\top |\boldsymbol{x}_{-j}|^q)^{1/q} - |x_j|\right\}$.*

Example 3.9 ($\mathcal{L}^q$ balls restricted to $\mathbb{R}^m_+$)**.** *Consider the intersection between $\mathbb{R}^m_+$ and the $\mathcal{L}^q$ ball with radius $r > 0$ and $q \geq 1$ defined by $\mathcal{D} \equiv \{\boldsymbol{x} \in \mathbb{R}^m_+ : \|\boldsymbol{x}\|_q \leq r\}$. Given a point $\boldsymbol{x} \in \mathcal{D}$ and for $j = 1, \ldots, m$, the domain for x_j while fixing $\boldsymbol{x}_{-j}$ is simply $[0, (r^q - \mathbf{1}^\top |\boldsymbol{x}_{-j}|^q)^{1/q}]$, and so $\varphi_{C_j,\mathcal{D},j}(\boldsymbol{x}) = \min\left\{C_j, x_j, (r^q - \mathbf{1}^\top |\boldsymbol{x}_{-j}|^q)^{1/q} - x_j\right\}$.*

Example 3.10 (Complement of $\mathcal{L}^q$ balls)**.** *Consider the complement of the $\mathcal{L}^q$ ball with radius $r > 0$ and $q \geq 1$ defined by $\mathcal{D} \equiv \{\boldsymbol{x} \in \mathbb{R}^m : \|\boldsymbol{x}\|_q > r\}$. Given an $\boldsymbol{x} \in \mathcal{D}$ and j, the domain for x_j while fixing $\boldsymbol{x}_{-j}$ is $\mathbb{R}$ if $\mathbf{1}^\top_{m-1} |\boldsymbol{x}_{-j}|^q > r^q$, or $\left(-\infty, -(r^q - \mathbf{1}^\top_{m-1} |\boldsymbol{x}_{-j}|^q)^{1/q}\right) \cup \left((r^q - \mathbf{1}^\top_{m-1} |\boldsymbol{x}_{-j}|^q)^{1/q}, +\infty\right)$ otherwise. So*

$$\varphi_{C_j,\mathcal{D},j}(\boldsymbol{x}) = \begin{cases} C_j & \text{if } \mathbf{1}^\top_{m-1} |\boldsymbol{x}_{-j}|^q > r^q, \\ \min\left\{C_j, |x_j| - (r^q - \mathbf{1}^\top_{m-1} |\boldsymbol{x}_{-j}|^q)^{1/q}\right\} & \text{otherwise.} \end{cases}$$

Example 3.11 (Complement of $\mathcal{L}^q$ balls restricted to $\mathbb{R}^m_+$)**.** *Consider the intersection between $\mathbb{R}^m_+$ and the complement of the $\mathcal{L}^q$ ball with radius $r > 0$ and $q \geq 1$: $\mathcal{D} \equiv \{\boldsymbol{x} \in \mathbb{R}^m_+ : \|\boldsymbol{x}\|_q > r\}$. Given an $\boldsymbol{x} \in \mathcal{D}$ and j, the domain for x_j while fixing $\boldsymbol{x}_{-j}$ is $\mathbb{R}_+$ if $\mathbf{1}^\top_{m-1} |\boldsymbol{x}_{-j}|^q > r^q$, or*

$\big((r^q - \mathbf{1}_{m-1}^\top |\boldsymbol{x}_{-j}|^q)^{1/q}, +\infty\big)$ *otherwise. So*

$$\varphi_{C_j,\mathcal{D},j}(\boldsymbol{x}) = \begin{cases} \min\{C_j, x_j\} & \text{if } \mathbf{1}_{m-1}^\top |\boldsymbol{x}_{-j}|^q > r^q, \\ \min\left\{C_j, x_j - (r^q - \mathbf{1}_{m-1}^\top |\boldsymbol{x}_{-j}|^q)^{1/q}\right\} & \text{otherwise.} \end{cases}$$

Example 3.12 (Complicated domains defined by inequality constraints)**.** *A domain $\mathcal{D}$ may be determined by a series of intersections/unions of regions determined by inequality constraints, e.g. $\mathcal{D} = \{\boldsymbol{x} \in \mathbb{R}^m : (f_1(\boldsymbol{x}) \le c_1 \wedge f_2(\boldsymbol{x}) \le c_2) \vee f_3(\boldsymbol{x}) \ge c_3\}$. In this case, for example, to calculate $\boldsymbol{\varphi}_{\boldsymbol{C},\mathcal{D}}$ one plug in $\boldsymbol{x}_{-j}$ as given and solve numerically $f_i(x_j; \boldsymbol{x}_{-j}) = c_i$ for $i = 1, 2, 3$, and obtain the boundary points for x_j using simple algorithms for interval unions/intersections. This is implemented in the package* `genscore` *for some types of polynomial f_i and arbitrary intersections/unions.*

Example 3.13 (Standard Simplices)**.** *Consider $\mathcal{D} \equiv \left\{\boldsymbol{x} \in \mathbb{R}_+^m \middle| \boldsymbol{x} \succ \mathbf{0},\, \mathbf{1}_m^\top \boldsymbol{x} = 1\right\}$, the standard simplex. Since this has zero Lebesgue measure in $\mathbb{R}^m$, we drop the last coordinate and consider instead $\mathcal{D}_{-m} \equiv \left\{\boldsymbol{x}_{-m} \in \mathbb{R}_+^{m-1} \middle| \boldsymbol{x}_{-m} \succ \mathbf{0},\, \mathbf{1}_{m-1}^\top \boldsymbol{x}_{-m} < 1\right\}$ as a set in $\mathbb{R}^{m-1}$. For $j = 1, \ldots, m-1$, the domain of x_j given $\boldsymbol{x}_{-j,-m}$ for $\boldsymbol{x}_{-m} \in \mathcal{D}_{-m}$ is thus $\big(0, 1 - \mathbf{1}_{m-2}^\top \boldsymbol{x}_{-j,-m}\big)$, and so $\varphi_{C_j,\mathcal{D}_{-m},j}(\boldsymbol{x}) = \min\left\{C_j, x_j, 1 - \mathbf{1}_{m-1}^\top \boldsymbol{x}_{-m}\right\} = \min\{C_j, x_j, x_m\}$. In practice since the standard simplex is naturally bounded by the unit cube, it is natural to drop the truncation by C_j and simply use $\varphi_{\mathcal{D}_{-m},j}(\boldsymbol{x}) = \min\left\{x_j, 1 - \mathbf{1}_{m-1}^\top \boldsymbol{x}_{-m}\right\} = \min\{x_j, x_m\}$.*

3.3.5 The Empirical Generalized Score Matching Loss

From this section, we drop the dependence of $\boldsymbol{\varphi}$ on $\boldsymbol{C}$ and $\mathcal{D}$ in notation for simplicity.

Lemma 12. *Suppose $\boldsymbol{C} \succ \mathbf{0}$, $p_0(\boldsymbol{x}) > 0$ for almost every $\boldsymbol{x} \in \mathcal{D}$ and $h_1(y), \ldots, h_m(y) > 0$ for all $y > 0$. Then $J_{\boldsymbol{h},\boldsymbol{C},\mathcal{D}}(P) = 0$ if and only if $p_0 = p$ for a.e. $\boldsymbol{x} \in \mathcal{D}$.*

Proof of Lemma 12. By the assumption on $\mathcal{D}$ and definition of $\boldsymbol{\varphi}$, $\boldsymbol{\varphi}(\boldsymbol{x}) \succ \mathbf{0}$ for almost every $\boldsymbol{x} \in \mathcal{D}$, and thus $(\boldsymbol{h} \circ \boldsymbol{\varphi})(\boldsymbol{x}) \succ \mathbf{0}$ for a.e. $\boldsymbol{x} \in \mathcal{D}$. So we have $J_{\boldsymbol{h},\boldsymbol{C},\mathcal{D}}(P) = 0$ if and only if $\nabla \log p_0(\boldsymbol{x}) = \nabla \log p(\boldsymbol{x})$ for a.e. $\boldsymbol{x} \in \mathcal{D}$, i.e. $\log p_0(\boldsymbol{x}) = \log p(\boldsymbol{x}) + c_0$ for a.e. $\boldsymbol{x} \in \mathcal{D}$ for some constant c_0, or $p_0(\boldsymbol{x}) = c_1 \cdot p(\boldsymbol{x})$ for a.e. $\boldsymbol{x} \in \mathcal{D}$ for some non-zero constant $c_1 \equiv \exp(c_0)$. Since p_0 and p both integrate to 1 over $\mathcal{D}$, we have $c_1 = 1$ and $p_0 = p$ for a.e. $\boldsymbol{x} \in \mathcal{D}$. $\square$

Lemma 13. *Similar to (A0.1) and (A0.2) in Section 3.2.2, assume the following assumptions hold for all $p \in \mathcal{P}(\mathcal{D})$,*

(A.1) $p_0(x_j; \boldsymbol{x}_{-j}) h_j(\varphi_j(\boldsymbol{x})) \partial_j \log p(x_j; \boldsymbol{x}_{-j}) \Big|_{x_j \nearrow b_k(\boldsymbol{x}_{-j})^-}^{x_j \searrow a_k(\boldsymbol{x}_{-j})^+} = 0$ *for all* $k = 1, \ldots, K_j(\boldsymbol{x}_{-j})$ *and* $\boldsymbol{x}_{-j} \in \mathcal{S}_{-j}$ *for all* j*;*

(A.2) $\int_{\mathcal{D}} p_0(\boldsymbol{x}) \left\| \nabla \log p(\boldsymbol{x}) \odot (\boldsymbol{h} \circ \boldsymbol{\varphi})^{1/2}(\boldsymbol{x}) \right\|_2^2 \mathrm{d}\boldsymbol{x} < +\infty$,
$\int_{\mathcal{D}} p_0(\boldsymbol{x}) \left\| [\nabla \log p(\boldsymbol{x}) \odot (\boldsymbol{h} \circ \boldsymbol{\varphi})(\boldsymbol{x})]' \right\|_1 \mathrm{d}\boldsymbol{x} < +\infty$.

Also assume that

(A.3) $\forall j = 1, \ldots, m$ *and a.e.* $\boldsymbol{x}_{-j} \in \mathcal{S}_{-j}$, h_j *is absolutely continuous in any bounded subinterval of* $\mathcal{C}_j(\boldsymbol{x}_{-j})$.

(This implies the same for $(h_j \circ \varphi_j)$ *and also that* $\partial_j (h_j \circ \varphi_j)$ *exists a.e.) Then*

$$J_{\boldsymbol{h},\boldsymbol{C},\mathcal{D}}(P) \equiv \frac{1}{2} \sum_{j=1}^{m} \int_{\mathcal{D}} p_0(\boldsymbol{x}) \cdot (h_j \circ \varphi_j)(\boldsymbol{x}) \cdot [\partial_j \log p(\boldsymbol{x})]^2 \, \mathrm{d}\boldsymbol{x} + \sum_{j=1}^{m} \int_{\mathcal{D}} p_0(\boldsymbol{x}) \cdot \partial_j \left[(h_j \circ \varphi_j)(\boldsymbol{x}) \cdot \partial_j \log p(\boldsymbol{x}) \right] \mathrm{d}\boldsymbol{x} \quad (3.10)$$

plus a constant depending on p_0 *only and independent of* p.

The proof of the lemma is in Appendix. The lemma enables us to estimate the population loss using the empirical loss

$$\hat{J}_{\boldsymbol{h},\boldsymbol{C},\mathcal{D}}(P) = \frac{1}{2} \sum_{j=1}^{m} \sum_{i=1}^{n} \frac{1}{2} (h_j \circ \varphi_j) \left(\boldsymbol{x}^{(i)} \right) \cdot \left[\partial_j \log p \left(\boldsymbol{x}^{(i)} \right) \right]^2 + \partial_j \left[(h_j \circ \varphi_j) \left(\boldsymbol{x}^{(i)} \right) \cdot \partial_j \log p \left(\boldsymbol{x}^{(i)} \right) \right]. \quad (3.11)$$

As the canonical choices of $\boldsymbol{h}$ are power functions in $\boldsymbol{x}$, we give the following sufficient conditions for the assumptions in the lemma.

Proposition 14. *Suppose for all* $j = 1, \ldots, m$, $h_j(x_j) = x_j^{\alpha_j}$ *for some* $\alpha_j > 0$. *Suppose in addition that for all* j *and* $\boldsymbol{x}_{-j} \in \mathcal{S}_{-j}$ *and all* $p \in \mathcal{P}$ *we have*

(1) $p_0(x_j; \boldsymbol{x}_{-j})\partial_j \log p(x_j; \boldsymbol{x}_{-j}) = o\left(1/(x_j - c_{k,j})^\alpha\right)$ *as* $x_j \nearrow c_{k,j} \equiv b_{k,j} \neq +\infty$ *or as* $x_j \searrow c_{k,j} \equiv a_{k,j} \neq -\infty$ *for all* k*, and*

(2) $p_0(x_j; \boldsymbol{x}_{-j})\partial_j \log p(x_j; \boldsymbol{x}_{-j}) \to 0$ *as* $x_j \nearrow +\infty$ *if* $\mathcal{C}_j(\boldsymbol{x}_{-j})$ *is unbounded from above, and as* $x_j \searrow -\infty$ *if* $\mathcal{C}_j(\boldsymbol{x}_{-j})$ *is unbounded from below.*

Then (A.1) and (A.3) are satisfied.

Proof of Proposition 14. (A.3) is satisfied by the property of h_j. (A.1) is trivial since by construction $(h_j \circ \varphi_j)(\boldsymbol{x})$ becomes $|x_j - c_{k,j}|^\alpha$ as $x_j \to c_{k,j} \in \cup_{k=1}^{K_j}\{a_{k,j}, b_{k,j}\}$ and also $(h_j \circ \varphi_j)$ is bounded by C^α as $x_j \nearrow +\infty$ or $x_j \searrow -\infty$, if applicable. □

According to Lemma 12 the method only makes sense if the domain $\mathcal{D}$ has a positive Lebesgue measure in $\mathbb{R}^m$. However, for some Lebesgue null sets one may apply appropriate transforms under which $\mathcal{D}$ becomes a set with positive Lebesgue measure in lower dimensions with dimensionality reduction. One important example is the $(m-1)$-dimensional standard simplex $\left\{\boldsymbol{x} \in \mathbb{R}_+^m \middle| \boldsymbol{x} \succ \mathbf{0},\, \mathbf{1}_m^\top \boldsymbol{x} = 1\right\}$, a null set since it is a subset of the $(m-1)$-dimensional hyperplane $\{\mathbf{1}_m^\top \boldsymbol{x} = 1\}$. Simply dropping one of the coordinate, e.g. x_m, and substituting $x_m = 1 - \mathbf{1}_{m-1}^\top \boldsymbol{x}_{-m}$ in p_0 and p we can thus apply the method to $\left\{\boldsymbol{x}_{-m} \in \mathbb{R}_+^{m-1} \middle| \boldsymbol{x}_{-m} \succ \mathbf{0},\, \mathbf{1}_{m-1}^\top \boldsymbol{x}_{-m} < 1\right\}$, a set with positive measure in $\mathbb{R}^{m-1}$. We further discuss this example in Section 3.6.

3.3.6 Extension of g_0 in Liu and Kanamori (2019)

We extend the method in Liu and Kanamori (2019) introduced in Section 3.2.3 from bounded domains to any *componentwise countable* domain $\mathcal{D}$ defined in Definition 5 by defining the loss of the same form as (3.4):

$$J_{g_0,C,\mathcal{D}}(P) \equiv \sup_{g \in \mathcal{G}} \frac{1}{2} \int_{\mathbb{R}_+^m} g(\boldsymbol{x}) p_0(\boldsymbol{x}) \left\| \nabla \log p(\boldsymbol{x}) - \nabla \log p_0(\boldsymbol{x}) \right\|_2^2 \mathrm{d}\boldsymbol{x},$$

but with $\mathcal{G} \equiv \{g | g(\boldsymbol{x}) = 0, \forall \boldsymbol{x} \in \partial\mathcal{D},\, g$ is L-Lipschitz continuous and $g \leq C\}$ instead, for the same Lipschitz constant $L > 0$ but with an extra truncation constant $C > 0$. Here the

boundary set $\partial\mathcal{D}$ is as defined in (3.6). Following the same proof as for their Proposition 1 and dropping the Lipschitz constant L by replacing C with C/L (or equivalently choosing $L = 1$ without loss of generality), it is easy to see that the maximum is obtained at

$$g_0(\boldsymbol{x}) \equiv \min\left\{C, \inf_{\boldsymbol{x}' \in \partial\mathcal{D}} \|\boldsymbol{x} - \boldsymbol{x}'\|_2\right\}, \tag{3.12}$$

the ℓ_2 distance of $\boldsymbol{x}$ to $\partial\mathcal{D}$ truncated above by C, which naturally extends the method in Liu and Kanamori (2019) to unbounded domains. In the special case where $\partial\mathcal{D} = \varnothing$, we must have $\mathcal{D} \equiv \mathbb{R}^m$, and (3.12) becomes constant C by the convention $\inf \varnothing = +\infty$, which coincides with the original score matching in Hyvärinen (2005).

Assuming assumptions (A.1) and (A.2) in Lemma 13 hold when replacing $(\boldsymbol{h} \circ \boldsymbol{\varphi})(\boldsymbol{x})$ by $g_0(\boldsymbol{x})\mathbf{1}_m$ and $h_j(\varphi_j(\boldsymbol{x}))$ by $g_0(\boldsymbol{x})$, the same conclusion there applies, i.e.

$$J_{g_0,C,\mathcal{D}}(P) \equiv \frac{1}{2}\sum_{j=1}^{m}\int_{\mathcal{D}} p_0(\boldsymbol{x})g_0(\boldsymbol{x})\left[\partial_j \log p(\boldsymbol{x})\right]^2 \mathrm{d}\boldsymbol{x} + \sum_{j=1}^{m}\int_{\mathcal{D}} p_0(\boldsymbol{x})\partial_j\left[g_0(\boldsymbol{x})\partial_j \log p(\boldsymbol{x})\right]\mathrm{d}\boldsymbol{x}$$

plus a constant depending on p_0 only and independent of p, the same empirical loss as (13) in Liu and Kanamori (2019) with an additional truncation from above applied to g_0. The proof is in the same spirit as that for Lemma 13 and is thus omitted.

3.4 Exponential Families and a-b Models

3.4.1 Exponential Families

Consider the case where $\mathcal{P}(\mathcal{D}) \equiv \{p_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^r\}$ for some $r = 1, 2, \ldots$ is an exponential family with continuous distributions supported on $\mathcal{D}$ with densities of the form

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{\theta}^\top \boldsymbol{t}(\boldsymbol{x}) - \psi(\boldsymbol{\theta}) + b(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathcal{D}, \tag{3.13}$$

where $\boldsymbol{\theta} \in \mathbb{R}^r$ is the unknown canonical parameter of interest, $\boldsymbol{t}(\boldsymbol{x}) \in \mathbb{R}^r$ are the sufficient statistics, $\psi(\boldsymbol{\theta})$ is the normalizing constant, and $b(\boldsymbol{x})$ is the base measure. The empirical loss $\hat{J}_{\boldsymbol{h},\boldsymbol{C},\mathcal{D}}$ (3.11) can then be written as a quadratic function in the canonical parameter:

$$\hat{J}_{\boldsymbol{h},\boldsymbol{C},\mathcal{D}}(p_{\boldsymbol{\theta}}) = \frac{1}{2}\boldsymbol{\theta}^\top \boldsymbol{\Gamma}(\mathbf{x})\boldsymbol{\theta} - \boldsymbol{g}(\mathbf{x})^\top \boldsymbol{\theta} + \text{const}, \quad \text{with} \tag{3.14}$$

$$\boldsymbol{\Gamma}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}(h_j \circ \varphi_j)\left(\boldsymbol{X}^{(i)}\right)\partial_j \boldsymbol{t}(\boldsymbol{X}^{(i)})\left(\partial_j \boldsymbol{t}\left(\boldsymbol{X}^{(i)}\right)\right)^\top \quad \text{and} \tag{3.15}$$

$$\begin{aligned}\boldsymbol{g}(\mathbf{x}) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\big[&(h_j \circ \varphi_j)\left(\boldsymbol{X}^{(i)}\right)\partial_j b\left(\boldsymbol{X}^{(i)}\right)\partial_j \boldsymbol{t}\left(\boldsymbol{X}^{(i)}\right) \\ &+(h_j \circ \varphi_j)\left(\boldsymbol{X}^{(i)}\right)\partial_{jj}\boldsymbol{t}\left(\boldsymbol{X}^{(i)}\right) + \partial_j(h_j \circ \varphi_j)\left(\boldsymbol{X}^{(i)}\right)\partial_j \boldsymbol{t}\left(\boldsymbol{X}^{(i)}\right)\big],\end{aligned} \tag{3.16}$$

where $\partial_j \boldsymbol{t}(\boldsymbol{x}) = (\partial_j t_1(\boldsymbol{x}), \ldots, \partial_j t_r(\boldsymbol{x}))^\top \in \mathbb{R}^r$. Note that (3.15) and (3.16) are sample averages of functions in the data matrix $\mathbf{x}$ only. These forms are an exact analog of those in Theorem 3. One thus have the following consistency result similar to 4:

Theorem 15. *Suppose the true density is $p_0 \equiv p_{\boldsymbol{\theta}_0}$ and that*

(C1) $\boldsymbol{\Gamma}$ *is almost surely invertible, and*

(C2) $\boldsymbol{\Gamma}_0 \equiv \mathbb{E}_{p_0}\boldsymbol{\Gamma}(\mathbf{x})$, $\boldsymbol{\Gamma}_0^{-1}$, $\boldsymbol{g}_0 \equiv \mathbb{E}_{p_0}\boldsymbol{g}(\mathbf{x})$, *and* $\boldsymbol{\Sigma}_0 \equiv \mathbb{E}_{p_0}\left[(\boldsymbol{\Gamma}(\mathbf{x})\boldsymbol{\theta}_0 - \boldsymbol{g}(\mathbf{x}))(\boldsymbol{\Gamma}(\mathbf{x})\boldsymbol{\theta}_0 - \boldsymbol{g}(\mathbf{x}))^\top\right]$ *exist and are component-wise finite.*

Then the minimizer of (3.14) is a.s. unique with closed form solution $\hat{\boldsymbol{\theta}} \equiv \boldsymbol{\Gamma}(\mathbf{x})^{-1}\boldsymbol{g}(\mathbf{x})$ with

$$\hat{\boldsymbol{\theta}} \to_{a.s.} \boldsymbol{\theta}_0 \quad \textit{and} \quad \sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \to_d \mathcal{N}_r\left(\mathbf{0}, \boldsymbol{\Gamma}_0^{-1}\boldsymbol{\Sigma}_0\boldsymbol{\Gamma}_0^{-1}\right) \quad \textit{as } n \to \infty.$$

3.4.2 Pairwise Interaction Power a-b Models

As an important example of graphical models from exponential family distributions, in this chapter we focus on the class of the pairwise interaction power models introduced in (1.1):

$$p_{\boldsymbol{\eta},\mathbf{K}}(\boldsymbol{x}) \propto \exp\left(-\frac{1}{2a}\boldsymbol{x}^{a\top}\mathbf{K}\boldsymbol{x}^a + \frac{1}{b}\boldsymbol{\eta}^\top\boldsymbol{x}^b\right)\mathbb{1}_{\mathcal{D}}(\boldsymbol{x}) \tag{3.17}$$

for which we treat $\boldsymbol{x}^0 \equiv \log \boldsymbol{x}$ and $1/0 \equiv 1$, on a domain $\mathcal{D} \subseteq \mathbb{R}^m$ with a positive measure. Here $a \geq 0$ and $b \geq 0$ are known constants and the interaction matrix $\mathbf{K} \in \mathbb{R}^{m\times m}$ and the linear vector $\boldsymbol{\eta} \in \mathbb{R}^m$ are unknown parameters of interest. As in Chapter 2, our focus will be on the support of $\mathbf{K}$, $S(\mathbf{K}) = \{(i,j) : \kappa_{ij} \neq 0\}$, that defines the conditional independence

graph of $\boldsymbol{X} \sim p_{\boldsymbol{\eta},\mathbf{K}}$, while simultaneously estimating the nuisance parameter $\boldsymbol{\eta}$ unless it is known or assumed to be $\mathbf{0}$.

When $a = b = 1$ (3.17) is simply a truncated Gaussian model. When $a = b = 1/2$ we obtain the exponential square-root graphical model in Inouye et al. (2016). The gamma model considered in Chapter 2 corresponds to $a = 1/2$ and $b = 0$. As we discuss in Section 3.6.2, A^{m-1} models in Aitchison (1985) are covered by $a = b = 0$ with a simplex domain $\mathcal{D}$.

3.4.3 Regularized Score Matching

In high-dimensional settings where the number of parameters r to be estimated is larger than the sample size n, one usually add some form of regularization for consistent estimation. For exponential family distributions, as in Chapter 2, we add an ℓ_1 penalty on $\boldsymbol{\theta}$ to the loss in (3.14), while multiplying the diagonals of the $\boldsymbol{\Gamma}$ by a *diagonal multiplier* $\delta > 1$:

Definition 8. *Let $\delta > 1$, and $\boldsymbol{\Gamma}_\delta(\mathbf{x})$ be $\boldsymbol{\Gamma}(\mathbf{x})$ with diagonal entries multiplied by δ. For exponential family distributions in (3.13), the* regularized generalized $(\boldsymbol{h}, \boldsymbol{C}, \mathcal{D})$-score matching loss *is defined as*

$$\hat{J}_{\boldsymbol{h},\boldsymbol{C},\mathcal{D},\lambda,\delta}(p_{\boldsymbol{\theta}}) \equiv \frac{1}{2}\boldsymbol{\theta}^\top \boldsymbol{\Gamma}_\delta(\mathbf{x})\boldsymbol{\theta} - \boldsymbol{g}(\mathbf{x})^\top \boldsymbol{\theta} + \lambda\|\boldsymbol{\theta}\|_1. \tag{3.18}$$

The multiplier $\delta > 1$, together with the ℓ_1 penalty, resembles an elastic net penalty and prevents the loss in (3.18) from being unbounded from below for smaller λ, in which case there can be infinitely many minimizers. For a more detailed discussion on this, refer to Section 2.4. The minimization of (3.18) can be efficiently done using coordinate-descent with warm starts, along with other computational details discussed in Section 2.5.3.

3.5 a-b Models on Domains with Positive Measure

Throughout this section, we assume that $\mathcal{D}$ has positive Lebesgue measure in $\mathbb{R}^m$.

3.5.1 Finite Normalizing Constant and Validity of Score Matching

The following theorem gives detailed sufficient conditions for the a-b density $p_{\boldsymbol{\eta},\mathbf{K}}$ (3.17) to be a proper density on $\mathcal{D} \subseteq \mathbb{R}^m$ with positive Lebesgue measure.

Theorem 16 (Sufficient conditions for finite normalizing constant)**.** *Denote the closure of the range of x_j in the domain $\mathcal{D}$ as $\rho_j(\mathcal{D}) \equiv \overline{\{x_j : \boldsymbol{x} \in \mathcal{D}\}}$. If any of the following conditions holds, the density in (3.17) is a proper density, i.e. the right-hand of (3.17) is integrable over $\mathcal{D}$:*

(CC1) $a > 0$, $b > 0$, $\mathcal{D}$ *bounded;*

(CC2) $a > 0$, $b > 0$, $\boldsymbol{v}^{a\top}\mathbf{K}\boldsymbol{v}^a > 0\ \forall \boldsymbol{v} \in \mathcal{D}\backslash\{\mathbf{0}\}$, *and either* $2a > b$ *or* $\boldsymbol{\eta}^\top \boldsymbol{v}^b \leq 0\ \forall \boldsymbol{v} \in \mathcal{D}$;

(CC3) $a > 0$, $b = 0$, $\eta_j > -1$ *for all* j *s.t.* $0 \in \rho_j(\mathcal{D})$, *and one of the following:*

- *(i)* $\mathcal{D}$ *bounded;*
- *(ii)* $\mathcal{D}$ *is unbounded and* $\boldsymbol{v}^{a\top}\mathbf{K}\boldsymbol{v}^a > 0\ \forall \boldsymbol{v} \in \mathcal{D}\backslash\{\mathbf{0}\}$;
- *(iii)* $\mathcal{D}$ *is unbounded,* $\boldsymbol{v}^{a\top}\mathbf{K}\boldsymbol{v}^a \geq 0\ \forall \boldsymbol{v} \in \mathcal{D}$ *and* $\eta_j < -1$ *for all* j *s.t.* $\rho_j(\mathcal{D})$ *is unbounded* ($\Rightarrow \rho_j(\mathcal{D}) = (0, +\infty)$ *is not allowed for any* j*);*

(CC4) $a = 0$, $\mathcal{D}$ *is bounded and* $0 \notin \rho_j(\mathcal{D})$ *for all* j*;*

(CC5) $a = 0$, $b = 0$, $\log(\boldsymbol{x})^\top \mathbf{K} \log(\boldsymbol{x}) > 0\ \forall \boldsymbol{x} \in \mathcal{D}$;

(CC6) $a = 0$, $b > 0$, $\log(\boldsymbol{x})^\top \mathbf{K} \log(\boldsymbol{x}) > 0\ \forall \boldsymbol{x} \in \mathcal{D}$ *and* $\eta_j \leq 0$ *for all* j *s.t.* $\rho_j(\mathcal{D})$ *is unbounded;*

(CC7) $a = 0$, $b > 0$, $\log(\boldsymbol{x})^\top \mathbf{K} \log(\boldsymbol{x}) \geq 0\ \forall \boldsymbol{x} \in \mathcal{D}$ *and* $\eta_j < 0$ *for all* j *s.t.* $\rho_j(\mathcal{D})$ *is unbounded.*

In the centered case where $\boldsymbol{\eta} = \mathbf{0}$ is known, any condition in terms of b and $\boldsymbol{\eta}$ can be ignored.

To simplify our discussion, the following corollary gives a simpler set of sufficient conditions for integrability of the density.

Corollary 17 (Sufficient conditions for finite normalizing constant; simplified)**.** *Suppose*

(CC0)* $\mathbf{K}$ *is positive definite*

and one of the following conditions holds, then the right-hand side of (3.17) is integrable over $\mathcal{D}$*.*

(CC1)* $2a > b > 0$ *or* $a = b = 0$*;*

(CC2)* $a > 0$*,* $b = 0$*,* $\boldsymbol{\eta} \succ -\mathbf{1}_m$*;*

(CC3)* $a = 0$*,* $b > 0$*,* $\eta_j \leq 0$ *for any* j *such that* x_j *is unbounded in* $\mathcal{D}$*.*

In the non-centered case where $\boldsymbol{\eta} \equiv \mathbf{0}$*, (CC0*) is sufficient.*

For simplicity, we use the set of conditions (CC0*) through (CC3*) throughout the rest of the chapter. The following theorem gives sufficient conditions on $\boldsymbol{h}$ such that conditions (A.1)–(A.3) in Lemma 13 for score matching are satisfied.

Theorem 18 (Sufficient conditions that satisfy assumptions for score matching)**.** *Suppose (CC0*) and one of (CC1*) through (CC3*) holds, and* $\boldsymbol{h}(\boldsymbol{x}) = (x_1^{\alpha_1}, \ldots, x_m^{\alpha_m})$*, where*

(1) if $a > 0$ *and* $b > 0$*,* $\alpha_j > \max\{0, 1-a, 1-b\}$*;*

(2) if $a > 0$ *and* $b = 0$*,* $\alpha_j > 1 - \eta_{0,j}$*;*

(3) if $a = 0$*,* $\alpha_j \geq 0$*.*

Then conditions (A.1), (A.2) and (A.3) in Lemma 13 are satisfied and the equivalent form of the generalized score matching loss (3.10) holds, and the empirical loss (3.11) is valid. In the centered case with $\boldsymbol{\eta} \equiv \mathbf{0}$*, it suffices to have* $a > 0$ *and* $\alpha_j > \max\{0, 1-a\}$ *or* $a = 0$ *and* $\alpha_j \geq 0$*.*

3.5.2 Estimation

Let $\boldsymbol{\Psi} \equiv \left[\mathbf{K}^\top \ \boldsymbol{\eta}\right]^\top \in \mathbb{R}^{(m+1)\times m}$. In this section, we give the form of $\boldsymbol{\Gamma} \in \mathbb{R}^{(m+1)m\times(m+1)m}$ and $\boldsymbol{g} \in \mathbb{R}^{(m+1)m}$ in the unpenalized loss $\frac{1}{2}\mathrm{vec}(\boldsymbol{\Psi})^\top \boldsymbol{\Gamma}\mathrm{vec}(\boldsymbol{\Psi}) - \boldsymbol{g}^\top \mathrm{vec}(\boldsymbol{\Psi})$ following (3.15)–(3.16). $\boldsymbol{\Gamma}$ is block-diagonal, with the j-th $\mathbb{R}^{(m+1)\times(m+1)}$ block

$$\begin{aligned}\boldsymbol{\Gamma}_j(\mathbf{x}) &\equiv \begin{bmatrix} \boldsymbol{\Gamma}_{\mathbf{K},j} & \boldsymbol{\gamma}_{\mathbf{K},\boldsymbol{\eta},j} \\ \boldsymbol{\gamma}_{\mathbf{K},\boldsymbol{\eta},j}^\top & \gamma_{\boldsymbol{\eta},j} \end{bmatrix} \\ &\equiv \frac{1}{n}\sum_{i=1}^{n} \begin{bmatrix} (h_j \circ \varphi_j)\left(\boldsymbol{X}^{(i)}\right) {X_j^{(i)}}^{2a-2} {\boldsymbol{X}^{(i)}}^a {\boldsymbol{X}^{(i)}}^\top & -(h_j \circ \varphi_j)\left(\boldsymbol{X}^{(i)}\right) {X_j^{(i)}}^{a+b-2} {\boldsymbol{X}^{(i)}}^a \\ -(h_j \circ \varphi_j)\left(\boldsymbol{X}^{(i)}\right) {X_j^{(i)}}^{a+b-2} {\boldsymbol{X}^{(i)}}^\top & (h_j \circ \varphi_j)\left(\boldsymbol{X}^{(i)}\right) {X_j^{(i)}}^{2b-2} \end{bmatrix},\end{aligned}$$

and $\boldsymbol{g} \equiv \mathrm{vec}\left(\left[\mathbf{g}_\mathbf{K}^\top \ \ \boldsymbol{g}_{\boldsymbol{\eta}}\right]^\top\right) \in \mathbb{R}^{(m+1)m}$, where $\mathbf{g}_\mathbf{K} \in \mathbb{R}^{m\times m}$ and $\boldsymbol{g}_{\boldsymbol{\eta}} \in \mathbb{R}^m$ correspond to $\mathbf{K}$ and $\boldsymbol{\eta}$, respectively. The j-th column of $\mathbf{g}_\mathbf{K}$ is

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} \Big(\partial_j(h_j \circ \varphi_j)\left(\boldsymbol{X}^{(i)}\right) {X_j^{(i)}}^{a-1} + (a-1)(h_j \circ \varphi_j)\left(\boldsymbol{X}^{(i)}\right) {X_j^{(i)}}^{a-2}\Big) {\boldsymbol{X}^{(i)}}^a \\ + a(h_j \circ \varphi_j)\left(\boldsymbol{X}^{(i)}\right) {X_j^{(i)}}^{2a-2} \boldsymbol{e}_{j,m}, \quad (3.19)\end{aligned}$$

where $\boldsymbol{e}_{j,m} \in \mathbb{R}^m$ has 1 at the j-th composition and 0 elsewhere, and the j-th entry of $\boldsymbol{g}_2$ is

$$\frac{1}{n}\sum_{i=1}^{n} -\partial_j(h_j \circ \varphi_j)\left(\boldsymbol{X}^{(i)}\right) {X_j^{(i)}}^{b-1} - (b-1)(h_j \circ \varphi_j)\left(\boldsymbol{X}^{(i)}\right) {X_j^{(i)}}^{b-2}. \quad (3.20)$$

(If $a = 0$, set the coefficients $(a-1) \leftarrow -1$ and $a \leftarrow 1$ in (3.19); for $b = 0$ let $(b-1) \leftarrow -1$ in the second term of (3.20).)

As in Chapter 2 we only apply the diagonal multiplier δ to the diagonals of $\boldsymbol{\Gamma}_{\mathbf{K},j} \in \mathbb{R}^{m\times m}$, not $\gamma_{\boldsymbol{\eta},j} \in \mathbb{R}$. Note that each block of $\boldsymbol{\Gamma}$ and $\boldsymbol{g}$ correspond to each column of $\boldsymbol{\Psi}$, i.e. $(\boldsymbol{\kappa}_{,j}, \eta_j) \in \mathbb{R}^{m+1}$. In the penalized generalized score-matching loss (3.18), the penalty λ for $\mathbf{K}$ and $\boldsymbol{\eta}$ can be different, $\lambda_\mathbf{K}$ and $\lambda_{\boldsymbol{\eta}}$, respectively, as long as the ratio $\lambda_{\boldsymbol{\eta}}/\lambda_\mathbf{K}$ is fixed. For $\mathbf{K}$ we follow the convention that we penalize its off-diagonal entries only. That is,

$$\frac{1}{2}\mathrm{vec}(\boldsymbol{\Psi})^\top \boldsymbol{\Gamma}_\delta(\mathbf{x})\mathrm{vec}(\boldsymbol{\Psi}) - \boldsymbol{g}(\mathbf{x})^\top \mathrm{vec}(\boldsymbol{\Psi}) + \lambda_\mathbf{K}\|\mathbf{K}_{\mathrm{off}}\|_1 + \lambda_{\boldsymbol{\eta}}\|\boldsymbol{\eta}\|_1. \quad (3.21)$$

In the case where we do not penalize $\boldsymbol{\eta}$, i.e. $\lambda_{\boldsymbol{\eta}} = 0$, as in Chapter 2 we can simply profile out $\boldsymbol{\eta}$, solve for $\hat{\boldsymbol{\eta}} = \boldsymbol{\Gamma}_{\boldsymbol{\eta}}^{-1}\left(\boldsymbol{g}_{\boldsymbol{\eta}} - \boldsymbol{\Gamma}_{\mathbf{K},\boldsymbol{\eta}}^{\top}\mathrm{vec}(\hat{\mathbf{K}})\right)$, plug this back in and rewrite the loss in $\mathbf{K}$ only. Let $\boldsymbol{\Gamma}_{\delta,\mathbf{K}} \in \mathbb{R}^{m^2\times m^2}$ be the block-diagonal matrix with blocks $\boldsymbol{\Gamma}_{\mathbf{K},j}$ and diagonal multiplier δ, and let $\boldsymbol{\Gamma}_{\mathbf{K},\boldsymbol{\eta}} \in \mathbb{R}^{m^2\times m}$ and $\boldsymbol{\Gamma}_{\boldsymbol{\eta}} \in \mathbb{R}^{m\times m}$ be the (block-)diagonal matrices with blocks $\boldsymbol{\gamma}_{\mathbf{K},\boldsymbol{\eta},j}$ and $\gamma_{\boldsymbol{\eta},j}$, respectively. Further denote $\boldsymbol{\Gamma}_{\delta,\star}$ as the Schur complement of $\boldsymbol{\Gamma}_{\delta,\mathbf{K}}$ of $\left[\begin{smallmatrix}\boldsymbol{\Gamma}_{\delta,\mathbf{K}} & \boldsymbol{\Gamma}_{\mathbf{K},\boldsymbol{\eta}}\\ \boldsymbol{\Gamma}_{\mathbf{K},\boldsymbol{\eta}}^{\top} & \boldsymbol{\Gamma}_{\boldsymbol{\eta}}\end{smallmatrix}\right]$, i.e. $\boldsymbol{\Gamma}_{\delta,\mathbf{K}} - \boldsymbol{\Gamma}_{\mathbf{K},\boldsymbol{\eta}}\boldsymbol{\Gamma}_{\boldsymbol{\eta}}^{-1}\boldsymbol{\Gamma}_{\mathbf{K},\boldsymbol{\eta}}^{\top}$, which is guaranteed to be positive definite for $\delta > 1$. Then the profiled loss is

$$\hat{J}_{\boldsymbol{h},\boldsymbol{C},\mathcal{D},\boldsymbol{\lambda},\delta,\star}(p_{\mathbf{K}}) \equiv \frac{1}{2}\mathrm{vec}(\mathbf{K})^{\top}\boldsymbol{\Gamma}_{\delta,\star}\mathrm{vec}(\mathbf{K}) - \left(\boldsymbol{g}_{\mathbf{K}} - \boldsymbol{\Gamma}_{\mathbf{K},\boldsymbol{\eta}}\boldsymbol{\Gamma}_{\boldsymbol{\eta}}^{-1}\boldsymbol{g}_{\boldsymbol{\eta}}\right)^{\top}\mathrm{vec}(\mathbf{K}) + \lambda_{\mathbf{K}}\|\mathbf{K}\|_1. \quad (3.22)$$

3.5.3 Univariate Examples

In this section we present univariate Gaussian models on general domains as an example of our generalized score matching estimator. In particular, we estimate one of μ_0 and σ_0^2 assuming the other is known, given that the true density is

$$p_{\mu_0,\sigma_0^2}(x) \propto \exp\left\{-\frac{(x-\mu_0)^2}{2\sigma_0^2}\right\}, \quad x \in \mathcal{D}$$

with $\mu_0 \in \mathbb{R}$, $\sigma_0^2 > 0$ and $\mathcal{D} \subset \mathbb{R}$ *piecewise decomposable*, i.e. a union of at most countably many intervals and with Lebesgue positive measure.

If we do not impose ℓ_1 or ℓ_2 penalty and assume the true σ_0^2 is known, given i.i.d. samples $X^{(1)},\ldots,X^{(n)} \sim p_{\mu_0,\sigma_0^2}$, similar to Example 3.1 in Section 2.3, we have the estimator for μ_0

$$\hat{\mu} \equiv \frac{\sum_{i=1}^n (h\circ\varphi_C)\left(X^{(i)}\right)\cdot X^{(i)} - \sigma_0^2(h\circ\varphi_C)'\left(X^{(i)}\right)}{\sum_{i=1}^n (h\circ\varphi_C)\left(X^{(i)}\right)}.$$

By Theorem 18, it suffices to choose $h(x) = x^{\alpha}$ with $\alpha > 0$. Similar to Section 2.3, we have

$$\sqrt{n}\left(\hat{\mu} - \mu_0\right) \to_d \mathcal{N}\left(0, \frac{\mathbb{E}_0\left[\sigma_0^2(h\circ\varphi_C)^2(X) + \sigma_0^4(h\circ\varphi_C)'^2(X)\right]}{\mathbb{E}_0^2\left[(h\circ\varphi_C)(X)\right]}\right), \quad (3.23)$$

if the expectations exist. On the other hand, similar to Example 3.2 in Section 2.3, assuming the true μ_0 is known, the estimator for σ^2 is,

$$\hat{\sigma}^2 \equiv \frac{\sum_{i=1}^n (h\circ\varphi_C)\left(X^{(i)}\right)\cdot\left(X^{(i)}-\mu_0\right)^2}{\sum_{i=1}^n (h\circ\varphi_C)\left(X^{(i)}\right) + (h\circ\varphi_C)'\left(X^{(i)}\right)\cdot\left(X^{(i)}-\mu_0\right)}$$

with limiting distribution

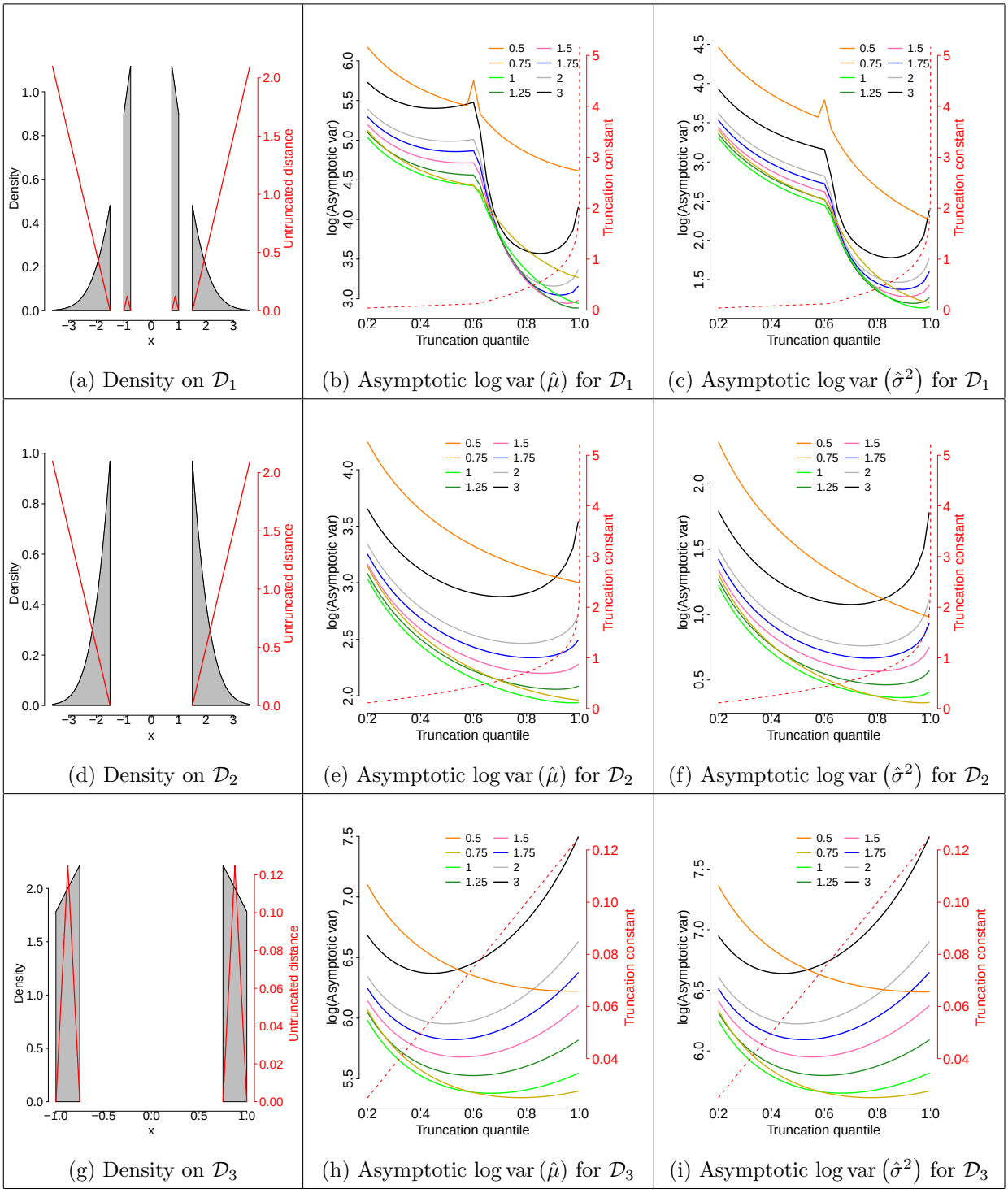
$$\sqrt{n}\left(\hat{\sigma}^2 - \sigma_0^2\right) \to_d \mathcal{N}\left(0, \frac{\mathbb{E}_0\left[2\sigma_0^6 (h \circ \varphi_C)^2(X) \cdot (X-\mu_0)^2 + \sigma_0^8 (h \circ \varphi_C)'^2(X) \cdot (X-\mu_0)^2\right]}{\mathbb{E}_0^2\left[(h \circ \varphi_C)(X) \cdot (X-\mu_0)^2\right]}\right). \tag{3.24}$$

In Figure 3.3 we show an example of standard normal $\mathcal{N}(0,1)$ restricted to three univariate domains, namely $\mathcal{D}_2 \equiv (-\infty, -3/2] \cup [3/2, +\infty)$, $\mathcal{D}_3 \equiv [-1, -3/4] \cup [3/4, 1]$, and their union $\mathcal{D}_1 \equiv (-\infty, -3/2] \cup [-1, -3/4] \cup [3/4, 1] \cup [3/2, +\infty)$. Note that $\mathcal{D}_1$ is the union of $\mathcal{D}_2$ and $\mathcal{D}_3$, where $\mathcal{D}_2$ comprises two unbounded intervals, and $\mathcal{D}_3$ consists of two bounded intervals. The endpoints are chosen so that the probability of the variable lying in each interval is roughly the same ($\mathbb{P}(\mathcal{N}(0,1) \in (-\infty, -3/2]) \approx 0.0668$, $\mathbb{P}(\mathcal{N}(0,1) \in [-1, -3/4]) \approx 0.0680$). To pick the truncation point C in the definition of the truncated distance $\varphi_{C,\mathcal{D}}$, we choose a $\pi \in (0, 1]$, and let C be the π quantile of $\varphi_{+\infty,\mathcal{D}}(X)$ with $X \sim \mathcal{N}(0,1)$ on $\mathcal{D}$, that is, the C such $\mathbb{P}(\varphi_{+\infty,\mathcal{D}}(X) \leq C) = \pi$. Here, $\varphi_{+\infty,\mathcal{D}_1}(X) = |X| - 3/2$ if $|X| > 3/2$, or $\min(|X| - 3/4, 1 - |X|)$ otherwise, $\varphi_{+\infty,\mathcal{D}_2}(X) = |X| - 3/2$, and $\varphi_{+\infty,\mathcal{D}_3}(X) = \min(|X| - 3/4, 1 - |X|)$.

The first subfigure in each row of Figure 3.3 shows the density on each domain, along with the corresponding $\varphi_{+\infty}(X)$ in red, whose y axis is on the right. The second in each row shows the log of the asymptotic variance for the corresponding $\hat{\mu}$, as on the right-hand side of (3.23), and the third shows that for $\hat{\sigma}^2$ as in (3.24). Each curve represents a different power α in $h(x) = x^\alpha$, and the x axis represents the quantiles π associated with the truncation point C, as mentioned above. Finally, the red dotted curve shows the truncation point C versus π for each domain. The "bumps" in the variance for $x^{0.5}$ are due to numerical instability in integration.

As we show in Section 3.7, for the purpose of edge recovery for graphical models, we recommend using $\boldsymbol{h}(\boldsymbol{x}) = (x_1^{\alpha_1}, \ldots, x_m^{\alpha_m})$ with $\alpha \geq 1$ for $\mathcal{D}$ that is a finite disjoint union of convex subsets of $\mathbb{R}^m$. Although minimizing the asymptotic variance in the univariate case is a different task, $\alpha = 1$ also seems to be consistently the best performing choice.

For $\mathcal{D}_2$ and $\mathcal{D}_3$, all variance curves are U-shaped, while for $\mathcal{D}_1 = \mathcal{D}_2 \cup \mathcal{D}_3$ each curve is visually piecewise connected by two parabolae, and the turning point corresponds to



(a) Density on $\mathcal{D}_1$ (b) Asymptotic log var $(\hat{\mu})$ for $\mathcal{D}_1$ (c) Asymptotic log var $(\hat{\sigma}^2)$ for $\mathcal{D}_1$

(d) Density on $\mathcal{D}_2$ (e) Asymptotic log var $(\hat{\mu})$ for $\mathcal{D}_2$ (f) Asymptotic log var $(\hat{\sigma}^2)$ for $\mathcal{D}_2$

(g) Density on $\mathcal{D}_3$ (h) Asymptotic log var $(\hat{\mu})$ for $\mathcal{D}_3$ (i) Asymptotic log var $(\hat{\sigma}^2)$ for $\mathcal{D}_3$

Figure 3.3: Univariate Gaussian example.

$C_0 = \max \varphi_{+\infty,\mathcal{D}_3}(x) = (1-3/4)/2 = 0.125$. To the right of this C_0, the truncation is applied to the two unbounded intervals (i.e. $\mathcal{D}_2$) only. The first segment of most $\mathrm{var}(\hat{\mu})$ curves for $\mathcal{D}_1$, as well as most curves for $\mathcal{D}_2$ suggest there may still be benefit from truncating the distances φ within the bounded intervals, although the $\mathrm{var}\left(\hat{\sigma}^2\right)$ curves for $\mathcal{D}_1$ as well as both curves for $x^{0.75}$ on $\mathcal{D}_2$ suggest otherwise. On the other hand, the curves for $\mathcal{D}_1$ and $\mathcal{D}_2$ imply that a truncation constant larger than C_0 is favorable; by referring to the ticks on the right-hand side, note that the curves for $\mathcal{D}_2$ reach their minimum at $C \geq 0.5$. This suggests the choice of a separate truncation point C for each connected component of $\mathcal{D}$, especially for unbounded versus bounded ones. Although a further generalization of our proposed method, this becomes infeasible for $m \gg 1$ and involves tuning multiple parameters, which we do not further examine in this paper.

3.6 a-b Models on Standard Simplices

In this section we consider a-b models on the most important example of $\mathcal{D}$ with zero Lebesgue measure in $\mathbb{R}^m$, the $(m-1)$-dimensional standard simplex $\mathcal{D} \equiv \left\{\boldsymbol{x} \in \mathbb{R}^m_+ \middle| \, \boldsymbol{x} \succ \mathbf{0},\, \mathbf{1}_m^\top \boldsymbol{x} = 1\right\}$. As in Example 3.13, in the density we profile out the last component of $\boldsymbol{x}$ using $x_m = 1 - \mathbf{1}_{m-1}^\top \boldsymbol{x}_{-m}$:

$$p_{\boldsymbol{\eta},\mathbf{K}}(\boldsymbol{x}_{-m}) \propto \exp\left[-\frac{1}{2a}{\boldsymbol{x}^a_{-m}}^\top \mathbf{K}_{-m,-m}\boldsymbol{x}^a_{-m} - \frac{1}{a}{\boldsymbol{x}^a_{-m}}^\top \boldsymbol{\kappa}_{-m,m}\left(1 - \mathbf{1}_{m-1}^\top \boldsymbol{x}_{-m}\right)^a \right.$$
$$\left. -\frac{1}{2a}\kappa_{m,m}\left(1 - \mathbf{1}_{m-1}^\top \boldsymbol{x}_{-m}\right)^{2a} + \frac{1}{b}\boldsymbol{\eta}_{-m}^\top \boldsymbol{x}^b_{-m} + \frac{\eta_m}{b}\left(1 - \mathbf{1}_{m-1}^\top \boldsymbol{x}_{-m}\right)^b\right] \quad (3.25)$$

on $\mathcal{D}_{-m} \equiv \left\{\boldsymbol{x}_{-m} \in \mathbb{R}^{m-1}_+ \middle| \, \boldsymbol{x}_{-m} \succ \mathbf{0},\, \mathbf{1}_{m-1}^\top \boldsymbol{x}_{-m} < 1\right\} \subseteq \mathbb{R}^{m-1}_+$, where for simplifying the notation of this density we again define that $1/0 \equiv 1$ and $\boldsymbol{x}^0_{-m} \equiv \log(\boldsymbol{x}_{-m})$.

Sufficient conditions on a and b for the density to be proper as discussed in Theorem 16 and Corollary 17 also hold in the simplex cases. Additional sufficient conditions for the case where $a = b = 0$ are extensively studied in Section 3.6.2. The following theorem states that the sufficient conditions on h_j for a general $\mathcal{D} \subseteq \mathbb{R}^m$ are also sufficient for $\mathcal{D}_{-m} \subseteq \mathbb{R}^{m-1}$, and the equivalent definition (3.10) of the generalized score matching loss and its empirical loss (3.11) are again valid.

Theorem 19. *Suppose the conditions in Theorem 18 hold, then (A.1), (A.2) and (A.3) in Lemma 13 are also satisfied for a-b models on $\mathcal{D}_{-m}$ in (3.25).*

By the nature of the simplex domain, sparsity pattern of $\mathbf{K}$ does not directly relate to the conditional independence between two components, as any two components are perfectly correlated when all others are fixed, due to the sum-to-one condition. However, one may still be interested in the pattern as a measure of how two components interact, as well as estimation of $\mathbf{K}$ and $\boldsymbol{\eta}$ with minimized error for the sole purpose of recovering the density itself. As a prerequisite, the following theorem gives the conditions for the identifiability of $\mathbf{K}$ and $\boldsymbol{\eta}$ from a given *a-b* density on the simplex. In short, the parameters are identifiable if $2a = b > 0$ does not hold and $a \neq 1$, and in these cases the goal of recovering the underlying true parameters $\mathbf{K}_0$ and $\boldsymbol{\eta}_0$ is well-defined.

Theorem 20. *Suppose there exist $\mathbf{K}_1$, $\mathbf{K}_2$, $\boldsymbol{\eta}_1$, $\boldsymbol{\eta}_2$ such that*

$$\exp\left(-\frac{1}{2a}\boldsymbol{x}^{a\top}\mathbf{K}_1\boldsymbol{x}^a + \frac{1}{b}\boldsymbol{\eta}_1^\top\boldsymbol{x}^b\right) = \exp\left(-\frac{1}{2a}\boldsymbol{x}^{a\top}\mathbf{K}_2\boldsymbol{x}^a + \frac{1}{b}\boldsymbol{\eta}_2^\top\boldsymbol{x}^b\right)$$

for all $\boldsymbol{x} \in \mathcal{D}$, where by $\boldsymbol{x}^0$ we mean $\log(\boldsymbol{x})$ and $1/0 \equiv 1$. Then $\mathbf{K}_1 = \mathbf{K}_2$ and $\boldsymbol{\eta}_1 = \boldsymbol{\eta}_2$, or else one of the following must hold: (1) $a = b = 1$, (2) $a = 1$, $b = 2$, (3) $a = 1$ and $\boldsymbol{\eta}_1 = \boldsymbol{\eta}_2$, (4) $2a = b > 0$ and $\mathbf{K}_1 - \mathbf{K}_2 = 2\boldsymbol{\eta}_1 - 2\boldsymbol{\eta}_2$.

3.6.1 Estimation for General a and b

In the equations in this section, for $a = 0$ (and similarly for $b = 0$) we substitute the coefficients "a" with "1" and "$(a-1)$" with "-1", and let $x^a \equiv \log x$.

By substituting $x_m \equiv 1 - \mathbf{1}_{m-1}^\top\boldsymbol{x}_{-m}$ and working on $\mathcal{D}_{-m}$ instead, $\partial_j \log p(\boldsymbol{x}_{-m})$ now depends on both $(\boldsymbol{\kappa}_{,j}, \eta_j)$ and $(\boldsymbol{\kappa}_{,m}, \eta_m)$. Thus, unlike in Section 3.5.2, $(\boldsymbol{\kappa}_{,j}, \eta_j)$ and $(\boldsymbol{\kappa}_{,m}, \eta_m)$ are no longer isolated in the score-matching loss. In particular,

$$\partial_j \log p(\boldsymbol{x}_{-m}) = -\left(\boldsymbol{\kappa}_{,j}^\top\boldsymbol{x}^a\right)x_j^{a-1} + \left(\boldsymbol{\kappa}_{,m}^\top\boldsymbol{x}^a\right)x_m^{a-1} + \eta_j x_j^{b-1} - \eta_m x_m^{b-1} \tag{3.26}$$

$$\partial_{jj} \log p(\boldsymbol{x}_{-m}) = -(a-1)\left(\boldsymbol{\kappa}_{,j}^\top\boldsymbol{x}^a\right)x_j^{a-2} - (a-1)\left(\boldsymbol{\kappa}_{,m}^\top\boldsymbol{x}^a\right)x_m^{a-2} - a\kappa_{jj}x_j^{2a-2} - a\kappa_{mm}x_m^{2a-2} \tag{3.27}$$

$$+ (b-1)\eta_j x_j^{b-2} + (b-1)\eta_m x_m^{b-2} + 2a\kappa_{jm} x_j^{a-1} x_m^{a-1}$$

and the unpenalized loss becomes

$$\frac{1}{2}\left(\mathrm{vec}(\mathbf{K}),\, \boldsymbol{\eta}\right)^\top \boldsymbol{\Gamma}\left(\mathrm{vec}(\mathbf{K}),\, \boldsymbol{\eta}\right) - \boldsymbol{g}^\top\left(\mathrm{vec}(\mathbf{K}),\, \boldsymbol{\eta}\right), \quad \text{where}$$

$$\boldsymbol{\Gamma} \equiv \begin{bmatrix} \boldsymbol{\Gamma}_{\mathbf{K}} & \boldsymbol{\Gamma}_{\mathbf{K},\boldsymbol{\eta}} \\ \boldsymbol{\Gamma}_{\mathbf{K},\boldsymbol{\eta}}^\top & \boldsymbol{\Gamma}_{\boldsymbol{\eta}} \end{bmatrix} \in \mathbb{R}^{m(m+1)\times m(m+1)}, \quad \boldsymbol{\Gamma}_{\mathbf{K}} \in \mathbb{R}^{m^2\times m^2}, \quad \boldsymbol{\Gamma}_{\mathbf{K},\boldsymbol{\eta}} \in \mathbb{R}^{m^2\times m}, \quad \boldsymbol{\Gamma}_{\boldsymbol{\eta}} \in \mathbb{R}^{m\times m}, \tag{3.28}$$

$$\boldsymbol{g} \equiv (\mathrm{vec}(\mathbf{g}_{\mathbf{K}}),\, \boldsymbol{g}_{\boldsymbol{\eta}}) \in \mathbb{R}^{m(m+1)}, \quad \mathbf{g}_{\mathbf{K}} \in \mathbb{R}^{m\times m}, \quad \boldsymbol{g}_{\boldsymbol{\eta}} \in \mathbb{R}^{m}. \tag{3.29}$$

The components $\boldsymbol{\Gamma}_{\mathbf{K}} \in \mathbb{R}^{m^2\times m^2}$, $\boldsymbol{\Gamma}_{\mathbf{K},\boldsymbol{\eta}} \in \mathbb{R}^{m(m+1)}$ and $\boldsymbol{\Gamma}_{\boldsymbol{\eta}} \in \mathbb{R}^{m\times m}$ are no longer block-diagonal with m blocks as for general $\mathcal{D} \subseteq \mathbb{R}^m$ in Section 3.4.2. Instead, each of them can be partitioned into $m \times m$ blocks with m "block-columns" and m "block-rows", and each matrix not only has m blocks on the diagonal, but also have the entire right-most "block-column" and bottom-most "block-row" filled, representing the interaction between the $(\boldsymbol{\kappa}_{,j}, \eta_j)$ and $(\boldsymbol{\kappa}_{,m}, \eta_m)$. In particular,

$$\boldsymbol{\Gamma}_{\mathbf{K}} \equiv \begin{bmatrix} \boldsymbol{\Gamma}_{\mathbf{K},1} & \mathbf{0} & \cdots & \mathbf{0} & \boldsymbol{\Gamma}_{\mathbf{K},(1,m)} \\ \mathbf{0} & \boldsymbol{\Gamma}_{\mathbf{K},2} & \cdots & \mathbf{0} & \boldsymbol{\Gamma}_{\mathbf{K},(2,m)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Gamma}_{\mathbf{K},m-1} & \boldsymbol{\Gamma}_{\mathbf{K},(m-1,m)} \\ \boldsymbol{\Gamma}_{\mathbf{K},(1,m)}^\top & \boldsymbol{\Gamma}_{\mathbf{K},(2,m)}^\top & \cdots & \boldsymbol{\Gamma}_{\mathbf{K},(m-1,m)}^\top & \boldsymbol{\Gamma}_{\mathbf{K},m} \end{bmatrix} \in \mathbb{R}^{m^2\times m^2}, \text{ each block } \in \mathbb{R}^{m\times m},$$

$$\boldsymbol{\Gamma}_{\mathbf{K},\boldsymbol{\eta}} \equiv \begin{bmatrix} \boldsymbol{\gamma}_{\mathbf{K},\boldsymbol{\eta},1} & \mathbf{0} & \cdots & \mathbf{0} & \boldsymbol{\gamma}_{\mathbf{K},\boldsymbol{\eta},(1,m)} \\ \mathbf{0} & \boldsymbol{\gamma}_{\mathbf{K},\boldsymbol{\eta},2} & \cdots & \mathbf{0} & \boldsymbol{\gamma}_{\mathbf{K},\boldsymbol{\eta},(2,m)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\gamma}_{\mathbf{K},\boldsymbol{\eta},m-1} & \boldsymbol{\gamma}_{\mathbf{K},\boldsymbol{\eta},(m-1,m)} \\ \boldsymbol{\gamma}_{\mathbf{K},\boldsymbol{\eta},(m,1)} & \boldsymbol{\gamma}_{\mathbf{K},\boldsymbol{\eta},(m,2)} & \cdots & \boldsymbol{\gamma}_{\mathbf{K},\boldsymbol{\eta},(m,m-1)} & \boldsymbol{\gamma}_{\mathbf{K},\boldsymbol{\eta},m} \end{bmatrix} \in \mathbb{R}^{m^2\times m}, \text{ each } \in \mathbb{R}^{m},$$

$$\boldsymbol{\Gamma}_{\boldsymbol{\eta}} \equiv \begin{bmatrix} \gamma_{\boldsymbol{\eta},1} & 0 & \cdots & 0 & \gamma_{\boldsymbol{\eta},(1,m)} \\ 0 & \gamma_{\boldsymbol{\eta},2} & \cdots & 0 & \gamma_{\boldsymbol{\eta},(2,m)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \gamma_{\boldsymbol{\eta},m-1} & \gamma_{\boldsymbol{\eta},(m-1,m)} \\ \gamma_{\boldsymbol{\eta},(1,m)} & \gamma_{\boldsymbol{\eta},(2,m)} & \cdots & \gamma_{\boldsymbol{\eta},(m-1,m)} & \gamma_{\boldsymbol{\eta},m} \end{bmatrix} \in \mathbb{R}^{m \times m}, \text{ each block } \in \mathbb{R},$$

where for $j = 1, \ldots, m-1$,

$$\begin{aligned} \boldsymbol{\Gamma}_j &\equiv \begin{bmatrix} \boldsymbol{\Gamma}_{\mathbf{K},j} & \boldsymbol{\gamma}_{\mathbf{K},\boldsymbol{\eta},j} \\ \boldsymbol{\gamma}_{\mathbf{K},\boldsymbol{\eta},j}^\top & \gamma_{\boldsymbol{\eta},j} \end{bmatrix} \\ &\equiv \frac{1}{n} \sum_{i=1}^{n} (h_j \circ \varphi_j)\left(\boldsymbol{X}^{(i)}\right) \begin{bmatrix} {X_j^{(i)}}^{a-1} \boldsymbol{X}^{(i)^a} \\ -{X_j^{(i)}}^{b-1} \end{bmatrix} \begin{bmatrix} {X_j^{(i)}}^{a-1} {\boldsymbol{X}^{(i)}}^{a\top} & -{X_j^{(i)}}^{b-1} \end{bmatrix}, \\ \boldsymbol{\Gamma}_m &\equiv \begin{bmatrix} \boldsymbol{\Gamma}_{\mathbf{K},m} & \boldsymbol{\gamma}_{\mathbf{K},\boldsymbol{\eta},m} \\ \boldsymbol{\gamma}_{\mathbf{K},\boldsymbol{\eta},m}^\top & \gamma_{\boldsymbol{\eta},m} \end{bmatrix} \\ &\equiv \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{m-1} (h_k \circ \varphi_k)\left(\boldsymbol{X}^{(i)}\right) \begin{bmatrix} {X_m^{(i)}}^{a-1} \boldsymbol{X}^{(i)^a} \\ -{X_m^{(i)}}^{b-1} \end{bmatrix} \begin{bmatrix} {X_m^{(i)}}^{a-1} {\boldsymbol{X}^{(i)}}^{a\top} & -{X_m^{(i)}}^{b-1} \end{bmatrix}, \\ \boldsymbol{\Gamma}_{(j,m)} &\equiv \begin{bmatrix} \boldsymbol{\Gamma}_{\mathbf{K},(j,m)} & \boldsymbol{\gamma}_{\mathbf{K},\boldsymbol{\eta},(j,m)} \\ \boldsymbol{\gamma}_{\mathbf{K},\boldsymbol{\eta},(m,j)} & \gamma_{\boldsymbol{\eta},(j,m)} \end{bmatrix} \\ &\equiv -\frac{1}{n} \sum_{i=1}^{n} (h_j \circ \varphi_j)\left(\boldsymbol{X}^{(i)}\right) \begin{bmatrix} {X_j^{(i)}}^{a-1} \boldsymbol{X}^{(i)^a} \\ -{X_j^{(i)}}^{b-1} \end{bmatrix} \begin{bmatrix} {X_m^{(i)}}^{a-1} {\boldsymbol{X}^{(i)}}^{a\top} & -{X_m^{(i)}}^{b-1} \end{bmatrix}. \end{aligned}$$

In addition,

$$\begin{aligned} \boldsymbol{g}_{\mathbf{K},j} &\equiv \frac{1}{n} \sum_{i=1}^{n} \left[\partial_j (h_j \circ \varphi_j)\left(\boldsymbol{X}^{(i)}\right) {X_j^{(i)}}^{a-1} + (a-1)(h_j \circ \varphi_j)\left(\boldsymbol{X}^{(i)}\right) {X_j^{(i)}}^{a-2} \right] \boldsymbol{X}^{(i)^a} \\ &\qquad + a(h_j \circ \varphi_j)\left(\boldsymbol{X}^{(i)}\right) {X_j^{(i)}}^{2a-2} \boldsymbol{e}_{j,m} - a(h_j \circ \varphi_j)\left(\boldsymbol{X}^{(i)}\right) {X_j^{(i)}}^{a-1} {X_m^{(i)}}^{a-1} \boldsymbol{e}_{m,m} \\ \boldsymbol{g}_{\mathbf{K},m} &\equiv \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{m-1} \left[-\partial_k (h_k \circ \varphi_k)\left(\boldsymbol{X}^{(i)}\right) {X_m^{(i)}}^{a-1} + (a-1)(h_k \circ \varphi_k)\left(\boldsymbol{X}^{(i)}\right) {X_m^{(i)}}^{a-2} \right] \boldsymbol{X}^{(i)^a} \\ &\qquad + a(h_k \circ \varphi_k)\left(\boldsymbol{X}^{(i)}\right) {X_m^{(i)}}^{2a-2} \boldsymbol{e}_{m,m} - a(h_k \circ \varphi_k)\left(\boldsymbol{X}^{(i)}\right) {X_k^{(i)}}^{a-1} {X_m^{(i)}}^{a-1} \boldsymbol{e}_{k,m}, \\ \boldsymbol{g}_{\boldsymbol{\eta},j} &\equiv \frac{1}{n} \sum_{i=1}^{n} -\partial_j (h_j \circ \varphi_j)\left(\boldsymbol{X}^{(i)}\right) {X_j^{(i)}}^{b-1} - (b-1)(h_j \circ \varphi_j)\left(\boldsymbol{X}^{(i)}\right) {X_j^{(i)}}^{b-2}, \end{aligned}$$

$$\boldsymbol{g}_{\boldsymbol{\eta},m} \equiv \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{m-1} \partial_k(h_k \circ \varphi_k)\left(\boldsymbol{X}^{(i)}\right) {X_m^{(i)}}^{b-1} - (b-1)(h_k \circ \varphi_k)\left(\boldsymbol{X}^{(i)}\right) {X_m^{(i)}}^{b-2}.$$

Instead of m diagonal blocks for a general $\mathcal{D} \subseteq \mathbb{R}^m$ of positive measure, in the simplex case each matrix has $(3m-2) = \mathcal{O}(m)$ blocks, so the computational complexity is the same as the general case.

We note that the profiled estimator in (3.22) for $\lambda_{\boldsymbol{\eta}} = 0$ is not practical in this case: The profiled quadratic matrix $\boldsymbol{\Gamma}_{\delta,\star} = \boldsymbol{\Gamma}_{\delta,\mathbf{K}} - \boldsymbol{\Gamma}_{\mathbf{K},\boldsymbol{\eta}}\boldsymbol{\Gamma}_{\boldsymbol{\eta}}^{-1}\boldsymbol{\Gamma}_{\mathbf{K},\boldsymbol{\eta}}^{\top}$ is no longer block-diagonal; in fact, all of its $m \times m$ blocks can be nonzero due to the new structure of $\boldsymbol{\Gamma}_{\delta,\mathbf{K}}$, $\boldsymbol{\Gamma}_{\mathbf{K},\boldsymbol{\eta}}$ and $\boldsymbol{\Gamma}_{\boldsymbol{\eta}}$. The inversion of $\boldsymbol{\Gamma}_{\boldsymbol{\eta}}$ also increases the computational burden and introduces more numerical error. Instead of profiling out $\boldsymbol{\eta}$ and solve for $\mathbf{K}$ only, it is thus recommended to directly perform coordinate-descent on both $\mathbf{K}$ and $\boldsymbol{\eta}$ without any penalty on $\boldsymbol{\eta}$.

3.6.2 log–log *Models on Standard Simplices*

In this section we consider the special case of $a = 0$ and $b = 0$, namely models with density proportional to $\exp\left(-\log \boldsymbol{x}^{\top}\mathbf{K}\log \boldsymbol{x}/2 + \boldsymbol{\eta}^{\top}\log \boldsymbol{x}\right)$ supported on the m-dimensional standard simplex. Central to these models, consider the A^{m-1} class of distributions formulated in (2.7) of Aitchison (1985) with parameters $\boldsymbol{\beta} \equiv (\beta_j)_{j=1,\ldots,m}$ and $(\gamma_{jk})_{1\le j\ne k\le m}$, where $\gamma_{jk} = \gamma_{kj}$, and density proportional to

$$\exp\left(-\frac{1}{2}\sum_{\substack{j\ne k\\ j,k=1,\ldots,m}} \gamma_{jk}(\log x_j - \log x_k)^2 + (\boldsymbol{\beta} - \mathbf{1}_m)^{\top}\log \boldsymbol{x}\right)\mathbb{1}_{\mathcal{D}}(\boldsymbol{x}).$$

Expanding the expression this can be rewritten as

$$\exp\left(-\sum_{j=1}^{m}(\log x_j)^2\left(\sum_{k\in\{1,\ldots,m\}\setminus\{j\}}\gamma_{jk}\right) + \sum_{\substack{j\ne k\\ j,k=1,\ldots,m}}\gamma_{jk}\log x_j \log x_k + (\boldsymbol{\beta}-\mathbf{1}_m)^{\top}\log \boldsymbol{x}\right)$$

and thus belong to our *a*-*b* model

$$\exp\left(-\frac{1}{2}\log \boldsymbol{x}^{\top}\mathbf{K}\log \boldsymbol{x} + \boldsymbol{\eta}^{\top}\log \boldsymbol{x}\right)$$

with $\boldsymbol{\eta} \equiv \boldsymbol{\beta} - \mathbf{1}_m$ as well as $\kappa_{jj} = 2\sum_{i\in\{1,\ldots,m\}\setminus\{j\}} \gamma_{ji}$ and $\kappa_{kj} = \kappa_{jk} = -2\gamma_{kj}$ for $j \neq k = 1, \ldots, m$. The A^{m-1} model thus translates to the a-b model with $a = b = 0$ and $\mathbf{K}\mathbf{1}_m = \mathbf{0}_m$ and $\mathbf{K} = \mathbf{K}^\top$.

In Theorem 21 we show that under simple conditions the density is again proper and the equivalent definition (3.10) of the generalized score matching loss is still valid. For estimation, we again consider $\mathcal{D}_{-m} \equiv \left\{\boldsymbol{x}_{-m} \in \mathbb{R}_+^{m-1} \middle| \, \boldsymbol{x}_{-m} \succ \mathbf{0},\, \mathbf{1}_{m-1}^\top \boldsymbol{x}_{-m} < 1\right\} \subseteq \mathbb{R}_+^{m-1}$ and profile out the last coordinate x_m in the density.

Theorem 21. *Suppose* $\mathbf{K}$ *is symmetric, and one of the following holds:*

(1) $\mathbf{K}$ *is positive definite, or*

(2) $\mathbf{K}\mathbf{1}_m = \mathbf{0}_m$, $\mathbf{K}_{-k,-k}$ *is positive definite for some* $k = 1, \ldots, m$, *and* $\mathbf{1}_m^\top \boldsymbol{\eta} + m \geq 0$, *or*

(3) $\mathbf{K}\mathbf{1}_m = \mathbf{0}_m$, $\mathbf{K}$ *is positive semi-definite, and* $\boldsymbol{\eta} \succ -\mathbf{1}_m$.

Then the density has a finite normalizing constant. Note that (2) implies that $\mathbf{K}$ *is positive semi-definite and (3) implies that* $\mathbf{K}_{-k,-k}$ *is positive semi-definite for all* k *(but not necessarily positive definite).*

For all $j = 1, \ldots, m-1$, *assume* $h_j(x) = x^{\alpha_j}$ *with* $\alpha_j > 0$ *and for (3) also assume* $\alpha_j > \max\{1 - \eta_{0,j}, 1 - \eta_{0,m}\}$, *then (A.1)–(A.3) in Lemma 13 are satisfied.*

Inference

As discussed in the beginning of Section 3.6, conditional dependence between two components cannot be inferred from $\mathbf{K}$ due to the nature of the simplex domain. However, one can still be interested in recovering the sparsity of $\mathbf{K}$ to see how different components interact, and recovering $\mathbf{K}$ and $\boldsymbol{\eta}$ to estimate the density itself. For this purpose, we have the following corollary of Theorem 20, i.e. $\mathbf{K}$ and $\boldsymbol{\eta}$ are exactly identifiable from the density (assuming $\mathbf{K}\mathbf{1}_m = \mathbf{0}_m$ or not):

Corollary 22. *Suppose there exist* $\mathbf{K}_1$, $\mathbf{K}_2$, $\boldsymbol{\eta}_1$, $\boldsymbol{\eta}_2$ *such that*

$$\exp\left(-\frac{1}{2}\log(\boldsymbol{x})^\top \mathbf{K}_1 \log(\boldsymbol{x}) + \boldsymbol{\eta}_1^\top \log(\boldsymbol{x})\right) = \exp\left(-\frac{1}{2}\log(\boldsymbol{x})^\top \mathbf{K}_2 \log(\boldsymbol{x}) + \boldsymbol{\eta}_2^\top \log(\boldsymbol{x})\right)$$

for all $\boldsymbol{x} \in \mathcal{D}$, *the standard simplex. Then* $\mathbf{K}_1 = \mathbf{K}_2$ *and* $\boldsymbol{\eta}_1 = \boldsymbol{\eta}_2$.

Now define the *additive log-ratio transformation* $\boldsymbol{y}_{-m} \equiv \text{alt}(\boldsymbol{x}) \equiv \log \boldsymbol{x}_{-m} - (\log x_m)\mathbf{1}_{m-1} = (\log(x_1/x_m), \ldots, \log(x_{m-1}/x_m))$. The A^{m-1} model, corresponding to (2) and (3) in Theorem 21, is proposed in Aitchison (1985) as a generalization of both the *additive logistic normal model* with a normal density in $\boldsymbol{y}_{-m}$ with inverse covariance $\mathbf{K}_{-m,-m}$ and mean $\mathbf{K}_{-m,-m}^{-1}\boldsymbol{\eta}_{-m}$, corresponding to (2) with $\mathbf{1}_m^\top \boldsymbol{\eta} = -m$, and the Dirichlet distribution with parameters $\boldsymbol{\eta} + \mathbf{1}_m$, corresponding to (3) with $\mathbf{K} = \mathbf{0}$. This generalization requires only one additional parameter than the additive logistic normal model, namely $\mathbf{1}_m^\top \boldsymbol{\eta}$ is no longer assumed to be $-m$. This brings in more flexibility to the two basic models over the simplex.

By nature of the simplex domains, any X_j and X_k are perfectly conditionally correlated given all other $\boldsymbol{X}_{-j,-k}$. On the other hand, under the additive logistic normal model, $Y_j = \log(X_j/X_m)$ and $Y_k = \log(X_k/X_m)$ are conditionally independent given all other $\log(X_\ell/X_m)$, $\ell \neq j, k, m$ if and only if $\kappa_{jk} = \kappa_{kj} = 0$; conversely, as we make clear in the proof of Theorem 21, this is true only for the additive logistic normal model ($\mathbf{1}_m^\top \boldsymbol{\eta} = -m$). We note that the statement in this paragraph still holds when we replace the ratios with respect to X_m by those w.r.t. any X_ℓ.

Assuming the more general A^{m-1} model, we can thus perform the following two-step test of conditional independence between X_j/X_ℓ and X_k/X_ℓ for triplet (j, k, ℓ) with j, k, ℓ all different. First test if $\mathbf{1}_m^\top \boldsymbol{\eta} + m$ is significantly different from 0 by comparing the BIC/eBIC for fitted A^{m-1} and additive logistic normal models with regularization discussed in Section 3.4.3; if we reject the null, then we cannot establish conditional independence for any such triplets. Otherwise, we claim that for all $\ell \neq j, k$, X_j/X_ℓ and X_k/X_ℓ are conditionally independent given all other X_i/X_ℓ ($i \neq j, k, \ell$) if and only if $\kappa_{jk} = \kappa_{kj} = 0$ in either fitted model. While the additive logistic normal model can be fitted as a Gaussian to the additive

log-ratio transformed data, the A^{m-1} model on the simplex can also be easily fitted with the help of generalized score matching, as discussed below.

Estimation for A^{m-1} Models

In Section 3.6.1 we formulate the elements for estimating $\mathbf{K}$ and $\boldsymbol{\eta}$ under the log–log model ($a = b = 0$) assuming (1) $\mathbf{K}$ is positive definite only. But for the A^{m-1} models (Aitchison, 1985) which assume the additional constraint $\mathbf{K}\mathbf{1}_m = \mathbf{0}_m$ and $\mathbf{K} = \mathbf{K}^\top$ as in (2) and (3) of Theorem 21, we need the following modification. We marginalize out the diagonals of $\mathbf{K}$ with $\kappa_{jj} = -\boldsymbol{\kappa}_{-j,j}^\top \mathbf{1}_{m-1}$ and estimate all off-diagonal elements. Under the assumption, for any matrices $\mathbf{A}$ with m rows and $\mathbf{B}$ with m columns, we can write

$$\begin{aligned}
\boldsymbol{\kappa}_{,j}^\top \mathbf{A} &= \boldsymbol{\kappa}_{-j,j}^\top \mathbf{A}_{-j,} + \kappa_{jj} \boldsymbol{a}_{j,}^\top = \boldsymbol{\kappa}_{-j,j}^\top \left(\mathbf{A}_{-j,} - \mathbf{1}_{m-1} \boldsymbol{a}_{j,}^\top \right) = \boldsymbol{\kappa}_{-j,j}^\top \left(\mathbf{C}^j \mathbf{A} \right), \\
\mathbf{B}\boldsymbol{\kappa}_{,j} &= \mathbf{B}_{,-j} \boldsymbol{\kappa}_{-j,j} + \boldsymbol{b}_{,j} \kappa_{jj} = \left(\mathbf{B}_{,-j} - \boldsymbol{b}_{,j} \mathbf{1}_{m-1}^\top \right) \boldsymbol{\kappa}_{-j,j} = \left(\mathbf{B} {\mathbf{C}^j}^\top \right) \boldsymbol{\kappa}_{-j,j},
\end{aligned}$$

where $\mathbf{C}^j \in \mathbb{R}^{(m-1)\times m}$ has its j-th column all equal to -1, and $(1,1), \ldots, (j-1, j-1), (j, j+1), \ldots, (m-1, m)$-th positions equal to 1, and 0 everywhere else. Let $\mathbf{C} \in \mathbb{R}^{(m-1)m \times m^2}$ be the block-diagonal matrix with blocks $\mathbf{C}^1, \ldots, \mathbf{C}^m$. The unpenalized generalized score-matching loss $\frac{1}{2}\mathrm{vec}\left(\left[\mathbf{K}^\top \ \boldsymbol{\eta}\right]\right)^\top \boldsymbol{\Gamma} \mathrm{vec}\left(\left[\mathbf{K}^\top \ \boldsymbol{\eta}\right]\right) - \boldsymbol{g}^\top \mathrm{vec}\left(\left[\mathbf{K}^\top \ \boldsymbol{\eta}\right]\right)$ thus becomes

$$\frac{1}{2}\mathrm{vec}\left(\begin{bmatrix} \mathbf{K}_{\text{off}} \\ \boldsymbol{\eta}^\top \end{bmatrix}\right)^\top \begin{bmatrix} \mathbf{C}\boldsymbol{\Gamma}_{\mathbf{K}}\mathbf{C}^\top & \mathbf{C}\boldsymbol{\Gamma}_{\mathbf{K},\boldsymbol{\eta}} \\ \boldsymbol{\Gamma}_{\mathbf{K},\boldsymbol{\eta}}^\top \mathbf{C}^\top & \boldsymbol{\Gamma}_{\boldsymbol{\eta}} \end{bmatrix} \mathrm{vec}\left(\begin{bmatrix} \mathbf{K}_{\text{off}} \\ \boldsymbol{\eta}^\top \end{bmatrix}\right) - \begin{bmatrix} \mathbf{C}\mathrm{vec}(\boldsymbol{g}_{\mathbf{K}}) \\ \boldsymbol{g}_{\boldsymbol{\eta}} \end{bmatrix}^\top \mathrm{vec}\left(\begin{bmatrix} \mathbf{K}_{\text{off}} \\ \boldsymbol{\eta}^\top \end{bmatrix}\right). \tag{3.30}$$

Denote the new components as $\tilde{\boldsymbol{\Gamma}} \in \mathbb{R}^{m^2 \times m^2}$ and $\tilde{\boldsymbol{g}} \in \mathbb{R}^{m^2}$. For any m^2-dimensional nonzero $\boldsymbol{v} \equiv (\boldsymbol{v}_{\mathbf{K},1}, \ldots, \boldsymbol{v}_{\mathbf{K},m}, \boldsymbol{v}_{\boldsymbol{\eta}}) \in \mathbb{R}^{m^2}$ with $\boldsymbol{v}_{\mathbf{K},j} \in \mathbb{R}^{m-1}$ and $\boldsymbol{v}_{\boldsymbol{\eta}} \in \mathbb{R}^m$, form $\tilde{\mathbf{V}} \in \mathbb{R}^{m \times m}$ whose j-th column has $\tilde{\boldsymbol{v}}_{-j,j} = \boldsymbol{v}_{\mathbf{K},j}$ and $\tilde{v}_{jj} = 1 - \mathbf{1}_{m-1}^\top \boldsymbol{v}_{\mathbf{K},j}$ (analogous to the inverse operation of $\mathbf{K} \mapsto \mathrm{vec}(\mathbf{K}_{\text{off}})$), and let $\tilde{\boldsymbol{v}} \equiv (\mathrm{vec}(\tilde{\mathbf{V}}), \boldsymbol{v}_{\boldsymbol{\eta}})$. Then by definition $\boldsymbol{v}^\top \tilde{\boldsymbol{\Gamma}} \boldsymbol{v} = \tilde{\boldsymbol{v}}^\top \boldsymbol{\Gamma} \tilde{\boldsymbol{v}}$. Since $\tilde{\boldsymbol{v}} \neq \mathbf{0}_{m(m+1)}$ if and only if $\boldsymbol{v} \neq \mathbf{0}_{m^2}$, $\tilde{\boldsymbol{\Gamma}}$ is positive (semi-)definite if and only if $\boldsymbol{\Gamma}$ is positive (semi-)definite.

Since $\tilde{\boldsymbol{\Gamma}}$ is merely a linear transformation of $\boldsymbol{\Gamma}$, we may apply diagonal multipliers directly to $\tilde{\boldsymbol{\Gamma}}$ for simplicity, and the concentration properties of the original $\boldsymbol{\Gamma}$ (high probability bounds

for deviation of $\mathbf{\Gamma}$ from its expectation) will be inherited, subject to different constants. The penalized generalized score-matching loss for the A^{m-1} models is thus defined as

$$\begin{aligned}&\hat{J}_{\boldsymbol{h},\boldsymbol{C},\mathcal{D},\lambda,\delta}(p_{\boldsymbol{\theta}})\\ &\equiv \frac{1}{2}\mathrm{vec}\left(\begin{bmatrix}\mathbf{K}_{\mathrm{off}}\\ \boldsymbol{\eta}^\top\end{bmatrix}\right)^\top \tilde{\mathbf{\Gamma}}_\delta(\mathbf{x})\mathrm{vec}\left(\begin{bmatrix}\mathbf{K}_{\mathrm{off}}\\ \boldsymbol{\eta}^\top\end{bmatrix}\right) - \tilde{\boldsymbol{g}}(\mathbf{x})^\top \mathrm{vec}\left(\begin{bmatrix}\mathbf{K}_{\mathrm{off}}\\ \boldsymbol{\eta}^\top\end{bmatrix}\right) + \lambda_{\mathbf{K}}\|\mathbf{K}_{\mathrm{off}}\|_1 + \lambda_{\boldsymbol{\eta}}\|\boldsymbol{\eta}\|_1,\end{aligned} \tag{3.31}$$

$$\tilde{\mathbf{\Gamma}}_\delta \equiv \begin{bmatrix}\left(\mathbf{C}\mathbf{\Gamma}_{\mathbf{K}}\mathbf{C}^\top\right)_\delta & \mathbf{C}\mathbf{\Gamma}_{\mathbf{K},\boldsymbol{\eta}}\\ \mathbf{\Gamma}_{\mathbf{K},\boldsymbol{\eta}}^\top\mathbf{C}^\top & \mathbf{\Gamma}_{\boldsymbol{\eta}}\end{bmatrix}, \qquad \tilde{\boldsymbol{g}} \equiv \begin{bmatrix}\mathbf{C}\mathrm{vec}(\boldsymbol{g}_{\mathbf{K}})\\ \boldsymbol{g}_{\boldsymbol{\eta}}\end{bmatrix},$$

where $(\cdot)_\delta$ denotes the operation of multiplying the diagonals of a matrix by $\delta > 1$.

3.7 Theoretical Properties

In this section we present theoretical guarantees for our generalized method applied to the pairwise interaction power *a-b* models. In particular, similar to Section 2.6, with high probability we bound the deviation of our estimates $\hat{\mathbf{K}}$ and $\hat{\boldsymbol{\eta}}$ from their true values $\mathbf{K}_0$ and $\boldsymbol{\eta}_0$. We first restate Definition 4 as below.

Definition 9. *Let* $\mathbf{\Gamma}_0 \equiv \mathbb{E}_0\mathbf{\Gamma}(\mathbf{x})$ *and* $\boldsymbol{g}_0 \equiv \mathbb{E}_0\boldsymbol{g}(\mathbf{x})$ *be the population versions of* $\mathbf{\Gamma}(\mathbf{x})$ *and* $\boldsymbol{g}(\mathbf{x})$ *under the distribution given by a true parameter matrix* $\mathbf{\Psi}_0 \equiv [\mathbf{K}_0, \boldsymbol{\eta}_0]^\top \in \mathbb{R}^{m(m+1)}$*, or* $\mathbf{\Psi}_0 \equiv \mathbf{K}_0 \in \mathbb{R}^{m^2}$ *in the centered case if we assume* $\boldsymbol{\eta}_0 \equiv \mathbf{0}$*. The support of a matrix* $\mathbf{\Psi}$ *is* $S(\mathbf{\Psi}) \equiv \{(i,j) : \psi_{ij} \neq 0\}$*, and we let* $S_0 = S(\mathbf{\Psi}_0)$*. For a matrix* $\mathbf{\Psi}_0$*, we define* $d_{\mathbf{\Psi}_0}$ *to be the maximum number of non-zero entries in any column, and* $c_{\mathbf{\Psi}_0} \equiv |||\mathbf{\Psi}_0|||_{\infty,\infty}$*. Writing* $\mathbf{\Gamma}_{0,AB}$ *for the* $A \times B$ *submatrix of* $\mathbf{\Gamma}_0$*, we define*

$$c_{\mathbf{\Gamma}_0} \equiv |||(\mathbf{\Gamma}_{0,S_0S_0})^{-1}|||_{\infty,\infty}. \tag{3.32}$$

Finally, $\mathbf{\Gamma}_0$ *satisfies* the irrepresentability condition with incoherence parameter $\omega \in (0,1]$ and edge set S_0 *if*

$$|||\mathbf{\Gamma}_{0,S_0^cS_0}(\mathbf{\Gamma}_{0,S_0S_0})^{-1}|||_{\infty,\infty} \leq (1-\omega). \tag{3.33}$$

3.7.1 Truncated Gaussian Graphical Models on A Finite Disjoint Union of Convex Sets with Positive Measure

Truncated Gaussian graphical models are covered by our *a-b* models described in Section 3.4.2 with $a = b = 1$. When the domain $\mathcal{D}$ is a finite disjoint union of convex sets with a positive Lebesgue measure, we have the following theorem similar to Theorem 10 in Chapter 2, which bounds the errors as long as one uses finite truncation points $\boldsymbol{C}$ for $\boldsymbol{\varphi}_{\boldsymbol{C},\mathcal{D}}$ and each component in $\boldsymbol{h}(\boldsymbol{x})$ is a power function with a positive exponent.

Theorem 23. *Let $\mathcal{D} \subset \mathbb{R}^m$ be a componentwise decomposable set (Def 5) with positive Lebesgue measure, and assume it is a finite disjoint union of convex sets $\boldsymbol{\Delta} \equiv \left\{\mathcal{D}_1, \ldots, \mathcal{D}_{|\boldsymbol{\Delta}|}\right\}$, i.e. $\mathcal{D} \equiv \mathcal{D}_1 \sqcup \cdots \sqcup \mathcal{D}_{|\boldsymbol{\Delta}|}$. Suppose the data matrix holds n i.i.d. copies of $\boldsymbol{X}$ following a truncated Gaussian distribution on $\mathcal{D}$ with inverse covariance parameter $\mathbf{K}_0 \in \mathbb{R}^{m\times m}$ and mean parameter $\boldsymbol{\mu}_0$, namely with density*

$$p_{\boldsymbol{\eta}_0,\mathbf{K}_0}(\boldsymbol{x}) \propto \exp\left(-\frac{1}{2}\boldsymbol{x}^\top \mathbf{K}_0 \boldsymbol{x} + \boldsymbol{\eta}_0^\top \boldsymbol{x}\right) \mathbb{1}_{\mathcal{D}}(\boldsymbol{x})$$

with $\mathbf{K}_0$ positive definite and $\boldsymbol{\eta}_0 \equiv \mathbf{K}_0\boldsymbol{\mu}_0$. Let $\boldsymbol{\Psi}_0 \equiv [\mathbf{K}_0, \boldsymbol{\eta}_0]^\top$.

Assume that (A.1)–(A.3) in Lemma 13 hold, and in addition that $\boldsymbol{h}$ and the truncation points $\boldsymbol{C}$ in the truncated componentwise distance $\boldsymbol{\varphi}_{\boldsymbol{C},\mathcal{D}}$ satisfy $0 \le \left(h_j \circ \varphi_{C_j,\mathcal{D},j}\right)(\boldsymbol{x}) \le M$ and $0 \le \partial_j \left(h_j \circ \varphi_{C_j,\mathcal{D},j}\right)(\boldsymbol{x}) \le M'$ almost surely for all $j = 1, \ldots, m$ for some constants $0 < M, M' < +\infty$. Note that $\boldsymbol{h}(\boldsymbol{x}) = (x_1^{\alpha_1}, \ldots, x_m^{\alpha_m})$ with $\alpha_1, \ldots, \alpha_m \ge 1$ satisfies all these assumptions, according to Theorem 18.

Let the diagonal multiplier δ introduced in Section 3.4.3 satisfy

$$1 < \delta < C(n,m) \equiv 2 - \left(1 + 4e \max\left\{\left(6\log m + 2\log|\boldsymbol{\Delta}|\right)/n, \sqrt{\left(6\log m + 2\log|\boldsymbol{\Delta}|\right)/n}\right\}\right)^{-1}$$

and suppose further that $\boldsymbol{\Gamma}_{0,S_0S_0}$ is invertible and satisfies the irrepresentability condition (2.24) with $\omega \in (0,1]$. Define $c_{\boldsymbol{X}} \equiv 2\max_{\mathcal{D}'\in\boldsymbol{\Delta}} \max_j \left|2\sqrt{(\mathbf{K}_0^{-1})_{jj}} + \sqrt{e}\,\mathbb{E}_0 X_j \mathbb{1}_{\mathcal{D}'}(\boldsymbol{X})\right|$. Suppose for $\tau > 3$ the sample size and the regularization parameter satisfy

$$n > \mathcal{O}\left(\left(\tau\log m + \log|\boldsymbol{\Delta}|\right)\max\left\{\frac{M^2 c_{\boldsymbol{\Gamma}_0}^2 c_{\boldsymbol{X}}^4 d_{\boldsymbol{\Psi}_0}^2}{\omega^2}, \frac{M c_{\boldsymbol{\Gamma}_0} c_{\boldsymbol{X}}^2 d_{\boldsymbol{\Psi}_0}}{\omega}\right\}\right), \tag{3.34}$$

$$\lambda > \mathcal{O}\left[(Mc_{\mathbf{\Psi}_0}c_{\boldsymbol{X}}^2 + M'c_{\boldsymbol{X}} + M)\left(\sqrt{\frac{\tau\log m + \log|\mathbf{\Delta}|}{n}} + \frac{\tau\log m + \log|\mathbf{\Delta}|}{n}\right)\right]. \quad (3.35)$$

Then the following statements hold with probability $1 - m^{3-\tau}$*:*

(a) The regularized generalized $\boldsymbol{h}$*-score matching estimator* $\hat{\mathbf{\Psi}}$ *that minimizes (3.18) is unique, has its support included in the true support,* $\hat{S} \equiv S(\hat{\mathbf{\Psi}}) \subseteq S_0$*, and satisfies*

$$\begin{aligned}
\|\hat{\mathbf{K}} - \mathbf{K}_0\|_\infty &\leq \frac{c_{\mathbf{\Gamma}_0}}{2-\omega}\lambda, & \|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0\|_\infty &\leq \frac{c_{\mathbf{\Gamma}_0}}{2-\omega}\lambda,\\
|||\hat{\mathbf{K}} - \mathbf{K}_0|||_F &\leq \frac{c_{\mathbf{\Gamma}_0}}{2-\omega}\lambda\sqrt{|S_0|}, & |||\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0|||_F &\leq \frac{c_{\mathbf{\Gamma}_0}}{2-\omega}\lambda\sqrt{|S_0|},\\
|||\hat{\mathbf{K}} - \mathbf{K}_0|||_2 &\leq \frac{c_{\mathbf{\Gamma}_0}}{2-\omega}\lambda\min\left(\sqrt{|S_0|}, d_{\mathbf{\Psi}_0}\right),\\
|||\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0|||_2 &\leq \frac{c_{\mathbf{\Gamma}_0}}{2-\omega}\lambda\min\left(\sqrt{|S_0|}, d_{\mathbf{\Psi}_0}\right).
\end{aligned}$$

(b) Moreover, if

$$\min_{j,k:(j,k)\in S_0}|\kappa_{0,jk}| > \frac{c_{\mathbf{\Gamma}_0}}{2-\omega}\lambda \qquad \textit{and} \qquad \min_{j:(m+1,j)\in S_0}|\eta_{0,j}| > \frac{c_{\mathbf{\Gamma}_0}}{2-\omega}\lambda,$$

then $\hat{S} = S_0$ *and* $\operatorname{sign}(\hat{\kappa}_{jk}) = \operatorname{sign}(\kappa_{0,jk})$ *for all* $(j,k) \in S_0$ *and* $\operatorname{sign}(\hat{\eta}_j) = \operatorname{sign}(\eta_{0j})$ *for* $(m+1, j) \in S_0$*.*

In the centered setting, the same bounds hold by removing the dependencies on $\hat{\boldsymbol{\eta}}$ *and* $\boldsymbol{\eta}_0$*.*

The detailed proof of the theorem is omitted, since it is analogous to that for Theorem 10 of Chapter 2 with two modifications: (i) using the triangle inequality, split the concentration bounds (A.6) (A.10) (A.11) in the Appendix for Chapter 2 into one for each $\mathcal{D}_1 \ldots, \mathcal{D}_{|\mathbf{\Delta}|}$ and use a union bound; (ii) in the proof of Lemma 36.1, replace $\mathcal{D} \equiv \mathbb{R}_+^m$ by any convex $\mathcal{D}' = \mathcal{D}_1, \ldots, \mathcal{D}_{|\mathbf{\Delta}|}$ and replace X_1 by $X_1\mathbb{1}_{\mathcal{D}'}(\boldsymbol{X})$, as the proof there only uses the convexity of the domain.

3.7.2 Bounded Domains in $\mathbb{R}_+^m$ *with Positive Measure*

In this section we present results for general *a-b* models on bounded domains with positive measure.

Theorem 24. *(1) Suppose $a > 0$ and $b \geq 0$. Let $\mathcal{D}$ be a bounded subset of $\mathbb{R}_+^m$ with positive Lebesgue measure with $\mathcal{D} \subseteq [u_1, v_1] \times \cdots \times [u_m, v_m]$ for finite nonnegative constants $u_1, v_1, \ldots, u_m, v_m$, and suppose that the true parameters $\mathbf{K}_0$ and $\boldsymbol{\eta}_0$ satisfy the conditions in Corollary 17 (so that the density is proper). Assume $\boldsymbol{h}(\boldsymbol{x}) \equiv (x_1^{\alpha_1}, \ldots, x_m^{\alpha_m})$ with $\alpha_1, \ldots, \alpha_m \geq \max\{1, 2-a, 2-b\}$, and suppose $\boldsymbol{\varphi}_{\boldsymbol{C},\mathcal{D}}$ has truncation points $\boldsymbol{C} = (C_1, \ldots, C_m)$ with $0 < C_j < +\infty$ for $j = 1, \ldots, m$. Define*

$$\zeta(C_j, u_j, v_j, \alpha_j, p_j) \equiv \begin{cases} \min\{C_j, (v_j - u_j)/2\}^{\alpha_j} (u_j + v_j)^{p_j} / 2^{p_j}, & p_j < 0, v_j - u_j \leq 2C_j, \\ \min\{C_j, (v_j - u_j)/2\}^{\alpha_j} (u_j + C_j)^{p_j}, & p_j < 0, v_j - u_j > 2C_j, \\ \min\{C_j, (v_j - u_j)/2\}^{\alpha_j} v_j^{p_j}, & p_j \geq 0, \end{cases}$$

$$\varsigma_{\boldsymbol{\Gamma}} \equiv \max_{j,k=1,\ldots,m} \max\left\{\zeta(C_j, u_j, v_j, \alpha_j, 2a-2) v_k^{2a},\, \zeta(C_j, u_j, v_j, \alpha_j, 2b-2)\right\},$$

$$\varsigma_{\boldsymbol{g}} \equiv \max_{j,k=1,\ldots,m} \max\big\{\alpha_j \zeta(C_j, u_j, v_j, \alpha_j - 1, a-1) v_k^a + |a-1| \zeta(C_j, u_j, v_j, \alpha_j, a-2) v_k^a + a\zeta(C_j, u_j, v_j, \alpha_j, 2a-2),$$

$$\alpha_j \zeta(C_j, u_j, v_j, \alpha_j - 1, b-1) + |b-1| \zeta(C_j, u_j, v_j, \alpha_j, b-2)\big\}.$$

Suppose that $\boldsymbol{\Gamma}_{0,S_0S_0}$ is invertible and satisfies the irrepresentability condition (3.33) with $\omega \in (0, 1]$. Suppose for $\tau > 0$ the sample size, the regularization parameter and the diagonal multiplier δ from Section 3.4.3 satisfy

$$n > 72 c_{\boldsymbol{\Gamma}_0}^2 d_{\boldsymbol{\Psi}_0}^2 \varsigma_{\boldsymbol{\Gamma}}^2 (\tau \log m + \log 2)/\omega^2, \tag{3.36}$$

$$\lambda > \frac{3(2-\omega)}{\omega} \max\left\{c_{\boldsymbol{\Psi}_0} \varsigma_{\boldsymbol{\Gamma}} \sqrt{2(\tau \log m + \log 4)/n}, \varsigma_{\boldsymbol{g}} \sqrt{(\tau \log m + \log 4)/(2n)}\right\}, \tag{3.37}$$

$$1 < \delta < C_{bounded}(n, m, \tau) \equiv 1 + \sqrt{(\tau \log m + \log 4)/(2n)}. \tag{3.38}$$

Then the statements (a) and (b) in Theorem 23 hold with probability at least $1 - m^{-\tau}$.

(2) For $a = 0$ and $b \geq 0$, if $u_j > 0$ for all $k = 1, \ldots, m$, the above holds with

$$\varsigma_{\boldsymbol{\Gamma}} \equiv \max_{j,k=1,\ldots,m} \max\{\zeta(C_j, u_j, v_j, \alpha_j, -2) \max\{|\log u_k|, |\log v_k|\}^2,\, \zeta(C_j, u_j, v_j, \alpha_j, 2b-2)\},$$

$$\varsigma_{\boldsymbol{g}} \equiv \max_{j,k=1,\ldots,m} \max\{\alpha_j \zeta(C_j, u_j, v_j, \alpha_j - 1, -1) \max\{|\log u_k|, |\log v_k|\}$$

$$+ |a-1|\zeta(C_j, u_j, v_j, \alpha_j, -2)\max\{|\log u_k|, |\log v_k|\},$$
$$\alpha_j\zeta(C_j, u_j, v_j, \alpha_j - 1, b-1) + |b-1|\zeta(C_j, u_j, v_j, \alpha_j, b-2)\}.$$

We note that the requirement on $\alpha_j \geq 1$ is only for bounding the two $\partial_j(h_j \circ \varphi_j)$ terms in $\boldsymbol{g}(\mathbf{x})$ and might not be necessary in practice as we see in the simulation studies.

3.7.3 Unbounded Domains in $\mathbb{R}_+^m$ with Positive Measure

For unbounded domains $\mathcal{D}$ in the non-negative orthant, we also have consistency results, albeit introducing an additional unknown constant factor that may depend on m to the sample complexity. For simplicity we only show the results for $a > 0$ and those for $a = 0$ should be similar. The following lemma enables us to bound each row of the data matrix $\mathbf{x}$ by a finite cube with high probability and then apply the same proof as for Theorem 24.

Lemma 25. *Suppose $\mathcal{D}$ has positive measure and the true parameters $\mathbf{K}_0$ and $\boldsymbol{\eta}_0$ satisfy the conditions in Corollary 17. Then for all $j = 1, \ldots, m$, X_j^{2a} is sub-exponential if $a > 0$ and $\log X_j$ is sub-exponential if $a = 0$.*

We have the following corollary of Theorem 24. The results involve an unknown constant, namely the sub-exponential norm $\|\cdot\|_{\psi_1}$ of X_j^{2a}, and we have roughly and asymptotically $n = \Omega(\log m)\max_j \mathcal{O}\left(\left\|X_j^{2a}\right\|_{\psi_1}\right)^{(\alpha_j+\max\{4a,2b\}-2)/a}$. We suspect the sub-exponential norm to scale like $\Omega\left((\log m)^c\right)$ for some c small, and the exact dependency on m is left for further research.

Corollary 26. *Suppose $a > 0$ and $\mathcal{D}$ is a subset of $\mathbb{R}_+^m$ with positive measure and suppose that the true parameters $\mathbf{K}_0$ and $\boldsymbol{\eta}_0$ satisfy the conditions in Corollary 17. Let $\rho_j(\mathcal{D}) \equiv \overline{\{x_j : \boldsymbol{x} \in \mathcal{D}\}}$ and $\rho_{\mathcal{D}}^* \equiv \{j = 1, \ldots, m : \sup \rho_j(\mathcal{D}) < +\infty\}$, and suppose $\rho_j(\mathcal{D}) \subseteq [u_j, v_j]$ for $j \in \rho_{\mathcal{D}}^*$. Then Theorem 24 holds with $\log 4$ replaced by $\log 6$ in (3.36)–(3.38), and $u_j = \max\left\{\mathbb{E}_0 X_j^{2a} - \epsilon_{3,j}, 0\right\}^{1/(2a)}$ and $v_j = \left(\mathbb{E}_0 X_j^{2a} + \epsilon_{3,j}\right)^{1/(2a)}$ for $j \notin \rho_{\mathcal{D}}^*$, where*

$$\epsilon_{3,j} \equiv \max\Bigg\{2\sqrt{2}e\left\|X_j^{2a}\right\|_{\psi_1}\sqrt{\log 3 + \log n + \tau\log m + \log\left(m - |\rho_{\mathcal{D}}^*|\right)},$$

$$4e \left\|X_j^{2a}\right\|_{\psi_1} (\log 3 + \log n + \tau \log m + \log (m - |\rho_{\mathcal{D}}^*|))\Big\},$$

$$\left\|X_j^{2a}\right\|_{\psi_1} \equiv \sup_{q \geq 1} \left(\mathbb{E}_0 |X_j|^{2aq}\right)^{1/q} / q \geq \mathbb{E}_0 X_j^{2a}.$$

3.7.4 Models on the Standard Simplices

For models with $a > 0$ on the simplex, since each coordinate is in $[0, 1]$, we have the following corollary of Theorem 24.

Corollary 27. *Suppose $a > 0$ and $b \geq 0$, and let $\mathcal{D} \equiv \left\{\boldsymbol{x} \in \mathbb{R}_+^m \middle| \boldsymbol{x} \succ \mathbf{0}, \mathbf{1}_m^\top \boldsymbol{x} = 1\right\}$ be the simplex domain in $\mathbb{R}_+^m$. Suppose that the true parameters $\mathbf{K}_0$ and $\boldsymbol{\eta}_0$ satisfy the conditions in Corollary 17. Then for the minimizer of (3.18) using the $\boldsymbol{\Gamma}$ and $\boldsymbol{g}$ formulated in Section 3.6.1, Theorem 24 holds with $\varsigma_{\boldsymbol{\Gamma}}$ and $\varsigma_{\boldsymbol{g}}$ replaced by $\varsigma_{\boldsymbol{\Gamma}} \equiv 1$ and $\varsigma_{\boldsymbol{g}} \equiv \max_{j=1,\ldots,m} \alpha_j + \max\{|a - 1| + 2a, |b - 1|\}$.*

In the proof of Corollary 27 we show that we can give tighter constant bounds $\varsigma_{\boldsymbol{\Gamma}}$ and $\varsigma_{\boldsymbol{g}}$ for entries in $\boldsymbol{\Gamma}$ and $\boldsymbol{g}$, respectively, which may be a lot smaller but have rather complicated forms.

For models with $a = 0$ on simplex domains, including the A^{m-1} models discussed in Section 3.6.2 we first use the following lemma to bound $\log X_j$ with high probability, just as in Lemma 25.

Lemma 28. *Suppose $\mathcal{D}$ is the standard simplex, and the true parameters $\mathbf{K}_0$ and $\boldsymbol{\eta}_0$ satisfy the conditions in Corollary 17 for $a > 0$ or $b > 0$, or in Theorem 21 for $a = b = 0$. Then for all $j = 1, \ldots, m$, X_j^{2a} is sub-exponential for $a > 0$, and $\log X_j$ is sub-exponential for $a = 0$.*

We then have the following corollary of Theorem 24.

Corollary 29. *Suppose $\mathcal{D}$ is the simplex domain in $\mathbb{R}_+^m$ and $a = 0$. Suppose $b = 0$ and the conditions for $\mathbf{K}_0$ and $\boldsymbol{\eta}_0$ in Theorem 21 hold, or $b > 0$ and the condition in Corollary 17 hold. Let $\boldsymbol{h}(\boldsymbol{x}) \equiv (x_1^{\alpha_1}, \ldots, x_m^{\alpha_m})$ with $\alpha_1, \ldots, \alpha_m \geq 2$. Then Theorem 24 holds with $\log 4$ replaced by $\log 6$ in (3.36)–(3.38), and*

$$\varsigma_{\boldsymbol{\Gamma}} \equiv \max\left\{1, c_{\log, \mathbf{K}_0, \boldsymbol{\eta}_0}^2\right\},$$

$$\varsigma_{\boldsymbol{g}} \equiv \max\left\{\left(\max_{j=1,\ldots,m}\alpha_j+1\right)c_{\log,\mathbf{K}_0,\boldsymbol{\eta}_0}+2, \max_j \alpha_j+|b-1|\right\}, \quad \textit{where}$$

$$c_{\log,\mathbf{K}_0,\boldsymbol{\eta}_0} \equiv \max_j \mathbb{E}_0 \log X_j + \max\left\{2\sqrt{2}e \max_{j=1,\ldots,m}\|\log X_j\|_{\psi_1}\sqrt{\log 3+\log n+(\tau+1)\log m},\right.$$
$$\left. 4e \max_{j=1,\ldots,m}\|\log X_j\|_{\psi_1}(\log 3+\log n+(\tau+1)\log m)\right\},$$

$$\|\log X_j\|_{\psi_1} \equiv \sup_{q\geq 1}\left(\mathbb{E}_0|\log X_j|^q\right)^{1/q}/q \geq -\mathbb{E}_0 \log X_j.$$

The results are written in terms of the max of the sub-exponential norms of $\log X_1$, …, $\log X_m$, an unknown constant, and we have roughly $n = \Omega(\log m)\,\mathcal{O}\left(\max_j \|\log X_j\|_{\psi_1}\right)^2$. We expect the sub-exponential norm to scale like $\Omega((\log m)^c)$ for some c small, and the exact dependency on m is left for further research.

3.8 Numerical Experiments

In this section we present results from numerical experiments using our method formulated in Section 3.3.2 and 3.3.5, as well as our extension to Liu and Kanamori (2019) presented in Section 3.3.6. The plots we report are in a similar fashion to those in Section 2.7.

3.8.1 Estimation — Choice of $\boldsymbol{h}$ and $\boldsymbol{C}$

Recall that multiplication of $\nabla \log p(\boldsymbol{x})$ with an $(\boldsymbol{h}\circ\boldsymbol{\varphi}_{\boldsymbol{C},\mathcal{D}})^{1/2}(\boldsymbol{x})$ is key to our method, in which the j-th component of $\boldsymbol{\varphi}_{\boldsymbol{C},\mathcal{D}}(\boldsymbol{x}) = (\varphi_{C_1,\mathcal{D},1}(\boldsymbol{x}),\ldots,\varphi_{C_m,\mathcal{D},m}(\boldsymbol{x}))$ is the distance of x_j to the boundary of its domain holding $\boldsymbol{x}_{-j}$ fixed, with this distance truncated from above by some constant $C_j > 0$.

We use a uniform h for all components ($\boldsymbol{h}(\boldsymbol{x}) = (h(x_1),\ldots,h(x_m))$) and compare the performance of our method using various $h(x)$ of the form x^c for some power $c \geq 0$ along with various truncation points $\boldsymbol{C}$. In particular, we choose $c = i/4$ for $i = 0, 1, \ldots, 8$. Instead of prespecifying constants $\boldsymbol{C}$, given $0 < \pi \leq 1$ we choose each C_j to be the π sample quantiles of $\varphi_{+\infty,\mathcal{D},j}$ applied to each row of the data matrix $\mathbf{x}$, namely the π quantile of

$\{\varphi_{+\infty,\mathcal{D},j}(\boldsymbol{x}^{(1)}), \ldots, \varphi_{+\infty,\mathcal{D},j}(\boldsymbol{x}^{(n)})\}$, assuming there are n samples in the data; infinite $\varphi_{+\infty,\mathcal{D},j}$ values are ignored. If all untruncated distances for the j-th component in the sample are infinite, i.e. $\varphi_{+\infty,\mathcal{D},j}(\boldsymbol{x}^{(1)}) = \cdots = \varphi_{+\infty,\mathcal{D},j}(\boldsymbol{x}^{(n)}) = +\infty$, we set $\varphi_j \equiv 1$. This dynamic way of choosing the truncation points allows us to automatically adapt to the scale of data, and is more informative than fixing the constant to a grid from 0.5 to 3 as done in Section 2.7. In our experiments, we choose $\pi = 0.2, 0.4, 0.6, 0.8, 1$, where $\pi = 1$ means no truncation for all finite φ_j values.

We remind that with power $c = 0$, $(\boldsymbol{h} \circ \boldsymbol{\varphi}_{\boldsymbol{C},\mathcal{D}})(\boldsymbol{x}) \equiv 1$ and it corresponds to the original score-matching for $\mathbb{R}^m$ of Hyvärinen (2005), while with $c = 2$, $\boldsymbol{C} = +\infty^m$ and $\mathcal{D} \equiv \mathbb{R}^m_+$, $(\boldsymbol{h} \circ \boldsymbol{\varphi}_{\boldsymbol{C},\mathcal{D}})(\boldsymbol{x}) \equiv \boldsymbol{x}^2$ corresponds to the estimator of Hyvärinen (2007) and Lin et al. (2016). The case where $\mathcal{D} \equiv \mathbb{R}^m_+$ corresponds to Section 2.7.

Finally, we also include results for our extension to the method proposed in Liu and Kanamori (2019) using $g_0(\boldsymbol{x})$ as opposed to $(\boldsymbol{h} \circ \boldsymbol{\varphi}_{\boldsymbol{C},\mathcal{D}})^{1/2}(\boldsymbol{x})$, taking the ℓ_2 distance to the boundary $\partial\mathcal{D}$ upper truncated by a constant C; see Section 3.3.6. The constant C in this case is also determined using quantiles of the untruncated ℓ_2 distances of the given data sample to $\partial\mathcal{D}$. For $C = +\infty$ ($\pi = 1$) there is no truncation and the estimator corresponds to Liu and Kanamori (2019).

3.8.2 Numerical Experiments for Domains with Positive Measure

In this section we present results for general a-b models restricted to domains with positive Lebesgue measure.

Experiment Setup

Throughout our experiments we choose dimension $m = 100$ and sample sizes $n = 80$ and 1000. For simplicity and given the length of this dissertation, we only present results for the centered case (assuming $\boldsymbol{\eta} \equiv \mathbf{0}$) where the b power does not come into play in the distribution, i.e. the density is proportional to $\exp\left\{-\boldsymbol{x}^{a\top}\mathbf{K}\boldsymbol{x}^a/(2a)\right\}$ for $a > 0$ or $\exp\left(-\log \boldsymbol{x}^\top \mathbf{K} \log \boldsymbol{x}/2\right)$ for $a = 0$. Supported by the exhaustive experiments of the same kind carried out in Section

2.7, we expect that the results for non-centered settings are similar, meaning that the best choice of h mainly depends on a but not b.

We consider six settings for a: (1) $a = 0$ (log), (2) $a = 1/2$ (exponential square root; Inouye et al. (2016)), (3) $a = 1$ (Gaussian), (4) $a = 3/2$ as well as some extreme cases (5) $a = 2$ and (6) $a = 3$. For all settings, we consider the following subsets of $\mathbb{R}_+^m$ as our domain $\mathcal{D}$:

i) non-negative ℓ_2 ball $\{\boldsymbol{x} \in \mathbb{R}_+^m : \|\boldsymbol{x}\|_2 \leq c_1\}$, which we call ℓ_2-nn ("non-negative"),

ii) complement of ℓ_2 ball in $\mathbb{R}_+^m$: $\{\boldsymbol{x} \in \mathbb{R}_+^m : \|\boldsymbol{x}\|_2 \geq c_1\}$, which we call $\ell_2^{\complement}$-nn, and

iii) $[c_1, +\infty)^m$, which we call *unif-nn*,

for some $c_1 > 0$. For the Gaussian ($a = 1$) case consider in addition the following subsets of $\mathbb{R}^m$:

iv) the entire ℓ_2 ball $\{\boldsymbol{x} \in \mathbb{R}^m : \|\boldsymbol{x}\|_2 \leq c_1\}$, which we call ℓ_2,

v) the complement of ℓ_2 ball in $\mathbb{R}^m$: $\{\boldsymbol{x} \in \mathbb{R}^m : \|\boldsymbol{x}\|_2 \geq c_1\}$, which we call $\ell_2^{\complement}$, and

vi) $((-\infty, c_1] \cup [c_1, +\infty))^m$, which we call *unif*.

The constant c_1 in each setting above is determined in the following way. We first generate n samples from the corresponding untruncated distribution on $\mathbb{R}_+^m$ for i)–iii) or $\mathbb{R}^m$ for iv)–vi), then determine the c_1 so that exactly half of the samples would fall inside the truncated boundary. We believe that the c_1 chosen adaptively in this way would be more meaningful and the truncation would not be too aggressive, similar to data one might see in practice.

The underlying true interaction matrices $\mathbf{K}_0$ are selected as in Lin et al. (2016) and Section 2.7. Each graph has 10 disconnected subgraphs, each containing $m/10 = 10$ nodes, making $\mathbf{K}_0$ block-diagonal. In each block, each lower-triangular element is set to 0 with probability $1 - \rho$ for some $\rho \in (0, 1)$, and is otherwise drawn from Uniform$[0.5, 1]$. The upper

triangular elements are determined by symmetry. The diagonal elements of $\mathbf{K}_0$ are chosen as a common positive value such that the minimum eigenvalue of $\mathbf{K}_0$ is 0.1. We generate 5 different true precision matrices $\mathbf{K}_0$, and run 10 trials using each of them. We choose $\rho = 0.8$ for $n = 1000$ and $\rho = 0.2$ for $n = 80$. This way $n/(d^2_{\mathbf{K}_0} \log m)$ is roughly constant, coinciding with our theory in Section 3.7.

Results

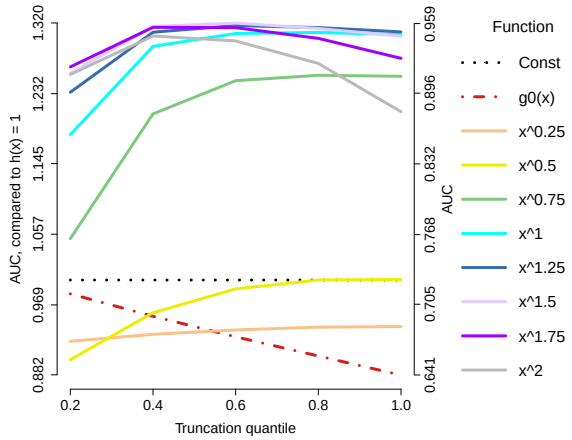
We use the area under the ROC curve (AUC) as the measure of performance of edge recovery of $\mathbf{K}_0$. Specifically, writing the estimate of $\mathbf{K}_0$ as $\hat{\mathbf{K}}$, the ROC curve plots the true positive rate (TPR) against the false positive rate (FPR), with

$$\text{FPR} \equiv \frac{|\hat{S}_{\text{off}} \backslash S_{0,\text{off}}|}{m(m-1) - |S_{0,\text{off}}|} \qquad \text{and} \qquad \text{TPR} \equiv \frac{|\hat{S}_{\text{off}} \cap S_{0,\text{off}}|}{|S_{0,\text{off}}|},$$

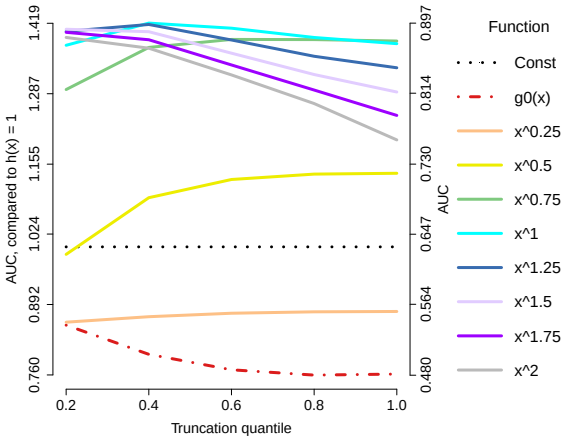
where $S_{0,\text{off}} \equiv \{(i,j) : i \neq j \wedge \kappa_{0,i,j} \neq 0\}$, and $\hat{S}_{\text{off}} \equiv \{(i,j) : i \neq j \wedge \hat{\kappa}_{i,j} \neq 0\}$. We plot the AUC averaged over 50 trials for each setting against $\pi = 0.2, 0.4, 0.6, 0.8, 1$, whose sample quantiles are used as the truncation points $\boldsymbol{C}$ of $\boldsymbol{\varphi}_{\boldsymbol{C},\mathcal{D}}$ (c.f. Section 3.8.1), while each curve represents $h(x) = 1$ (Hyvärinen, 2005), $g_0(\boldsymbol{x})$ from Section 3.3.6, or $h(x) = x^c$ with $c = 1/4, 1/2, \ldots, 2$. The y-ticks on the right-hand side are the original AUC values, whereas those on the left are the AUCs divided by the AUC for $h(x) = 1$, measuring the relative performance of each method compared to the original score matching in Hyvärinen (2005). (Note that $h(x) = 1$ is invariant to the truncation points and is thus constant in each plot.)

From the plots we conclude that in most settings our method using $h(x) = x^c$ with a c close to $\max\{2-a, 0\}$ works the best, as we observed in Section 2.7, and in most settings the truncated g_0 function does not work well (Liu and Kanamori (2019) corresponds to $\pi = 1$). The only notable exception is the domains iv)–vi) above, namely Gaussian models on subsets of $\mathbb{R}^m$ not restricted to $\mathbb{R}^m_+$, shown in Figure 3.7. In particular, the original score matching in Hyvärinen (2005) designed for densities in the entire $\mathbb{R}^m$ seems to work the best in these settings, suggesting that estimation of GGMs for domains with such truncations might not be challenging enough for one to switch to the more complex generalized methods. The

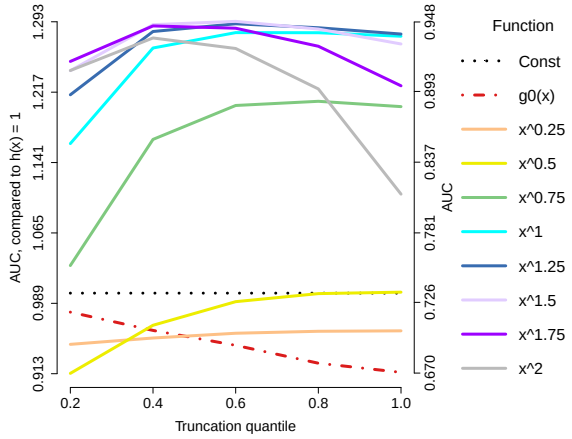
reason behind this is left for future research. On the other hand, by reading the y ticks on the left side of the plots in Figure 3.7, one should however note that for the iv) ℓ_2 and v) $\ell_2^{\mathbb{C}}$ domains, the difference in the performance of all estimators is insignificant.



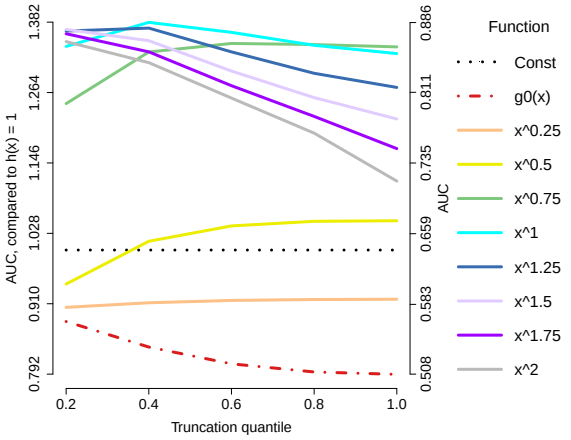
(a) $n = 80$, ℓ_2-nn domain



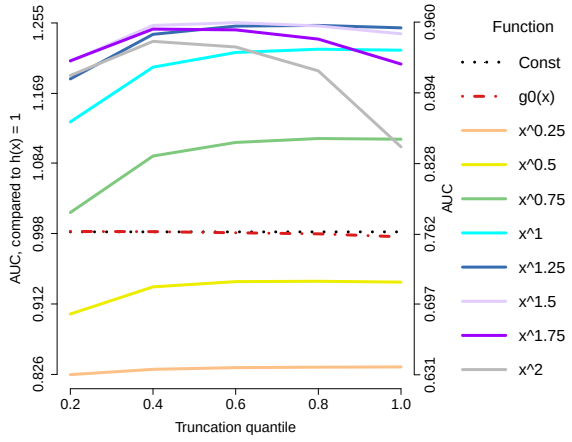
(b) $n = 1000$, ℓ_2-nn domain



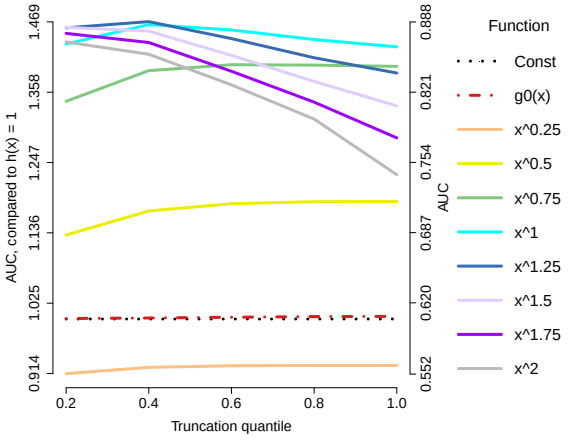
(c) $n = 80$, $\ell_2^{\mathbf{C}}$-nn domain



(d) $n = 1000$, $\ell_2^{\mathbf{C}}$-nn domain

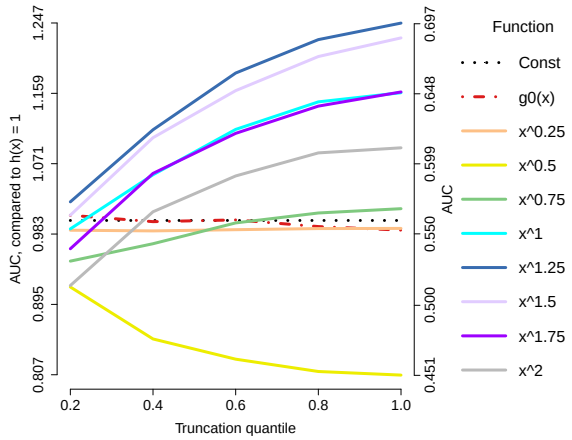


(e) $n = 80$, unif-nn domain

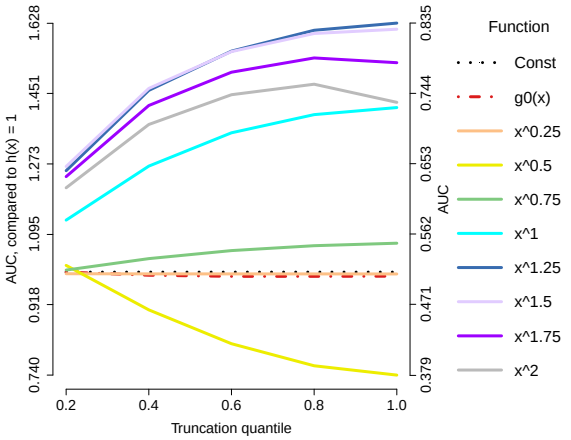


(f) $n = 1000$, unif-nn domain

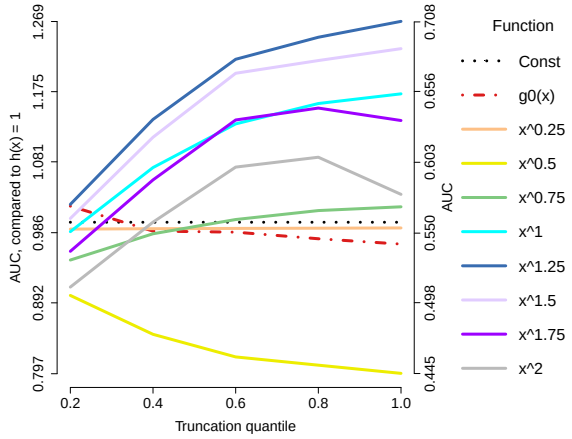
Figure 3.4: AUCs averaged over 50 trials for edge recovery using generalized score matching for the log models ($a = 0$). Each curve represents using either our extension to $g_0(\boldsymbol{x})$ proposed in Liu and Kanamori (2019) or a different choice of power function $h(x) = x^c$, and the x axis marks the π values whose columnwise sample quantiles are used as the truncation points $\boldsymbol{C}$ for the truncated componentwise distances. The colors are sorted by the power c.



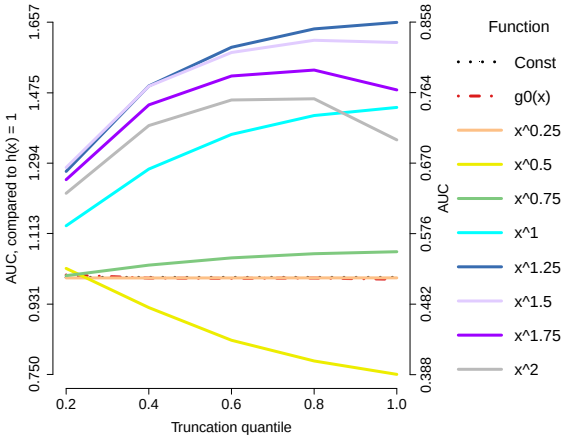
(a) $n = 80$, ℓ_2-nn domain



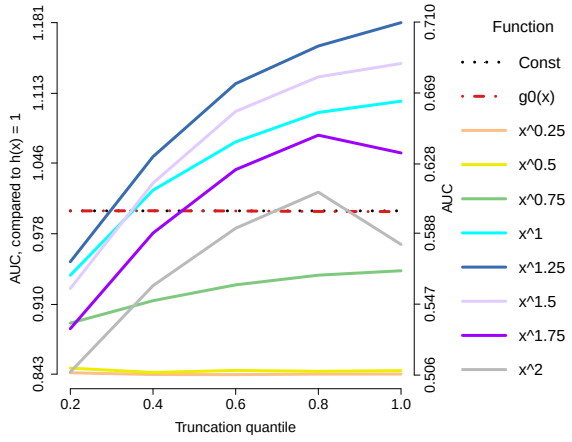
(b) $n = 1000$, ℓ_2-nn domain



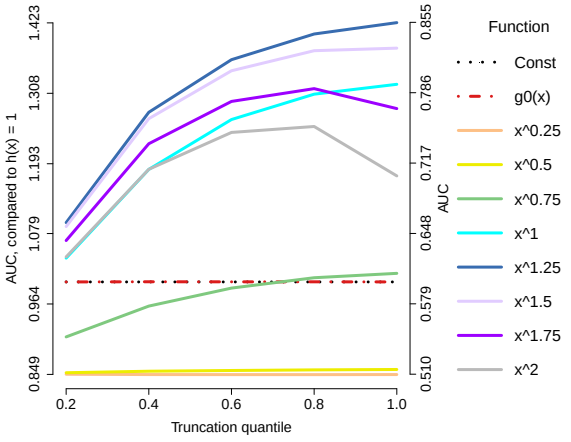
(c) $n = 80$, $\ell_2^{\complement}$-nn domain



(d) $n = 1000$, $\ell_2^{\complement}$-nn domain

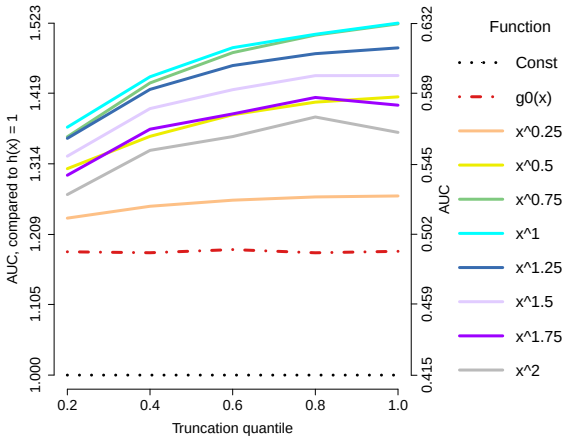


(e) $n = 80$, unif-nn domain

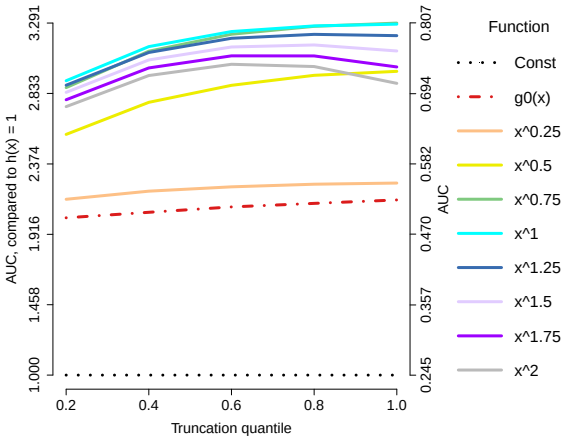


(f) $n = 1000$, unif-nn domain

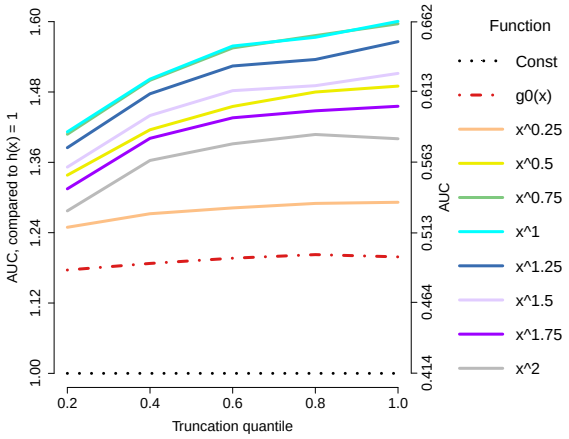
Figure 3.5: AUCs averaged over 50 trials for edge recovery using generalized score matching for the exponential square-root models ($a = 1/2$).



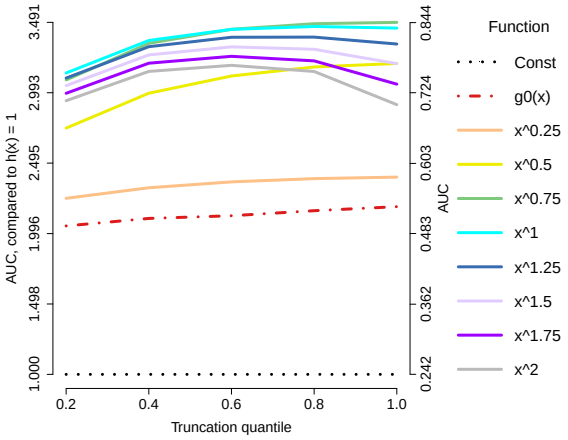
(a) $n = 80$, ℓ_2-nn domain



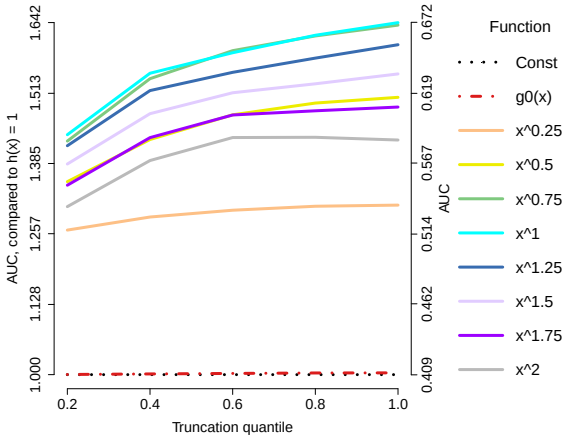
(b) $n = 1000$, ℓ_2-nn domain



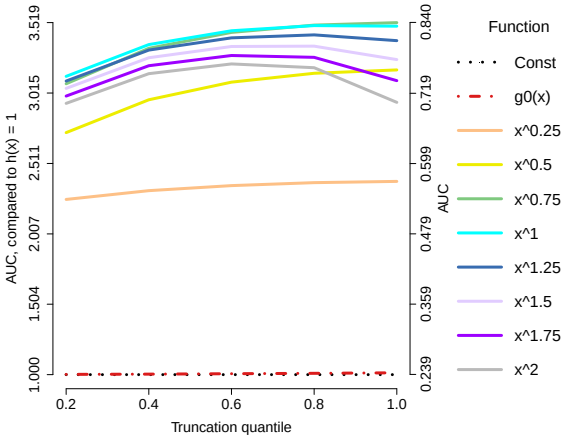
(c) $n = 80$, $\ell_2^{\mathbb{C}}$-nn domain



(d) $n = 1000$, $\ell_2^{\mathbb{C}}$-nn domain

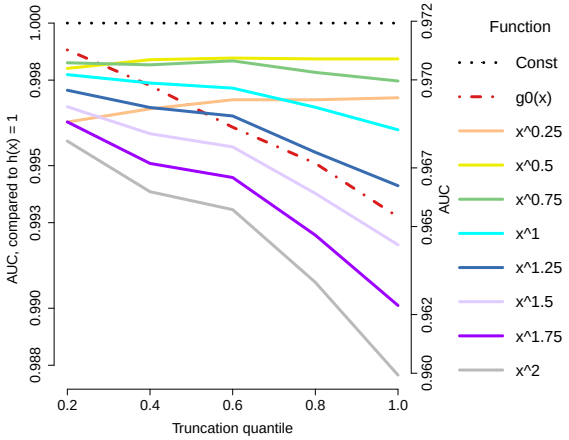


(e) $n = 80$, unif-nn domain

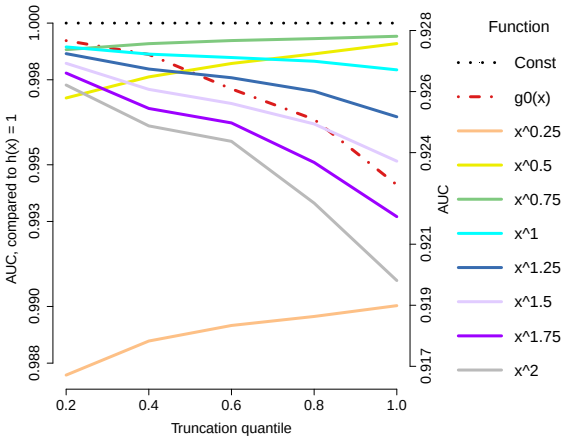


(f) $n = 1000$, unif-nn domain

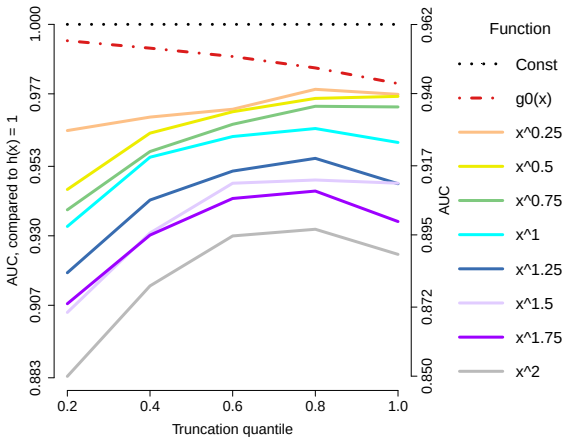
Figure 3.6: AUCs averaged over 50 trials for edge recovery using generalized score matching for the Gaussian models ($a = 1$) on domains being subsets of $\mathbb{R}_+^m$.



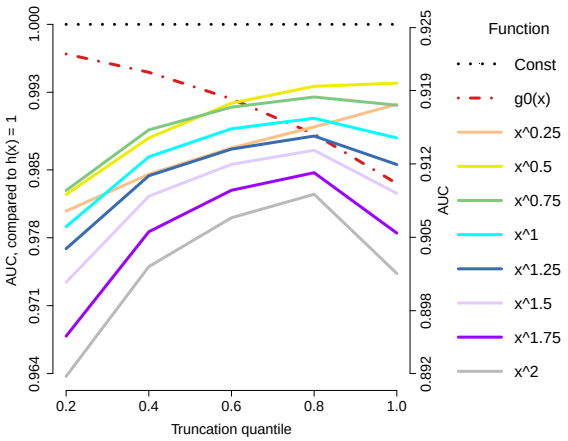
(a) $n = 80$, ℓ_2 domain



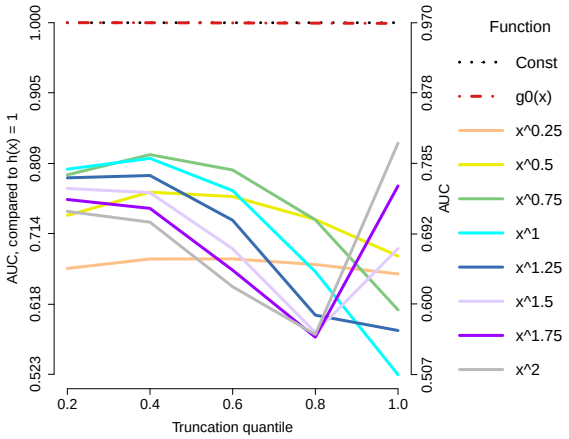
(b) $n = 1000$, ℓ_2 domain



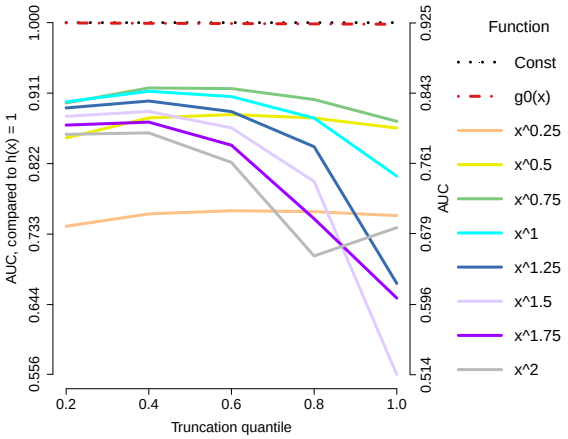
(c) $n = 80$, $\ell_2^{\mathbb{C}}$ domain



(d) $n = 1000$, $\ell_2^{\mathbb{C}}$ domain

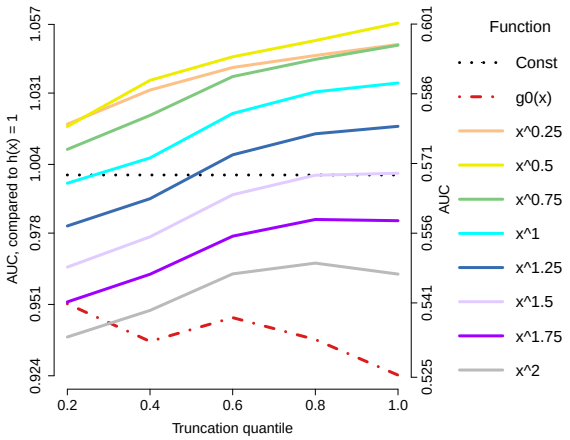


(e) $n = 80$, unif domain

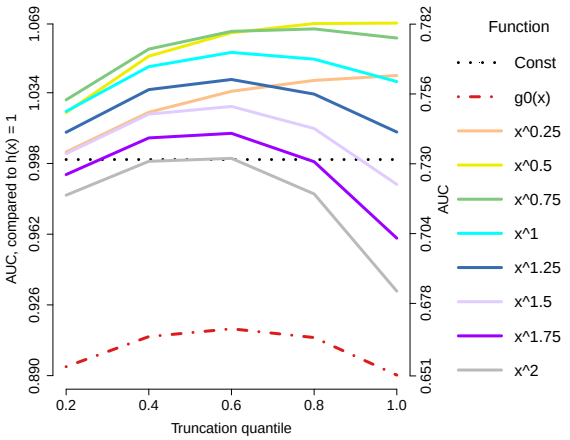


(f) $n = 1000$, unif domain

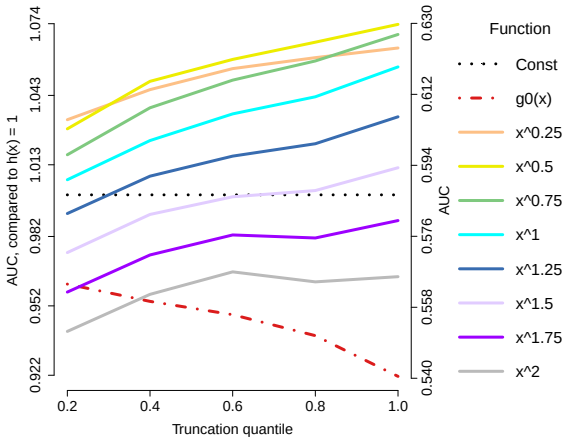
Figure 3.7: AUCs averaged over 50 trials for edge recovery using generalized score matching for the Gaussian models ($a = 1$) on domains being subsets of $\mathbb{R}^m$ (not restricted to $\mathbb{R}^m_+$).



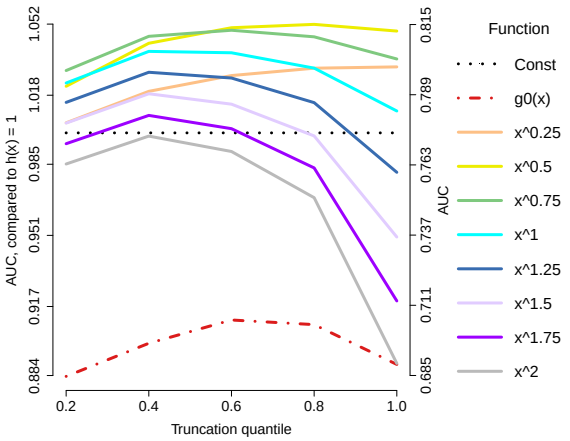
(a) $n = 80$, ℓ_2-nn domain



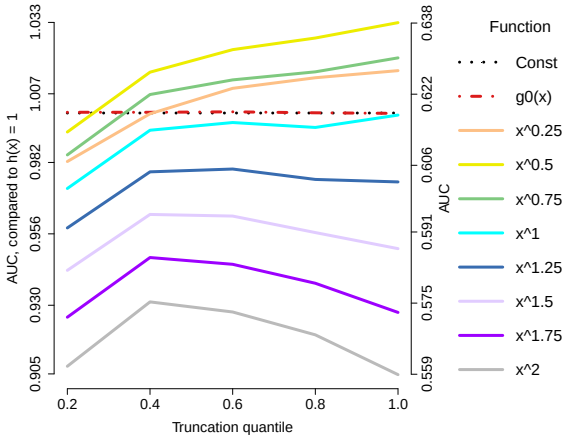
(b) $n = 1000$, ℓ_2-nn domain



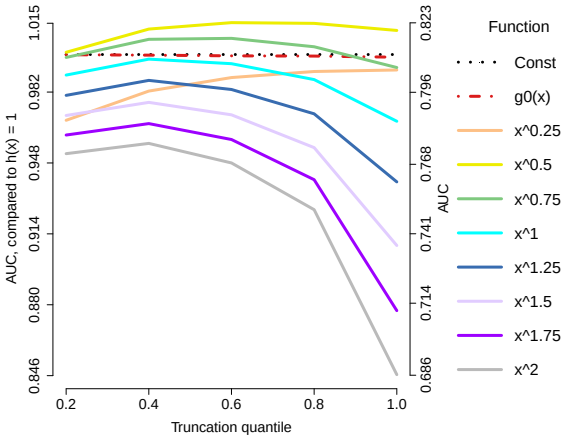
(c) $n = 80$, $\ell_2^{\complement}$-nn domain



(d) $n = 1000$, $\ell_2^{\complement}$-nn domain

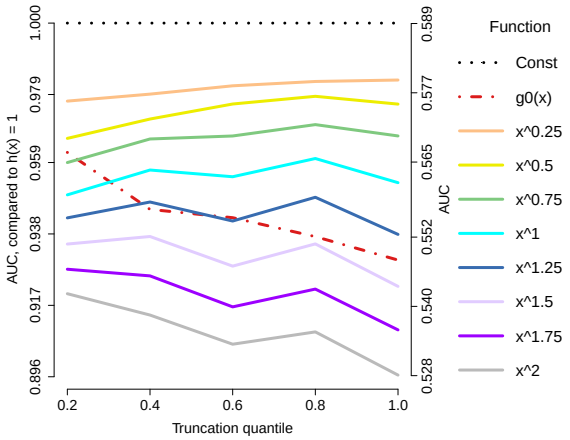


(e) $n = 80$, unif-nn domain

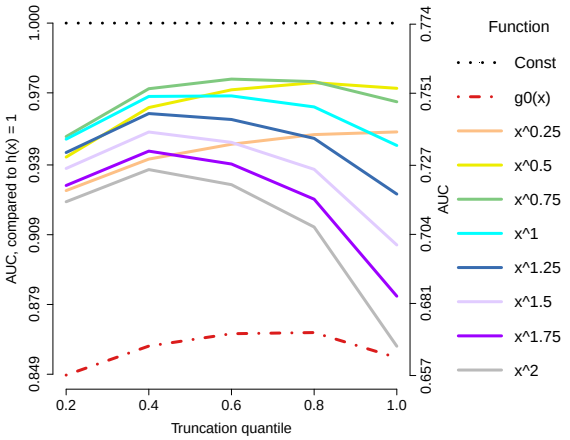


(f) $n = 1000$, unif-nn domain

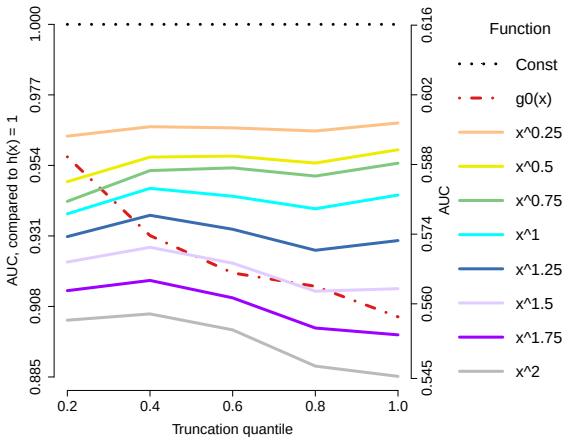
Figure 3.8: AUCs averaged over 50 trials for edge recovery using generalized score matching for the $a = 3/2$ models.



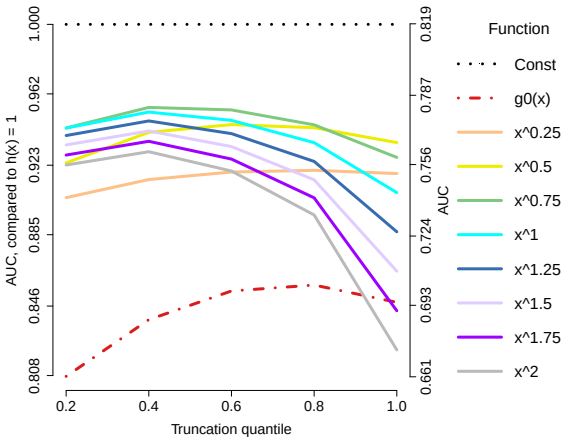
(a) $n = 80$, ℓ_2-nn domain



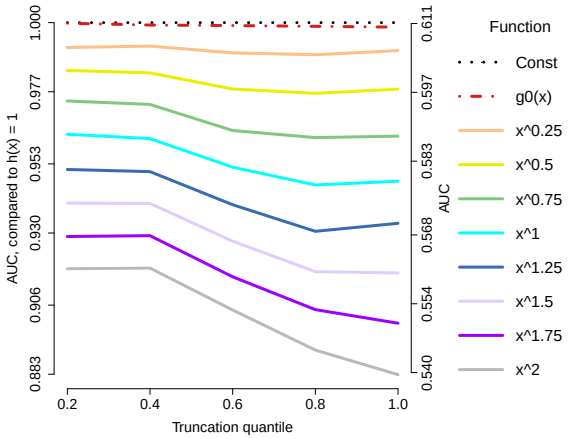
(b) $n = 1000$, ℓ_2-nn domain



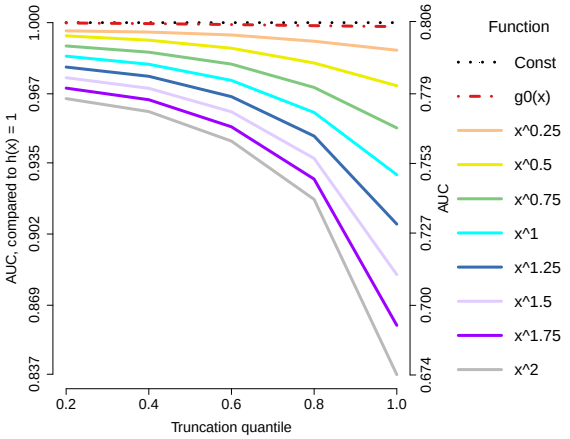
(c) $n = 80$, $\ell_2^{\mathbb{C}}$-nn domain



(d) $n = 1000$, $\ell_2^{\mathbb{C}}$-nn domain

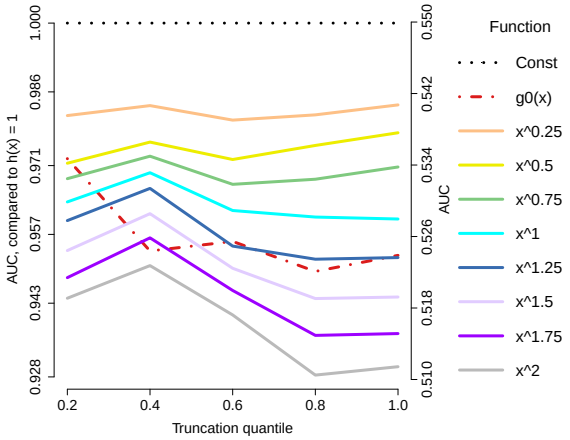


(e) $n = 80$, unif-nn domain

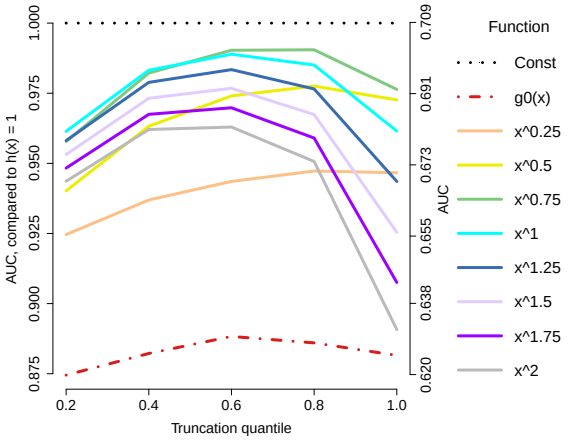


(f) $n = 1000$, unif-nn domain

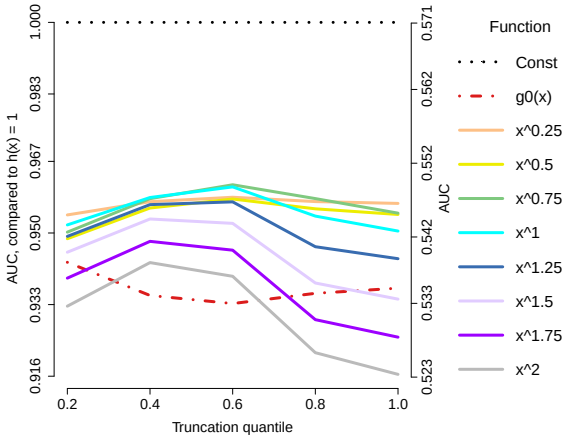
Figure 3.9: AUCs averaged over 50 trials for edge recovery using generalized score matching for the $a = 2$ models.



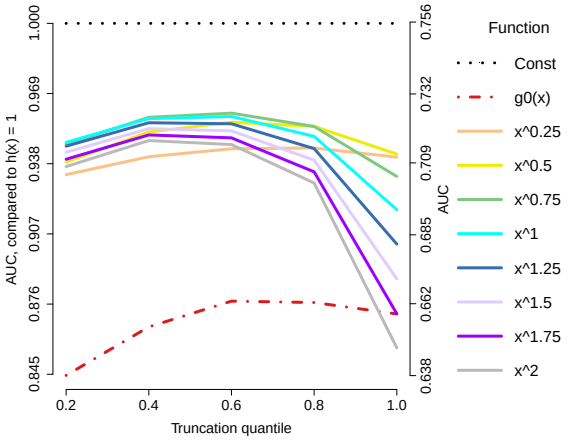
(a) $n = 80$, ℓ_2-nn domain



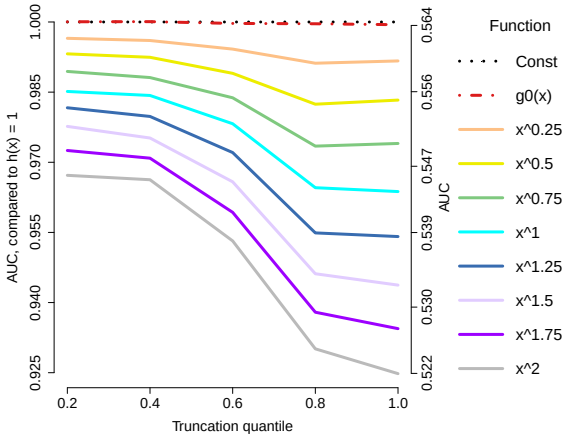
(b) $n = 1000$, ℓ_2-nn domain



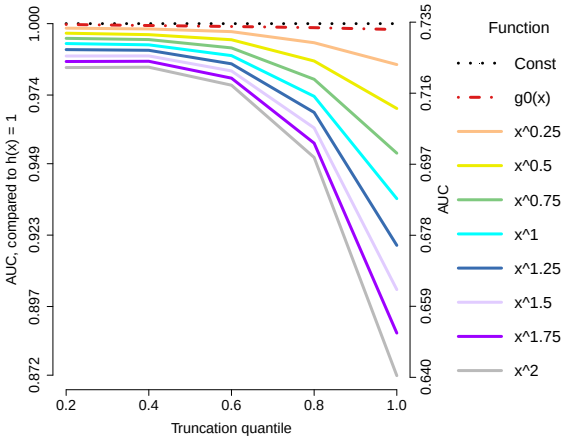
(c) $n = 80$, $\ell_2^{\mathbf{C}}$-nn domain



(d) $n = 1000$, $\ell_2^{\mathbf{C}}$-nn domain



(e) $n = 80$, unif-nn domain



(f) $n = 1000$, unif-nn domain

Figure 3.10: AUCs averaged over 50 trials for edge recovery using generalized score matching for the $a = 3$ models.

3.8.3 Numerical Experiments for A^{m-1} Models on the Simplex

Experiment Setup

As the most important example of models on the simplex $\mathcal{D} \equiv \left\{\boldsymbol{x} \in \mathbb{R}_+^m \middle| \boldsymbol{x} \succ \mathbf{0},\, \mathbf{1}_m^\top \boldsymbol{x} = 1\right\}$, in this section, we consider A^{m-1} models discussed in Section 3.6.2, i.e. when $a = b = 0$ and $\mathbf{K}\mathbf{1}_m = \mathbf{0}_m$ with $\mathbf{K} = \mathbf{K}^\top$. As in Section 3.8.2 we consider dimension $m = 100$, sample sizes $n = 80$ and $n = 1000$, and assume $\boldsymbol{\eta}_0 \equiv \mathbf{0}$ is known for simplicity; the density is proportional to $\exp\left(-\log \boldsymbol{x}^\top \mathbf{K} \log \boldsymbol{x}/2\right)$.

For the true interaction matrix $\mathbf{K}_0$ we use band matrices with bandwidths $s = 7$ for $n = 1000$ and $s = 2$ for $n = 80$, where the bandwidth is defined as $\max\{|i-j| : \kappa_{0,i,j} > 0\}$. We set $\kappa_{0,i,j}$ to $1 - |i-j|/(s+1)$ for $1 \le |i-j| \le s$, and set the diagonals so that $\mathbf{K}_0\mathbf{1}_m = \mathbf{0}_m$. Each n is thus associated with only one graph, for which we run 50 trials. Note that the bandwidth is chosen so that $n/(d_{\mathbf{K}_0}^2 \log m)$ is roughly constant, where d is the maximum node degree, as this quantity is suggested to be linked to the probability of successful edge recovery, according to our consistency theory in Section 3.7.

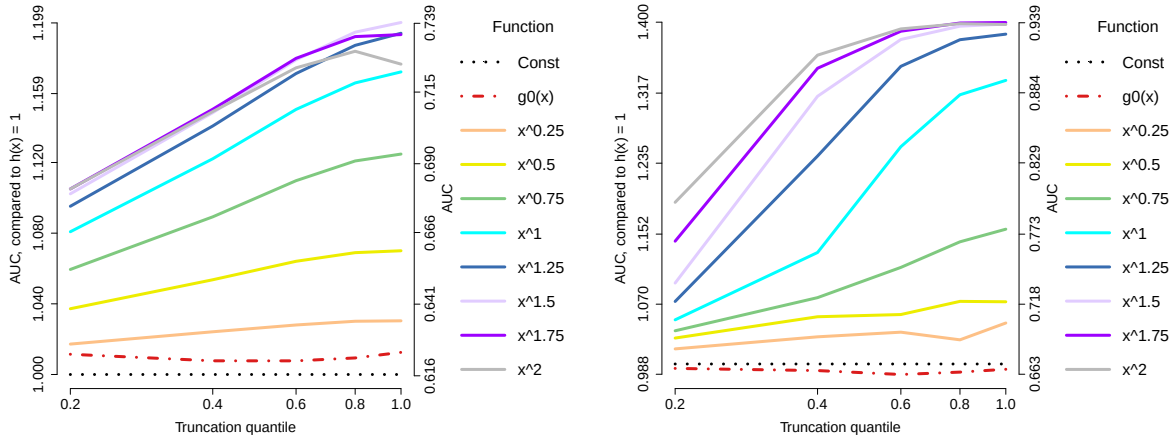
Results

As discussed in Section 3.6.2, although conditional dependence between two components does not correspond to the zero/nonzero pattern of the entries in $\mathbf{K}$, one may still be interested in successful recovery of such pattern as well as minimization of the estimation error of $\mathbf{K}$ and $\boldsymbol{\eta}$. This is backed up by Corollary 22, according to which $\mathbf{K}$ and $\boldsymbol{\eta}$ are exactly identifiable from the distribution.

As in Section 3.8.2, in Figure 3.11 we plot the AUC averaged over 50 trials against the $\pi = 0.2, 0.4, 0.6, 0.8, 1$ whose columnwise sample quantiles are used as the truncation points $\boldsymbol{C}$ of $\boldsymbol{\varphi}_{\boldsymbol{C},\mathcal{D}}$ (c.f. Section 3.8.1), and each curve represents $h(x) = 1$ (Hyvärinen, 2005), $g_0(\boldsymbol{x})$ from Section 3.3.6, or $h(x) = x^c$ with $c = 1/4, 1/2, \ldots, 2$. In addition, in Figure 3.12 we plot the estimation error in spectral and Frobenius norms, namely $|||\hat{\mathbf{K}} - \mathbf{K}_0|||_2$ and $|||\hat{\mathbf{K}} - \mathbf{K}_0|||_F$, against π. The y-ticks on the right-hand side are the errors, and those on the left are the

errors divided by the error for $h(x) = 1$, measuring the relative performance of each method compared to Hyvärinen (2005). Opposite to the earlier plots, curves closer to the bottom of each plot indicate better performance.

Based on the plots, for both edge recovery and error measured by the Frobenius norm, $h(x) = x^2$ is among the best performers, supporting our previous conclusion of the choice of $h(x) = x^{\max\{2-a,0\}}$ for general a-b models. When the error is measured in the spectral norm, $h(x) = x^2$ has the largest error for $n = 80$ but shows better improvements over other estimators as n is increased.



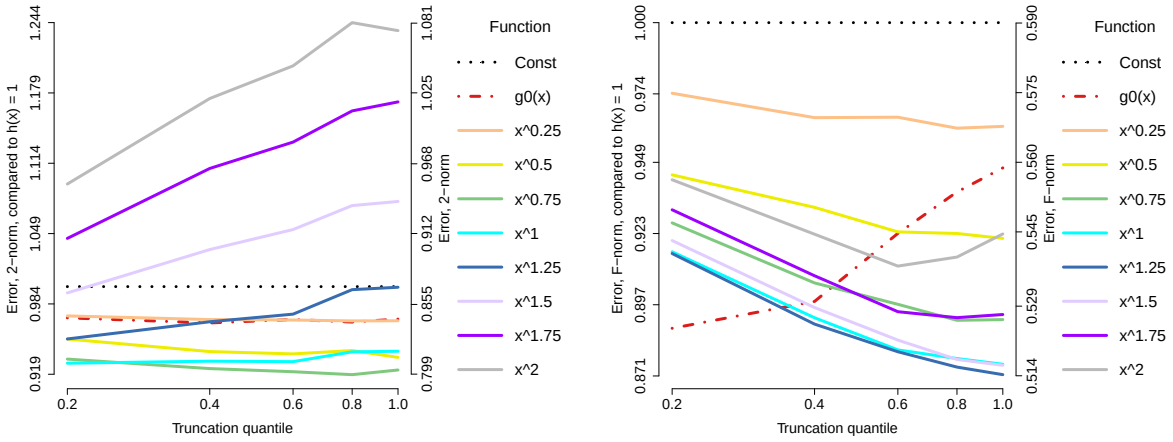
(a) $n = 80$, ℓ_2-nn domain

(b) $n = 1000$, ℓ_2-nn domain

Figure 3.11: AUCs averaged over 50 trials for edge recovery for the A^{m-1} models on the simplex.

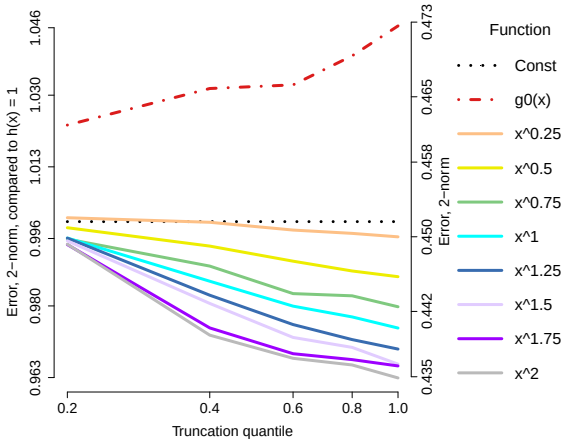
3.9 DNA Methylation Data

In this section, we apply our generalized score matching to a DNA methylation dataset. The dataset contains methylation levels of CpG islands for 500 patients associated with head and neck cancer from the Cancer Genome Atlas (TCGA), where CpG islands are regions of DNA where a guanine nucleotide very frequently follows a cytosine nucleotide along its 5' $\rightarrow$ 3' direction. Methylation levels are associated with epigenetic regulation of genes (Du

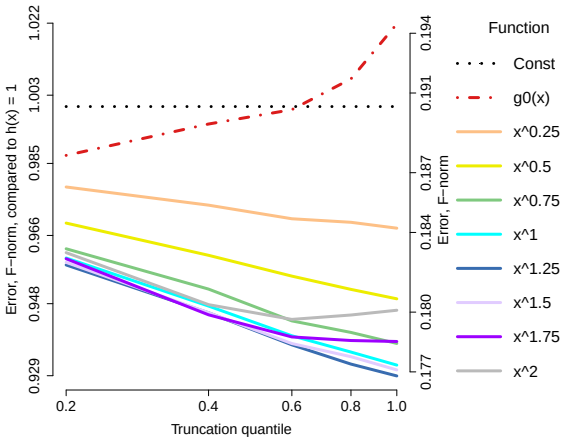


(a) $n = 80$, error in matrix 2-norm

(b) $n = 1000$, error in matrix 2-norm



(c) $n = 80$, error in matrix F-norm



(d) $n = 1000$, error in matrix F-norm

Figure 3.12: Averaged error in spectral and F norms over 50 trials normalized by the corresponding norms of the true $\mathbf{K}_0$; sparsity chosen by cross validation for the A^{m-1} models on the simplex.

et al., 2010). According to Du et al. (2010), to measure the methylation level, one can use the Beta-values, a value in $[0, 1]$ as the ratio of the methylated probe intensity and the sum of methylated and unmethylated probe intensities, or the M-values, which is defined as $\log_2(\text{Beta}/1 - \text{Beta})$ taking values in $\mathbb{R}$, the logit of the Beta-values with base 2. We are interested in applying our method and the *a*-*b* model framework to a subset of the data using Beta-values and M-values to estimate the network/undirected graph of the CpG sites or the corresponding genes.

Since the dataset contains 500 samples and 93715 sites, we choose a subset of sites corresponding to genes known to belong to the pathway for Thyroid cancer according to the Kyoto Encyclopedia of Genes and Genomes (KEGG). With the *a*-*b* model assumption, we remove the sites that are clearly bimodal; specifically, for each site, we apply `Mclust` to its non-outlier Beta-value observations and test if it is a Gaussian mixture having more than one component. This results in 500 samples and 478 sites, corresponding to 36 genes.

To estimate the graph using the M-values, we assume a Gaussian model on $\mathbb{R}^m$, i.e. *a*-*b* model with $a = b = 1$, and use the profiled estimator in (3.22), and choose the upper-bound diagonal multiplier $2 - \left(1 + 80\sqrt{\log m/n}\right)^{-1}$ as suggested in Chapter 2.6.2. Note that the $(\boldsymbol{h} \circ \boldsymbol{\varphi})$ is unnecessary in this setting and we simply use the original score matching with $(\boldsymbol{h} \circ \boldsymbol{\varphi})(\boldsymbol{x}) = \mathbf{1}_m$. For Beta-values, we assume a log-log model ($a = b = 0$) on $[0, 1]^m$, and use the profiled estimator with the upper-bound diagonal multiplier $1 + \sqrt{(\tau \log m + \log 4)/(2n)}$ as in (3.38) with the choice of $\tau = 3$. We use $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{x}^2$ as suggested by our theory, and choose the truncation points in $\boldsymbol{\varphi}$ to be the 40th sample percentile, as suggested by the simulation results in Figure 3.4. As an illustration, the λ parameter that defines the ℓ_1 penalty on $\mathbf{K}$ is chosen so that the number of edges is equal to 478, the number of sites, following Section 2.7.4 and Lin et al. (2016).

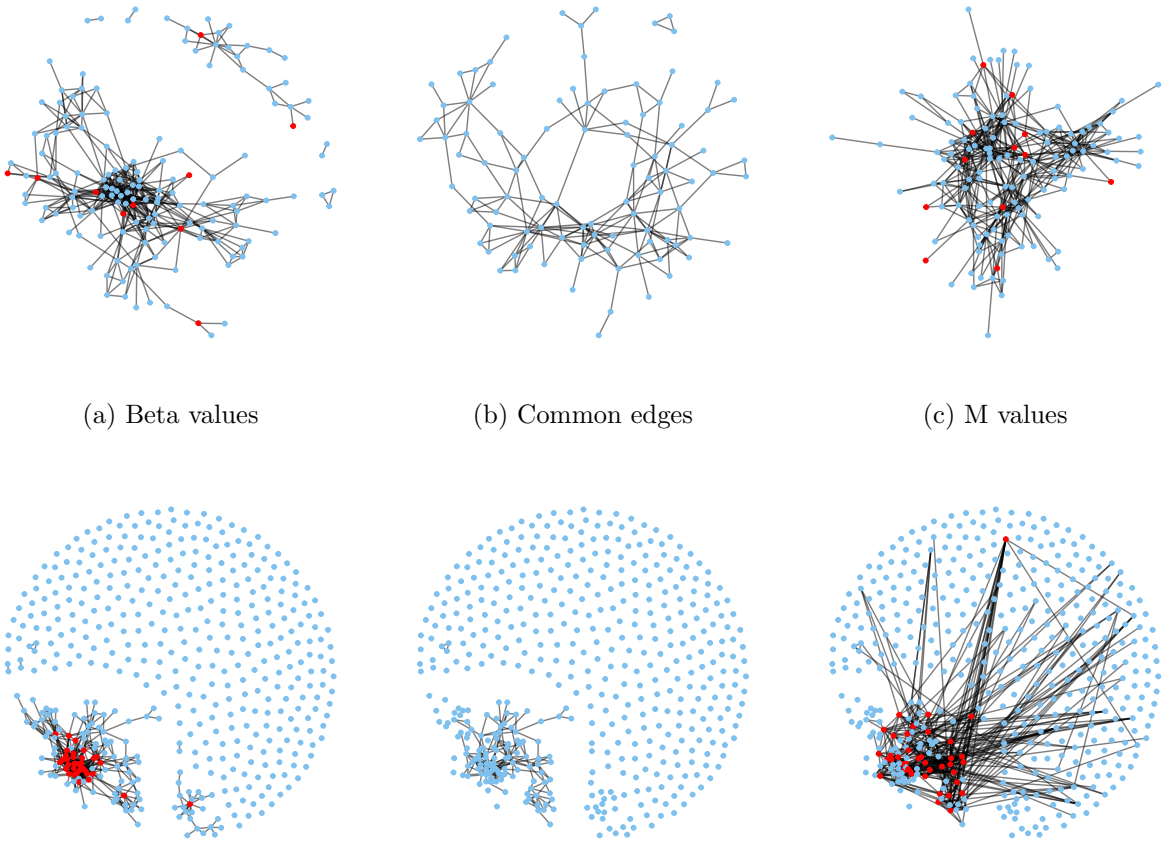
The estimated graphs are presented in Figure 3.13, where graphs (a) and (d) are the estimated graphs for Beta values, graphs (c) and (f) are those for M values, and (b) and (e) are the common edges, i.e. intersection graphs. In particular, graphs (a) (b) (c) exclude isolated nodes that have no edges and the layout is optimized for each graph, while for (d)

(e) (f) the layout is optimized for the Beta graph. Figure 3.14 shows graphs in Figure 3.13 aggregated by the genes associated with the sites. In (a), (c), (d), (f), red points indicate nodes with degree at least 10 (i.e. genes connected to at least 10 other genes in the case of Figure 3.14). Sites with highest node degrees are listed in Table 3.1, where those shared by the two graphs are highlighted in bold.

Beta values	M values
CDH1—4 (28)	RXRB—24 (25)
TCF7L1—18 (22)	**MAPK3—8 (22)**
RXRA—19 (21)	PAX8—6 (21)
RXRA—22 (21)	CCND1—19 (20)
RET—22 (21)	RXRA—10 (20)
RXRB—82 (21)	**RXRA—19 (20)**
NTRK1—40 (21)	RXRB—18 (20)
LEF1—2 (20)	PAX8—9 (20)
TCF7L1—13 (20)	TCF7—3 (18)
CDKN1A—10 (20)	TCF7L1—9 (18)
CDKN1A—6 (19)	**TCF7L1—18 (18)**
MAPK3—8 (17)	TCF7L2—63 (18)
PAX8—28 (17)	TPM3—12 (18)
PAX8—29 (17)	**PAX8—29 (17)**

Table 3.1: List of sites with the highest node degrees in each estimated graph.

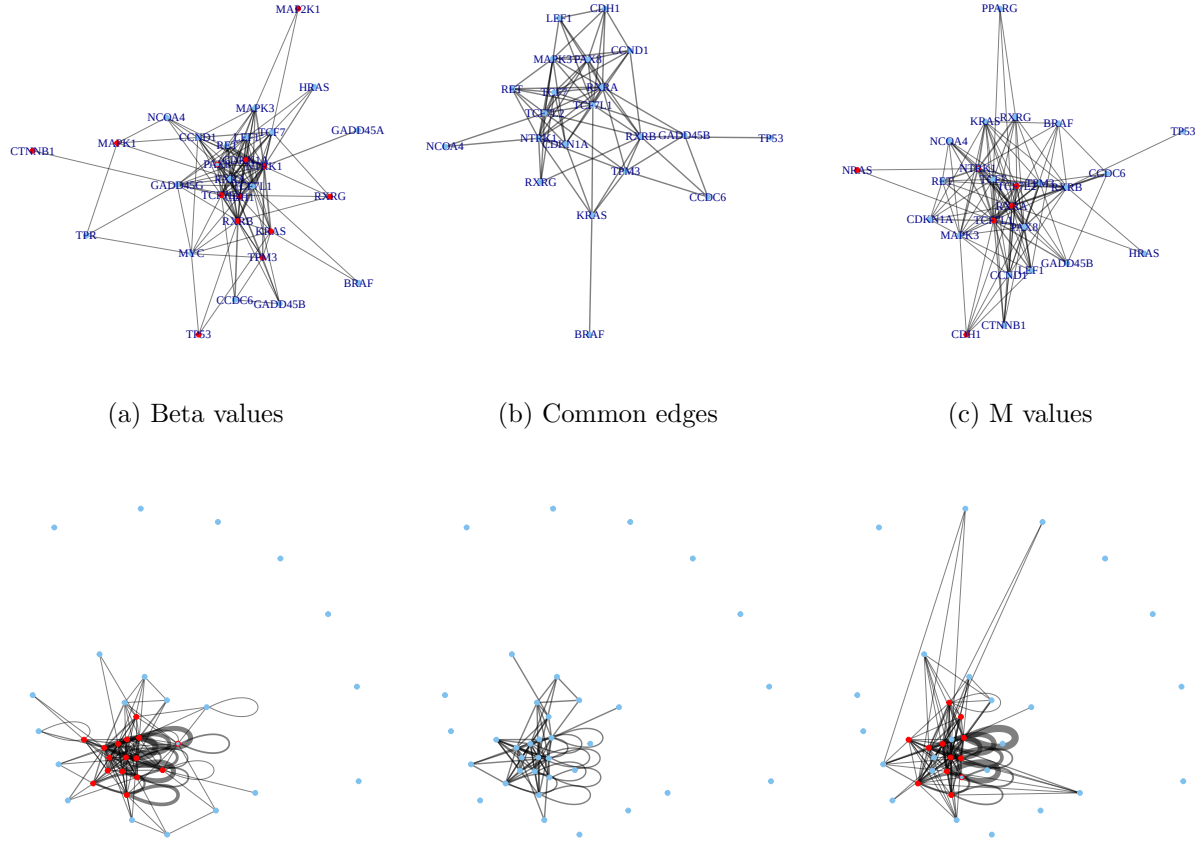
To quantify the similarity between the two graphs, we calculate their Hamming distance, namely the sum of absolute differences in their adjacency matrices divided by 2. The Hamming distance is 568, which is a lot smaller than 936, the minimal Hamming distance between the graph for Beta values and 10000 randomly generated graphs with the same number of edges, and 940, that value using the graph for M values. Furthermore, the discrete one-



(a) Beta values (b) Common edges (c) M values

(d) Beta values (e) Common edges (f) M values

Figure 3.13: Graphs for CpG sites estimated by regularized generalized score matching estimator using Beta values (a, d) and M values (c, f), and their intersection graph (b, e). In (a), (b) and (c), isolated nodes with no edges are removed, and the layout is optimized for each plot. In (d), (e) and (f), isolated nodes are included and the layout is optimized for the graph for Beta values. In (a), (c), (d), (f), red points indicate nodes with degree at least 10 ("hub nodes").

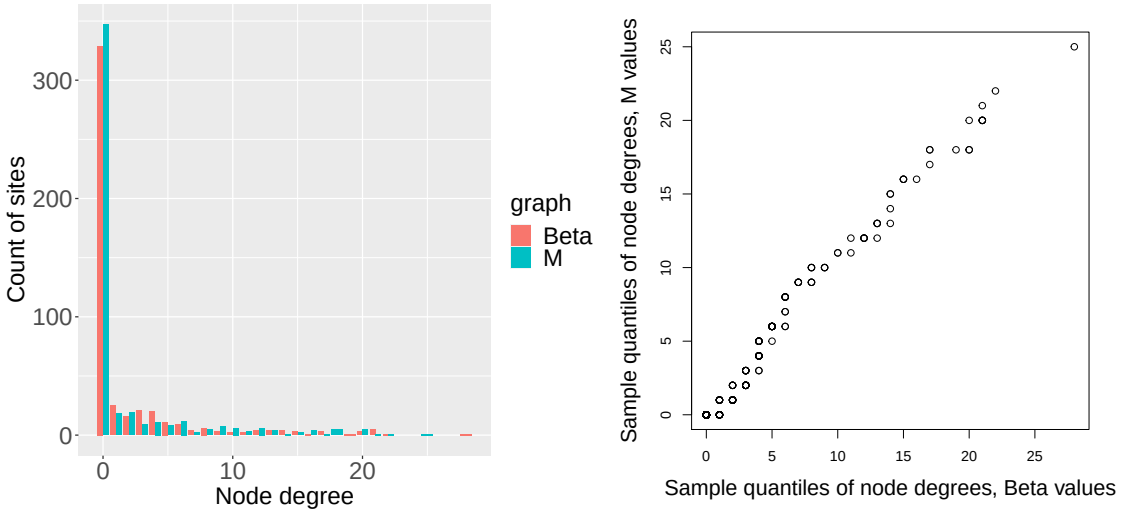


(a) Beta values (b) Common edges (c) M values

(d) Beta values (e) Common edges (f) M values

Figure 3.14: Graphs in Figure 3.13 aggregated by the genes associated with the CpG sites, with Beta values (a, d) and M values (c, f), and their intersection graph (b, e). In (a), (b) and (c), isolated nodes and self loops (edges between sites for the same gene) are removed, and the layout is optimized for each plot. In (d), (e) and (f), isolated nodes and self loops are included and the layout is optimized for the graph for Beta values. In (a), (c), (d), (f), red points indicate genes connected to at least 10 other genes. Thickness of edges scales with square root of the total number of edges for sites between each pair of genes (or edges between sites within the same group in the case of self loops).

sample two-sided Kolmogorov-Smirnov test on the sample distribution of node degrees of one graph, while treating that of the other graph as the null distribution, gives p-values 0.1208 and 0.1397, respectively. In Figure 3.15 we compare the distribution of node degrees for the site graphs estimated using Beta and M values, with interlaced histogram on the left and Q-Q plot on the right. All these results suggest that the two estimated graphs are sufficiently similar to each other.



(a) Beta values

(b) Common edges

Figure 3.15: Interlaced histogram (left) and Q-Q plot (right) showing the node degree distributions for both site graphs.

3.10 Discussion

The generalized score matching loss proposed in Chapter 2 (Yu et al., 2019b) estimates densities supported on $\mathbb{R}_+^m$ using a generalized score matching loss, whose definition involves componentwise multiplication of $\nabla \log p(\boldsymbol{x})$ with an $\boldsymbol{h}(\boldsymbol{x})$ function, generalizing the choice of $\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{x}^2$ in Hyvärinen (2007). As in Hyvärinen (2007), the estimator has a closed-form solution in the low-dimensional settings without any regularization and avoids calculating the often computationally intensive normalizing constant.

In this chapter, we extend our work from distributions on $\mathbb{R}_+^m$ to those on *componentwise countable* domains as defined in Definition 5; informally these are the domains for which, fixing all other components $\boldsymbol{x}_{-j}$, the induced domain of one component x_j is a union of at most countably many intervals in $\mathbb{R}$. We accomplish this by composing the $\boldsymbol{h}$ with a distance function $\boldsymbol{\varphi_C} = (\varphi_{C_1,1}, \ldots, \varphi_{C_m,m}) : \mathcal{D} \to \mathbb{R}_+^m$, with $\varphi_{C_j,j}(\boldsymbol{x})$ being the distance of x_j to its induced boundary holding $\boldsymbol{x}_{-j}$ fixed, truncated from above by some prespecified C_j. We show that the loss can again be approximated by an empirical loss when we substitute $\boldsymbol{h}$ in Chapter 2 with $(\boldsymbol{h} \circ \boldsymbol{\varphi_C})$, which is quadratic in the canonical parameter for exponential families, thus having closed-form solutions in low-dimensional settings. For domains with zero Lebesgue measure, we take the m-simplex domain as a canonical example and defined the loss by profiling out the last component.

As in Chapter 2, we focus on a-b pairwise interaction models with density proportional to $\exp\left\{-\boldsymbol{x}^{a\top}\mathbf{K}\boldsymbol{x}^a/(2a) + \boldsymbol{\eta}^\top \boldsymbol{x}^b/b\right\}$, where for $a = 0$ we let $\boldsymbol{x}^{a\top}\mathbf{K}\boldsymbol{x}^a/(2a) \equiv \log \boldsymbol{x}^\top \mathbf{K} \log \boldsymbol{x}/2$ and for $b = 0$, $\boldsymbol{\eta}^\top \boldsymbol{x}^b/b \equiv \boldsymbol{\eta}^\top \log \boldsymbol{x}$. We formulate our estimators for a-b models on domains with positive Lebesgue measures as well as on simplex domains, for the latter of which we extensively discuss the A^{m-1} models (Aitchison, 1985) as the most important example.

We extend the consistency theory for edge recovery in Chapter 2 to Gaussian graphical models on domains that are finite disjoint unions of convex sets, as well as general a-b models on simplex domains with $a > 0$ and on bounded domains with positive Lebesgue measure, requiring the sample size to be $n = \Omega(\log m)$. For unbounded domains with $a > 0$ and

the simplex domains with $a = 0$, we require an additional multiplicative factor that may weakly depend on m. Finally, through simulation studies we confirm that the choice of $\boldsymbol{h}(\boldsymbol{x}) = (x_1^c, \ldots, x_m^c)$ with $c = \max\{2 - a, 0\}$ performs the best in most settings in terms of edge recovery, similar to the conclusion in Chapter 2.

In the simulations we adaptively select the truncation points $\boldsymbol{C}$ of $\boldsymbol{\varphi}$ using the sample quantiles of the untruncated distances. Developing a method to choose the best truncation points remain further research. For the theory part, it is also interesting to investigate into the real sample complexity in terms of dimension m for unbounded domains with $a > 0$ or simplex domains with $a = 0$.

Chapter 4

DIRECTED GRAPHICAL MODELS AND CAUSAL DISCOVERY FOR ZERO-INFLATED DATA

4.1 Introduction

Graphical models specify conditional independence relations among variables in a random vector $\boldsymbol{Y}$ indexed by the nodes $\mathcal{V}$ of a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with edge set $\mathcal{E}$ (Maathuis et al., 2019). Models based on undirected graphs may be used to explore conditional independence between any two variables Y_V and Y_U given all others $(Y_W)_{W \neq U,V}$, as represented by the absence of an edge between V and U in $\mathcal{E}$. Models based on directed acyclic graphs (DAGs), for which $\mathcal{E}$ is comprised of directed edges, capture conditional independence structure that naturally arises from cause-effect relationships between the variables.

In biology and genetics, graphical models have been applied to infer the structure of gene regulatory networks based on measurements of gene expression (Maathuis et al., 2019, Sections 20-21). Traditional technologies produce expression levels aggregated over hundreds or thousands of individual cells, and these bulk measurements are frequently modeled using the assumption of Gaussianity. In directed GGMs the exact structure of the underlying DAG cannot be identified from purely observational data, and the target of inference becomes an equivalence class of DAGs. For instance, one cannot differentiate between $V \to U$ and $U \to V$ when the variables are assumed bivariate normal. In the Gaussian case, directed graphical models posit linear functional relationships between the variables coupled with additive Gaussian noise. A more recent line of work emphasizes that directed graphical models that alter this assumption to nonlinear functional relationships and additive noise (Peters et al., 2014), or linear relations and non-Gaussian noise (Shimizu et al., 2006; Wang and Drton, 2020), or linear relations with homoscedastic Gaussian noise (Peters and Bühlmann, 2013; Chen et al., 2019) are amenable to causal discovery in the sense that different DAGs are no longer equivalent.

More recent technology obtains sequencing measurements of mRNA present in single cells. This new technology, as well as the larger sample sizes it provides, promise to give more information than bulk measurements, but at the same time bring in a unique new challenge. At the single cell level, genes appear as "on" with positive single cell gene expression levels,

or as "off" with the recorded measurements zero or negligible (McDavid et al., 2019).

Figure 4.1 shows pairwise scatter plots of four genes from a T helper single-cell dataset with 1951 measurements from eight healthy donors, which we analyze in Section 4.6. It is a superset of the single-cell T-follicular helper data considered in McDavid et al. (2019), which is similarly plotted in their Figure 1. The lower panels show the pairwise scatter plots along with a fitted linear regression curve, and the diagonal panels show the univariate smoothed kernel density estimates for each gene. As we can see, each gene has a large number of zero values and a linear regression model is not sufficient for modeling the pairwise relationships.

A novel undirected graphical model that deals with this zero-inflation was introduced by McDavid et al. (2019). Their approach considers Hurdle density models, where for a random vector of dimension m, the joint probability density function has the form

$$f(\boldsymbol{y}; \mathbf{A}, \mathbf{B}, \mathbf{K}) \propto \exp\left(\mathbb{1}_{\boldsymbol{y}}^{\top}\mathbf{A}\mathbb{1}_{\boldsymbol{y}} + \mathbb{1}_{\boldsymbol{y}}^{\top}\mathbf{B}\boldsymbol{y} - \frac{1}{2}\boldsymbol{y}^{\top}\mathbf{K}\boldsymbol{y}\right), \qquad \boldsymbol{y} \in \mathbb{R}^m, \tag{4.1}$$

with $\mathbb{1}_{\boldsymbol{y}}$ being the elementwise indicators of nonzero entries in $\boldsymbol{y} \in \mathbb{R}^m$. The dominating measure for this density is the m-fold product of the sum of the Lebesgue measure and a point mass at zero. However, since more information can be inferred from single-cell sequencing data, one would hope that the data can also be analyzed using more informative directed graphical models, and that we can infer which variables (genes) are the causes of change in other variables (expression levels of other genes). In this chapter, we formulate such directed graphical models for zero-inflated data, and prove that under a weak assumption one can recover the exact DAG from the joint distribution. In contrast to the setting of McDavid et al. (2019), the distributions in our models are not merely zero-inflated Gaussian as we allow variables that are "on" to be non-linear polynomial functions of other variables and stochastic noise.

In DAG models, the joint distribution can be factorized into the product of conditional distributions of each variable given parent variables. For simplicity we call these conditional distributions the *node conditionals*. In our DAG model for zero-inflated data, we form the node conditionals by taking the conditional distribution of one variable given the others

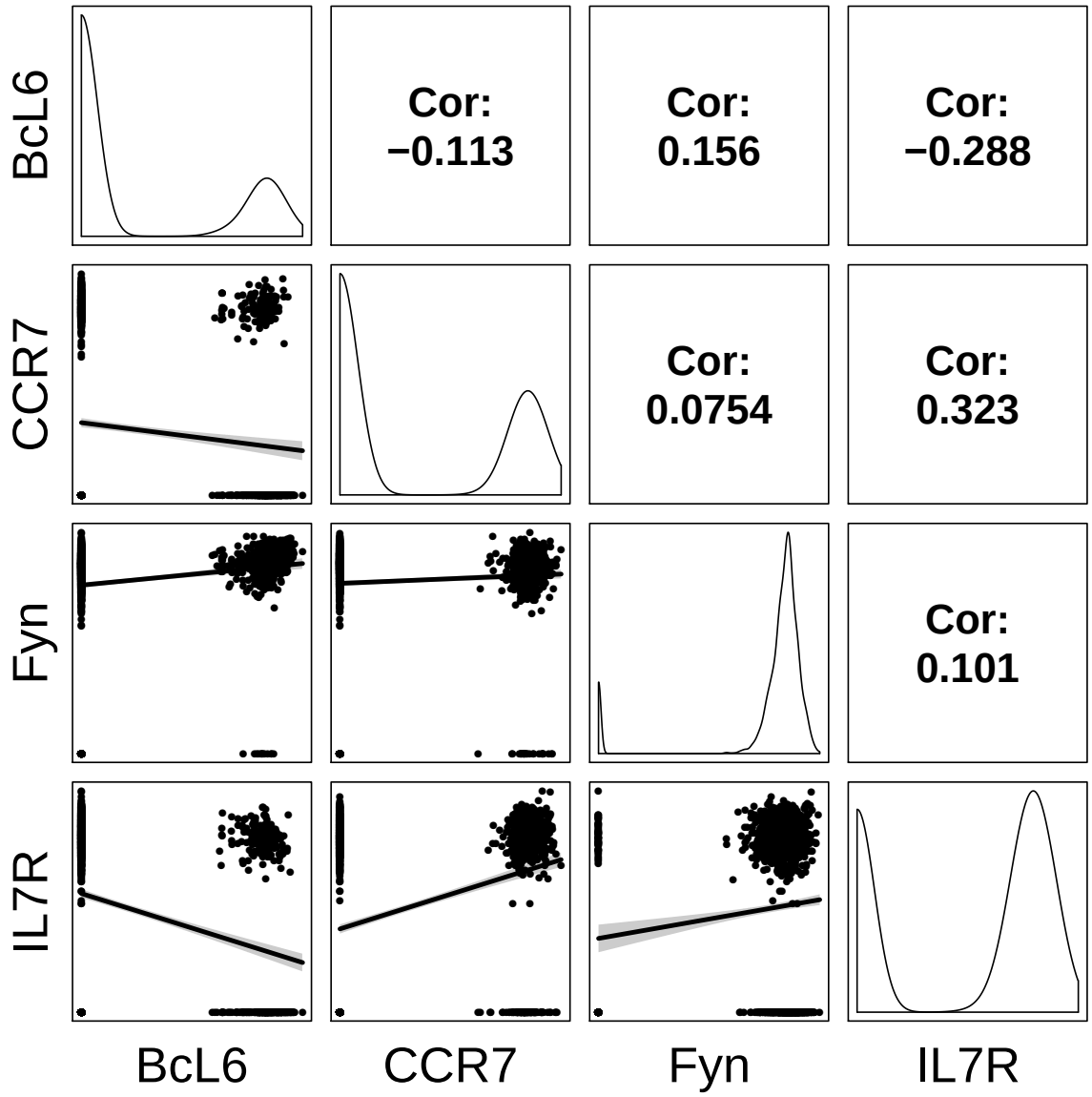


Figure 4.1: Pairwise scatter plots and kernel densities on four genes from the T helper cell data analyzed in Section 4.6.

in the joint model from (4.1). We refer to the resulting graphical model as the model in (α, β, k)-parametrization, or in *canonical* parametrization. Here, the α and β parameters in the conditional distribution are derived from the matrix parameters $\mathbf{A}$ and $\mathbf{B}$ in (4.1). The α and β are polynomials in the parent variables and their 0/1 indicators being zero/nonzero. An alternative second type of model is obtained by directly specifying each node conditional as a mixture of a point mass at zero and a Gaussian distribution, with the log odds of being nonzero ($\log(p/(1-p))$) and the mean in the Gaussian part being polynomials in the parent variables and their indicators. The Gaussian variance is taken constant in the parents. We call this second formulation the model in (p, μ, σ^2)- or *moment* parametrization, since the parameters directly correspond to the (conditional) moments. The detailed specification of both model types is developed in Section 4.2.

In Section 4.3, we show that under our models, the distributions that can be represented by two different DAGs must be distributions of *two-Gaussian type* (Definition 15). We then prove that such distributions do not exist for dimension $m = 2$ and $m = 3$; we also conjecture they do not exist for $m > 3$. Moreover, we are able to prove that under a natural and practical assumption, we have full identifiability in the sense of being able to identify the exact DAG underlying the model. This assumption specifies that for each node, $\alpha + \beta^2/(2k)$ or equivalently $\log(p/(1-p))$ has a separate univariate term for each parent (e.g. $y_1 + y_2 + y_1y_2 + y_1^2$ instead of $y_1 + y_1y_2 + y_1^2$, which does not have a separate term for y_2). Note that this assumption is natural in the context of linear regression.

In Section 4.4, we introduce different methods for estimation of the DAG. Simulation studies supporting the use of these methods are given in Section 4.5, and they are then applied to the T-follicular helper cell dataset (Section 4.6).

The work in this chapter is presented in Yu et al. (2020).

4.2 Directed Graphical Models for Zero-Inflated Data

In this section we motivate and formally define our models for zero-inflated data based on directed acyclic graphs (DAGs).

4.2.1 Hurdle Joint Distributions for Zero-Inflated Continuous Observations

McDavid et al. (2019) proposed a *Hurdle joint distribution* with density

$$f(\boldsymbol{y}; \mathbf{A}, \mathbf{B}, \mathbf{K}) \propto \exp\left(\mathbb{1}_{\boldsymbol{y}}^\top \mathbf{A} \mathbb{1}_{\boldsymbol{y}} + \mathbb{1}_{\boldsymbol{y}}^\top \mathbf{B} \boldsymbol{y} - \frac{1}{2} \boldsymbol{y}^\top \mathbf{K} \boldsymbol{y}\right), \qquad \boldsymbol{y} \in \mathbb{R}^m, \tag{4.2}$$

with respect to λ^m, where λ is the sum of a point mass at 0 and the Lebesgue measure on $\mathbb{R}$, and $\mathbf{A} = (\alpha_{ij})_{i,j}, \mathbf{B} = (\beta_{ij})_{i,j}, \mathbf{K} = (k_{ij})_{i,j} \in \mathbb{R}^{m \times m}$ are matrices of interaction parameters with $\mathbf{K}$ positive definite. The indicator vector $\mathbb{1}_{\boldsymbol{y}} \equiv (\mathbb{1}_{\{y_1 \neq 0\}}, \ldots, \mathbb{1}_{\{y_m \neq 0\}}) \in \{0, 1\}^m$ captures which components of $\boldsymbol{y}$ are non-zero.

Consider a random vector $\boldsymbol{Y} \in \mathbb{R}^m$ that follows the Hurdle joint distribution. Intuitively, the density in (4.2) is obtained by combining an Ising model for the indicator vector $\mathbb{1}_{\boldsymbol{Y}}$ and a conditional normal distribution for $\boldsymbol{Y}$ given its nonzero pattern $\mathbb{1}_{\boldsymbol{Y}}$. The Ising model postulates a probability mass function proportional to $\exp\left(\mathbb{1}_{\boldsymbol{y}}^\top \mathbf{A} \mathbb{1}_{\boldsymbol{y}}\right)$. The conditional normal distribution has density $p\left(\boldsymbol{Y} = \boldsymbol{y} | \mathbb{1}_{\boldsymbol{Y}} = \mathbb{1}_{\boldsymbol{y}}; \mathbf{B}, \mathbf{K}\right) \propto \exp\left(\mathbb{1}_{\boldsymbol{y}}^\top \mathbf{B} \boldsymbol{y} - \frac{1}{2} \boldsymbol{y}^\top \mathbf{K} \boldsymbol{y}\right)$ with respect to the Lebesgue measure restricted to the subspace of $\mathbb{R}^m$ compatible with $\mathbb{1}_{\boldsymbol{y}}$.

The exponential specification in (4.2) entails that conditional independence between two variables is equivalent to the corresponding entries in all interaction matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{K}$ being 0. In other words, $\alpha_{ij} = \alpha_{ji} = \beta_{ij} = \beta_{ji} = k_{ij} = k_{ji} = 0$ if and only if Y_i and Y_j are conditionally independent given all other variables. Indeed, it is easy to see that the induced conditional distribution of Y_i given all other variables $\boldsymbol{Y}_{-i}$ in $\boldsymbol{Y}$, has density

$$p(Y_i = y_i | \boldsymbol{Y}_{-i} = \boldsymbol{y}_{-i}) = f(y_i; \alpha_{ii} + \boldsymbol{\alpha}_{i,-i}^\top \mathbb{1}_{\boldsymbol{y}_{-i}} + \boldsymbol{\beta}_{i,-i}^\top \boldsymbol{y}_{-i}, \beta_{ii} + \boldsymbol{\beta}_{-i,i}^\top \mathbb{1}_{\boldsymbol{y}_{-i}} - \boldsymbol{k}_{i,-i}^\top \boldsymbol{y}_{-i}, k_{ii}), \tag{4.3}$$

that is, the distribution is a Hurdle distribution in $m = 1$ dimension with parameters α, β, and k being linear functions in $\boldsymbol{Y}_{-i}$ and $\mathbb{1}_{\boldsymbol{Y}_{-i}}$; here f is the univariate version of (4.2).

4.2.2 Hurdle Conditionals

The observation in (4.3) above gives rise to the following definition. Recall that λ is the sum of a point mass at 0 and the Lebesgue measure on $\mathbb{R}$.

Definition 10 ((α, β, k)-Hurdle conditionals)**.** *Given a scalar random variable X and an m-dimensional random vector $\boldsymbol{Z}$, we say that the conditional distribution of X given $\boldsymbol{Z}$ is of* (α, β, k)-Hurdle type *if it admits conditional densities with respect to λ of the form*

$$p(X = x|\boldsymbol{Z} = \boldsymbol{z}) = f^{(m)}_{\alpha,\beta,k}(X|\boldsymbol{Z}) \equiv \frac{\exp\left(\alpha(\boldsymbol{z})\mathbb{1}_x + \beta(\boldsymbol{z})x - kx^2/2\right)}{\sqrt{2\pi/k}\exp\left(\alpha(\boldsymbol{z}) + \beta^2(\boldsymbol{z})/(2k)\right) + 1}. \tag{4.4}$$

Here, α and β are functions of $\boldsymbol{Z}$ (and its indicator vector).

Reparametrizing we give another intuitive formulation of Hurdle conditionals that clearly exhibits their nature of a mixture between a point mass at 0 and a conditional Gaussian distribution.

Definition 11 ((p, μ, σ^2)-Hurdle conditionals)**.** *Given a scalar random variable X and an m-dimensional random vector $\boldsymbol{Z}$, we say that the conditional distribution of X given $\boldsymbol{Z}$ is of* (p, μ, σ^2)-Hurdle type *if it admits conditional densities with respect to λ of the form*

$$\begin{aligned} p(X = x|\boldsymbol{Z} = \boldsymbol{z}) = f^{(m)}_{p,\mu,\sigma^2}(X|\boldsymbol{Z}) \equiv (1 - p(\boldsymbol{z}))(1 - \mathbb{1}_x) \\ + p(\boldsymbol{z})\mathbb{1}_x \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu(\boldsymbol{z}))^2}{2\sigma^2}\right). \end{aligned} \tag{4.5}$$

Here, p and μ are functions of $\boldsymbol{Z}$ (and its indicator vector).

It is easy to show that the two parametrizations (4.4) and (4.5) are connected through

$$\log\frac{p}{1-p} = \alpha + \frac{\beta^2}{2k} - \frac{1}{2}\log\left(\frac{k}{2\pi}\right), \qquad \mu = \frac{\beta}{k}, \qquad \sigma^2 = \frac{1}{k}. \tag{4.6}$$

That is, the conditional log odds of being nonzero is linear in α and quadratic in β, and the conditional Gaussian mean is proportional to β.

We note that while the (α, β, k)-parametrization takes canonical parameters $\alpha(\boldsymbol{Z})$, $\beta(\boldsymbol{Z})$ and k using a representation as exponential family, the moment parametrization directly models the conditional mixing probability $p(\boldsymbol{Z})$, and the mean $\mu(\boldsymbol{Z})$ and variance σ^2 parameters of the conditional Gaussian distribution. We thus refer to (4.4) as the *canonical parametrization*, and (4.5) as the *moment parametrization*.

4.2.3 Directed Graphical Models for Zero-Inflation Data

Consider an m-dimensional random vector $\boldsymbol{Y}$ whose components are indexed by the vertices of a DAG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and whose distribution is dominated by a product measure on $\mathbb{R}^m$. A graphical model based on $\mathcal{G}$ requires that the density of the joint distribution admits a factorization as

$$f(\boldsymbol{y}) = \prod_{V \in \mathcal{V}} f_V \left(y_V | \boldsymbol{y}_{\mathrm{pa}(V)} \right), \tag{4.7}$$

where each factor $f_V \left(y_V | \boldsymbol{y}_{\mathrm{pa}(V)} \right)$ is a conditional density for y_V given its parent variables $\boldsymbol{y}_{\mathrm{pa}(V)}$. The set of parents is defined to be $\mathrm{pa}(V) \equiv \{U : U \to V \in \mathcal{E}\}$.

In Section 4.2.1 we observed that, for the Hurdle joint distributions from (4.2), the conditional distribution of one variable Y_i given the others is an (α, β, k)-Hurdle with k constant, and α and β linear functions of those variables (and their indicators) that are conditionally dependent on Y_i; see (4.3). Motivated by this fact, we specify directed graphical models for zero-inflated data by assuming the conditional densities in the factorization in (4.7) to be (α, β, k)- or (p, μ, σ^2)-Hurdle conditionals. We then assume the parameters in these conditionals to be *Hurdle polynomials* in its parents, as defined now.

Definition 12 (Hurdle polynomials)**.** *Let $\boldsymbol{Y} = (Y_V)_{V \in \mathcal{V}}$ be an m-dimensional random vector indexed by a set $\mathcal{V}$, and suppose $\mathcal{S} \subseteq \mathcal{V}$. If $\mathcal{S} \neq \varnothing$, define the space of* Hurdle polynomials *in $\boldsymbol{y}_{\mathcal{S}}$ as*

$$\mathcal{H}(\boldsymbol{Y}; \mathcal{S}) \equiv \Bigg\{ c_0 + \sum_{j=1}^{T} c_j \prod_{U \in \mathcal{U}_j} Y_U^{d_{j,U}} \prod_{V \in \mathcal{V}_j} \mathbb{1}_{Y_V}, \quad c_0 \in \mathbb{R},\, T \in \mathbb{N},\, c_j \neq 0, \\ \mathcal{U}_j \subseteq \mathcal{S},\, \mathcal{V}_j \subseteq \mathcal{S} \backslash \mathcal{U}_j,\, d_{j,U} \in \mathbb{N}\ \forall U \in \mathcal{U}_j\ \forall j = 1, \ldots, T \Bigg\}, \tag{4.8}$$

where $\mathbb{N} = \{1, 2, \ldots\}$. This is the set of polynomials in values and indicators of nodes in $\mathcal{S}$. If $\mathcal{S} = \varnothing$, define $\mathcal{H}(\boldsymbol{Y}; \mathcal{S}) \equiv \mathbb{R}$. The degree *of a hurdle polynomial as specified in* (4.8) *is* $\max_{j=1,\ldots,T} \sum_{U \in \mathcal{U}_j} d_{j,U} + |\mathcal{V}_j|$. *Here $|\cdot|$ denotes the set cardinality.*

We are now ready to formally define our models.

Definition 13 (DAG models for zero-inflated data)**.** *Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a DAG with $|\mathcal{V}| = m$ nodes. A zero-inflated conditional Gaussian DAG model associated with $\mathcal{G}$ is a set of joint distributions on $\mathbb{R}^m$ that admit a density (with respect to λ^m) that factors as in* (4.7) *with each conditional density $f_V\left(y_V|\boldsymbol{y}_{\mathrm{pa}(V)}\right)$ being a Hurdle conditional*

(1) in the (α, β, k)-parametrization with parameters α_V, β_V and k_V, where k_V is constant, α_V and β_V are Hurdle polynomials in $\boldsymbol{y}_{\mathrm{pa}(V)}$; or

(2) in the (p, μ, σ^2)-parametrization with parameters p_V, μ_V and σ_V^2, where σ_V^2 is constant, $\log(p_V/(1-p_V))$ and μ_V are Hurdle polynomials in $\boldsymbol{y}_{\mathrm{pa}(V)}$.

It is apparent from the relationship (4.6) that if we allow the relevant parameters to be Hurdle polynomials of *any* degree, the two parametrizations are equivalent, meaning that given an underlying DAG, they share the same space of all possible joint distributions. However for computational convenience it is useful to bound the degree. In later applications, we will only consider degrees up to three.

4.3 Identifiability

4.3.1 Strong Identifiability

As we show next, the directed graphical models from Definition 13 are amenable to causal discovery in the sense that the DAG underlying the model is uniquely identifiable from a given joint distribution. More precisely, we prove identifiability under an explicit mild assumption on the Hurdle conditionals determining the considered joint distribution.

Let $\pi(\boldsymbol{y}_{\mathcal{S}}) \in \mathcal{H}(\boldsymbol{Y}; \mathcal{S})$ be a Hurdle polynomial for a subset $\mathcal{S} \subseteq \mathcal{V}$. For $U \in \mathcal{S}$, let $\pi_U(y_U) \equiv \pi(y_U, \mathbf{0})$ be the restriction of $\pi(\boldsymbol{y}_{\mathcal{S}})$ obtained by setting all entries other than y_U to zero. Then $\pi_U(y_U) \in \mathcal{H}(\boldsymbol{Y}; \{U\})$ is a univariate Hurdle polynomial.

Definition 14 (Strong Hurdle polynomials)**.** *Let $\pi(\boldsymbol{y}_{\mathcal{S}}) \in \mathcal{H}(\boldsymbol{Y}; \mathcal{S})$. We call $\pi(\boldsymbol{y}_{\mathcal{S}})$ a* strong Hurdle polynomial *if all of its restrictions $\pi_U(y_U)$ take at least three different values. In*

other words, for each $U \in \mathcal{S}$, the Hurdle polynomial $\pi(\boldsymbol{y}_{\mathcal{S}})$ contains at least one term of the form $c_j y_U^d$ with $c_j \neq 0$ and $d \geq 1$.

Theorem 30 (DAG identifiability with strong Hurdle polynomials)**.** *Let $f(\boldsymbol{y})$ be a joint density with respect to λ^m that factors according to a DAG $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, as in* (4.7)*. Suppose for each $V \in \mathcal{V}$, the conditional $f_V\left(y_V|\boldsymbol{y}_{\mathrm{pa}(V)}\right)$ is of Hurdle type with parameters (α_V, β_V, k_V) or (p_V, μ_V, σ_V^2). If for each V, $\alpha_V + \beta_V^2/(2k_V)$, or equivalently $\log(p_V/(1-p_V))$, is a strong Hurdle polynomial, then $f(\boldsymbol{y})$ does not factor with respect to any other DAG $\mathcal{G}' \neq \mathcal{G}$.*

In the proof in Appendix C we show that the given restriction on the parameters of the Hurdle conditionals is actually stronger than what we need for identifiability. However, the assumption of *strong* Hurdle polynomials is very natural in that it specifies a weak form of hierarchy among interactions by requiring that the conditional distributions are parametrized to include at least one univariate power term in every parent variable and not just indicators or interaction terms with other parents.

4.3.2 Weak Identifiability

Without assuming the Hurdle polynomials for the conditional distributions to be *strong*, we can still offer a weaker identifiability result that shows that the distributions in the intersection between the models obtained from two Markov equivalent DAGs with Hurdle polynomial parameters always have to be of what we call *two-Gaussian type*. In our definition of this concept, we write $\phi(\,\cdot\,; \mu, \nu)$ for the univariate normal density function with mean μ and inverse variance ν.

Definition 15. *Let $\boldsymbol{Y} = (Y_V)_{V \in \mathcal{V}}$ be a random vector, and let $W, U \in \mathcal{V}$ be the indices for two of its components. Further, let $\mathcal{P} \subseteq \mathcal{V} \backslash \{W, U\}$ be a set of additional indices. Then the joint distribution of $\boldsymbol{Y}$ is of* two-Gaussian type w.r.t. *$(W, U, \mathcal{P})$ if the following holds for both $V = W$ and $V = U$: There exists a constant ν_1^V, polynomials $\mu_1^V(\boldsymbol{y}_{\mathcal{P}})$, $\mu_2^V(\boldsymbol{y}_{\mathcal{P}})$, $\nu_2^V(\boldsymbol{y}_{\mathcal{P}})$, and functions $c_1^V(\boldsymbol{y}_{\mathcal{P}})$ and $c_2^V(\boldsymbol{y}_{\mathcal{P}})$ such that for almost every $\boldsymbol{y}_{\mathcal{P}} \in \mathbb{R}^{|\mathcal{P}|}$, $c_1^V(\boldsymbol{y}_{\mathcal{P}}) > 0$, $c_2^V(\boldsymbol{y}_{\mathcal{P}}) > 0$, either $\mu_1^V(\boldsymbol{y}_{\mathcal{P}}) \neq \mu_2^V(\boldsymbol{y}_{\mathcal{P}})$ or $\nu_1^V \neq \nu_2^V(\boldsymbol{y}_{\mathcal{P}})$, and the conditional density*

$$\mathbb{P}\left(Y_V = y \,|Y_V \neq 0, \boldsymbol{Y}_{\mathcal{P}} = \boldsymbol{y}_{\mathcal{P}}\right) = c_1^V(\boldsymbol{y}_{\mathcal{P}})\phi\left(y; \mu_1^V(\boldsymbol{y}_{\mathcal{P}}), \nu_1^V\right) + c_2^V(\boldsymbol{y}_{\mathcal{P}})\phi\left(y; \mu_2^V(\boldsymbol{y}_{\mathcal{P}}), \nu_2^V(\boldsymbol{y}_{\mathcal{P}})\right)$$

is a mixture of exactly two distinct Gaussian distributions with means polynomial in $\boldsymbol{y}_{\mathcal{P}}$*, one with an absolute constant inverse variance parameter and the other polynomial in* $\boldsymbol{y}_{\mathcal{P}}$*.*

If $\mathcal{P} = \varnothing$*, then two-Gaussian type w.r.t.* $(W, U, \varnothing)$ *requires that both* $\mathbb{P}(Y_W|Y_W \neq 0)$ *and* $\mathbb{P}(Y_U|Y_U \neq 0)$ *are mixtures of exactly two distinct univariate Gaussian distributions with constant parameters, respectively.*

We next recall the following observation that appears as Proposition 29(ii) in Peters et al. (2014); see Sections 1.8 and 15.3.2 of Maathuis et al. (2019) for definitions of the Markov property and faithfulness.

Proposition 31. *Suppose the distribution of* $\boldsymbol{Y}$ *is Markov and faithful with respect to two distinct Markov equivalent graphs* $\mathcal{G}$ *and* $\mathcal{G}'$*. Then, there must exist nodes* W *and* U *such that* $W \to U$ *in* $\mathcal{G}$ *and* $U \to W$ *in* $\mathcal{G}'$*, while* $\mathcal{P} \equiv \mathrm{pa}_{\mathcal{G}}(U)\backslash\{W\} = \mathrm{pa}_{\mathcal{G}'}(W)\backslash\{U\}$*.*

Remark 4. *Proposition 31 is at the heart of many proofs of DAG identifiability, which combine it with suitable probabilistic conditioning to reduce the comparison of two DAG models to bivariate problems involving the two graphs* $W \to U$ *and* $W \leftarrow U$*. However, in our setting, a key new challenge arises because the form of the Hurdle conditionals precludes us from applying conditioning to form sets of bivariate distributions that are of the considered Hurdle type. Indeed, conditioning on descendants of the considered variables (i.e., other variables that in the graph can be reached along directed paths) generally gives conditional distributions that are no longer of the Hurdle type used in the definition of our model class.*

We claim that the intersection of sets of joint distributions represented by two distinct Markov equivalent $\mathcal{G}$ and $\mathcal{G}'$ must be a subset of 2-Gaussian type distributions with respect to a triplet $(W, U, \mathcal{P})$ obtained from Proposition 31.

Theorem 32 (General Identifiability)**.** *Let* $\boldsymbol{Y}$*,* $\mathcal{G}$*,* $\mathcal{G}'$*,* W*,* U*,* $\mathcal{P}$ *be as in Proposition 31. Let* $\boldsymbol{Y}$ *have a* λ^m*-density that factors w.r.t. both graphs* $\mathcal{G}$ *and* $\mathcal{G}'$*. For each* $\mathcal{H} = \mathcal{G}, \mathcal{G}'$*, let the*

node conditionals in the factorization be Hurdle conditionals with the parameters $(\alpha_V^{\mathcal{H}})_{V\in\mathcal{V}}$ *and* $(\beta_V^{\mathcal{H}})_{V\in\mathcal{V}}$ *from (4.4), or equivalently* $(p_V^{\mathcal{H}})_{V\in\mathcal{V}}$ *and* $(\mu_V^{\mathcal{H}})_{V\in\mathcal{V}}$ *from (4.5), that are Hurdle polynomials of the form (4.8), where for* $(V, T, \mathcal{H}) = (U, W, \mathcal{G})$ *and* $(V, T, \mathcal{H}) = (W, U, \mathcal{G}')$ *it holds that*

(i) $\beta_V^{\mathcal{H}}(y_T, \boldsymbol{y}_{\mathcal{P}})$ *(or* $\mu_V^{\mathcal{H}}(y_T, \boldsymbol{y}_{\mathcal{P}})$*) depends on at least one of* $\mathbb{1}_{y_T}$ *and* y_T*, or*

(ii) $\alpha_V^{\mathcal{H}}(y_T, \boldsymbol{y}_{\mathcal{P}})$ *(or* $p_V^{\mathcal{H}}(y_T, \boldsymbol{y}_{\mathcal{P}})$*) depends on the value of* y_T *(and maybe additionally on* $\mathbb{1}_{y_T}$*).*

Then the distribution of $\boldsymbol{Y}$ *must be of two-Gaussian type w.r.t.* $(W, U, \mathcal{P})$*. In this case we also say the distribution is of two-Gaussian type w.r.t.* $\mathcal{G}$ *and* $\mathcal{G}'$*.*

Note that the assumption of faithfulness in Proposition 31 implies that we have (i) or (ii) or a condition (iii) that states that $\alpha_V^{\mathcal{H}}(y_T, \boldsymbol{y}_{\mathcal{P}})$ (or $p_V^{\mathcal{H}}(y_T, \boldsymbol{y}_{\mathcal{P}})$) depends on $\mathbb{1}_{y_T}$ only and $\beta_V^{\mathcal{H}}(y_T, \boldsymbol{y}_{\mathcal{P}})$ (or $\mu_V^{\mathcal{H}}(y_T, \boldsymbol{y}_{\mathcal{P}})$) is constant in y_T. It is case (iii) that we rule out in our assumption of Theorem 32.

The result is proved in Appendix C. It is easy to show that the result also holds if we make modifications such as restricting the maximum degree of the polynomial or excluding interactions between the discrete and continuous components.

In the two- and three-dimensional cases (i.e., $m = 2, 3$) we show in Appendix C that there does not exist a joint distribution for $\boldsymbol{Y}$ that is of two-Gaussian type with respect to two distinct Markov equivalent graphs. We thus have the following result on full identifiability for graphs with two or three nodes.

Corollary 33 (Identifiability in two and three dimensions)**.** *If* $|\mathcal{V}| \leq 3$*, i.e., in a binary/triary setting, there does not exist a joint distribution that is of two-Gaussian type w.r.t. two distinct Markov equivalent DAGs* $\mathcal{G}$ *and* $\mathcal{G}'$*. Thus, strong identifiability is guaranteed as in Theorem 30, meaning that the sets of Markov and faithful distributions associated to* $\mathcal{G}$ *and* $\mathcal{G}'$ *must be disjoint.*

Theorem 30 and Corollary 33 state that the DAGs are perfectly identifiable from the distributions if $m = 2, 3$ or if we assume the Hurdle polynomials to be *strong*; Theorem 32 claims that without assuming *strong* Hurdle polynomials, the distributions for $m > 3$ from which the graph is not identifiable must be a subset of the *two-Gaussian type* distributions. We conjecture that in general, with $m > 3$, the set of two-Gaussian type distributions with respect to any two graphs is an empty set.

In Section 4.5.1 we show scatter plots of simulated data that give some indication of how Markov equivalent graphs may be differentiated under our models.

4.4 Estimation of DAGs from Zero-Inflated Data

Suppose now that we are given an i.i.d. sample $\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(n)}$ comprised of m-variate observations. The log-likelihood function ℓ of any DAG model can be decomposed into the sum of conditional (or nodewise) log-likelihood functions ℓ^V for the V-th variable conditional on its parent variables. Let $y_V^{(1)}, \ldots, y_V^{(n)}$ be the n observations of the V-th variable. For the canonical (α, β, k)-parametrization from (4.4), the nodewise log-likelihood function is

$$\ell^V\left(\alpha_V, \beta_V, k_V \,|\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(n)}\right) = \sum_{i=1}^{n} \Big(\alpha_V\left(\boldsymbol{y}^{(i)}_{\mathrm{pa}(V)}\right) \mathbb{1}_{y_V^{(i)}} + \beta_V\left(\boldsymbol{y}^{(i)}_{\mathrm{pa}(V)}\right) y_V^{(i)} - k_V {y_V^{(i)}}^2/2 \\ - \log\left[\sqrt{2\pi/k_V}\exp\left\{\alpha_v\left(\boldsymbol{y}^{(i)}_{\mathrm{pa}(V)}\right) + \beta_V^2\left(\boldsymbol{y}^{(i)}_{\mathrm{pa}(V)}\right)/(2k_V)\right\} + 1\right]\Big);$$

for the moment (p, μ, σ^2)-parametrization from (4.5) it is

$$\ell^V\left(p_V, \mu_V, \sigma_V^2 \,|\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(n)}\right) = \sum_{i: y_V^{(i)}=0} \log\left\{1 - p_V\left(\boldsymbol{y}^{(i)}_{\mathrm{pa}(V)}\right)\right\} \\ + \sum_{i: y_V^{(i)}\neq 0}\left[\log p_V\left(\boldsymbol{y}^{(i)}_{\mathrm{pa}(V)}\right) - \frac{1}{2}\log(2\pi\sigma_V^2) - \left\{y_V^{(i)} - \mu_V\left(\boldsymbol{y}^{(i)}_{\mathrm{pa}(V)}\right)\right\}^2/(2\sigma_V^2)\right].$$

In the latter case, we see the sum of the log-likelihood functions from the logistic regression model for p_V and the linear regression for μ_V restricted to the observations with $y_V^{(i)} \neq 0$. Here we recall that the parameters $\alpha_V, \beta_V, p_V, \mu_V$ are themselves polynomials in $\boldsymbol{y}_{\mathrm{pa}(V)}$ and

their indicators, and we are using them as a shorthand notation on the left-hand sides where we really mean ℓ^V as a function of the parameters (i.e., coefficients) in those polynomials.

4.4.1 Fitting Hurdle Conditionals

Estimation of the graphical models amounts to fitting the conditional distribution of one node given a set of others. For the canonical (α, β, k)-parametrization, the log-likelihood function is convex in α_V, β_V and k_V. Moreover, α_V and β_V are linear in the polynomial coefficients. Therefore, the log-likelihood is convex in the coefficients to estimate and can be maximized by standard methods; e.g., coordinate descent. Estimation for the moment (p, μ, σ^2)-parametrization (4.5), on the other hand, can be easily solved by separately fitting a logistic regression to p_V and a linear regression to μ_V. Recall again that the two parametrizations, canonical and moment, are equivalent when assuming a full polynomial model, i.e., when the degree and structure of the polynomials is unrestricted. However, when restricting, for instance, the degree the two parametrizations yield different models.

The (α, β, k)-parametrization with linear Hurdle polynomials (i.e., degree 1) is interesting as it naturally comes from conditional distributions of the joint distribution defined for undirected graphical models in McDavid et al. (2019). However, at least for higher degree, the (p, μ, σ^2)-parametrization may be more intuitive and useful in practice as it leads to a decomposition into a logistic regression and a linear regression. This decomposition enables one to use optimized standard regression solvers for model fitting. The (p, μ, σ^2)-parametrization also makes it easy to apply available routines to incorporate regularization on the coefficients/parameters into our loss, which is helpful when the number of samples is small compared to the number of parameters. Such higher dimensionality of the models arises in particular when assuming a higher degree for the Hurdle polynomials. The regularization is automatically applied in the implementation in our R package `ZiDAG` available on `GitHub`.

For estimation of our models, we assume a highest degree of the Hurdle polynomials. To select the degree from data we adopt the Bayesian information criterion (BIC). This

functionality is incorporated in `ZiDAG`.

4.4.2 Graph Search

For estimation of the DAG underlying the graphical model, we mainly consider two state-of-the-art methods: (A) exhaustive score-based search and (B) greedy search. Both methods rely on a model score which we take to be the BIC defined as $\nu \log n - 2\ell$, where ν is the total number of parameters in the model, n is the sample size, and ℓ is the log-likelihood as introduced in the beginning of Section 4.4.

(A) **Exhaustive search**: Optimizing the BIC over the set of all DAGs is possible for moderately small m using the dynamic programming algorithm of Silander and Myllymäki (2006). This approach is justified by the asymptotic consistency of the BIC as well as the identifiability of our model (recall Section 4.3). The experiments of Silander and Myllymäki (2006) suggest that for Gaussian models the search is practical for $m < 32$. Estimation of our models is computationally more challenging but exhaustive search is feasible at least for $m < 16$.

(B) **Greedy search**: Instead of optimizing BIC over all DAGs, we may apply a greedy search that iteratively improves BIC by moving to a neighboring DAG that provides the largest improvement. The neighborhood is defined using edge additions, deletions, and reversals; compare Chickering (2003). While Chickering (2003) discusses consistency of graph recovery in terms of equivalence classes, in our case the algorithm determines individual graphs. For faster estimation in sparse settings, we consider restricting the maximum node in-degree (i.e., the maximum number of parents).

Remark 5. *We have also experimented with a version of the PC algorithm, which is not easily applicable since it relies on a suitable conditional independence test between two variables given any potential parent set. Indeed, by the nature of our models if the potential parent set is misspecified, the conditional distributions may no longer be Hurdle. Another*

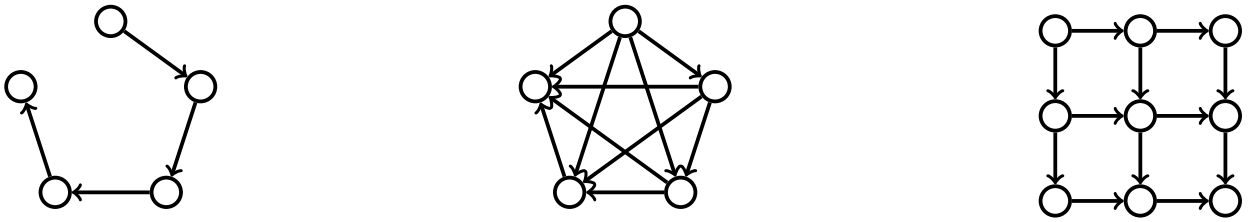
possible approach starts with greedily estimating the topological ordering of nodes by iteratively picking the node that maximizes the conditional likelihood given nodes already chosen, followed by a variable selection problem using, for example, a Wald test or ℓ_1 regularization techniques; this method relies on very subtle features of the distributions. Neither the PC algorithm we designed nor the approach focusing on the topological ordering were competitive in our experiments.

4.4.3 Stability Selection

In our application to single-cell gene expression data, we seek to also achieve some control of the false discovery rate (FDR). To this end, we apply stability selection in graph estimation. In particular, we take up the approach outlined in Shah and Samworth (2013). We randomly choose $B = 50$ subsets of the data (each of size $\lfloor n/2 \rfloor$), and obtain B other sets as their complements of equal size, randomly throwing out one sample if n is odd. We then estimate the graph using $\lfloor n/2 \rfloor$ subsamples indexed by each of these $2B$ sets of equal size, and obtain $2B$ estimated DAGs. Given the desired FDR, we compute a frequency threshold using the formula from Shah and Samworth (2013, Eqn. (8)) with number of total parameters $m(m-1)/2$. We then keep all edges that occur more often than the frequency threshold and produce a graph as our final estimate. In our implementation in `ZiDAG`, if a graph estimated this way is not acyclic, the user can choose to return it as is, or the function will increase the threshold up to the point where the resulting graph is a DAG, even though the resulting graph might be empty in extreme cases.

4.5 Numerical Experiments

In this section we present numerical experiments for exact DAG recovery using simulated zero-inflated conditional Gaussian data. The main goal is to verify identifiability using exhaustive search, and examine how accurately greedy search can recover the true graph.



Chain, $m = 5$ Complete, $m = 5$ Lattice, $m = 9$

Figure 4.2: Graph structures used in our experiments.

4.5.1 True Underlying DAGs and Distributions

We consider three DAG structures, i) chain graph with $m = 5$, ii) complete graph with $m = 5$, iii) lattice graph with $m = 9$, as illustrated in Figure 4.2. We keep m small for tractability of repeated application of Silander and Myllymäki (2006) with quadratic Hurdle polynomials. In practice, we suggest applying the directed graph models to the connected components inferred from undirected graphs, estimated using the joint Hurdle distribution (4.2) as in McDavid et al. (2019). This often results in considerably smaller sizes m. In particular, sizes of m in this section are similar to the largest component in our data analysis in Section 4.6.

For each structure, we consider true generating conditional distributions using the following parametrizations: a) (α, β, k)-(canonical) parametrization with *linear* Hurdle polynomials, b) (p, μ, σ^2)-(moment) parametrization with *linear* Hurdle polynomials, and c) (p, μ, σ^2)-(moment) parametrization with *quadratic* Hurdle polynomials. We note that the distributions represented by c) is a superset of those by a) and b). By (4.6), distributions represented by a) and b) are disjoint because $\log(p/(1-p))$ is a weighted sum of α and β^2.

Recall the definition of Hurdle conditionals in (4.4) and (4.5) in Section 4.2.2. In our experiments, whenever $\text{pa}(V) = \varnothing$, we generate $y_V \sim f_0$ such that $f_0(x) = \frac{1}{2}(1 - \mathbb{1}_x) + \frac{1}{2}\phi(x; 0, 1)$, where ϕ is the standard normal density. Otherwise, for parametrization a),

we use Hurdle conditionals with parameters $k_V = 1$, $\alpha_V(\boldsymbol{y}_{\mathrm{pa}(V)}) = \sum_{U\in\mathrm{pa}(V)} \mathbb{1}_{y_U} + y_U$ and $\beta_V(\boldsymbol{y}_{\mathrm{pa}(V)}) = \sum_{U\in\mathrm{pa}(V)} (\mathbb{1}_{y_U} - y_U)$; similarly for parametrization b) we take $\sigma_V^2 = 1$, $\log \frac{p_V}{1-p_V}(\boldsymbol{y}_{\mathrm{pa}(V)}) = \sum_{U\in\mathrm{pa}(V)} (\mathbb{1}_{y_U} + y_U)$ and $\mu_V(\boldsymbol{y}_{\mathrm{pa}(V)}) = \sum_{U\in\mathrm{pa}(V)} (\mathbb{1}_{y_U} - y_U)$; finally, for parametrization c) we take $\sigma_V^2 = 1$ and

$$\log \frac{p_V}{1-p_V}(\boldsymbol{y}_{\mathrm{pa}(V)}) = \sum_{U\in\mathrm{pa}(V)} \left(\mathbb{1}_{y_U} + y_U + \frac{y_U^2}{10}\right) + \frac{1}{10} \sum_{\substack{U,W\in\mathrm{pa}(V)\\ U\neq W}} (\mathbb{1}_{y_U} + y_U)(\mathbb{1}_{y_V} + y_V), \text{ and}$$

$$\mu_V(\boldsymbol{y}_{\mathrm{pa}(V)}) = \sum_{U\in\mathrm{pa}(V)} \left(\mathbb{1}_{y_U} - y_U - \frac{y_U^2}{10}\right) + \frac{1}{10} \sum_{\substack{U,W\in\mathrm{pa}(V)\\ U\neq W}} (\mathbb{1}_{y_U}\mathbb{1}_{y_V} - y_U\mathbb{1}_{y_V} - y_V\mathbb{1}_{y_U} - y_V y_U).$$

We then normalize the coefficients in the above expressions $(\pm 1, \pm 1/10)$ such that α_V, β_V, $\log p_V/(1-p_V)$ and μ_V have means 0 and 1, respectively, across the samples. This normalization ensures that the marginal probability of being nonzero, the marginal mean, and the marginal variance for each node are stabilized, in order to show that the DAGs are truly recovered based on the conditional dependency structure instead of additional signals from these marginal quantities. In fact, in the generated samples the marginal probability is about 0.5 and the marginal mean is about 0 for all nodes, and the marginal variance for the nonzero part only is about the same for all except the source node.

In Figure 4.3, we present pairwise scatter plots of one instance of data generated with the chain graph (upper row) and the complete graph (lower row), respectively, both with (p, μ, k)-linear parametrization. Since the true topological ordering is $1 \to 2 \to 3 \to 4 \to 5$, for clarity we exclude the source and sink nodes (1 and 5) and only include nodes 2, 3 and 4. Plots on the left are plotted in the order $2, 3, 4$ and those on the right are reversed. In the histograms on the diagonals we only plot the continuous part.

The scatter plots indicate a slight difference in the respective marginal distributions of nodes 2 and 4 conditioned on node 3 being 0 (and vice versa). This difference intuitively explains how the orientation $2 \to 3 \to 4$ versus $4 \to 3 \to 2$ can be identified. Note that other than this difference, the marginal statistics for the three nodes are indistinguishable and there is little noticeable difference between plots on the left and on the right.

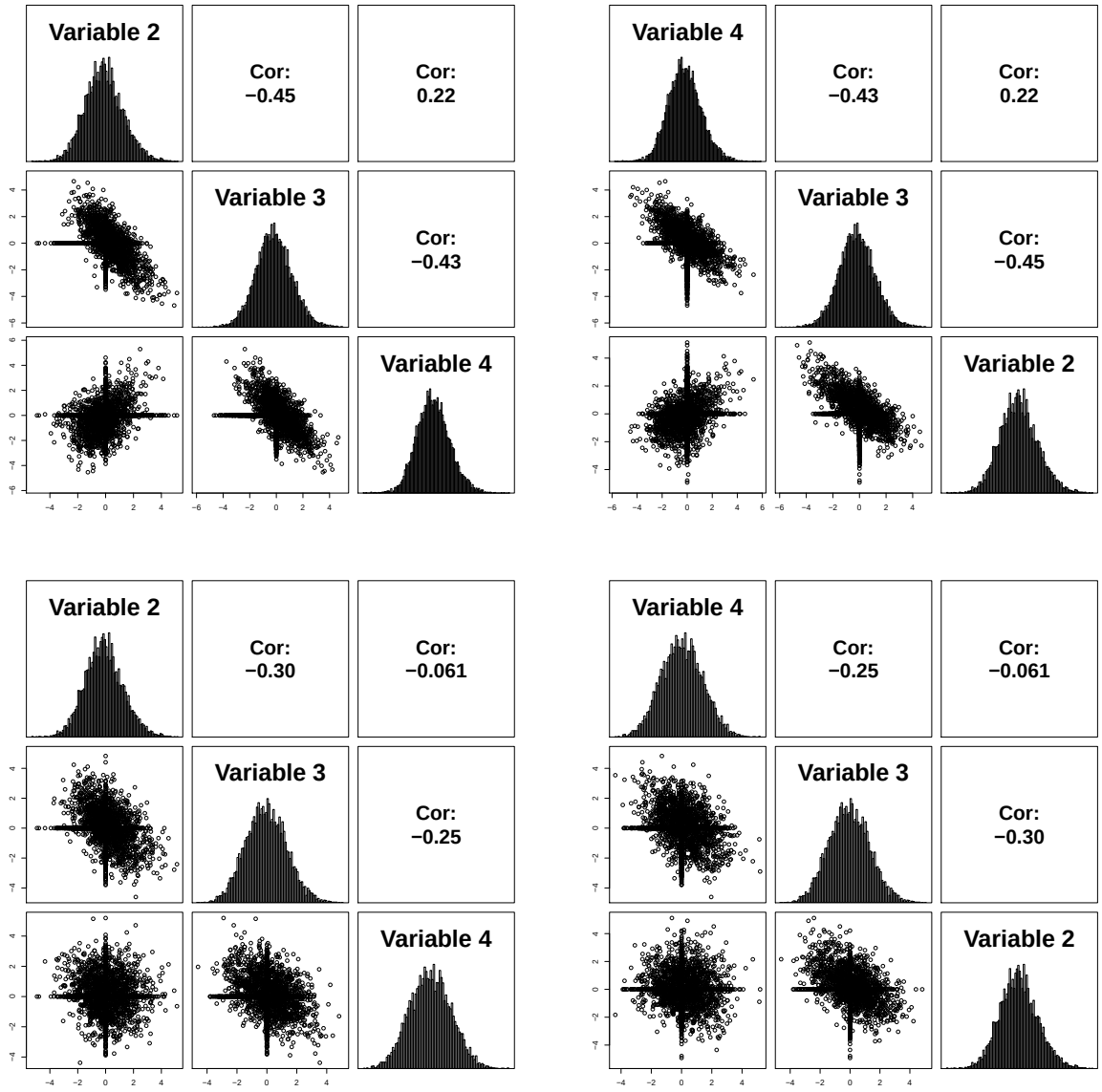


Figure 4.3: Pairwise scatter plots of zero-inflated data generated using chain graphs (upper row) and complete graphs (lower row), both with topological ordering $1 \to 2 \to 3 \to 4 \to 5$; only nodes 2, 3 and 4 are plotted. Plots on the left are plotted in the order $2, 3, 4$, and $4, 3, 2$ on the right. Only the continuous part is plotted in the histograms on the diagonals. There is little noticeable difference between the histograms and scatter plots when we reverse the graph order, yet our model can still determine the correct topological ordering.

4.5.2 Estimation

We use the two graph estimation methods described in Section 4.4.2, namely our self-implemented greedy search (GDS) (Chickering, 2003) with BIC score, and exhaustive search with dynamic programming (Silander and Myllymäki, 2006). Details on fitting the Hurdle conditionals themselves were presented in Section 4.4.1.

In our simulation, we aim to assess the performance of the different estimation procedures for correctly specific and misspecified parametrizations. To this end, for each combination of true DAG and true data generating parametrization—(α, β, k)-linear and (p, μ, σ^2)-linear and quadratic—we estimate the DAG using all three parametrizations for generating data. For simplicity and given that the simulation results are presented over $B = 100$ iterations, stability selection is not used in these experiments.

4.5.3 Results on Graph Recovery

Results are shown in Figures 4.4–4.6. Each figure has one true underlying DAG from those shown in Figure 4.2. In all figures, each row indicates one choice of true data generating parametrization—(α, β, k)-linear, and (p, μ, σ^2)-linear and quadratic—and each column shows the results using each estimating parametrization. Thus, plots on the diagonal (with bold titles) correspond to correct parametrizations, where the estimating parametrization agrees with the truth. Off-diagonal plots, in contrast, correspond to cases where the model parametrization is misspecified.

The solid lines with '$\times$' points show the success rates (exact recovery) out of 100 trials versus n for greedy search (GDS), and the solid lines with '$\circ$' points represent exhaustive search (Silander). The gray dotted lines signify success rates measured by recovery of the equivalence class.

Since exhaustive search compares all possible DAGs for m nodes, for n large enough it provides an indicator of identifiability. Indeed, the results indicate that in all settings, exhaustive search with correct parametrization almost always identifies the exact DAG for

large n. In fact, since the (p, μ, σ^2)-quadratic parametrization covers the other two, in all cases the graphs can be perfectly recovered using the quadratic estimating parametrization. In contrast, when the underlying truth is quadratic, the graph may not be easily identified from estimates that use the other two parametrizations. This is especially the case for the lattice graph. Comparing the linear parametrizations themselves, (p, μ, σ^2) seems less prone to model misspecification and has the advantage of faster estimation with the help of standard softwares for logistic and linear regressions.

Overall, our simulation studies confirm the identifiability theory (Theorem 30). In particular, our experiments indicate that exhaustive search performs well. They also indicate that GDS works reasonably well for sparse graphs but may require larger samples for recovering the structure of complete, or very dense, graphs. While exhaustive search often succeeds with high probability even with small samples, it may not be scalable for large m. In such cases, the greedy and faster GDS method, which shows promising results, provides a viable alternative. Utilizing the stability selection method of Section 4.4.3 can further improve the GDS results.

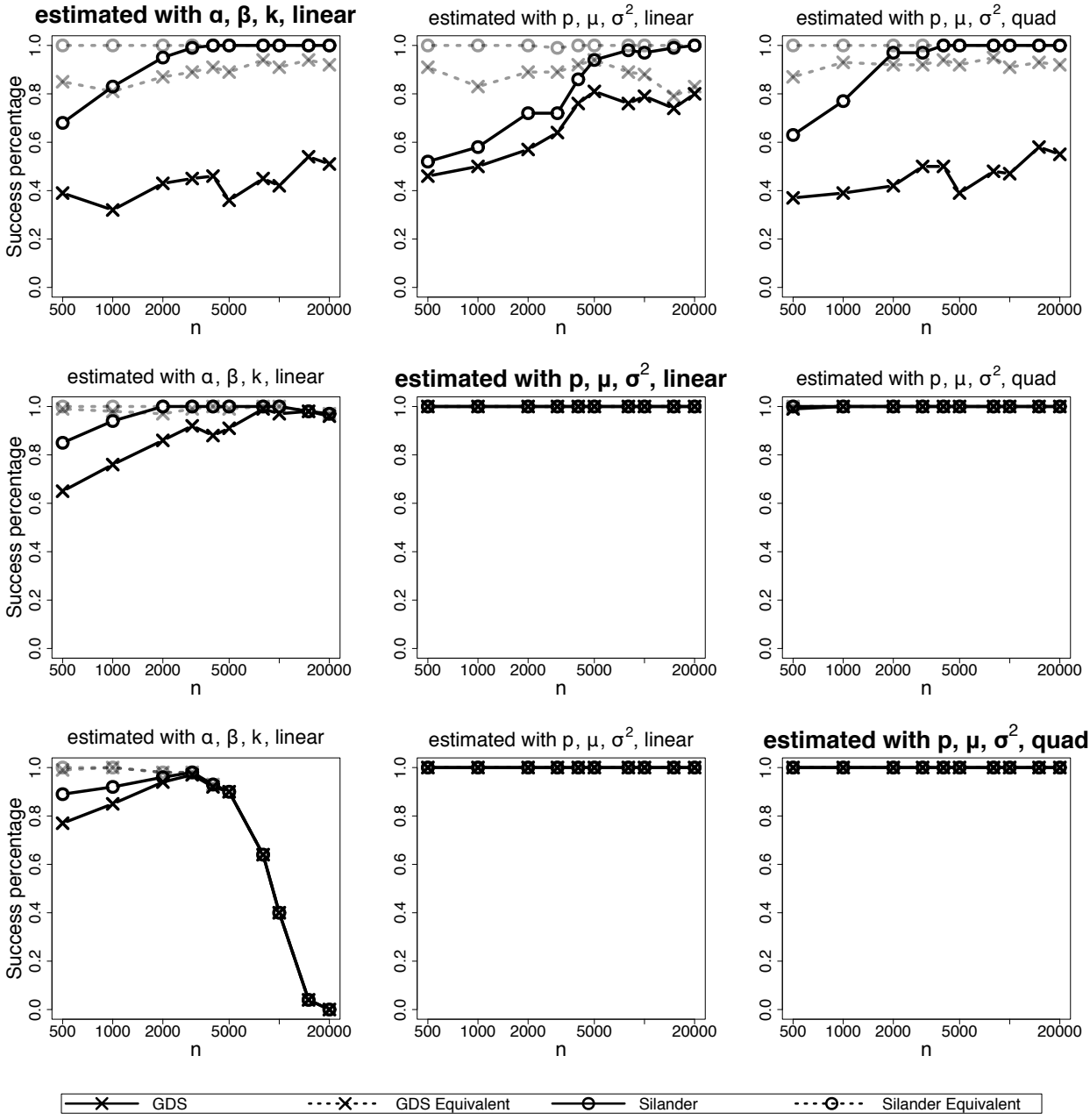


Figure 4.4: Chain graph, $m = 5$. Each row corresponds to a different generating parametrization, and each column a different estimating parametrization. Generating and estimating parametrizations agree on the diagonal. Solid '$\times$': success rates of exact DAG recovery for greedy search; solid '$\circ$': success rates of exact DAG recovery for exhaustive search; gray dotted lines: success rates for recovery of equivalence class.

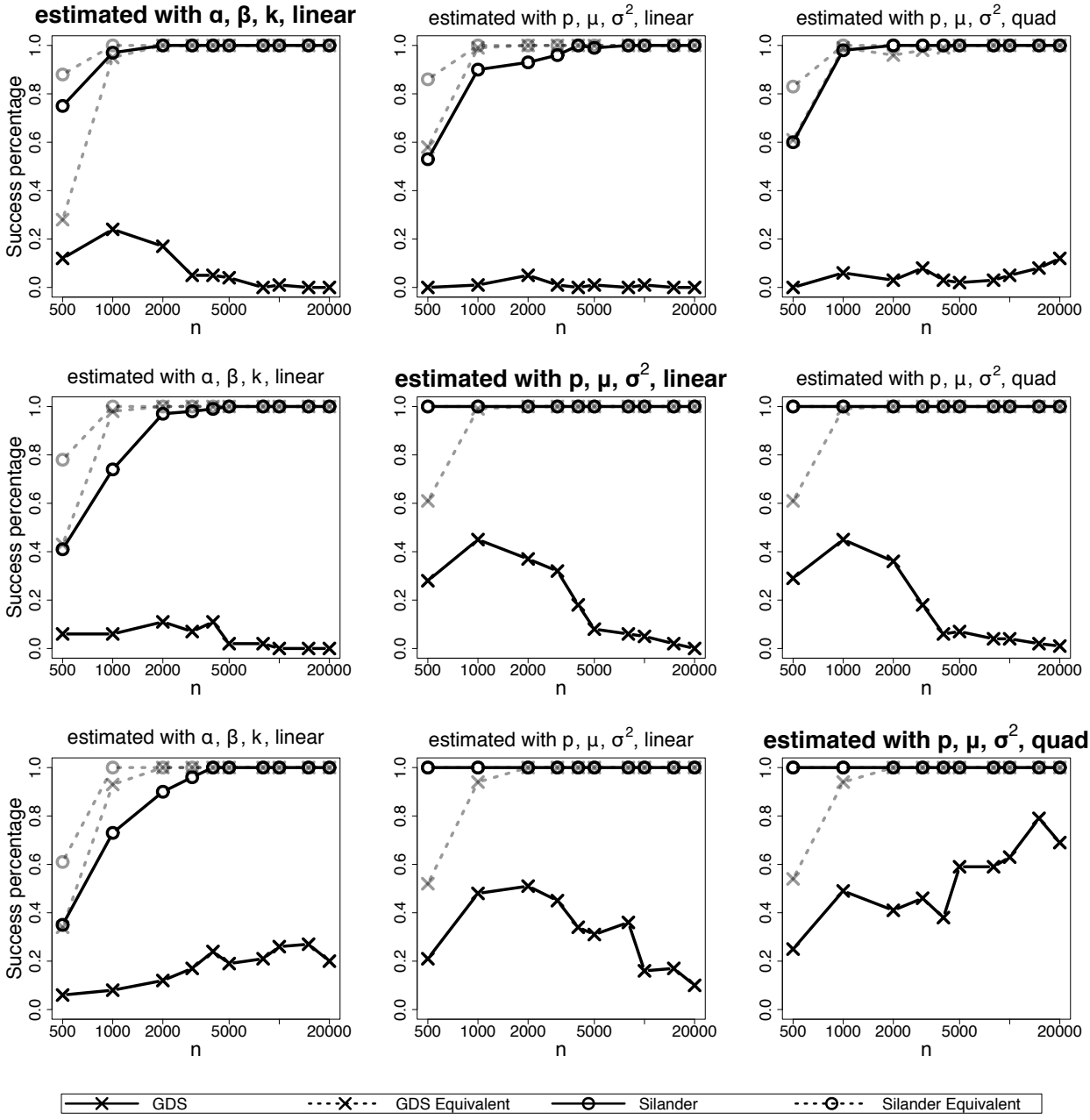


Figure 4.5: Complete graph, $m = 5$.

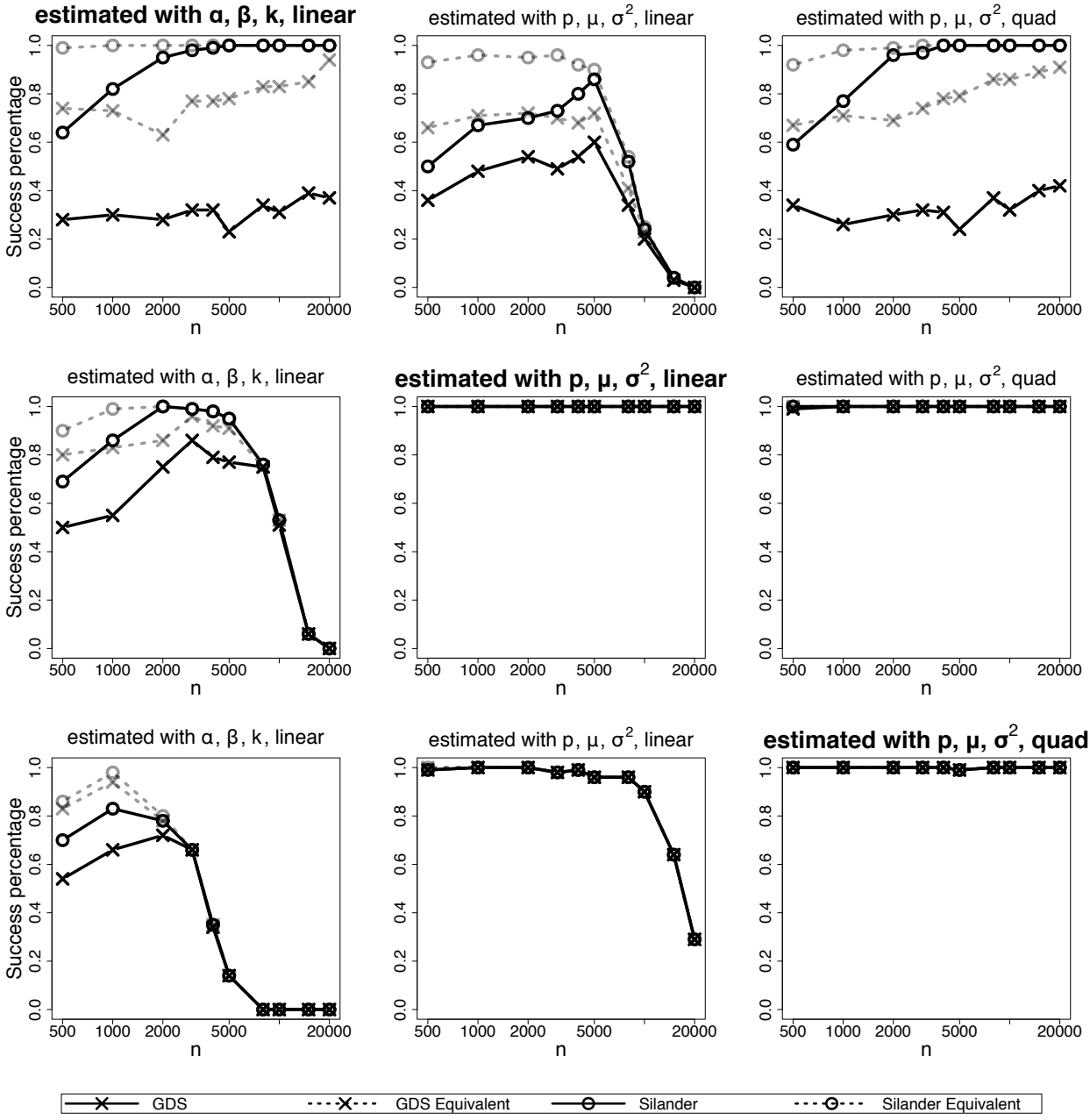


Figure 4.6: Lattice graph, $m = 9$.

4.6 T Helper Cell Data

In this section we present the results of applying our model to a T helper cell expression dataset. Specifically, the dataset is considered in McDavid et al. (2019) and contains both single cell and 10-cell expression measurements for T helper cells for 80 genes in eight healthy donors. We use all 1951 single cell measurements for these donors (a superset of the 465 measurements in McDavid et al. (2019)) to ensure we have a large enough sample size to produce reliable estimates. In particular, McDavid et al. (2019) consider only the T-follicular ($CXCR5^+PD1^+$) cells that produce high levels of proteins CXCR5 and PD1, while we do not make this restriction. Instead, we add the indicators of $CXCR5^{+/-}$ and $PD1^{+/-}$ as regressors when fitting the conditional distributions. Following McDavid et al. (2019), we choose the 61 genes that have at least 5% zero and 5% nonzero values.

While the measurements are all nonnegative, the minimum, mean, and standard deviation of the nonzero values in the dataset are 7.89, 18.53, and 1.91, respectively. We thus assume zero-inflated conditional Gaussianity without considering the effect of truncation from below at 0.

To estimate the DAG structure, we first use the procedure of McDavid et al. (2019) to identify the connected components in an estimated undirected graph with the same sparsity as the graph therein. We then estimate the directed edges in each connected component using our method. This procedure is justified by the fact that theoretically the connected components for the underlying true undirected and directed graphs coincide. Thus, we generally recommend this strategy in practice, as the connected components can be much more efficiently obtained from the undirected graph.

We use the (p, μ, σ^2)-parametrization as it is more flexible than the (α, β, k), and extra fixed covariates and controlling factors can be easily added, since fitting the conditionals only involves linear and logistic regressions. As discussed in Section 4.5, the (p, μ, σ^2) is also more robust than (α, β, k). We use polynomials up to degree three and data-adaptively choose the optimal degree by BIC.

To estimate the DAG, we use the greedy search (GDS) algorithm, which showed promising performance in the simulations of Section 4.5. We also use the stability selection procedure of Section 4.4.3, with the goal of controlling the FDR at 10% for each connected component. For smaller connected components, if controlling the FDR at 10% is not possible, we pick the sparsest graph that maximally maintains the connectivity. Finally, we restrict the node in-degrees to five, in order to both speed up estimation and to constrain the search space. This constraint is motivated by the fact that in gene regulatory networks, each gene is only expected to be regulated by a small number of other genes (Albert, 2005). In contrast, since genetic networks often involve hub genes that regulate many others, we do not restrict the out-degree.

The estimated undirected graph using the procedure of McDavid et al. (2019) is plotted in the upper half of Figure 4.7, with edge width and saturation representing the edge strength. In the lower half of the figure, we plot the estimated DAG using our method; the estimate with stability selection and FDR control is shown on the left and the one without stability selection is on the right. Examining the estimates, we find that CD3E is a hub node with degree five in both estimated DAGs, while it has four neighbors in the estimated undirected graph. On the other hand, in the estimate with stability selection, three genes in the largest connected components, namely CD28, JAK1 and STATS5B, are isolated. This is reasonable as they are each associated with only one weak edge in the undirected graph. Moreover, the undirected and directed graph estimates have different thresholds for determining whether an edge is present. For the other nodes, the estimated DAG structures are very similar to the undirected graph estimate.

4.7 Discussion

Motivated by the recent advent of single-cell RNA-seq data, in this paper we develop new methods for learning DAGs from zero-inflated data. Our procedures take advantage of two key features of single-cell RNA-seq data, namely, the zero-inflated nature of the data, and the large number of observations from individual samples.

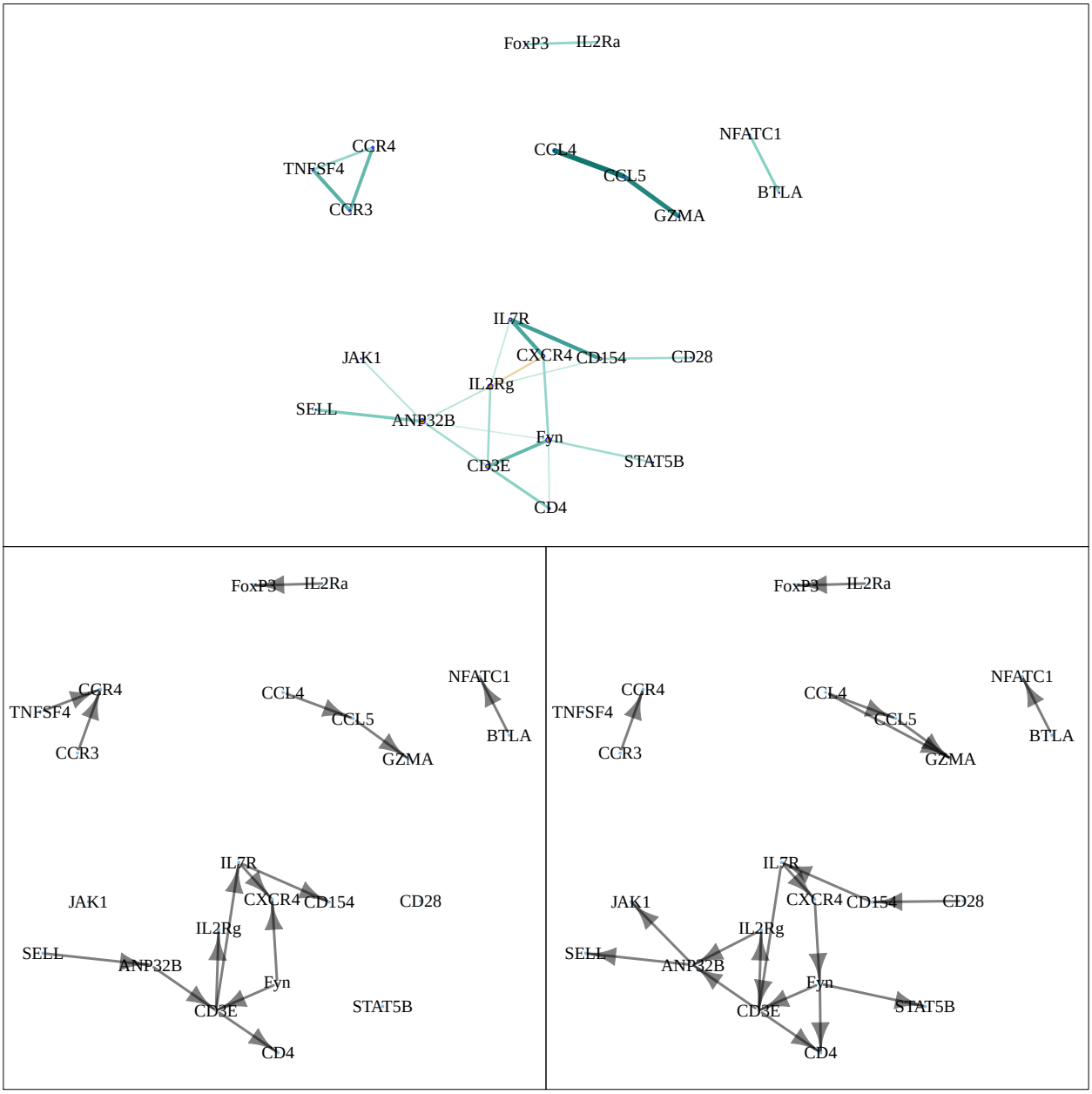


Figure 4.7: Upper: Graph estimated for T helper cell data using undirected zero-inflated graphical models; similar to Figure 6 from McDavid et al. (2019). Lower: Directed graph estimated for T helper cell data with (left) and without (right) stability selection and FDR control.

Our key contribution is establishing identifiability of DAGs from observational zero-inflated data. Specifically, we prove that the exact DAG can be recovered from the joint distribution under reasonable assumptions. We also show that in the most general case, the distributions from which the DAGs are not identifiable only form a small subset, which we prove to be empty in the bivariate and trivariate cases. While our proof uses a very general result on DAGs from Peters et al. (2014) as its first step, our models do not fit into the framework in that paper; we thus take a different approach that considers the zero-inflation and polynomial structures directly.

Our approach is based on factorizing the joint distribution into zero-inflated conditional Gaussian distributions with parameters polynomial in the parents and their indicators of having nonzero values. We present models in terms of two parametrizations, one called (α, β, k) that is linked to the undirected graphs studied in McDavid et al. (2019), and the other called (p, μ, σ^2) that directly models the conditional moments. Both approaches have computational appeal. In particular, the (α, β, k)-parametrization leads to convex loss functions in the parameters to be estimated, while the (p, μ, σ^2)-parametrization offers the additional benefit of allowing one to utilize standard software for logistic and linear regression. We combine these models with two state-of-the-art estimation procedures, namely greedy DAG search (GDS) and exhaustive search with dynamic programming. We also validate our identifiability theory using extensive numerical studies. These experiments indicate that the exhaustive search algorithm is effective in correctly identifying DAGs with small number of nodes. For moderate to large DAGs, the GDS algorithm offers a reasonable alternative, with performance comparable to the exhaustive search when the sample size is large enough.

Our work opens the door to multiple future research directions and extensions. The first is to prove our conjecture that the sets of distributions from which the DAG is not identifiable are empty also for graphs with more than 3 nodes. The second direction of future research is proving the consistency and investigating finite sample properties of the proposed estimation procedures. Finally, an interesting extension of our model would be to consider zero-inflated distributions under a truncation to the nonnegative orthant $\mathbb{R}^m_+$, which

would be of interest for nonnegative *omics* data. The main challenge in this case would be the normalizing constant as a function of the parents in the conditional distributions, since it would not have a closed-form expression. While this may be resolved by generalizing the *score matching* loss (Hyvärinen, 2005, 2007; Lyu, 2009; Yu et al., 2019b) to data of mixed type, the additional difficulty would lie in proving identifiability and addressing estimation from observational data.

Chapter 5

DISCUSSION AND FUTURE WORK

In this dissertation, we discuss both undirected and directed graphical models and their estimation. In Chapter 2 we propose the generalized score matching estimator for estimating continuous distributions supported on $\mathbb{R}^m_+$ with applications to pairwise interaction power models; the estimator enjoys the advantages of avoiding calculations of normalizing constants as well as being quadratic in the canonical parameters of exponential families, while significantly outperforming the original estimator proposed in Hyvärinen (2007). The generalization is done by replacing the $\nabla \log p(\boldsymbol{x}) \otimes \boldsymbol{x}$ term in Hyvärinen (2007) with $\nabla \log p(\boldsymbol{x}) \otimes \boldsymbol{h}^{1/2}(\boldsymbol{x})$ for some general $\boldsymbol{h} : \mathbb{R}^m_+ \to \mathbb{R}^m_+$. In Chapter 3 we extend the estimator to distributions supported on more general domains $\mathcal{D}$, which should cover most applications in practice, but only involves little extra work from calculating the *truncated componentwise distance* $\boldsymbol{\varphi}$ of the data to the boundary of $\mathcal{D}$, truncated above by prespecified constants $\boldsymbol{C}$. Finally, in Chapter 4, inspired by the recent technology of single-cell RNA-seq data, we propose a class of directed graphical models for zero-inflated data, for which we establish identifiability of directed acyclic graphs (DAGs) under reasonable assumptions, and present graph estimation methods that prove to be successful in practice.

In Chapters 2 and 3, for high-dimensional exponential family graphical models, following Lin et al. (2016) we add an ℓ_1 penalty, and in addition we introduce an ℓ_2-type penalty using *diagonal multipliers* to overcome the issue in Lin et al. (2016) that the regularized loss may be unbounded if the ℓ_1 penalty is small. We consider pairwise interaction power models, which are exponential families in which the log density is the sum of pairwise interactions between $\boldsymbol{x}^a$ ($\log \boldsymbol{x}$ when $a = 0$) plus linearly weighted effects $\boldsymbol{x}^b$ ($\log \boldsymbol{x}$ when $b = 0$). This covers a wide range of models, such as truncated normal distributions with $a = b = 1$, exponential square root models (Inouye et al., 2016) with $a = b = 1/2$, a class of multivariate gamma distributions $a = 1/2$ and $b = 0$, as well as A^{m-1} models (Aitchison, 1985) with $a = b = 0$ and $\mathcal{D}$ being the m-dimensional simplex. We show that for Gaussian graphical models on $\mathcal{D} \equiv \mathbb{R}^m_+$, we can reduce the sample size required for consistency from $\Omega(d^2 \log^8 m)$ in Lin et al. (2016) to $\Omega(d^2 \log m)$, where d is the max degree of the graph. For general domains $\mathcal{D}$, we have similar theory requiring $n = \Omega(d^2 \log m)$ for Gaussian graphical models on domains

that are finite disjoint unions of convex sets, as well as general a-b models on simplex domains assuming $a > 0$ or on bounded subsets of $\mathbb{R}_+^m$ with positive Lebesgue measure and $a \geq 0$. For other domains $\mathcal{D}$ we require an additional multiplicative factor that may weakly depend on m. Through simulation studies, we recommend the choice of $\boldsymbol{h}(\boldsymbol{x}) = (x_1^c, \ldots, x_m^c)$ with $c = \max\{2 - a, 0\}$, and adaptively choose the truncation points using sample quantiles.

In Chapter 4, we present our DAG models with two different parametrizations, which take advantage of the zero-inflated nature of the data and the large number of observations from individual samples. We prove that under reasonable assumptions, one can recover the exact DAG from the joint distribution, and that in the most general case, the distributions from which the exact DAGs are not identifiable only form a small subset, which is shown to be empty in bivariate and trivariate cases and conjectured to be empty for $m > 3$. The identifiability is supported by simulation studies using exhaustive search with BIC score. We use simulated experiments to show that the greedy DAG search (GDS) also works well when the sample size is large enough, a reasonable assumption for single-cell RNA-seq data.

Possible future extensions to our work in Chapters 2 and 3 could include developing a method that adaptively chooses the $\boldsymbol{h}$ function along with the truncation points $\boldsymbol{C}$ for the truncated distances $\boldsymbol{\varphi}$, as well as probing into the dependence of the additional multiplicative constant factor on $\log m$ that appears in the consistency results for certain domains $\mathcal{D}$. For the DAG models in Chapter 4, one may be interested in proving our conjecture above and the consistency and finite sample properties of the proposed estimation procedures, as well as extending our models to truncated zero-inflated distributions, possibly utilizing our work in Chapters 2 and 3.

BIBLIOGRAPHY

John Aitchison. A general class of distributions on the simplex. *Journal of the Royal Statistical Society: Series B (Methodological)*, 47(1):136–146, 1985.

Reka Albert. Scale-free networks in cell biology. *Journal of Cell Science*, 118(21):4947–4957, 2005.

Murilo P. Almeida and Basilis Gidas. A variational method for estimating the parameters of MRF from complete or incomplete data. *Ann. Appl. Probab.*, 3(1):103–136, 1993.

Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

Rina Foygel Barber and Mathias Drton. High-dimensional Ising model selection with Bayesian information criteria. *Electron. J. Stat.*, 9(1):567–607, 2015.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities*. Oxford University Press, Oxford, 2013.

Igor Brikun, Deborah Nusskern, Daniel Gillen, Amy Lynn, Daniel Murtagh, John Feczko, William G Nelson, and Diha Freije. A panel of DNA methylation markers reveals extensive methylation in histologically benign prostate biopsy cores from cancer patients. *Biomarker Research*, 2(1):25, 2014.

Peter Bühlmann and Sara van de Geer. *Statistics for High-dimensional Data: Methods, Theory and Applications*. Springer Science & Business Media, 2011.

V. V. Buldygin and Yu. V. Kozachenko. *Metric Characterization of Random Variables and Random Processes*, volume 188 of *Translations of Mathematical Monographs*. American

Mathematical Society, Providence, RI, 2000. Translated from the 1998 Russian original by V. Zaiats.

Scott L. Carter, Christian M. Brechbühler, Michael Griffin, and Andrew T. Bond. Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics*, 20(14):2242–2250, 2004.

Jiahua Chen and Zehua Chen. Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771, 2008.

Shizhe Chen, Daniela M. Witten, and Ali Shojaie. Selection and estimation for mixed graphical models. *Biometrika*, 102(1):47–64, 2015.

Wenyu Chen, Mathias Drton, and Y. Samuel Wang. On causal discovery with an equal-variance assumption. *Biometrika*, 106(4):973–980, 2019.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research (JMLR)*, 3(3):507–554, 2003.

Adrian Dobra and Alex Lenkoski. Copula Gaussian graphical models and their application to modeling functional disability data. *Ann. Appl. Stat.*, 5(2A):969–993, 2011.

Mathias Drton and Marloes H. Maathuis. Structure learning in graphical modeling. *Annual Review of Statistics and Its Application*, 4:365–393, 2017.

Mathias Drton and Michael D. Perlman. Model selection for Gaussian concentration graphs. *Biometrika*, 91(3):591–602, 2004.

Pan Du, Xiao Zhang, Chiang-Ching Huang, Nadereh Jafari, Warren A Kibbe, Lifang Hou, and Simon M Lin. Comparison of beta-value and m-value methods for quantifying methylation levels by microarray analysis. *BMC bioinformatics*, 11(1):587, 2010.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96(456):1348–1360, 2001.

Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006.

Bernd Fellinghauer, Peter Bühlmann, Martin Ryffel, Michael von Rhein, and Jan D. Reinhardt. Stable graphical model estimation with random forests for discrete, continuous, and mixed variables. *Comput. Statist. Data Anal.*, 64:132–152, 2013.

Rina Foygel and Mathias Drton. Extended Bayesian information criteria for Gaussian graphical models. In *Advances in Neural Information Processing Systems*, pages 604–612, 2010.

Jerome Friedman, Trevor Hastie, Holger Höfling, and Robert Tibshirani. Pathwise coordinate optimization. *Ann. Appl. Stat.*, 1(2):302–332, 2007.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Asish Ghoshal and Jean Honorio. Learning identifiable Gaussian Bayesian networks in polynomial time and sample complexity. In *Advances in Neural Information Processing Systems*, pages 6457–6466, 2017.

Jing-Dong J. Han, Nicolas Bertin, Tong Hao, Debra S. Goldberg, Gabriel F. Berriz, Lan V. Zhang, Denis Dupuy, Albertha JM. Walhout, Michael E. Cusick, Frederick P. Roth, et al. Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature*, 430(6995):88, 2004.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *J. Mach. Learn. Res.*, 6:695–709, 2005.

Aapo Hyvärinen. Some extensions of score matching. *Comput. Statist. Data Anal.*, 51(5): 2499–2512, 2007.

David Inouye, Pradeep Ravikumar, and Inderjit Dhillon. Square root graphical models: Multivariate generalizations of univariate exponential families that permit positive de-

pendencies. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2445–2453, 2016.

Hawoong Jeong, Sean P. Mason, A-L. Barabási, and Zoltan N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41, 2001.

Markus Kalisch and Peter Bühlmann. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8(Mar):613–636, 2007.

Kshitij Khare, Sang-Yun Oh, and Bala Rajaratnam. A convex pseudolikelihood framework for high dimensional partial correlation estimation with convergence guarantees. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 77(4):803–825, 2015.

Diederik P. Kingma and Yann L. Cun. Regularized estimation of image statistics by score matching. In *Advances in Neural Information Processing Systems*, pages 1126–1134, 2010.

Urs Köster and Aapo Hyvärinen. A two-layer ICA-like model estimated by score matching. *Artificial Neural Networks–ICANN 2007*, pages 798–807, 2007.

Steffen L. Lauritzen. *Graphical Models*, volume 17 of *Oxford Statistical Science Series*. The Clarendon Press, Oxford University Press, New York, 1996.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivariate Anal.*, 88(2):365–411, 2004.

Lina Lin, Mathias Drton, and Ali Shojaie. Estimation of high-dimensional graphical models using regularized score matching. *Electron. J. Stat.*, 10(1):806–854, 2016.

Han Liu, John Lafferty, and Larry Wasserman. The nonparanormal: semiparametric estimation of high dimensional undirected graphs. *J. Mach. Learn. Res.*, 10:2295–2328, 2009.

Han Liu, Fang Han, Ming Yuan, John Lafferty, and Larry Wasserman. High-dimensional semiparametric Gaussian copula graphical models. *Ann. Statist.*, 40(4):2293–2326, 2012.

Song Liu and Takafumi Kanamori. Estimating density models with complex truncation boundaries. *arXiv preprint arXiv:1910.03834*, 2019.

Weidong Liu and Xi Luo. Fast and adaptive sparse precision matrix estimation in high dimensions. *J. Multivariate Anal.*, 135:153–162, 2015.

Jun Luo, Thomas Dunn, Charles Ewing, Jurga Sauvageot, Yidong Chen, Jeffrey Trent, and William Isaacs. Gene expression signature of benign prostatic hyperplasia revealed by cDNA microarray analysis. *The Prostate*, 51(3):189–200, 2002.

Siwei Lyu. Interpretation and generalization of score matching. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 359–366. AUAI Press, 2009.

Marloes Maathuis, Mathias Drton, Steffen Lauritzen, and Martin Wainwright. *Handbook of graphical models*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, Boca Raton, FL, 2019. ISBN 978-1-4987-8862-5.

Andrew McDavid, Raphael Gottardo, Noah Simon, and Mathias Drton. Graphical models for zero-inflated single cell gene expression. *The Annals of Applied Statistics*, 13(2):848–873, 2019.

Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34(3):1436–1462, 2006.

Giuseppe Morgia, Mario Falsaperla, Grazia Malaponte, Massimo Madonia, Manuela Indelicato, Salvatore Travali, and Maria Clorinda Mazzarino. Matrix metalloproteinases as diagnostic (MMP-13) and prognostic (MMP-2, MMP-9) markers of prostate cancer. *Urological Research*, 33(1):44–50, 2005.

Preetam Nandy, Alain Hauser, Marloes H. Maathuis, et al. High-dimensional consistency in score-based and hybrid structure learning. *The Annals of Statistics*, 46(6A):3151–3183, 2018.

Shintaro Narita, Alan So, Susan Ettinger, Norihiro Hayashi, Mototsugu Muramaki, Ladan Fazli, Youngsoo Kim, and Martin E Gleave. GLI2 knockdown using an antisense oligonucleotide induces apoptosis and chemosensitizes cells to paclitaxel in androgen-independent prostate cancer. *Clinical Cancer Research*, 14(18):5769–5777, 2008.

Masashi Okamoto. Distinctness of the eigenvalues of a quadratic form in a multivariate sample. *The Annals of Statistics*, 1:763–765, 1973.

Matthew Parry. Extensive scoring rules. *Electron. J. Stat.*, 10(1):1098–1108, 2016.

Matthew Parry, A. Philip Dawid, and Steffen Lauritzen. Proper local scoring rules. *Ann. Statist.*, 40(1):561–592, 2012.

Jie Peng, Nengfeng Zhou, and Ji Zhu. Partial correlation estimation by joint sparse regression models. *J. Amer. Statist. Assoc.*, 104(486):735–746, 2009.

Jonas Peters and Peter Bühlmann. Identifiability of Gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2013.

Jonas Peters, Joris M. Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *The Journal of Machine Learning Research*, 15(1): 2009–2053, 2014.

Pradeep Ravikumar, Martin J. Wainwright, and John D. Lafferty. High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Ann. Statist.*, 38(3):1287–1319, 2010.

Pradeep Ravikumar, Martin J. Wainwright, Garvesh Raskutti, and Bin Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Stat.*, 5:935–980, 2011.

Rajen D. Shah and Richard J. Samworth. Variable selection with error control: another

look at stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(1):55–80, 2013.

Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7 (Oct):2003–2030, 2006.

Tomi Silander and Petri Myllymäki. A simple approach for finding the globally optimal bayesian network structure. In *Conference on Uncertainty in Artificial Intelligence*, pages 445–452, 2006.

Peter Spirtes, Clark Glymour, and Richard Scheines. Causation, prediction, and search. adaptive computation and machine learning, 2000.

Mona Stearns and Mark E. Stearns. Evidence for increased activated metalloproteinase 2 (MMP-2a) expression associated with human prostate cancer progression. *Oncology Research Featuring Preclinical and Clinical Cancer Therapeutics*, 8(2):69–75, 1996.

G. M. Tallis. The moment generating function of the truncated multi-normal distribution. *J. Roy. Statist. Soc. Ser. B*, 23:223–229, 1961.

Saravanan Thiyagarajan, Neehar Bhatia, Shannon Reagan-Shaw, Diana Cozma, Andrei Thomas-Tikhonenko, Nihal Ahmad, and Vladimir S Spiegelman. Role of GLI2 transcription factor in growth and tumorigenicity of prostate cells. *Cancer Research*, 67(22): 10642–10646, 2007.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.

Robert Tibshirani, Jacob Bien, Jerome Friedman, Trevor Hastie, Noah Simon, Jonathan Taylor, and Ryan J. Tibshirani. Strong rules for discarding predictors in lasso-type problems. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 74(2):245–266, 2012.

Ryan J. Tibshirani. The lasso problem and uniqueness. *Electron. J. Stat.*, 7:1456–1490, 2013.

Dominique Trudel, Yves Fradet, François Meyer, François Harel, and Bernard Têtu. Significance of MMP-2 expression in prostate cancer. *Cancer Research*, 63(23):8511–8515, 2003.

Hannu Väliaho. Criteria for copositive matrices. *Linear Algebra Appl.*, 81:19–34, 1986.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing*, pages 210–268. Cambridge Univ. Press, Cambridge, 2012.

Arend Voorman, Ali Shojaie, and Daniela Witten. Graph estimation with joint additive models. *Biometrika*, 101(1):85–101, 2014.

Martin J Wainwright. *High-dimensional statistics: A Non-Asymptotic Viewpoint*, volume 48. Cambridge University Press, 2019.

Y. Samuel Wang and Mathias Drton. High-dimensional causal discovery under non-Gaussianity. *Biometrika*, 107(1):41–59, 2020.

Susan R Wilson, Sandra Gallagher, Kate Warpeha, and Susan J Hawthorne. Amplification of MMP-2 and MMP-9 production by prostate cancer cell lines via activation of protease-activated receptors. *The Prostate*, 60(2):168–174, 2004.

Tiancheng Xie, Binbin Dong, Yangye Yan, Guanghui Hu, and Yunfei Xu. Association between MMP-2 expression and prostate cancer: A meta-analysis. *Biomedical Reports*, 4(2): 241–245, 2016.

Eunho Yang, Pradeep Ravikumar, Genevera I. Allen, and Zhandong Liu. Graphical models via univariate exponential family distributions. *J. Mach. Learn. Res.*, 16:3813–3847, 2015.

Ming Yu, Mladen Kolar, and Varun Gupta. Statistical inference for pairwise graphical models using score matching. In *Advances in Neural Information Processing Systems*, pages 2829–2837, 2016.

Ming Yu, Varun Gupta, and Mladen Kolar. Simultaneous inference for pairwise graphical models with generalized score matching. *arXiv preprint arXiv:1905.06261*, 2019a.

Shiqing Yu, Mathias Drton, and Ali Shojaie. Graphical models for non-negative data using generalized score matching. In *International Conference on Artificial Intelligence and Statistics*, pages 1781–1790, 2018.

Shiqing Yu, Mathias Drton, and Ali Shojaie. Generalized score matching for non-negative data. *Journal of Machine Learning Research*, 20(76):1–70, 2019b.

Shiqing Yu, Mathias Drton, and Ali Shojaie. Directed graphical models and causal discovery for zero-inflated data. *arXiv preprint arXiv:2004.04150*, 2020.

Ming Yuan and Yi Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

Małgorzata Żak-Szatkowska and Małgorzata Bogdan. Modified versions of the Bayesian information criterion for sparse generalized linear models. *Comput. Statist. Data Anal.*, 55(11):2908–2924, 2011.

Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, 38(2):894–942, 2010.

Teng Zhang and Hui Zou. Sparse precision matrix estimation via lasso penalized D-trace loss. *Biometrika*, 101(1):103–120, 2014.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(2):301–320, 2005.

Appendix A

APPENDICES TO CHAPTER 2

A.1 Proofs

A.1.1 Proof of Theorem 2

The following integration by parts lemma is used in the proof of Theorem 2.

Lemma 34. *Let $f, g : \mathbb{R}_+ \to \mathbb{R}$ be functions that are absolutely continuous in every bounded sub-interval of $\mathbb{R}_+$. Then*

$$\lim_{x \nearrow +\infty} f(x)g(x) - \lim_{x \searrow 0^+} f(x)g(x) = \int_0^\infty f(\boldsymbol{x}) \frac{\mathrm{d}g(x)}{\mathrm{d}x} \, \mathrm{d}x + \int_0^\infty g(\boldsymbol{x}) \frac{\mathrm{d}f(x)}{\mathrm{d}x} \, \mathrm{d}x.$$

Proof. This is an analog of Lemma 4 from Hyvärinen (2005) that can be proved by integrating the partial derivatives, and follows from the fundamental theorem of calculus for absolutely continuous functions and the product rule. In particular, we work on integrals in bounded $[0, c]$, where the product of two absolutely continuous functions in a bounded interval is again absolutely continuous, and the result is then obtained by letting $c \nearrow +\infty$. □

Proof of Theorem 2. Recall the following assumptions from Section 2.2.2:

(A1) $p_0(\boldsymbol{x}) h_j(x_j) \partial_j \log p(\boldsymbol{x}) \Big|_{x_j \searrow 0^+}^{x_j \nearrow +\infty} = 0 \,, \;\; \forall \boldsymbol{x}_{-j} \in \mathbb{R}_+^{m-1}, \quad \forall p \in \mathcal{P}_+;$

(A2) $\mathbb{E}_{p_0} \|\nabla \log p(\boldsymbol{X}) \circ \boldsymbol{h}^{1/2}(\boldsymbol{X})\|_2^2 < +\infty, \;\; \mathbb{E}_{p_0} \|(\nabla \log p(\boldsymbol{X}) \circ \boldsymbol{h}(\boldsymbol{X}))'\|_1 < +\infty, \;\; \forall p \in \mathcal{P}_+.$

Without explicitly writing the domains $\mathbb{R}_+$ or $\mathbb{R}_+^m$ in all integrals, by (2.4) we have

$$\begin{aligned} J_{\boldsymbol{h}}(p) &= \frac{1}{2} \int p_0(\boldsymbol{x}) \left[\|\nabla \log p(\boldsymbol{x}) \circ \boldsymbol{h}^{1/2}(\boldsymbol{x})\|_2^2 \right. \\ &\qquad \left. -2(\nabla \log p(\boldsymbol{x}) \circ \boldsymbol{h}^{1/2}(\boldsymbol{x}))^\top (\nabla \log p_0(\boldsymbol{x}) \circ \boldsymbol{h}^{1/2}(\boldsymbol{x})) + \|\nabla \log p_0(\boldsymbol{x}) \circ \boldsymbol{h}^{1/2}(\boldsymbol{x})\|_2^2 \right] \mathrm{d}\boldsymbol{x} \\ &= \underbrace{\frac{1}{2} \int p_0(\boldsymbol{x}) \sum_{j=1}^m h_j(x_j) \left(\frac{\partial \log p(\boldsymbol{x})}{\partial x_j} \right)^2 \mathrm{d}\boldsymbol{x}}_{\equiv A} + \underbrace{\frac{1}{2} \int p_0(\boldsymbol{x}) \sum_{j=1}^m h_j(x_j) \left(\frac{\partial \log p_0(\boldsymbol{x})}{\partial x_j} \right)^2 \mathrm{d}\boldsymbol{x}}_{\equiv C} \\ &\qquad \underbrace{- \int p_0(\boldsymbol{x}) \sum_{j=1}^m h_j(x_j) \frac{\partial \log p(\boldsymbol{x})}{\partial x_j} \frac{\partial \log p_0(\boldsymbol{x})}{\partial x_j} \, \mathrm{d}\boldsymbol{x}}_{\equiv B}, \end{aligned}$$

where A will simply appear in the final display as is, C is a constant as it only involves the true pdf p_0, and we wish to simplify B by integration by parts. We can split the integral into these three parts since A and C are assumed finite in the first part of (A2), and the integrand in B is integrable since $|2ab| \leq a^2 + b^2$. Thus, by linearity and Fubini's theorem, we can write

$$\begin{aligned} B &= -\sum_{j=1}^{m} \int p_0(\boldsymbol{x}) h_j(x_j) \frac{\partial \log p(\boldsymbol{x})}{\partial x_j} \frac{\partial \log p_0(\boldsymbol{x})}{\partial x_j} \, \mathrm{d}\boldsymbol{x} \\ &= -\sum_{j=1}^{m} \int \left[\int p_0(\boldsymbol{x}) h_j(x_j) \frac{\partial \log p(\boldsymbol{x})}{\partial x_j} \frac{\partial \log p_0(\boldsymbol{x})}{\partial x_j} \, \mathrm{d}x_j \right] \mathrm{d}\boldsymbol{x}_{-j}. \end{aligned}$$

By the fact that $\frac{\partial \log p_0(\boldsymbol{x})}{\partial x_j} = \frac{1}{p_0(\boldsymbol{x})} \frac{\partial p_0(\boldsymbol{x})}{\partial x_j}$, this can be simplified to

$$B = -\sum_{j=1}^{m} \int \left[\int \frac{\partial p_0(\boldsymbol{x})}{\partial x_j} h_j(x_j) \frac{\partial \log p(\boldsymbol{x})}{\partial x_j} \, \mathrm{d}x_j \right] \mathrm{d}\boldsymbol{x}_{-j}.$$

But, we assume p_0 and p are twice continuously differentiable, for every $j = 1, \ldots, m$ and fixed $\boldsymbol{x}_{-j} \in \mathbb{R}_+^{m-1}$. Hence, in every bounded sub-interval of $\mathbb{R}_+$, $p_0(\boldsymbol{x}_{-j}; x_j)$ is an absolutely continuous function of x_j, $\partial_j \log p(\boldsymbol{x}_{-j}, x_j) = \partial_j p(\boldsymbol{x}_{-j}, x_j)/p(\boldsymbol{x}_{-j}, x_j)$ is a continuously differentiable (and hence absolutely continuous) function of x_j by the quotient rule. Thus $h_j(x_j)\partial_j \log p(\boldsymbol{x}_{-j}; x_j)$ is also absolutely continuous by the absolute continuity assumption on h_j. Then, by Lemma 34, where we take $f \equiv p_0(\boldsymbol{x}_{-j}; x_j)$ and $g \equiv h_j(x_j)\partial_j \log p(\boldsymbol{x}_{-j}; x_j)$ as functions of x_j, followed by assumption (A1),

$$\begin{aligned} B = -\sum_{j=1}^{m} \int \Bigg[& \lim_{a \nearrow +\infty, b \searrow 0^+} \left[p_0(\boldsymbol{x}_{-j}; a) h_j(a) \partial_j \log p(\boldsymbol{x}_{-j}, a) - p_0(\boldsymbol{x}_{-j}; b) h_j(b) \partial_j \log p(\boldsymbol{x}_{-j}, b) \right] \\ & - \int p_0(\boldsymbol{x}) \frac{\partial \left(h_j(x_j) \partial_j \log p(\boldsymbol{x}) \right)}{\partial x_j} \, \mathrm{d}x_j \Bigg] \mathrm{d}\boldsymbol{x}_{-j} \\ = \sum_{j=1}^{m} \int & \left[\int p_0(\boldsymbol{x}) \frac{\partial (h_j(x_j) \partial_j \log p(\boldsymbol{x}))}{\partial x_j} \, \mathrm{d}x_j \right] \mathrm{d}\boldsymbol{x}_{-j}. \end{aligned}$$

Justified by the second half of (A2), by Fubini-Tonelli and linearity again

$$B = \sum_{j=1}^{m} \int p_0(\boldsymbol{x}) \frac{\partial (h_j(x_j) \partial_j \log p(\boldsymbol{x}))}{\partial x_j} \, \mathrm{d}\boldsymbol{x},$$

$$= \sum_{j=1}^{m} \int h_j'(x_j) \frac{\partial \log p(\boldsymbol{x})}{\partial x_j} p_0(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} + \sum_{j=1}^{m} \int h_j(x_j) \frac{\partial^2 \log p(\boldsymbol{x})}{\partial x_j^2} p_0(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}.$$

Thus,

$$\begin{aligned}
&J_{\boldsymbol{h}}(p) \\
&= B + A + C \\
&= \int_{\mathbb{R}_+^m} p_0(\boldsymbol{x}) \sum_{j=1}^{m} \left[h_j'(x_j) \frac{\partial \log p(\boldsymbol{x})}{\partial x_j} + h_j(x_j) \frac{\partial^2 \log p(\boldsymbol{x})}{\partial x_j^2} + \frac{1}{2} h_j(x_j) \left(\frac{\partial \log p(\boldsymbol{x})}{\partial x_j} \right)^2 \right] \mathrm{d}\boldsymbol{x} + C,
\end{aligned}$$

where C is a constant that does not depend on p. □

A.1.2 Proof of Theorems and Examples in Section 2.3

Proof of Theorem 3. For exponential families and under the assumptions, the empirical loss $\hat{J}_{\boldsymbol{h}}(p_{\boldsymbol{\theta}})$ in (2.6) becomes (up to an additive constant)

$$\begin{aligned}
&\hat{J}_{\boldsymbol{h}}(p_{\boldsymbol{\theta}}) \\
&= \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} \left[h_j'(X_j^{(i)}) \frac{\partial \log p_{\boldsymbol{\theta}}(\boldsymbol{X}^{(i)})}{\partial X_j^{(i)}} + h_j(X_j^{(i)}) \frac{\partial^2 \log p_{\boldsymbol{\theta}}(\boldsymbol{X}^{(i)})}{\partial (X_j^{(i)})^2} \right. \\
&\qquad \left. + \frac{1}{2} h_j(X_j^{(i)}) \left(\frac{\partial \log p_{\boldsymbol{\theta}}(\boldsymbol{X}^{(i)})}{\partial X_j^{(i)}} \right)^2 \right] \\
&= \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} \left[h_j'(X_j^{(i)}) (\boldsymbol{\theta}^\top \boldsymbol{t}_j'(\boldsymbol{X}^{(i)}) + b_j'(\boldsymbol{X}^{(i)})) + h_j(X_j^{(i)}) (\boldsymbol{\theta}^\top \boldsymbol{t}_j''(\boldsymbol{X}^{(i)}) + b_j''(\boldsymbol{X}^{(i)})) \right. \\
&\qquad \left. + \frac{1}{2} h_j(X_j^{(i)}) (\boldsymbol{\theta}^\top \boldsymbol{t}_j'(\boldsymbol{X}^{(i)}) + b_j'(\boldsymbol{X}^{(i)}))^2 \right] \\
&= \frac{1}{n} \left\{ \frac{1}{2} \boldsymbol{\theta}^\top \left[\sum_{i=1}^{n} \sum_{j=1}^{m} h_j(X_j^{(i)}) \boldsymbol{t}_j'(\boldsymbol{X}^{(i)}) \boldsymbol{t}_j'(\boldsymbol{X}^{(i)})^\top \right] \boldsymbol{\theta} + \right. \\
&\quad \left. \left[\sum_{i=1}^{n} \sum_{j=1}^{m} h_j(X_j^{(i)}) b_j'(\boldsymbol{X}^{(i)}) \boldsymbol{t}_j'(\boldsymbol{X}^{(i)}) + h_j(X_j^{(i)}) \boldsymbol{t}_j''(\boldsymbol{X}^{(i)}) + h_j'(X_j^{(i)}) \boldsymbol{t}_j'(\boldsymbol{X}^{(i)}) \right]^\top \boldsymbol{\theta} \right\} + \text{const},
\end{aligned}$$

which is quadratic in $\boldsymbol{\theta}$. Let

$$\boldsymbol{\Gamma}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} h_j(X_j^{(i)})\boldsymbol{t}_j'(\boldsymbol{X}^{(i)})\boldsymbol{t}_j'(\boldsymbol{X}^{(i)})^\top, \tag{A.1}$$

$$\boldsymbol{g}(\mathbf{x}) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\left[h_j(X_j^{(i)})b_j'(\boldsymbol{X}^{(i)})\boldsymbol{t}_j'(\boldsymbol{X}^{(i)}) + h_j(X_j^{(i)})\boldsymbol{t}_j''(\boldsymbol{X}^{(i)}) + h_j'(X_j^{(i)})\boldsymbol{t}_j'(\boldsymbol{X}_i)\right]. \tag{A.2}$$

Then we can write $\hat{J}_{\boldsymbol{h}}(p_{\boldsymbol{\theta}}) = \frac{1}{2}\boldsymbol{\theta}^\top\boldsymbol{\Gamma}(\mathbf{x})\boldsymbol{\theta} - \boldsymbol{g}(\mathbf{x})^\top\boldsymbol{\theta} + \text{const.}$ □

Proof of Theorem 4. By Theorem 3, $\hat{J}_{\boldsymbol{h}}(p_{\boldsymbol{\theta}}) = \frac{1}{2}\boldsymbol{\theta}^\top\boldsymbol{\Gamma}\boldsymbol{\theta} - \boldsymbol{g}^\top\boldsymbol{\theta} + \text{const.}$ The minimizer of $\hat{J}_{\boldsymbol{h}}(p_{\boldsymbol{\theta}})$ is thus available in the unique closed form $\hat{\boldsymbol{\theta}} \equiv \boldsymbol{\Gamma}(\mathbf{x})^{-1}\boldsymbol{g}(\mathbf{x})$ as long as $\boldsymbol{\Gamma}$ is invertible (C1). Since $\boldsymbol{\Gamma}$ and $\boldsymbol{g}$ are sample averages, the weak law of large numbers yields that $\boldsymbol{\Gamma} \to_p \mathbb{E}_{p_0}\boldsymbol{\Gamma} \equiv \boldsymbol{\Gamma}_0$ and $\boldsymbol{g} \to_p \mathbb{E}_{p_0}\boldsymbol{g} \equiv \boldsymbol{g}_0$, where existence of $\boldsymbol{\Gamma}_0$ and $\boldsymbol{g}_0$ is assumed in (C2). Since $J_{\boldsymbol{h}}(p_{\boldsymbol{\theta}}) = \mathbb{E}[\hat{J}_{\boldsymbol{h}}(p_{\boldsymbol{\theta}})] = \mathbb{E}[\frac{1}{2}\boldsymbol{\theta}^\top\boldsymbol{\Gamma}(\mathbf{x})\boldsymbol{\theta} - \boldsymbol{g}(\mathbf{x})^\top\boldsymbol{\theta}] = \frac{1}{2}\boldsymbol{\theta}^\top\boldsymbol{\Gamma}_0\boldsymbol{\theta} - \boldsymbol{g}_0\boldsymbol{\theta}$ and we know $\boldsymbol{\theta}_0$ minimizes $J_{\boldsymbol{h}}(p_{\boldsymbol{\theta}})$ by definition, by the first-order condition we must have $\boldsymbol{\Gamma}_0\boldsymbol{\theta}_0 = \boldsymbol{g}_0$. Then by the Lindeberg-Lévy central limit theorem,

$$\sqrt{n}(\boldsymbol{g}(\mathbf{x}) - \boldsymbol{\Gamma}(\mathbf{x})\boldsymbol{\theta}_0) \to_d \mathcal{N}_m(\mathbf{0}, \boldsymbol{\Sigma}_0),$$

where $\boldsymbol{\Sigma}_0 \equiv \mathbb{E}_{p_0}[(\boldsymbol{\Gamma}(\mathbf{x})\boldsymbol{\theta}_0 - \boldsymbol{g}(\mathbf{x}))(\boldsymbol{\Gamma}(\mathbf{x})\boldsymbol{\theta}_0 - \boldsymbol{g}(\mathbf{x}))^\top]$, as long as $\boldsymbol{\Sigma}_0$ exists (C2). Thus, by Slutsky's theorem,

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \equiv \sqrt{n}(\boldsymbol{\Gamma}(\mathbf{x})^{-1}(\boldsymbol{g}(\mathbf{x}) - \boldsymbol{\Gamma}(\mathbf{x})\boldsymbol{\theta}_0)) \to_d \mathcal{N}_r(\mathbf{0}, \boldsymbol{\Gamma}_0^{-1}\boldsymbol{\Sigma}\boldsymbol{\Gamma}_0^{-1}),$$

as long as $\boldsymbol{\Gamma}_0$ is invertible (C2).

For the second half of the theorem, (C2) $\mathbb{E}_{p_0}\boldsymbol{\Gamma}(\mathbf{x}) < \infty$ and $\mathbb{E}_{p_0}\boldsymbol{g}(\mathbf{x}) < \infty$ implies $\mathbb{E}_{p_0}|\boldsymbol{\Gamma}(\mathbf{x})| < \infty$ and $\mathbb{E}_{p_0}|\boldsymbol{g}(\mathbf{x})| < \infty$, so by strong law of large numbers (and a union bound on at most k^2 null sets)

$$\boldsymbol{\Gamma}(\mathbf{x}) \to_{\text{a.s.}} \boldsymbol{\Gamma}_0, \quad \boldsymbol{g}(\mathbf{x}) \to_{\text{a.s.}} \boldsymbol{g}_0.$$

Then outside a null set,

$$\hat{\boldsymbol{\theta}} \equiv \boldsymbol{\Gamma}(\mathbf{x})^{-1}\boldsymbol{g}(\mathbf{x}) \to_{\text{a.s.}} \boldsymbol{\Gamma}_0^{-1}\boldsymbol{g}_0 = \boldsymbol{\theta}_0.$$

□

Proof for Example 3.1. We choose to estimate $\theta \equiv \mu/\sigma^2$. Then by (2.9) and (2.10),

$$\begin{aligned}
\hat{\mu}_h &= \sigma^2\hat{\theta} \equiv \sigma^2\Gamma(\boldsymbol{x})^{-1}g(\boldsymbol{x}) \\
&= -\sigma^2\left[\sum_{i=1}^n h(X_i)t'(X_i)^2\right]^{-1}\left[\sum_{i=1}^n h(X_i)b'(X_i)t'(X_i) + h(X_i)t''(X_i) + h'(X_i)t'(X_i)\right] \\
&= -\sigma^2\left[\sum_{i=1}^n h(X_i)\right]^{-1}\left[\sum_{i=1}^n -h(X_i)\frac{X_i}{\sigma^2} + h'(X_i)\right].
\end{aligned}$$

By Theorem 4,

$$\begin{aligned}
\sqrt{n}(\hat{\mu}_h - \mu_0) &\to_d \mathcal{N}\left(0, \frac{\sigma^4\mathbb{E}_0\left[-h(X)\frac{X-\mu_0}{\sigma^2} + h'(X)\right]^2}{\mathbb{E}_0^2[h(X)]}\right) \\
&\sim \mathcal{N}\left(0, \frac{\mathbb{E}_0\left[-h(X)(X-\mu_0) + \sigma^2 h'(X)\right]^2}{\mathbb{E}_0^2[h(X)]}\right).
\end{aligned}$$

By integration by parts, (suppressing the dependence of p_{μ_0} on μ_0)

$$\begin{aligned}
&\mathbb{E}_0[h(X)h'(X)(X-\mu_0)] \\
&= \int_0^\infty h'(x)h(x)(x-\mu_0)p(x)\,\mathrm{d}x \\
&= \int_0^\infty h(x)(x-\mu_0)p(x)\,\mathrm{d}h(x) \\
&= h^2(x)(x-\mu_0)p(x)\big|_0^\infty - \int h(x)\,\mathrm{d}h(x)(x-\mu_0)p(x) \\
&= -\int h^2(x)p(x)\,\mathrm{d}x - \int h(x)h'(x)(x-\mu_0)p(x)\,\mathrm{d}x + \int h^2(x)\frac{(x-\mu_0)^2}{\sigma^2}p(x)\,\mathrm{d}x,
\end{aligned}$$

where the last step follows from the assumptions $\lim_{x\searrow 0^+} h(x) = 0$ and $\lim_{x\nearrow+\infty} h^2(x)(x-\mu_0)p_{\mu_0}(x) = 0$. So

$$\mathbb{E}_0[h(X)h'(X)(X-\mu_0)] = \frac{\mathbb{E}[h^2(X)((X-\mu_0)^2/\sigma^2 - 1)]}{2}. \tag{A.3}$$

The asymptotic variance is thus

$$\begin{aligned}
&\frac{\mathbb{E}_0\left[-h(X)(X-\mu_0) + \sigma^2 h'(X)\right]^2}{\mathbb{E}_0^2[h(X)]} \\
&= \frac{\mathbb{E}_0\left[h^2(X)(X-\mu_0)^2 - 2\sigma^2 h^2(X)\left((X-\mu_0)^2/\sigma^2 - 1\right)/2 + \sigma^4 h'^2(X)\right]}{\mathbb{E}_0^2[h(X)]}
\end{aligned}$$

$$= \frac{\mathbb{E}_0[\sigma^2 h^2(X) + \sigma^4 h'^2(X)]}{\mathbb{E}_0^2[h(X)]}.$$

The Cramér-Rao lower bound follows from taking the second derivative of $\log p_{\mu_0}$ with respect to μ_0. □

Proof for Example 3.2. We estimate $\theta \equiv 1/\sigma^2$. By (2.9) and (2.10),

$$\begin{aligned}
\hat{\theta} &\equiv \Gamma(\boldsymbol{x})^{-1} g(\boldsymbol{x}) \\
&= -\left[\sum_{i=1}^n h(X_i) t'(X_i)^2\right]^{-1} \left[\sum_{i=1}^n h(X_i) b'(X_i) t'(X_i) + h(X_i) t''(X_i) + h'(X_i) t'(X_i)\right] \\
&= \left[\sum_{i=1}^n h(X_i)(X_i - \mu)^2\right]^{-1} \left[\sum_{i=1}^n h(X_i) + h'(X_i)(X_i - \mu)\right].
\end{aligned}$$

By Theorem 4, $\sqrt{n}(\hat{\theta} - \theta) \to_d \mathcal{N}(0, \varsigma^2)$, where

$$\begin{aligned}
\varsigma^2 &\equiv \frac{\mathbb{E}_0\left[h(X)((X-\mu)^2/\sigma_0^2 - 1) - h'(X)(X-\mu)\right]^2}{\mathbb{E}_0^2[h(X)(X-\mu)^2]} \\
&= \frac{1}{\mathbb{E}_0^2[h(X)(X-\mu)^2]} \Big(\mathbb{E}_0[h^2(X)(X-\mu)^4/\sigma_0^4 - 2h^2(X)(X-\mu)^2/\sigma_0^2 + h^2(X) \\
&\qquad + h'^2(X)(X-\mu)^2 - 2h(X)h'(X)(X-\mu)^3/\sigma_0^2 + 2h(X)h'(X)(X-\mu) \Big).
\end{aligned}$$

By integration by parts, (suppressing the dependence of $p_{\sigma_0^2}$ on σ_0^2)

$$\begin{aligned}
&\mathbb{E}_0[h(X)h'(X)(X-\mu)^3] \\
&= \int_0^\infty h'(x)h(x)(x-\mu)^3 p(x)\,\mathrm{d}x \\
&= \int_0^\infty h(x)(x-\mu)^3 p(x)\,\mathrm{d}h(x) \\
&= h^2(x)(x-\mu)^3 p(x)\big|_0^\infty - \int h(x)\,\mathrm{d}h(x)(x-\mu)^3 p(x) \\
&= -\int h(x)h'(x)(x-\mu)^3 p(x)\,\mathrm{d}x - 3\int h^2(x)(x-\mu)^2 p(x)\,\mathrm{d}x + \int h^2(x)\frac{(x-\mu)^4}{\sigma_0^2} p(x)\,\mathrm{d}x,
\end{aligned}$$

where the last step follows from the assumptions $\lim_{x \searrow 0^+} h(x) = 0$ and $\lim_{x \nearrow +\infty} h^2(x)(x-\mu)^3 p_{\sigma_0^2}(x) = 0$. Combining this with (A.3) we get

$$\sqrt{n}(\hat{\theta} - \theta) \to_d \mathcal{N}(0, \varsigma^2) \sim \mathcal{N}\left(0, \frac{2\mathbb{E}_0[h^2(X)(X-\mu)^2/\sigma_0^2] + \mathbb{E}_0[h'^2(X-\mu)^2]}{\mathbb{E}_0^2[h(X)(X-\mu)^2]}\right),$$

and so by the delta method, for $\hat{\sigma}_k^2 \equiv \hat{\theta}^{-1}$,

$$\sqrt{n}(\hat{\sigma}_h^2 - \sigma_0^2) \to_d \mathcal{N}\left(0, \frac{2\sigma_0^6 \mathbb{E}_0[h^2(X)(X-\mu)^2] + \sigma_0^8 \mathbb{E}_0[h'^2(X-\mu)^2]}{\mathbb{E}_0^2[h(X)(X-\mu)^2]}\right).$$

The Cramér-Rao lower bound follows from taking the second derivative of $\log p_{\sigma_0^2}$ with respect to σ_0^2. □

A.1.3 Proof of Theorems in Section 2.5

Proof of Theorem 5.

Case $b \neq 0$: We use a strategy similar to that of Inouye et al. (2016). Let $\mathcal{V}_1 = \{\boldsymbol{v} : \|\boldsymbol{v}\|_1 = 1, \boldsymbol{v} \in \mathbb{R}_+^m\}$. Then by Fubini-Tonelli the normalizing constant is,

$$\begin{aligned}
&\int_{\mathbb{R}_+^m} \exp\left(\boldsymbol{\eta}^\top \frac{\boldsymbol{x}^b - \mathbf{1}_m}{b} - \frac{1}{2a}\boldsymbol{x}^{a\top}\mathbf{K}\boldsymbol{x}^a\right) \mathrm{d}\boldsymbol{x} \\
&= \int_{\mathcal{V}_1}\int_0^\infty \exp\left(\boldsymbol{\eta}^\top \frac{z^b\boldsymbol{v}^b - \mathbf{1}_m}{b} - \frac{1}{2a}z^{2a}\boldsymbol{v}^{a\top}\mathbf{K}\boldsymbol{v}^a\right) \mathrm{d}z\,\mathrm{d}\boldsymbol{v} \\
&\propto \int_{\mathcal{V}_1}\int_0^\infty \exp\Big(z^b(\boldsymbol{\eta}^\top\boldsymbol{v}^b)/b - z^{2a}(\boldsymbol{v}^{a\top}\mathbf{K}\boldsymbol{v}^a)/(2a)\Big) \mathrm{d}z\,\mathrm{d}\boldsymbol{v}.
\end{aligned}$$

Here $\mathcal{V}_1$ is compact and the inner integral, if finite, is continuous in $\boldsymbol{v}$. It thus suffices to show that the inner integral is finite at every single $\boldsymbol{v} \in \mathcal{V}_1$.

Fixing $\boldsymbol{v} \in \mathcal{V}_1$, write $A \equiv A(\boldsymbol{v}) \equiv \boldsymbol{v}^{a\top}\mathbf{K}\boldsymbol{v}^a/(2a)$ and $B \equiv B(\boldsymbol{v}) \equiv (\boldsymbol{\eta}^\top\boldsymbol{v}^b)/b$. We need to show that

$$N(A, B, a, b) \equiv \int_0^\infty \exp\left(-Az^{2a} + Bz^b\right) \mathrm{d}z < +\infty.$$

Recall that (CC1) $\boldsymbol{v}^\top\mathbf{K}\boldsymbol{v} > 0$ for all $\boldsymbol{v} \in \mathbb{R}_+^m \backslash \{\mathbf{0}\}$, so $A > 0$.

(i) Suppose $B \leq 0$. Then $N(A, B, a, b) \leq \int_0^\infty \exp(-Az^{2a})\,\mathrm{d}z = A^{-a/2}\Gamma(1 + 1/(2a))$, a finite constant since $A > 0$ and $a > 0$.

(ii) Suppose $B > 0$. We first want to bound $\exp(-Az^{2a} + Bz^b) \leq N_0 \exp(-Az^{2a}/2)$ by some finite constant $N_0 > 0$, so that $N(A, B, a, b) \leq N_0 \int_0^\infty \exp(-Az^{2a}/2)\,\mathrm{d}z$, a finite constant for $a > 0$. Thus, it remains to give conditions so that $\exp(-Az^{2a}/2 + Bz^b)$

is bounded by some finite constant N_0, which by continuity only requires a finite limit as $z \searrow 0$ and as $z \nearrow +\infty$. As $z \nearrow +\infty$, $Bz^b \nearrow +\infty$, while $-Az^{2a}/2 \searrow -\infty$. We thus need $b < 2a$ so that the sum of the two does not go to positive infinity. On the other hand, as $z \searrow 0$, $-Az^{2a}/2 \nearrow 0$, so we need $b > 0$, otherwise $z^b \nearrow +\infty$. In conclusion, we require that $2a > b > 0$.

It thus suffices to require (CC1) and (CC2) $2a > b > 0$ to eliminate restrictions on B, and hence on $\boldsymbol{\eta}$. That is, $\boldsymbol{\eta}$ can take value in the entirety of $\mathbb{R}^m$.

Case $b = 0$: Again in (CC1) we assume $\boldsymbol{v}^\top \mathbf{K} \boldsymbol{v} > 0$ for all $\boldsymbol{v} \in \mathbb{R}^m_+ \backslash \{\mathbf{0}\}$. Since $\mathcal{V}_2 \equiv \{\boldsymbol{v} : \|\boldsymbol{v}\|_2 = 1, \boldsymbol{v} \in \mathbb{R}^m_+\}$ is compact and $\boldsymbol{v}^\top \mathbf{K} \boldsymbol{v}$ is continuous in $\boldsymbol{v}$ and strictly positive on $\mathcal{V}_2$, the image of $\mathcal{V}_2$ under $\boldsymbol{v}^\top \mathbf{K} \boldsymbol{v}$ is a compact subset of $(0, \infty)$, i.e. $N_{\mathbf{K}} \equiv \inf_{\boldsymbol{v} \in \mathbb{R}^m_+ \backslash \{\mathbf{0}\}} \boldsymbol{v}^\top \mathbf{K} \boldsymbol{v} / \boldsymbol{v}^\top \boldsymbol{v} \equiv \inf_{\boldsymbol{v} \in \mathcal{V}_2} \boldsymbol{v}^\top \mathbf{K} \boldsymbol{v} > 0$. We thus have

$$
\begin{aligned}
&\int_{\mathbb{R}^m_+} \exp\left(\boldsymbol{\eta}^\top \log(\boldsymbol{x}) - \frac{1}{2a} \boldsymbol{x}^{a\top} \mathbf{K} \boldsymbol{x}^a\right) \mathrm{d}\boldsymbol{x} \\
\leq &\int_{\mathbb{R}^m_+} \exp\left(\boldsymbol{\eta}^\top \log(\boldsymbol{x}) - \frac{N_{\mathbf{K}}}{2a} \boldsymbol{x}^{a\top} \boldsymbol{x}^a\right) \mathrm{d}\boldsymbol{x} \\
= &\prod_{j=1}^m \int_0^\infty \exp\left(\eta_j \log(x_j) - \frac{N_{\mathbf{K}}}{2a} x_j^{2a}\right) \mathrm{d}x_j \\
= &\prod_{j=1}^m \left[\Gamma\left(\frac{\eta_j + 1}{2a}\right) \frac{(N_{\mathbf{K}}/2a)^{-\frac{\eta_j+1}{2a}}}{2a}\right],
\end{aligned}
$$

where the integration follows by change of variable and requires $a > 0$. Assuming $a > 0$, the last quantity is finite if and only if $\boldsymbol{\eta} \succ -\mathbf{1}_m$, by definition of the gamma function.

In conclusion, given conditions (CC1) $\inf_{\boldsymbol{v} \in \mathbb{R}^m_+ \backslash \{\mathbf{0}\}} \boldsymbol{v}^\top \mathbf{K} \boldsymbol{v} > 0$, (CC2) $2a > b > 0$, and (CC3) $a > 0$, $b = 0$ and $\boldsymbol{\eta} \succ -\mathbf{1}_m$, the unnormalized density (2.14) has a finite normalizing constant when (CC1) and (CC2) both hold, or (CC1) and (CC3) both hold.

The centered settings, where the term involving $\boldsymbol{x}^b$ is excluded, can be considered as a special case of both (1) and (2) with $\boldsymbol{\eta} \equiv \mathbf{0}$, and thus (CC1) and $a > 0$ are sufficient. $\square$

Proof of Theorem 6. Recall assumptions (A1) and (A2):

(A1) $p_0(\boldsymbol{x})h_j(x_j)\partial_j \log p(\boldsymbol{x}) \Big|_{x_j \searrow 0^+}^{x_j \nearrow +\infty} = 0\,, \quad \forall \boldsymbol{x}_{-j} \in \mathbb{R}_+^{m-1}, \quad \forall p \in \mathcal{P}_+;$

(A2) $\mathbb{E}_{p_0}\|\nabla \log p(\boldsymbol{X}) \circ \boldsymbol{h}^{1/2}(\boldsymbol{X})\|_2^2 < +\infty,\ \mathbb{E}_{p_0}\|(\nabla \log p(\boldsymbol{X}) \circ \boldsymbol{h}(\boldsymbol{X}))'\|_1 < +\infty,\ \forall p \in \mathcal{P}_+.$

Let $\mathbf{K}_0$ and $\boldsymbol{\eta}_0$ be the true parameters so that $p_0 \in \mathcal{P}_+$, with $\mathcal{P}_+$ corresponding to a parameter space in which all parameters satisfy the conditions for a finite normalizing constant. We now give sufficient conditions for h to satisfy (A1) and (A2).

Conditions for (A1): Fix $j = 1, \ldots, m$ and $\boldsymbol{x}_{-j} \in \mathbb{R}_+^{m-1}$. We show that the conditions on h_j imply that the limits go to 0 as $x_j \nearrow +\infty$ and as $x_j \searrow 0^+$, which is stronger than (A1); in fact, from (A.4) below, the limits cannot go to a nonzero finite constant assuming an h with polynomial tail, since $a > 0$ and $B_1 \equiv \kappa_{0,jj} > 0$ for all j. Now,

$$
\begin{aligned}
& p_0(\boldsymbol{x})h_j(x_j)\partial_j \log p(\boldsymbol{x}) \\
\propto\ & h_j(x_j) \exp\left(-\frac{1}{2a}\boldsymbol{x}^{a\top}\mathbf{K}_0\boldsymbol{x}^a + \boldsymbol{\eta}_0^\top \frac{\boldsymbol{x}^b - \mathbf{1}_m}{b}\right) \partial_j \left(-\frac{1}{2a}\boldsymbol{x}^{a\top}\mathbf{K}\boldsymbol{x}^a + \boldsymbol{\eta}^\top \frac{\boldsymbol{x}^b - \mathbf{1}_m}{b}\right) \\
\propto\ & h_j(x_j) \exp\left(-\frac{1}{a}(\boldsymbol{k}_{0,j,-j}^\top \boldsymbol{x}_{-j}^a)x_j^a - \frac{\kappa_{0,jj}}{2a}x_j^{2a} + \eta_{0,j}\frac{x_j^b - 1}{b}\right) \times \\
& \qquad \left(-\boldsymbol{k}_{j,-j}^\top \boldsymbol{x}_{-j}^a x_j^{a-1} - \kappa_{jj}x_j^{2a-1} + \eta_j x_j^{b-1}\right) \\
\equiv\ & h_j(x_j) \exp\left(\frac{A_1 x_j^a}{a} + \frac{B_1 x_j^{2a}}{2a} + C_1\frac{x_j^b - 1}{b}\right)\left(A_2 x_j^{a-1} + B_2 x_j^{2a-1} + C_2 x_j^{b-1}\right), \qquad \text{(A.4)}
\end{aligned}
$$

where $A_1 \equiv -\boldsymbol{k}_{0,j,-j}^\top \boldsymbol{x}_{-j}^a$, $A_2 \equiv -\boldsymbol{k}_{j,-j}^\top \boldsymbol{x}_{-j}^a$, $B_1 \equiv -\kappa_{0,jj} < 0$ and $B_2 \equiv -\kappa_{jj} < 0$ by condition (CC1). Finally $C_1 \equiv \eta_{0,j}$, $C_2 \equiv \eta_j$.

(1) Let $x_j \nearrow +\infty$. If $b > 0$, since $2a > b > 0$ and $B_1 < 0$, the exponential term in (A.4) decreases to 0 exponentially and its reciprocal dominates any polynomial functions. Thus, the entire product goes to 0 if $h_j(x_j)$ grows no faster than polynomially as $x_j \nearrow +\infty$. If $b = 0$, the $C_1 \log x_j$ term is again dominated by $B_1 x_j^{2a}/(2a)$, and the same conclusion holds.

(2) Let $x_j \searrow 0$.

(i) Let $b > 0$. Then the exponential term in (A.4) goes to constant $\exp(-C_1/b)$, and we only need

$$\lim_{x_j \searrow 0^+} h_j(x_j)(A_2 x_j^{a-1} + B_2 x_j^{2a-1} + C_2 x_j^{b-1}) = 0. \tag{A.5}$$

- If $a > 1$ and $b > 1$, the second term in (A.5) is a polynomial with three terms having powers $\geq \min\{a-1, b-1\}$. The product goes to zero if and only if $h_j(x_j) = o(x_j^{\max\{1-a,1-b\}})$ as $x_j \searrow 0$. Note that this is satisfied by any h_j that has a finite right limit at 0.
- If $a = 1$ and $b \geq 1$, or $a \geq 1$ and $b = 1$, then the second term in (A.5) is a polynomial of non-negative power plus a potentially nonzero constant. A sufficient condition for (A.5) is thus $\lim_{x_j \searrow 0} h_j(x_j) = 0$.
- If $a < 1$ or $b < 1$, then the second part in (A.5) is a polynomial having terms with negative degree $\geq \min\{a-1, b-1\}$. To counteract this a sufficient condition is $h_j(x_j) = o(x_j^{\max\{1-a,1-b\}})$.

In conclusion, $\lim_{x_j \searrow 0^+} p_0(\boldsymbol{x}) h_j(x_j) \partial_j \log p(\boldsymbol{x}) = 0$ if and only if

$$\lim_{x_j \searrow 0^+} h_j(x_j)/x_j^{\max\{1-a,1-b\}} = 0.$$

(ii) Now assume $b = 0$. Then, (A.4) now becomes

$$h_j(x_j) \exp\left(A_1 x_j^a/a + B_1 x_j^{2a}/(2a) + C_1 \log x_j\right)\left(A_2 x_j^{a-1} + B_2 x_j^{2a-1} + C_2/x_j\right).$$

With $C_1 \log x_j$ dominating, the exponential part scales as $x_j^{C_1}$. We thus require

$$\lim_{x_j \searrow 0^+} h_j(x_j)(A_2 x_j^{a-1+C_1} + B_2 x_j^{2a-1+C_1} + C_2 x_j^{C_1-1}) = 0,$$

which by the previous discussion on (A.5) holds if and only if

$$\lim_{x_j \searrow 0^+} h_j(x_j)/x_j^{1-C_1} = 0$$

since $1 - a - C_1 < 1 - C_1$.

In summary, (A1) is satisfied if $h_j(x_j)$ grows at most polynomially as $x_j \nearrow +\infty$, and $\lim_{x_j \searrow 0^+} h_j(x_j)/x_j^{\max\{1-a,1-b\}} = 0$ if $b > 0$, or $\lim_{x_j \searrow 0^+} h_j(x_j)/x_j^{1-\eta_{0,j}} = 0$ if $b = 0$.

Conditions for (A2): For (A2), we consider powers of x as the h functions for simplicity; conclusions for other functions that have the same tail behavior (big-O scaling) as $x \searrow 0$ and $x \nearrow +\infty$ follow similarly. Sufficiency results for piecewise power functions follow by partitioning, and similarly for other functions h whose function values and derivatives can be bounded by those of some piecewise power function (e.g. truncated powers), since (A2) is on integrability of products involving positive powers of h and h'.

Let $\mathbf{K}_0$ and $\boldsymbol{\eta}_0$ be the true parameters from the parameter space that satisfies the conditions for finite normalizing constant. By part (2) of the proof of Theorem 5, the assumption that $\mathbf{K}_0$ satisfies (CC1) implies that $\min_{\boldsymbol{v} \in \mathbb{R}_+^m \setminus \{\mathbf{0}\}} \boldsymbol{v}^\top \mathbf{K}_0 \boldsymbol{v} / \boldsymbol{v}^\top \boldsymbol{v} \equiv N_{\mathbf{K}_0} > 0$. Then we have the following decomposition

$$\begin{aligned} p_{\mathbf{K}_0, \boldsymbol{\eta}_0}(\boldsymbol{x}) &\equiv \exp\left(-\frac{1}{2a} {\boldsymbol{x}^a}^\top \mathbf{K}_0 \boldsymbol{x}^a + \boldsymbol{\eta}_0^\top \frac{\boldsymbol{x}^b - \mathbf{1}_m}{b} \right) \\ &\leq \prod_{j=1}^m \exp\left(-\frac{N_{\mathbf{K}_0}}{2a} x_j^{2a} + \eta_{0,j} \frac{x_j^b - 1}{b} \right). \end{aligned}$$

Then for any other $\mathbf{K}$ and $\boldsymbol{\eta}$ in the parameter space, for the first part of (A2) it suffices to show for any $j = 1, \ldots, m$ that $D < \infty$, where

$$\begin{aligned} D &\equiv \int_{\mathbb{R}_+^m} \prod_{j=1}^m \exp\left(-\frac{N_{\mathbf{K}_0}}{2a} x_j^{2a} + \eta_{0,j} \frac{x_j^b - 1}{b} \right) h_j(x_j) \times \\ &\qquad \left(-\kappa_{jj} x_j^{2a-1} - \sum_{i \neq j} \kappa_{ji} x_i^a x_j^{a-1} + \eta_j x_j^{b-1} \right)^2 \mathrm{d}\boldsymbol{x} \\ &\geq \int_{\mathbb{R}_+^m} p_0(\boldsymbol{x}) h_j(x_j) \left(\partial_j \log p(\boldsymbol{x}) \right)^2 \mathrm{d}\boldsymbol{x}. \end{aligned}$$

Note that

$$\left(-\kappa_{jj} x_j^{2a-1} - \sum_{i \neq j} \kappa_{ji} x_i^a x_j^{a-1} + \eta_j x_j^{b-1} \right)^2$$

$$= \kappa_{jj}^2 x_j^{4a-2} + \sum_{i\neq j,\ell\neq j} \kappa_{ji}\kappa_{j\ell} x_i^a x_\ell^a x_j^{2a-2} + \eta_j^2 x_j^{2b-2} + 2\sum_{i\neq j} \kappa_{jj}\kappa_{ji} x_i^a x_j^{3a-2} - 2\sum_{i\neq j} \kappa_{ji}\eta_j x_i^a x_j^{a+b-2} - 2\kappa_{jj}\eta_j x_j^{2a+b-2}.$$

Thus, plugging this back in the definition of D, we can split D into a sum of six terms D_1 through D_6, each of which is a sum of terms of the form

$$\begin{aligned}
&\int_{\mathbb{R}_+^m} \prod_{k=1}^m \exp\left(-\frac{N_{\mathbf{K}_0}}{2a} x_k^{2a} + \eta_{0,k}\frac{x_k^b - 1}{b}\right) h_j(x_j) x_i^{\mathrm{pow}_i} x_\ell^{\mathrm{pow}_\ell} x_j^{\mathrm{pow}_j}\,\mathrm{d}\boldsymbol{x} \\
=&\prod_{k\neq j} \int_0^\infty \exp\left(-\frac{N_{\mathbf{K}_0}}{2a} x_k^{2a} + \eta_{0,k}\frac{x_k^b - 1}{b}\right) x_k^{\mathrm{pow}_k}\,\mathrm{d}x_k \\
&\qquad\qquad \times \int_0^\infty \exp\left(-\frac{N_{\mathbf{K}_0}}{2a} x_j^{2a} + \eta_{0,j}\frac{x_j^b - 1}{b}\right) h_j(x_j) x_j^{\mathrm{pow}_j}\,\mathrm{d}x_j
\end{aligned}$$

times a constant involving $\mathbf{K}$ and η_j, where $\mathrm{pow}_k \geq 0$ for each $k \neq j$. We have thus decomposed the integral into a product of univariate integrals. Note that

$$\int_0^\infty \exp\left(-\frac{N_{\mathbf{K}_0}}{2a} x_i^{2a} + \eta_{0,i}\frac{x_i^b - 1}{b}\right) x_i^{\mathrm{pow}_i}\,\mathrm{d}x_i$$

is finite for all $\mathrm{pow}_i \geq 0$ regardless of whether b is nonzero, since we assumed $\mathbf{K}_0$ and $\boldsymbol{\eta}_0$ to lie in the parameter space with a finite normalizing constant. Indeed, if $b > 0$ then the terms in the exponential is a regular polynomial with positive degree and a negative leading term; if $b = 0$ then integrability follows from $\eta_{0,i} + \mathrm{pow}_i \geq \eta_{0,i} > -1$. Thus, we only need to consider the univariate integral that involve the x_j terms, namely

$$\int_0^\infty \exp\left(-\frac{N_{\mathbf{K}_0}}{2a} x_j^{2a} + \eta_{0,j}\frac{x_j^b - 1}{b}\right) h_j(x_j) x_j^{\mathrm{pow}_j}\,\mathrm{d}x_j,$$

where pow_j takes value in $\{4a-2, 2a-2, 2b-2, 3a-2, a+b-2, 2a+b-2\} \subseteq [2\min\{a,b\} - 2, 4a-2]$. We split the integral into two parts over $[0,1]$ and $[1,\infty]$, respectively.

- If $b > 0$, on $[0,1]$ the exponential part is bounded above and below by positive constants, and for (A1) we require $h_j(x) = o(x^{1-\min\{a,b\}})$ as $x \searrow 0^+$, so the integrand is $o(x^{\min\{a,b\}-1}) = o(x^{-1})$ and is thus integrable on $[0,1]$. The integrand on $[1,\infty)$ is integrable as in (A1) we assume h to grow at most polynomially.

- If $b = 0$, $\mathrm{pow}_j \in [-2, 4a-2]$ and the integrand becomes

$$\exp(-N_{\mathbf{K}_0} x_j^{2a}/(2a)) h_j(x_j) x_j^{\mathrm{pow}_j + \eta_{0,j}}.$$

  On $[0,1]$, (A1) requires $h_j(x) = o(x^{1-\min_j \eta_{0,j}})$, so $h_j(x_j) x_j^{\mathrm{pow}_j+\eta_{0,j}} = o(x^{-1})$ and the integrand is again integrable. Integrability on $[1,\infty)$ follows similarly to the case with $b > 0$.

Now consider the second part of (A2). By definition $\mathbb{E}_{p_0} \|(\nabla \log p(\boldsymbol{X}) \circ \boldsymbol{h}(\boldsymbol{X}))'\|_1$ equals

$$\begin{aligned}
&\int_{\mathbb{R}_+^m} p_0(\boldsymbol{x}) \sum_{j=1}^m \left| h_j'(X_j) \partial_j \log p(\boldsymbol{X}) + h_j(X_j) \partial_j^2 \log p(\boldsymbol{X}) \right| \mathrm{d}\boldsymbol{x} \\
\le &\sum_{j=1}^m \int_{\mathbb{R}_+^m} \exp\left(-\frac{N_{\mathbf{K}_0}}{2a} x_j^{2a} + \eta_{0,j} \frac{x_j^b - 1}{b} \right) \left| h_j'(x_j) \left(-\kappa_{jj} x_j^{2a-1} - \sum_{i \neq j} \kappa_{ji} x_i^a x_j^{a-1} + \eta_j x_j^{b-1} \right) \right. \\
&\qquad \left. + h_j(x_j) \left(-\kappa_{jj}(2a-1) x_j^{2a-2} - \sum_{i \neq j} \kappa_{ji}(a-1) x_i^a x_j^{a-2} + (b-1)\eta_j x_j^{b-2} \right) \right| \mathrm{d}\boldsymbol{x}.
\end{aligned}$$

By the triangle inequality and the fact that $h_j \ge 0$ and $h_j' \ge 0$, similar to the proof for the first part, for each j the integral can be bounded by a sum of six integrals, each of the form

$$\text{const} \times \int_{\mathbb{R}_+^m} \prod_{k=1}^m \exp\left(-\frac{N_{\mathbf{K}_0}}{2a} x_k^{2a} + \eta_{0,k} \frac{x_k^b - 1}{b} \right) h_j(x_j) x_i^{\mathrm{pow}_i} x_j^{\mathrm{pow}_j} \, \mathrm{d}\boldsymbol{x},$$

or with h_j replaced by h_j'. Finiteness thus follows from the same type of discussion by noting that $h_j(x) = o\left(x^{1-\min\{a,b\}}\right)$ and $h_j'(x) = o\left(x^{-\min\{a,b\}}\right)$.

We conclude that if the true and the proposed parameters give densities with finite normalizing constants, and if h satisfies assumption (A1), then (A2) is automatically satisfied.

In the centered case where we assume $\boldsymbol{\eta} \equiv \mathbf{0}$, we only need $\lim_{x_j \searrow 0^+} h_j(x_j)/x_j^{1-a} = 0$ as it is a special case with $b = 2a$. □

A.1.4 Proof of Theorems in Section 2.6

Proof of Corollary 8. By Theorem 7, under assumptions in that theorem, the support of $\hat{\boldsymbol{\Psi}}$ is a subset of the true support of $\boldsymbol{\Psi}_0$, and $\|\hat{\boldsymbol{\Psi}} - \boldsymbol{\Psi}_0\|_\infty \le \frac{c_{\boldsymbol{\Gamma}_0}}{2-\alpha}\lambda$. Since $\boldsymbol{\Psi}_0$ has $|S_0|$ nonzero

entries,

$$|||\hat{\mathbf{\Psi}} - \mathbf{\Psi}_0|||_F = \left[\sum_{\mathbf{\Psi}_{0,jk}\neq 0} (\hat{\mathbf{\Psi}}_{jk} - \mathbf{\Psi}_{0,jk})^2\right]^{1/2} \leq \sqrt{|S_0|}\|\hat{\mathbf{\Psi}} - \mathbf{\Psi}_0\|_\infty \leq \frac{c_{\mathbf{\Gamma}_0}}{2-\alpha}\lambda\sqrt{|S_0|}.$$

Similarly, by the definition of matrix ℓ_∞-ℓ_∞ norm,

$$|||\hat{\mathbf{\Psi}} - \mathbf{\Psi}_0|||_2 \leq |||\hat{\mathbf{\Psi}} - \mathbf{\Psi}_0|||_\infty = \max_{j=1,\ldots,m}\sum_{k=1}^{m}|\hat{\mathbf{\Psi}}_{jk} - \mathbf{\Psi}_{0,jk}| \leq \frac{c_{\mathbf{\Gamma}_0}}{2-\alpha}\lambda d_{\mathbf{\Psi}_0}.$$

The result follows by also noting that $|||\hat{\mathbf{\Psi}} - \mathbf{\Psi}_0|||_2 \leq |||\hat{\mathbf{\Psi}} - \mathbf{\Psi}_0|||_F$. □

Proof of Theorem 9. The proof is based on Theorem 7 and a probabilistic bound on $\|\mathbf{\Gamma}_{\boldsymbol{\gamma}} - \mathbf{\Gamma}_0\|_\infty$, where in the case of centered Gaussian $\mathbf{\Gamma} = \text{diag}(\mathbf{x}\mathbf{x}^\top, \ldots, \mathbf{x}\mathbf{x}^\top)$. Denote $\mathbf{\Sigma}_0 = \mathbf{K}_0^{-1}$. In particular, given $\tau > 2$ we wish to show that for $\epsilon = 80\sqrt{2}c_0 \max_j(\Sigma_{0,jj})$, assuming $c_0 \equiv \sqrt{(\tau\log m + \log 4)/n} < 1/\sqrt{2}$,

$$\mathbb{P}\left(\left|n^{-1}\sum_{i=1}^{n} X_j^{(i)}X_k^{(i)} + \gamma_{1j}\mathbb{1}_{\{j=k\}} - \mathbb{E}X_jX_k\right| > \epsilon\right) \leq m^{2-\tau},$$

and so the results follow from Theorem 7.

By Lemma 1 of Ravikumar et al. (2011), since $X_j/\sqrt{\Sigma_{0,jj}}$ is Gaussian with mean 0 and standard deviation 1,

$$\mathbb{P}\left(\left|n^{-1}\sum_{i=1}^{n} X_j^{(i)}X_k^{(i)} - \mathbb{E}X_jX_k\right| > t\right) \leq 4\exp\left(-\frac{nt^2}{3200\max_j(\Sigma_{0,jj})^2}\right)$$

for $t \in (0, 40\max_j(\Sigma_{0,jj}))$. Denote the event as $\mathcal{E}_{j,k}(t)$. Note that $\mathbb{E}X_j^2 \leq \max_j \Sigma_{0,jj} = \epsilon/(80\sqrt{2}c_0)$. Then letting $t = \epsilon/2$ and conditioning on the complement of $\mathcal{E}_{j,j}(\epsilon/2)$, we have

$$n^{-1}\sum_{i=1}^{n} {X_j^{(i)}}^2 \leq \mathbb{E}X_j^2 + \epsilon/2 \leq \frac{\epsilon}{2}\left(1 + \frac{1}{40\sqrt{2}c_0}\right).$$

Thus, choosing $\gamma_{\ell j} = (\delta - 1)\sum_{i=1}^{n} {X_j^{(i)}}^2/n$ for $\ell = 1, \ldots, m$ ($\mathbf{\Gamma}$ has m identical blocks) with $1 < \delta < 1 + \left(1 + 1/(40\sqrt{2}c_0)\right)^{-1}$, by the triangle inequality and a union bound we have

$$\mathbb{P}\left(\max_{j,k}\left|n^{-1}\sum_{i=1}^{n} X_j^{(i)}X_k^{(i)} + \gamma_{1j}\mathbb{1}_{\{j=k\}} - \mathbb{E}X_jX_k\right| > \epsilon\right) \leq \mathbb{P}(\mathcal{E}_{j,k}(\epsilon/2)) = m^{2-\tau}.$$

Since $\tau > 2$, it holds that $1 + \left(1 + 1/(40\sqrt{2}c_0)\right)^{-1} = 2 - (1 + 40\sqrt{2}c_0)^{-1}$ is larger than $2 - (1 + 80\sqrt{\log m/n})^{-1} \equiv C(n, m)$, so it is safe to choose any $\delta \in (1, C(n, m))$. Thus by the requirement on ϵ, the theorem statement holds when $n > \max(c^* c_1^2 d_{\mathbf{K}}^2, 2)(\tau \log m + \log 4)$ with $c^* = 12800 \max_j(\Sigma_{0,jj})^2$. □

Proof of Theorem 10. The proof of Theorem 7 from Lin et al. (2016) does not rely on the fact that the original $\mathbf{\Gamma}$ is an unbiased estimator for the population $\mathbf{\Gamma}_0$, but instead only requires one to bound $\|\mathbf{\Gamma} - \mathbf{\Gamma}_0\|_\infty$. Thus, for $\mathbf{\Gamma}_{\boldsymbol{\gamma}} = \mathbf{\Gamma} + \mathrm{diag}(\boldsymbol{\gamma})$, by Theorem 7 it suffices to prove that for any $\tau > 3$, we can bound $\|\mathbf{\Gamma}(\mathbf{x}) + \mathrm{diag}(\boldsymbol{\gamma}(\mathbf{x})) - \mathbf{\Gamma}_0\|_\infty$ by some ϵ_1 and $\|\boldsymbol{g}(\mathbf{x}) - \boldsymbol{g}_0\|_\infty$ by some ϵ_2, uniformly with probability $1 - m^{3-\tau}$. Recall from (2.19) that the j^{th} block of $\mathbf{\Gamma}_{\boldsymbol{\gamma}} \in \mathbb{R}^{m^2 \times m^2}$ has (k, ℓ)-th entry

$$n^{-1} \sum_{i=1}^{n} X_k^{(i)} X_\ell^{(i)} h_j\left(X_j^{(i)}\right) + \gamma_{kj} \cdot \mathbb{1}_{\{k=\ell\}}.$$

The entry in $\boldsymbol{g} \in \mathbb{R}^{m^2}$ (obtained by linearizing a $m \times m$ matrix) corresponding to (j, k) is

$$n^{-1} \sum_{i=1}^{n} X_k^{(i)} h_j'\left(X_j^{(i)}\right) + n^{-1} \mathbb{1}_{\{j=k\}} \sum_{i=1}^{n} h_j\left(X_j^{(i)}\right).$$

Denote $M \equiv \max_j \sup_{x>0} h_j(x)$, $M' \equiv \max_j \sup_{x>0} h_j'(x)$, and $c_{\boldsymbol{X}} \equiv 2 \max_j (2\sqrt{\Sigma_{jj}} + \sqrt{e}\, \mathbb{E}_0 X_j)$. Using results for sub-Gaussian random variables from Lemma 36.2 in Appendix A.2, we have for any $t_1 > 0$,

$$\mathbb{P}\left(\left| n^{-1} \sum_{i=1}^{n} X_k^{(i)} X_\ell^{(i)} h_j\left(X_j^{(i)}\right) - \mathbb{E}_0 X_k X_\ell h_j(X_j) \right| > t_1\right) \leq 2 \exp\left(-\min\left(\frac{n t_1^2}{2M^2 c_{\boldsymbol{X}}^4}, \frac{n t_1}{2M c_{\boldsymbol{X}}^2}\right)\right).$$

Thus, choosing $\epsilon_1 \equiv 2M c_{\boldsymbol{X}}^2 c_{n,m}$, where $c_{n,m} \equiv \max\left\{\frac{2(\log m^\tau + \log 6)}{n}, \sqrt{\frac{2(\log m^\tau + \log 6)}{n}}\right\}$, for $\gamma_{kj} \leq \epsilon_1/2$, we have

$$\begin{aligned} &\mathbb{P}\left(\left| n^{-1} \sum_{i=1}^{n} X_k^{(i)} X_\ell^{(i)} h_j\left(X_j^{(i)}\right) + \gamma_{kj} \mathbb{1}_{\{k=\ell\}} - \mathbb{E}_0 X_k X_\ell h_j(X_j) \right| > \epsilon_1\right) \\ \leq &\mathbb{P}\left(\left| n^{-1} \sum_{i=1}^{n} X_k^{(i)} X_\ell^{(i)} h_j\left(X_j^{(i)}\right) - \mathbb{E}_0 X_k X_\ell h_j(X_j) \right| > \epsilon_1/2\right) \end{aligned} \tag{A.6}$$

$$\leq 2\exp\left(-\min\left(\frac{n\epsilon_1^2}{8M^2c_{\boldsymbol{X}}^4}, \frac{n\epsilon_1}{4Mc_{\boldsymbol{X}}^2}\right)\right) \leq \frac{1}{3m^\tau}. \tag{A.7}$$

Denote the event inside the probability in (A.6) as $\mathcal{E}_{k,\ell,j}(\epsilon_1/2)$. By definition,

$$c_{\boldsymbol{X}}^2 = 4\max_k\left(4\Sigma_{kk} + 4\sqrt{e}\sqrt{\Sigma_{kk}}\,\mathbb{E}_0X_k + e(\mathbb{E}_0X_k)^2\right) \geq 4e\max_k\left(\Sigma_{kk} + (\mathbb{E}_0X_k)^2\right).$$

By Lemmas 35.2 and 36.1 from Appendix A.2, $\mathrm{var}(X_k) \leq \Sigma_{kk}$, so $c_{\boldsymbol{X}}^2 \geq 4e\max_k \mathbb{E}_0X_k^2 \geq 4e\mathbb{E}_0X_k^2h_j(X_j)/M$. Thus, setting $\epsilon_1 = 2Mc_{\boldsymbol{X}}^2c_{n,m}$, on the complement of $\mathcal{E}_{k,k,j}(\epsilon_1/2)$ we have

$$n^{-1}\sum_{i=1}^n {X_k^{(i)}}^2 h_j\left(X_j^{(i)}\right) \leq \mathbb{E}_0X_k^2h_j(X_j) + \epsilon_1/2 \leq \frac{\epsilon_1}{2}\left(1 + \frac{1}{4ec_{n,m}}\right).$$

Then

$$\frac{1}{1+1/(4ec_{n,m})}\frac{1}{n}\sum_{i=1}^n {X_k^{(i)}}^2 h_j\left(X_j^{(i)}\right) \leq \epsilon_1/2 \tag{A.8}$$

on the complement of $\mathcal{E}_{k,k,j}(\epsilon_1/2)$, again with $c_{n,m} \equiv \max\left\{\frac{2(\log m^\tau + \log 6)}{n}, \sqrt{\frac{2(\log m^\tau+\log 6)}{n}}\right\}$. Note that the multiplier on the left of (A.8) is increasing in $c_{n,m}$, and that $2(\log m^\tau + \log 6) > 6\log m$ by the assumption that $\tau > 3$. Thus, if we let

$$\gamma_{kj} \equiv \frac{1}{1+1/\left(4e\max\left\{6\log m/n, \sqrt{6\log m/n}\right\}\right)}\frac{1}{n}\sum_{i=1}^n {X_k^{(i)}}^2 h_j\left(X_j^{(i)}\right),$$

which is just a constant multiple of the (k,k)-th entry of $\boldsymbol{\Gamma}_j$ itself, with the constant explicitly calculable and a function of p and n only, then for $k = \ell$

$$\mathbb{P}\left(\left|n^{-1}\sum_{i=1}^n X_k^{(i)}X_\ell^{(i)}h_j\left(X_j^{(i)}\right) + \gamma_{kj} - \mathbb{E}_0X_kX_\ell h_j(X_j)\right| \geq \epsilon_1\right) \leq \mathbb{P}(\mathcal{E}_{k,\ell,j}(\epsilon_1/2)) \leq \frac{1}{3m^\tau}.$$

Since this also holds for $k \neq \ell$ without the γ_{kj} term, by a union bound over m^3 events,

$$\mathbb{P}\left(\max_{j,k,\ell}\left|n^{-1}\sum_{i=1}^n X_k^{(i)}X_\ell^{(i)}h_j\left(X_j^{(i)}\right) + \gamma_{kj}\mathbb{1}_{\{k=\ell\}} - \mathbb{E}_0X_kX_\ell h_j(X_j)\right| \geq \epsilon_1\right) \leq \frac{1}{3m^{\tau-3}}. \tag{A.9}$$

Now, on the other hand, Lemma 36.1 and Hoeffding's inequality give for any $t_{2,1}, t_{2,2} > 0$ that

$$\mathbb{P}\left(\left|n^{-1}\sum_{i=1}^n X_k^{(i)}h_j'\left(X_j^{(i)}\right) - \mathbb{E}_0X_kh_j'(X_j)\right| \geq t_{2,1}\right) \leq 2\exp\left(-\frac{nt_{2,1}^2}{2M'^2c_{\boldsymbol{X}}^2}\right),$$

$$\mathbb{P}\left(\left|n^{-1}\sum_{i=1}^{n} h_j\left(X_j^{(i)}\right) - \mathbb{E}_0 h_j(X_j)\right| \geq t_{2,2}\right) \leq 2\exp\left(-2nt_{2,2}^2/M^2\right).$$

Choosing $\epsilon_{2,1} \equiv \sqrt{2}M' c_{\boldsymbol{X}}\sqrt{\frac{\log m^{\tau-1}+\log 6}{n}}$, $\epsilon_{2,2} \equiv M\sqrt{\frac{\log m^{\tau-2}+\log 6}{2n}}$ and taking union bounds over m^2, and m events, respectively, we have

$$\mathbb{P}\left(\max_{j,k}\left|n^{-1}\sum_{i=1}^{n} X_k^{(i)} h_j'\left(X_j^{(i)}\right) - \mathbb{E}_0 X_k h_j'(X_j)\right| \geq \epsilon_{2,1}\right) \leq \frac{1}{3m^{\tau-3}}, \tag{A.10}$$

$$\mathbb{P}\left(\max_{j}\left|n^{-1}\sum_{i=1}^{n} h_j\left(X_j^{(i)}\right) - \mathbb{E}_0 h_j(X_j)\right| \geq \epsilon_{2,2}\right) \leq \frac{1}{3m^{\tau-3}}. \tag{A.11}$$

Hence, by (A.9) (A.10) (A.11), with probability at least $1 - m^{3-\tau}$, $\|\boldsymbol{\Gamma}_{\boldsymbol{\gamma}}(\mathbf{x}) - \boldsymbol{\Gamma}_0\|_\infty < \epsilon_1$ and $\|\boldsymbol{g}(\mathbf{x}) - \boldsymbol{g}_0\|_\infty < \epsilon_2 \equiv \epsilon_{2,1} + \epsilon_{2,2}$. Consider any $\tau > 3$, and let

$$\begin{aligned}
c_2 &\equiv \frac{6}{\alpha} c_{\boldsymbol{\Gamma}_0}, \\
n &> \max\left\{2M^2 c_{\boldsymbol{X}}^4 c_2^2 d_{\mathbf{K}_0}^2(\tau\log m + \log 6), 2Mc_{\boldsymbol{X}}^2 c_2 d_{\mathbf{K}_0}(\tau\log m + \log 6)\right\}, \\
\lambda &> \frac{3(2-\alpha)}{\alpha}\max\{c_{\mathbf{K}_0}\epsilon_1, \epsilon_2\} \\
&\equiv \frac{3(2-\alpha)}{\alpha}\max\left\{4Mc_{\mathbf{K}_0}c_{\boldsymbol{X}}^2\frac{(\log m^\tau + \log 6)}{n},\right. \\
&\qquad \left. 2Mc_{\mathbf{K}_0}c_{\boldsymbol{X}}^2\sqrt{\frac{2(\log m^\tau + \log 6)}{n}}, \sqrt{2}M' c_{\boldsymbol{X}}\sqrt{\frac{\log m^{\tau-1}+\log 6}{n}} + M\sqrt{\frac{\log m^{\tau-2}+\log 6}{2n}}\right\}.
\end{aligned}$$

Then $d_{\mathbf{K}_0}\epsilon_1 \leq \alpha/(6c_{\boldsymbol{\Gamma}_0})$ and the results follow from Theorem 7. □

Proof of Theorem 11. Similar to the proof of Theorem 10, by Theorem 7 it suffices to prove that for any $\tau > 3$, we can bound $\|\boldsymbol{\Gamma}_{\boldsymbol{\gamma}}(\mathbf{x}) - \boldsymbol{\Gamma}_0\|_\infty$ by some ϵ_1 and $\|\boldsymbol{g}(\mathbf{x}) - \boldsymbol{g}_0\|_\infty$ by some ϵ_2, uniformly with probability $1 - m^{3-\tau}$. Recall that $\boldsymbol{\Gamma} \in \mathbb{R}^{(m^2+m)\times(m^2+m)}$ is a rearrangement of $\boldsymbol{\Gamma}^{(*)}$, which is in turn formed by $\boldsymbol{\Gamma}_{11} \in \mathbb{R}^{m^2\times m^2}$, $\boldsymbol{\Gamma}_{12} \in \mathbb{R}^{m^2\times m}$, $\boldsymbol{\Gamma}_{12}^\top$ and $\boldsymbol{\Gamma}_{22} \in \mathbb{R}^{m\times m}$, all of which are block-diagonal with m blocks.

The j^{th} block of $\boldsymbol{\Gamma}_{11} \in \mathbb{R}^{m^2\times m^2}$ has (k,ℓ)-th entry

$$n^{-1}\sum_{i=1}^{n} X_k^{(i)} X_\ell^{(i)} h_j\left(X_j^{(i)}\right),$$

the k^{th} entry in the j^{th} block of $\mathbf{\Gamma}_{12}$ is

$$-n^{-1}\sum_{i=1}^{n} X_k^{(i)} h_j\left(X_j^{(i)}\right),$$

the j^{th} diagonal entry of $\mathbf{\Gamma}_{22}$ is

$$n^{-1}\sum_{i=1}^{n} h_j\left(X_j^{(i)}\right).$$

On the other hand, $\boldsymbol{g} \in \mathbb{R}^{(m^2+m)}$ is a rearrangement of $\boldsymbol{g}^{(*)} \equiv [\boldsymbol{g}_1^\top, \boldsymbol{g}_2^\top]^\top$, where the entry in $\boldsymbol{g}_1 \in \mathbb{R}^{m^2}$ (obtained by linearizing a $m \times m$ matrix) corresponding to (j,k), is

$$n^{-1}\sum_{i=1}^{n} X_k^{(i)} h_j'\left(X_j^{(i)}\right) + n^{-1}\mathbb{1}_{\{j=k\}}\sum_{i=1}^{n} h_j\left(X_j^{(i)}\right),$$

while the j-th component of $\boldsymbol{g}_2 \in \mathbb{R}^m$ is

$$-n^{-1}\sum_{i=1}^{n} h_j'\left(X_j^{(i)}\right).$$

Recalling that the bounds in Lemma 36 also hold when $\boldsymbol{\mu} \neq 0$, we may then use bounds similar to those in the proof of Theorem 10, and use union bounds to arrive at analogous consistency results, modulus different constants. The amplifiers $\boldsymbol{\gamma}$ can be incorporated analogously. □

A.2 Auxiliary Lemmas and Definitions

In this appendix, to simplify notation, when it is clear from the context, the operator $\mathbb{E}$ is defined as the expectation under the true distribution, unless otherwise noted.

Definition 16 (Sub-Gaussian and Sub-Exponential Variables)**.**
The sub-Gaussian ($r = 2$) *and* sub-exponential ($r = 1$) norms *of a random variable are*

$$\|X\|_{\psi_r} \equiv \sup_{q\geq 1} q^{-1/r}(\mathbb{E}|X|^{rq})^{1/(rq)} \equiv \sup_{q\geq 1} q^{-1/r}\|X\|_{rq}.$$

If $\|X\|_{\psi_2} < \infty$ *we say* X *is* sub-Gaussian*; if* $\|X\|_{\psi_1} < \infty$ *we call* X sub-exponential*. For a* zero-mean *sub-Gaussian random variable* X *also define the* sub-Gaussian parameter

$$\tau(X) = \inf\{\tau \geq 0 : \mathbb{E}\exp(tX) \leq \exp(\tau^2 t^2/2),\ \forall t \in \mathbb{R}\}.$$

The definition of sub-Gaussian norm here allows for a non-centered variable and differs from the one in Vershynin (2012), which uses $\|X\|_q$. Instead, it coincides with θ_2 in Buldygin and Kozachenko (2000). The sub-Gaussian parameter is defined as in Buldygin and Kozachenko (2000) and the sub-exponential norm as in Vershynin (2012).

Lemma 35 (Properties of Sub-Gaussian and Sub-Exponential Variables)**.**

1) For any X and $r = 1, 2$, $\|X - \mathbb{E}X\|_{\psi_r} \leq 2\|X\|_{\psi_r}$ and $\|X\|_{\psi_r} \leq \|X - \mathbb{E}X\|_{\psi_r} + |\mathbb{E}X|$, as long as the expectation and norms are finite.

2) (Buldygin and Kozachenko, 2000) $\tau(X)$ is a norm on the space of all zero-mean sub-Gaussian variables; so $\tau(X + Y) \leq \tau(X) + \tau(Y)$. If X is zero-mean sub-Gaussian, then $\mathrm{var}(X) \leq \tau^2(X)$, $\|X\|_{\psi_2} \leq 2\tau(X)/\sqrt{e}$, $\tau(X) \leq \sqrt{e}\|X\|_{\psi_2}$. If $X_1, \ldots, X_n$ are i.i.d. zero-mean sub-Gaussian, $\tau\left(n^{-1}\sum_{i=1}^n X_i\right) \leq n^{-1/2}\tau(X_i)$.

3) If X_1 and X_2 are sub-Gaussian (not necessarily independent) with $\|X_1\|_{\psi_2} \leq K_1$ and $\|X_2\|_{\psi_2} \leq K_2$, then X_1X_2 is sub-exponential with $\|X_1X_2\|_{\psi_1} \leq K_1K_2$.

4) (Buldygin and Kozachenko, 2000) If X is zero-mean sub-Gaussian and $q > 0$, then

$$\mathbb{E}|X|^q \leq 2(q/e)^{q/2}\tau^q(X).$$

5) (Buldygin and Kozachenko, 2000) If $X_1, \ldots, X_n$ are independent zero-mean, sub-Gaussian variables, then for any $\epsilon > 0$,

$$\mathbb{P}(|X_1| \geq \epsilon) \leq 2\exp\left(-\frac{\epsilon^2}{2\tau^2(X_1)}\right),$$
$$\mathbb{P}\left(\left|n^{-1}\sum_{i=1}^n X_i\right| > \epsilon\right) \leq 2\exp\left(-\frac{n\epsilon^2}{2\max_i \tau^2(X_i)}\right).$$

6) (Vershynin, 2012) If $X_1, \ldots, X_n$ are independent zero-mean sub-exponential random variables with $K \geq \max_i \|X_i\|_{\psi_1}$, then for any $\epsilon > 0$,

$$\mathbb{P}(|X_1| \geq \epsilon) \leq 2\exp\left(-\min\left(\frac{\epsilon^2}{8e^2K^2}, \frac{\epsilon}{4eK}\right)\right),$$

$$\mathbb{P}\left(\left|n^{-1}\sum_{i=1}^{n} X_i\right| \geq \epsilon\right) \leq 2\exp\left(-\min\left(\frac{n\epsilon^2}{8e^2K^2}, \frac{n\epsilon}{4eK}\right)\right).$$

7) (Boucheron et al., 2013) If for X_i i.i.d. there exists some $B > 0$ such that

$$\sup_{q\geq 2}\left(\frac{\mathbb{E}|X|^q}{q!}\right)^{1/q} \leq B/2$$

then for all $\epsilon > 0$,

$$\mathbb{P}\left(\left|n^{-1}\sum_{i=1}^{n}(X_i - \mathbb{E}X_i)\right| \geq \epsilon\right) \leq 2\exp\left(-\min\left(\frac{n\epsilon^2}{2B^2}, \frac{n\epsilon}{2B}\right)\right).$$

Proof. 1) For $r = 1, 2$, by the triangle inequality, $\|X - \mathbb{E}X\|_{\psi_r} \leq \|X\|_{\psi_r} + \|\mathbb{E}X\|_{\psi_r} = \|X\|_{\psi_r} + |\mathbb{E}X| \leq \|X\|_{\psi_r} + \mathbb{E}|X| \leq 2\|X\|_{\psi_r}$, where in the last step we used the definition of $\|\cdot\|_{\psi_r}$ with $q = 1$ for $r = 1$ and $\mathbb{E}|X| \leq (\mathbb{E}|X|^2)^{1/2}$ with $q = 2$ for $r = 2$. On the other hand, $\|X\|_{\psi_r} \leq \|X - \mathbb{E}X\|_{\psi_r} + \|\mathbb{E}X\|_{\psi_r} = \|X - \mathbb{E}X\|_{\psi_r} + |\mathbb{E}X|$.

2) These follow from Theorems 1.2 and 1.3 and Lemmas 1.2 and 1.7 from Buldygin and Kozachenko (2000), and $\sqrt[4]{3.1}e^{9/16}/\sqrt{2} \approx 1.6467 \leq 1.6487 \approx \sqrt{e}$.

3) By Hölder's inequality (or Cauchy-Schwarz),

$$\begin{aligned}
\|X_1X_2\|_{\psi_1} &= \sup_{q\geq 1} q^{-1}(\mathbb{E}|X_1X_2|^q)^{1/q} = \sup_{q\geq 1} q^{-1}(\mathbb{E}|X_1^qX_2^q|)^{1/q} \\
&\leq \sup_{q\geq 1} q^{-1}\left[(\mathbb{E}|X_1|^{2q})^{1/2}(\mathbb{E}|X_2|^{2q})^{1/2}\right]^{1/q} \\
&\leq \sup_{q\geq 1}\left[q^{-1/2}(\mathbb{E}|X_1|^{2q})^{1/2q}\right]\sup_{q\geq 1}\left[q^{-1/2}(\mathbb{E}|X_2|^{2q})^{1/2q}\right] \\
&= \|X_1\|_{\psi_2}\|X_2\|_{\psi_2} \leq K_1K_2.
\end{aligned}$$

4-6) These are Lemma 1.4 and Theorem 1.5 in Buldygin and Kozachenko (2000), and a consequence of Corollary 5.17 in Vershynin (2012).

7) By Theorem 2.10 of Boucheron et al. (2013) wherein we let $v \equiv nB^2/2$ and $c \equiv B/2$, we have

$$\mathbb{P}\left(\left|n^{-1}\sum_{i=1}^{n}(X_i - \mathbb{E}X_i)\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{n\epsilon^2}{B^2 + B\epsilon}\right)$$

for all $\epsilon > 0$. (Theorem 2.10 gives an one-sided bound; bound for the other side is obtained by taking $X_i = -X_i$). The inequality follows by splitting into cases $\epsilon \le B$ and $\epsilon > B$. $\square$

Lemma 36. *Suppose $\boldsymbol{X}$ follows a truncated normal distribution on $\mathbb{R}_+^m$ with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma} = \mathbf{K}^{-1} \succ \mathbf{0}$. Let $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(n)}$ be i.i.d. copies of $\boldsymbol{X}$, with j-th component of the i-th copy being $X_j^{(i)}$. Then*

1. *For $j = 1, \ldots, p$, $\tau(X_j - \mathbb{E}X_j) \le \sqrt{\Sigma_{jj}}$. That is, the sub-Gaussian parameter of any marginal distribution of $\boldsymbol{X}$, after centering, is bounded by the square root of its corresponding diagonal entry in the covariance parameter $\boldsymbol{\Sigma}$. Then, for any $\epsilon > 0$,*

$$\mathbb{P}\left(\left|n^{-1}\sum_{i=1}^n X_j^{(i)} - \mathbb{E}X_j\right| > \epsilon\right) \le 2\exp\left(-\frac{n\epsilon^2}{2\Sigma_{jj}}\right).$$

In particular, if h_0 is a function bounded by M_0, then for any $\epsilon > 0$,

$$\mathbb{P}\left(\left|n^{-1}\sum_{i=1}^n X_j^{(i)} h_0\left(X_k^{(i)}\right) - \mathbb{E}X_j h_0(X_k)\right| \ge \epsilon\right) \le 2\exp\left(-\frac{n\epsilon^2}{8M_0^2(2\sqrt{\Sigma_{jj}} + \sqrt{e}\,\mathbb{E}X_j)^2}\right),$$

$$\tau\left(n^{-1}\sum_{i=1}^n X_j^{(i)} h_0\left(X_k^{(i)}\right) - \mathbb{E}X_j h_0(X_k)\right) \le \frac{2M_0}{\sqrt{n}}\left(2\sqrt{\Sigma_{jj}} + \sqrt{e}\,\mathbb{E}X_j\right),$$

$$\left\|n^{-1}\sum_{i=1}^n X_j^{(i)} h_0\left(X_k^{(i)}\right) - \mathbb{E}X_j h_0(X_k)\right\|_{\psi_2} \le \frac{4M_0}{\sqrt{en}}\left(2\sqrt{\Sigma_{jj}} + \sqrt{e}\,\mathbb{E}X_j\right).$$

2. *For $j, k, \ell \in \{1, \ldots, p\}$, if h_0 is a function bounded by M_0, then*

$$\|X_j X_k h_0(X_\ell) - \mathbb{E}X_j X_k h_0(X_\ell)\|_{\psi_1} \le \frac{M_0}{2e} c_{\boldsymbol{X}}^2, \tag{A.12}$$

where $c_{\boldsymbol{X}} \equiv 2\max_j(2\sqrt{\Sigma_{jj}} + \sqrt{e}\,\mathbb{E}X_j)$. In particular, for any $\epsilon > 0$,

$$\mathbb{P}\left(\left|n^{-1}\sum_{i=1}^n X_j^{(i)} X_k^{(i)} h_0\left(X_\ell^{(i)}\right) - \mathbb{E}X_j X_k h_0(X_\ell)\right| > \epsilon\right)$$
$$\le 2\exp\left(-\min\left(\frac{n\epsilon^2}{2M_0^2 c_{\boldsymbol{X}}^4}, \frac{n\epsilon}{2M_0 c_{\boldsymbol{X}}^2}\right)\right).$$

Proof of Lemma 36. 1. Without loss of generality choose $j = 1$. By the definition of sub-Gaussian parameters, we need to show that for all $t \in \mathbb{R}$,

$$\mathbb{E}\exp(tX_1) \leq \exp\left(t^2\Sigma_{11}/2 + t\mathbb{E}X_1\right),$$

which is equivalent to

$$t^2\Sigma_{11}/2 + t\mathbb{E}X_1 - \log\mathbb{E}\exp(tX_1) \geq 0 \quad \forall t \in \mathbb{R}. \tag{A.13}$$

Since the left-hand side of (A.13) equals 0 at $t = 0$, it suffices to show that its derivative,

$$t\Sigma_{11} + \mathbb{E}X_1 - \frac{\mathrm{d}\log\mathbb{E}\exp(tX_1)}{\mathrm{d}t} = t\Sigma_{11} + \mathbb{E}X_1 - \frac{\frac{\mathrm{d}\mathbb{E}\exp(tX_1)}{\mathrm{d}t}}{\mathbb{E}\exp(tX_1)}, \tag{A.14}$$

is non-negative on $(0, \infty)$ and non-positive on $(-\infty, 0)$. By properties of moment-generating functions, $\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\exp(tX_1)$ evaluated at $t = 0$ equals $\mathbb{E}X_1$, so (A.14) equals 0 at $t = 0$. It in turn suffices to show the derivative of (A.14), namely

$$\Sigma_{11} - \frac{\mathrm{d}^2\log\mathbb{E}\exp(tX_1)}{\mathrm{d}t^2} \tag{A.15}$$

is non-negative in $t \in \mathbb{R}$.

Given any vector $\boldsymbol{v} \in \mathbb{R}^p$, define $\mathbb{R}^p_+ - \boldsymbol{v} \equiv \{\boldsymbol{u} - \boldsymbol{v} : \boldsymbol{u} \in \mathbb{R}^p_+\}$. By Tallis (1961), denoting the first column of $\boldsymbol{\Sigma}$ as $\boldsymbol{\Sigma}_1$, the moment-generating function of the marginal distribution of X_1 is

$$\frac{\int_{\mathbb{R}^p_+ - (\boldsymbol{\mu} + t\boldsymbol{\Sigma}_1)} \exp\left(-\frac{1}{2}\boldsymbol{x}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{x}\right)\mathrm{d}\boldsymbol{x}}{\int_{\mathbb{R}^p_+ - \boldsymbol{\mu}} \exp\left(-\frac{1}{2}\boldsymbol{x}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{x}\right)\mathrm{d}\boldsymbol{x}} \exp\left(t\mu_1 + \frac{1}{2}t^2\Sigma_{11}\right).$$

(A.15) thus becomes

$$-\frac{\mathrm{d}^2}{\mathrm{d}t^2}\log\int_{\mathbb{R}^p_+ - (\boldsymbol{\mu} + t\boldsymbol{\Sigma}_1)} \exp\left(-\frac{1}{2}\boldsymbol{x}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{x}\right)\mathrm{d}\boldsymbol{x}.$$

Showing this is non-negative in $t \in \mathbb{R}$ is equivalent to showing that the integral itself is log-concave in t. But

$$\int_{\mathbb{R}^p_+ - (\boldsymbol{\mu} + t\boldsymbol{\Sigma}_1)} \exp\left(-\frac{1}{2}\boldsymbol{x}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{x}\right)\mathrm{d}\boldsymbol{x} = \int_{\mathbb{R}^p} \exp\left(-\frac{1}{2}\boldsymbol{x}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{x}\right)\mathbb{1}_{\mathbb{R}^p_+ - \boldsymbol{\mu}}(\boldsymbol{x} + t\boldsymbol{\Sigma}_1)\,\mathrm{d}\boldsymbol{x}$$

with $\exp\left(-\frac{1}{2}\boldsymbol{x}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{x}\right)$ log-concave in $\boldsymbol{x}$ and $\mathbb{1}_{\mathbb{R}^p_+-\boldsymbol{\mu}}(\boldsymbol{x}+t\boldsymbol{\Sigma}_1)$ log-concave in $(\boldsymbol{x},t)$ since $\mathbb{R}^p_+-\boldsymbol{\mu}$ is a convex set. Since log-concavity is closed under multiplication and integration over $\mathbb{R}^p$, the integral is indeed log-concave, and our proof of the bound on the sub-Gaussian parameter of $X_j-\mathbb{E}X_j$ is complete. The tail bound follows from 5) of Lemma 35.

Now by 1) and 2) of Lemma 35,

$$\|X_j\|_{\psi_2}\le 2\sqrt{\Sigma_{jj}/e}+\mathbb{E}X_j.$$

If h_0 is a function bounded by M_0, then by definition

$$\|X_jh_0(X_k)\|_{\psi_2}\le M_0\left(2\sqrt{\Sigma_{jj}/e}+\mathbb{E}X_j\right).$$

By 1) and 2) of Lemma 35 again,

$$\begin{aligned}\tau(X_jh_0(X_k)-\mathbb{E}X_jh_0(X_k))&\le\sqrt{e}\|X_jh_0(X_k)-\mathbb{E}X_jh_0(X_k)\|_{\psi_2}\\&\le 2\sqrt{e}\|X_jh_0(X_k)\|_{\psi_2}\\&\le 2M_0(2\sqrt{\Sigma_{jj}}+\sqrt{e}\,\mathbb{E}X_j).\end{aligned}$$

The tail bound thus follows from the first inequality using 5) of Lemma 35. By 2) of the Lemma 35,

$$\begin{aligned}\tau\left(n^{-1}\sum_{i=1}^n X_j^{(i)}h_0\left(X_k^{(i)}\right)-\mathbb{E}X_jh_0(X_k)\right)&\le\frac{2M_0}{\sqrt{n}}\left(2\sqrt{\Sigma_{jj}}+\sqrt{e}\,\mathbb{E}X_j\right),\\\left\|n^{-1}\sum_{i=1}^n X_j^{(i)}h_0\left(X_k^{(i)}\right)-\mathbb{E}X_jh_0(X_k)\right\|_{\psi_2}&\le\frac{4M_0}{\sqrt{en}}\left(2\sqrt{\Sigma_{jj}}+\sqrt{e}\,\mathbb{E}X_j\right).\end{aligned}$$

2. By the proof of 1) of this lemma, $\|X_j\|_{\psi_2}\le 2\sqrt{\Sigma_{jj}/e}+\mathbb{E}X_j$, and by 3) of Lemma 35,

$$\|X_jX_k\|_{\psi_1}\le\left(2\sqrt{\Sigma_{jj}/e}+\mathbb{E}X_j\right)\left(2\sqrt{\Sigma_{kk}/e}+\mathbb{E}X_k\right)\le\max_j\left(2\sqrt{\Sigma_{jj}/e}+\mathbb{E}X_j\right)^2.$$

Since h_0 is a function bounded by M_0, by definition

$$\|X_jX_kh_0(X_\ell)\|_{\psi_1}\le M_0\max_j\left(2\sqrt{\Sigma_{jj}/e}+\mathbb{E}X_j\right)^2.$$

Then by 1) of Lemma 35 again,

$$\|X_j X_k h_0(X_\ell) - \mathbb{E}X_j X_k h_0(X_\ell)\|_{\psi_1} \leq 2M_0 \max_j \left(2\sqrt{\Sigma_{jj}/e} + \mathbb{E}X_j\right)^2.$$

The tail bound then follows from 6) of Lemma 35. □

Although not used for our consistency results, in the special case of $h_0 \equiv 1$, we also have the following lemma. The notable difference between bounds (A.16) below and (A.12) from Lemma 36.2 is in the constants and dependency on $\mathbb{E}X_j$: The constants in the denominator in the right-hand side of (A.12) is smaller and thus gives a tighter bound, but (A.16) is preferred when $\mathbb{E}X_j$ is notably large compared to $\sqrt{\Sigma_{jj}}$, since the constant is only linear in $\mathbb{E}X_j$.

Lemma 37. *Consider the setting in Lemma 36. Then for $j, k \in \{1, \ldots, p\}$, for any $\epsilon > 0$,*

$$\mathbb{P}\left(n^{-1}\left|\sum_{i=1}^n X_j^{(i)} X_k^{(i)} - \mathbb{E}X_j X_k\right| \geq \epsilon\right) \leq 4\exp\left(-\min\left(\frac{2n\epsilon^2}{C_1^2}, \frac{n\epsilon}{C_1}\right)\right), \tag{A.16}$$

where $C_1 \equiv 91 \max_j \Sigma_{jj} + 72 \max_j \mathbb{E}X_j \max_j \sqrt{\Sigma_{jj}}$.

Proof of Lemma 37. We use a proof similar to Lemma 1 in Ravikumar et al. (2011) (note that $\mathbb{E}X_j$ may be nonzero in our case). Define

$$U_{jk}^{(i)} \equiv X_j^{(i)} + X_k^{(i)}, \quad U_{jk} \equiv X_j + X_k, \quad V_{jk}^{(i)} \equiv X_j^{(i)} - X_k^{(i)}, \quad V_{jk} \equiv X_j - X_k.$$

Since $X_j^{(i)} X_k^{(i)} = \frac{1}{4}\left({U_{jk}^{(i)}}^2 - {V_{jk}^{(i)}}^2\right)$, by union bound we have

$$\begin{aligned}
&\mathbb{P}\left(\left|n^{-1}\sum_{i=1}^n X_j^{(i)} X_k^{(i)} - \mathbb{E}X_j X_k\right| \geq \epsilon\right) \\
\leq &\mathbb{P}\left(\left|n^{-1}\sum_{i=1}^n {U_{jk}^{(i)}}^2 - \mathbb{E}U_{jk}^2\right| \geq 2\epsilon\right) + \mathbb{P}\left(\left|n^{-1}\sum_{i=1}^n {V_{jk}^{(i)}}^2 - \mathbb{E}V_{jk}^2\right| \geq 2\epsilon\right).
\end{aligned}$$

We next define

$$Z_{jk}^{(i)} \equiv {U_{jk}^{(i)}}^2 - \mathbb{E}U_{jk}^2 = {A_{jk}^{(i)}}^2 + B_{jk}^{(i)} + C_{jk}, \qquad \overline{X}_j^{(i)} \equiv X_j^{(i)} - \mathbb{E}X_j,$$

$$A_{jk}^{(i)} \equiv \overline{X}_j^{(i)} + \overline{X}_k^{(i)}, \quad B_{jk}^{(i)} \equiv 2(\mathbb{E}X_j + \mathbb{E}X_k)(\overline{X}_j^{(i)} + \overline{X}_k^{(i)}), \quad C_{jk} \equiv -\mathbb{E}(\overline{X}_j^{(i)} + \overline{X}_k^{(i)})^2.$$

Then since τ is a norm by 2) of Lemma 35, A_{jk} is sub-Gaussian with parameter $\leq \sqrt{\Sigma_{jj}} + \sqrt{\Sigma_{kk}}$, and B_{jk} is sub-Gaussian with parameter $\leq 2(\mathbb{E}X_j + \mathbb{E}X_k)\left(\sqrt{\Sigma_{jj}} + \sqrt{\Sigma_{kk}}\right)$. Using 4) of Lemma 35 together with the inequality $(a+b+c)^q \leq (3\max\{a,b,c\})^q \leq 3^q(a^q + b^q + c^q)$ for all $a, b, c \geq 0$ and $q > 0$, we have for any $q \geq 2$

$$\begin{aligned}
(\mathbb{E}|Z_{jk}|^q)^{1/q} &\leq \left(3^q\left(\mathbb{E}|A_{jk}|^{2q} + \mathbb{E}|B_{jk}|^q + |C_{jk}|^q\right)\right)^{1/q} \\
&\leq 3^{1+1/q}\left((\mathbb{E}|A_{jk}|^{2q})^{1/q} + (\mathbb{E}|B_{jk}|^q)^{1/q} + |C_{jk}|\right) \\
&\leq 3^{1+1/q}\left(2^{1/q}(2q/e)\left(\sqrt{\Sigma_{jj}} + \sqrt{\Sigma_{kk}}\right)^2\right. \\
&\qquad \left. + 2^{1/q}\sqrt{q/e}2(\mathbb{E}X_j + \mathbb{E}X_k)(\sqrt{\Sigma_{jj}} + \sqrt{\Sigma_{kk}}) + \mathrm{var}(X_j + X_k)\right).
\end{aligned}$$

Using $\mathrm{var}(X+Y) \leq 2(\mathrm{var}(X) + \mathrm{var}(Y))$ and the fact that $\mathrm{var}(X_j) = \mathrm{var}(X_j - \mathbb{E}X_j) \leq \tau^2(X_j - \mathbb{E}X_j) \leq \Sigma_{jj}$ (by 2) of Lemma 35 and 1) of Lemma 36, we then have

$$\begin{aligned}
&\left(\frac{\mathbb{E}|Z_{jk}|^q}{q!}\right)^{1/q} \\
&\qquad \leq 3^{1+1/q}\frac{2^{3+1/q}(q/e)\max_j \Sigma_{jj} + 2^{3+1/q}\sqrt{q/e}\max_j \mathbb{E}X_j \cdot \max_j \sqrt{\Sigma_{jj}} + 4\max_j \Sigma_{jj}}{(q!)^{1/q}}.
\end{aligned}$$

Since all three coefficients involving q are decreasing in $q \geq 2$, we have

$$\sup_{q \geq 2}\left(\frac{\mathbb{E}|Z_{jk}|^q}{q!}\right)^{1/q} \leq \left(48\sqrt{3}/e + 6\sqrt{6}\right)\max_j \Sigma_{jj} + 24\sqrt{6/e}\max_j \mathbb{E}X_j \max_j \sqrt{\Sigma_{jj}}.$$

Thus by 7) of Lemma 35, letting $B \equiv \left(91\max_j \Sigma_{jj} + 72\max_j \mathbb{E}X_j \max_j \sqrt{\Sigma_{jj}}\right)$, we have for all $\epsilon > 0$:

$$\mathbb{P}\left(n^{-1}\left|\sum_{i=1}^n Z_{jk}^{(i)}\right| \geq 2\epsilon\right) \leq 2\exp\left(-\min\left(\frac{2n\epsilon^2}{B^2}, \frac{n\epsilon}{B}\right)\right).$$

A tail bound for the sample average of V_{jk}^2 can be similarly derived, and the result follows. □

A.3 Simulation Results for Erdös-Rényi Graphs

We revisit the simulations from Section 2.7 but use Erdös-Rényi (ER) graphs in which each possible edge is independently included with probability π. Independent uniform draws from

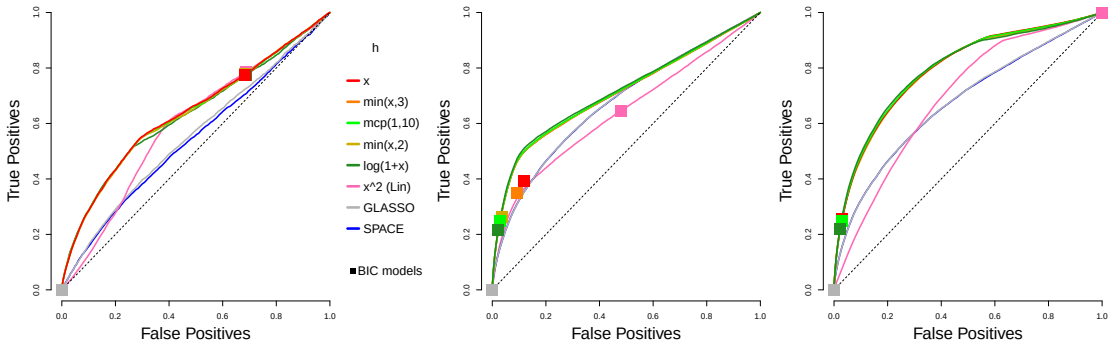
$[0.5, 1]$ are used to fill the non-zero off-diagonal entries of the symmetric matrix $\mathbf{K}_0$. The diagonal elements are set such that $\mathbf{K}_0$ has minimum eigenvalue 0.1. We choose $\pi = 0.08$ for $n = 1000$, and $\pi = 0.02$ for $n = 80$.

A.3.1 Truncated GGMs

In this section we present the results for truncated GGMs.

Choice of $\boldsymbol{h}$

The results for truncated centered GGMs are reported in Table A.1 and Figure A.1. Those for truncated non-centered GGMs using the profiled estimator are in Table A.2 and Figure A.2.



(a) $n = 80$, mult $= 1.8647$ (b) $n = 1000$, mult $= 1$ (c) $n = 1000$, mult $= 1.6438$

Figure A.1: Average ROC curves of our *centered* estimator for $m = 100$ variables and two sample sizes n under various choices of h, compared to SPACE and GLASSO, for the *truncated centered GGM* case. Squares indicate average true positive rate (TPR) and false positive rate (FPR) of models picked by eBIC with refitting for the estimator in the same color.

Centered, $n = 80$, multiplier 1.8647					
$\min(\log(1+x), c)$			$\min(x, c)$		
c	Mean	sd	c	Mean	sd
∞	0.632	0.036	∞	0.638	0.035
2	0.632	0.036	3	0.638	0.035
1	0.630	0.035	2	0.635	0.035
0.5	0.613	0.033	1	0.623	0.033
MCP$(1, c)$			SCAD$(1, c)$		
c	Mean	sd	c	Mean	sd
10	0.637	0.035	10	0.638	0.035
5	0.636	0.036	5	0.637	0.035
1	0.617	0.033	2	0.632	0.035
$x^{1.5}$: (0.627, 0.032)			x^2: (0.595,0.028)		
GLASSO (0.553,0.029)			SPACE: (0.544, 0.026)		
NS: (0.543,0.028)			SJ: (0.519,0.028)		

Centered, $n = 1000$, multiplier 1					
$\min(\log(1+x), c)$			$\min(x, c)$		
c	Mean	sd	c	Mean	sd
∞	0.716	0.016	2	0.710	0.016
2	0.716	0.016	3	0.710	0.016
1	0.715	0.016	1	0.710	0.017
0.5	0.694	0.017	∞	0.709	0.016
MCP$(1, c)$			SCAD$(1, c)$		
c	Mean	sd	c	Mean	sd
5	0.714	0.016	2	0.713	0.016
10	0.711	0.016	5	0.711	0.016
1	0.707	0.017	10	0.710	0.016
$x^{1.5}$: (0.678,0.016)			x^2: (0.64,0.017)		
GLASSO: (0.675,0.016)			SPACE: (0.675,0.016)		
NS: (0.675,0.016)			SJ: (0.624,0.017)		

Centered, $n = 1000$, multiplier 1.6438					
$\min(\log(1+x), c)$			$\min(x, c)$		
c	Mean	sd	c	Mean	sd
∞	0.796	0.014	∞	0.795	0.014
2	0.796	0.014	3	0.794	0.014
1	0.794	0.014	2	0.792	0.014
0.5	0.772	0.015	1	0.784	0.015
MCP$(1, c)$			SCAD$(1, c)$		
c	Mean	sd	c	Mean	sd
5	0.796	0.014	5	0.795	0.014
10	0.796	0.014	10	0.795	0.014
1	0.778	0.015	2	0.793	0.014
$x^{1.5}$: (0.757,0.015)			x^2: (0.693,0.016)		
GLASSO: (0.675,0.016)			SPACE: (0.675,0.016)		
NS: (0.675,0.016)			SJ: (0.624,0.017)		

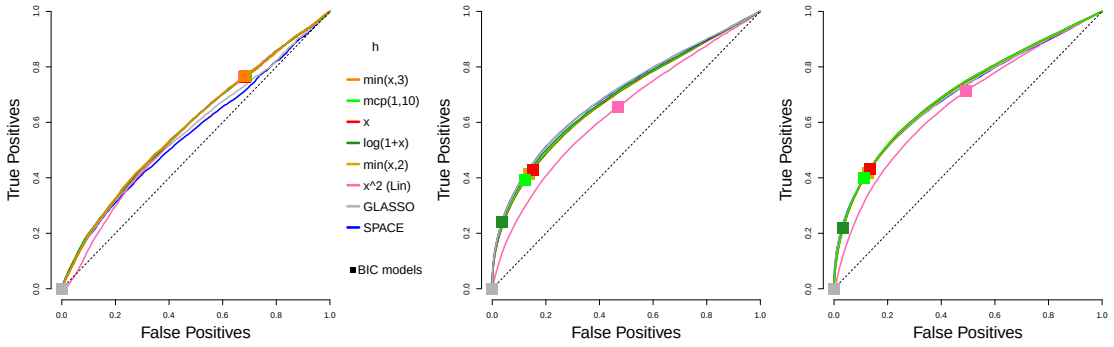
Table A.1: Mean and standard deviation of areas under the ROC curves (AUC) using different estimators in the centered setting, with $n = 80$ and multiplier 1.8647, or $n = 1000$ and multipliers 1 and 1.6438. Methods include our estimator with different choices of h, GLASSO, SPACE, neighborhood selection (NS), and Space JAM (SJ).

Non-centered profiled, $n = 80$, multiplier 1.8647					
$\min(\log(1+x), c)$			$\min(x, c)$		
c	Mean	sd	c	Mean	sd
1	0.588	0.034	3	0.588	0.033
∞	0.588	0.034	∞	0.588	0.033
2	0.588	0.034	2	0.588	0.033
0.5	0.576	0.033	1	0.583	0.033
MCP$(1, c)$			SCAD$(1, c)$		
c	Mean	sd	c	Mean	sd
5	0.588	0.033	5	0.588	0.033
10	0.588	0.033	10	0.588	0.033
1	0.581	0.033	2	0.587	0.033
$x^{1.5}$: (0.582,0.028)			x^2: (0.576,0.028)		
GLASSO: (0.572,0.033)			SPACE: (0.562,0.031)		
NS: (0.560,0.032)			SJ: (0.535,0.027)		

Non-centered profiled, $n = 1000$, multiplier 1					
$\min(\log(1+x), c)$			$\min(x, c)$		
c	Mean	sd	c	Mean	sd
2	0.692	0.022	1	0.687	0.022
∞	0.692	0.022	∞	0.686	0.022
1	0.691	0.022	3	0.685	0.022
0.5	0.684	0.02	2	0.685	0.022
MCP$(1, c)$			SCAD$(1, c)$		
c	Mean	sd	c	Mean	sd
5	0.689	0.022	2	0.687	0.022
1	0.689	0.020	5	0.687	0.022
10	0.687	0.022	10	0.686	0.022
$x^{1.5}$: (0.663,0.020)			x^2: (0.638,0.019)		
GLASSO (0.700,0.022)			SPACE: (0.699,0.022)		
NS: (0.699,0.022)			SJ: (0.655,0.021)		

Non-centered profiled, $n = 1000$, multiplier 1.6438					
$\min(\log(1+x), c)$			$\min(x, c)$		
c	Mean	sd	c	Mean	sd
2	0.705	0.021	∞	0.705	0.022
∞	0.705	0.021	3	0.705	0.021
1	0.703	0.021	2	0.702	0.022
0.5	0.683	0.019	1	0.695	0.021
MCP$(1, c)$			SCAD$(1, c)$		
c	Mean	sd	c	Mean	sd
5	0.706	0.021	10	0.705	0.022
10	0.706	0.022	5	0.705	0.022
1	0.690	0.019	2	0.703	0.022
$x^{1.5}$: (0.689,0.021)			x^2: (0.664,0.019)		
GLASSO (0.700,0.022)			SPACE: (0.699,0.022)		
NS: (0.699,0.022)			SJ: (0.655,0.021)		

Table A.2: Mean and standard deviation of AUC using different profiled estimators in the non-centered setting, with $n = 80$ and multiplier 1.8647, or $n = 1000$ and multipliers 1 and 1.6438. Methods as for Table A.1.

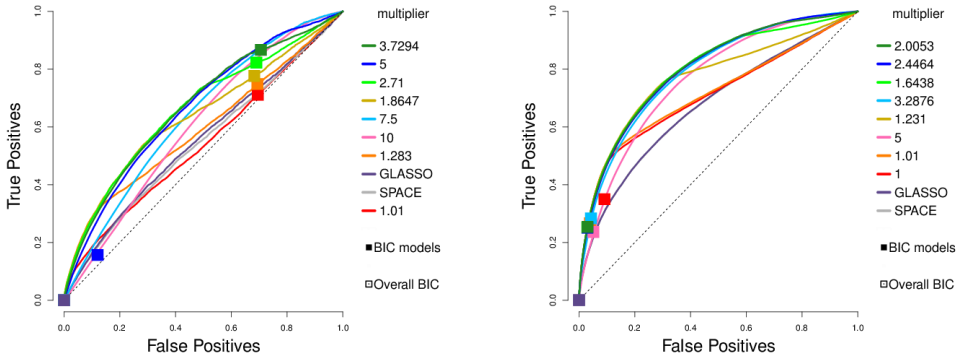


(a) $n = 80$, mult $= 1.8647$ (b) $n = 1000$, mult $= 1$ (c) $n = 1000$, mult $= 1.6438$

Figure A.2: Average for the truncated non-centered GGM case. $n = 80$ or 1000, $m = 100$.

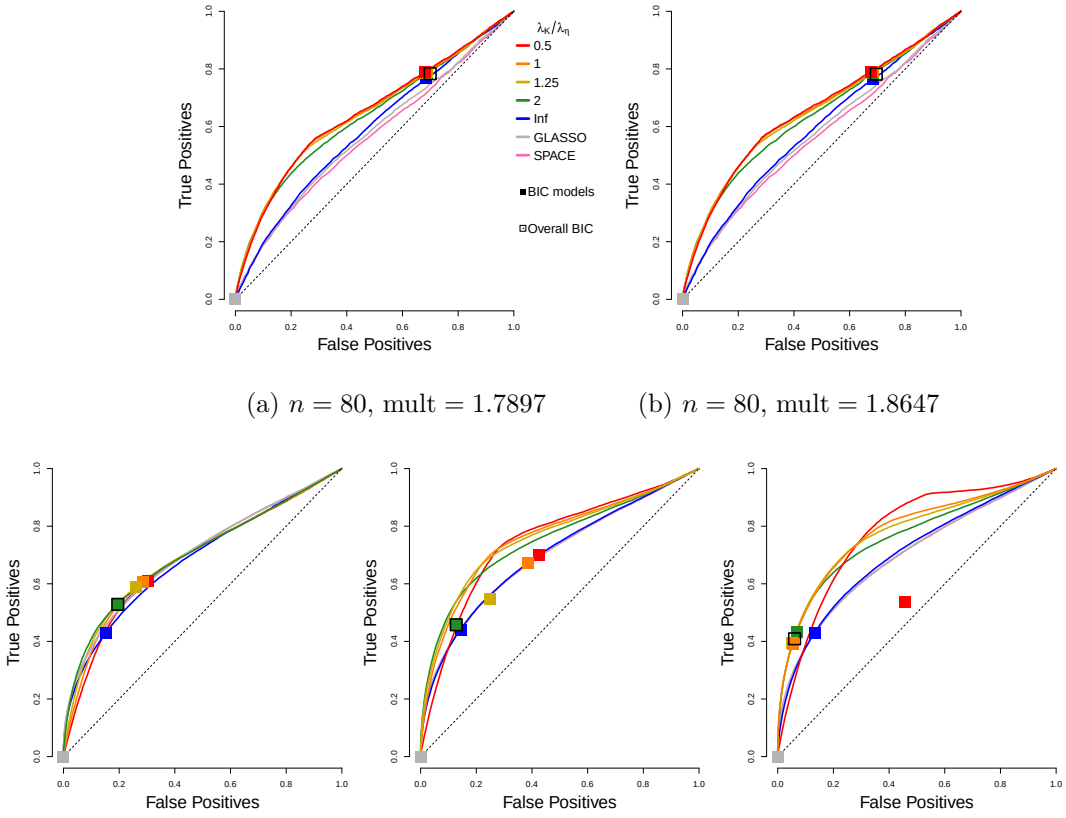
Choice of multiplier

The results for truncated centered GGMs where each curve represents a different multiplier are shown in Figure A.3, and those for truncated non-centered GGMs are in Figure A.4, where each curve corresponds to a different ratio $\lambda_{\mathbf{K}}/\lambda_{\boldsymbol{\eta}}$.



(a) $n = 80$ (b) $n = 1000$

Figure A.3: Performance of $\min(x, 3)$ for truncated centered GGMs with different multipliers, compared to GLASSO and SPACE, in the centered setting, $n = 80$ or 1000.



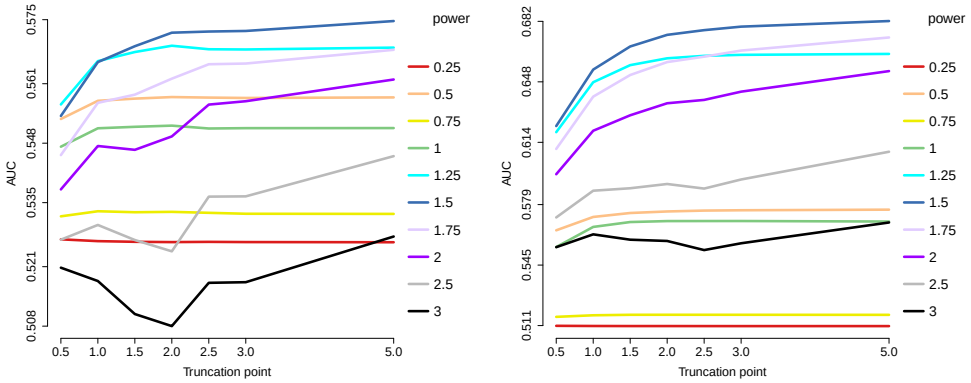
(a) $n = 80$, mult $= 1.7897$ (b) $n = 80$, mult $= 1.8647$

(c) $n = 1000$, mult $= 1$ (d) $n = 1000$, mult $= 1.2310$ (e) $n = 1000$, mult $= 1.6438$

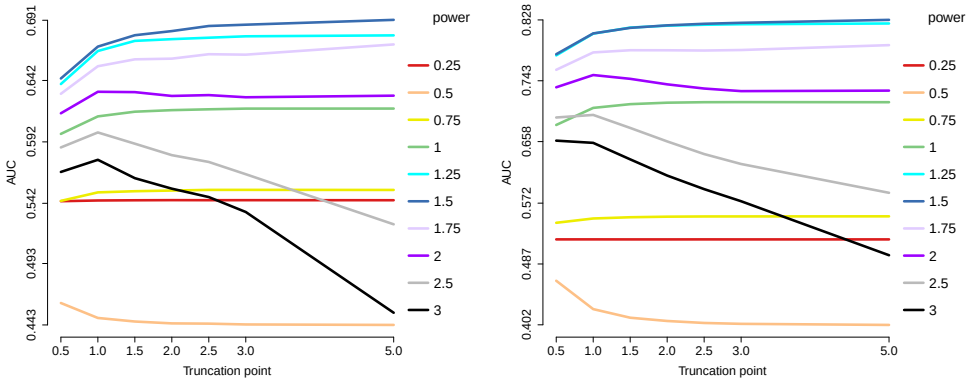
Figure A.4: Performance of the non-centered estimator with $h(x) = \min(x, 3)$. Each curve corresponds to a different choice of $\lambda_{\mathbf{K}}/\lambda_{\boldsymbol{\eta}}$. Squares indicate models picked by eBIC with refit. The square with black outline has the highest eBIC among all models (combinations of $\lambda_{\mathbf{K}}$, $\lambda\boldsymbol{\eta}$). The multipliers correspond to medium or high for $n = 80$, and low, medium and high for $n = 1000$, respectively.

A.3.2 Other a/b Models

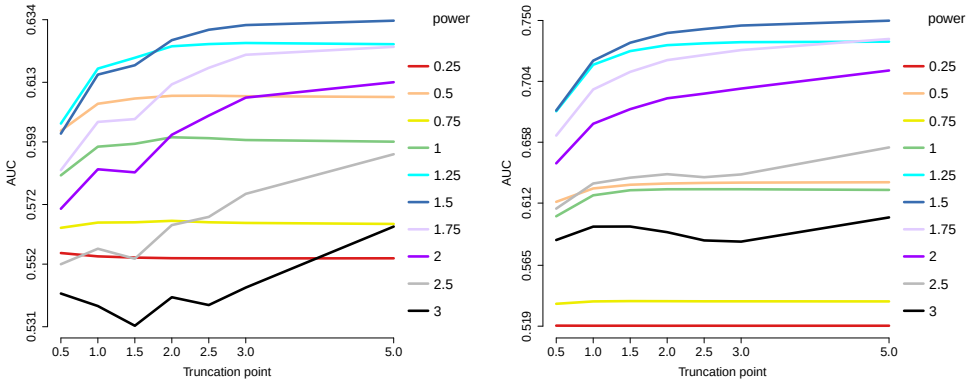
Figure A.5 exhibits the results for the exponential models.



(a) $n = 80$, $\boldsymbol{\eta} = -0.5\mathbf{1}_{100}$, profiled est (b) $n = 1000$, $\boldsymbol{\eta} = -0.5\mathbf{1}_{100}$, profiled est



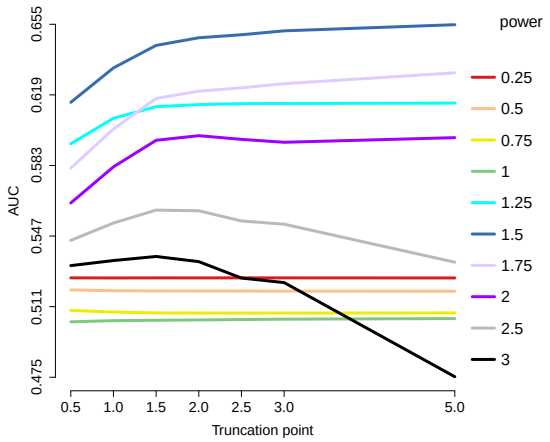
(c) $n = 80$, $\boldsymbol{\eta} = \mathbf{0}$, centered est (d) $n = 1000$, $\boldsymbol{\eta} = \mathbf{0}$, centered est



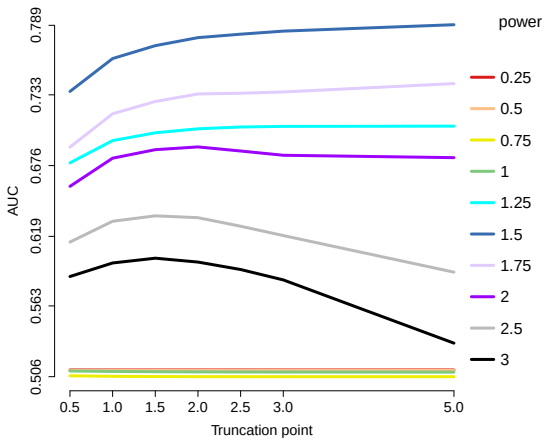
(e) $n = 80$, $\boldsymbol{\eta} = 0.5\mathbf{1}_{100}$, profiled est (f) $n = 1000$, $\boldsymbol{\eta} = 0.5\mathbf{1}_{100}$, profiled est

Figure A.5: AUCs for edge recovery using generalized score matching for exponential models.

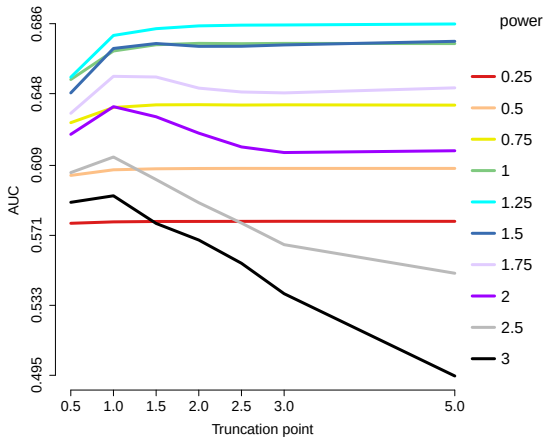
Figure A.6 displays the results for the gamma models.



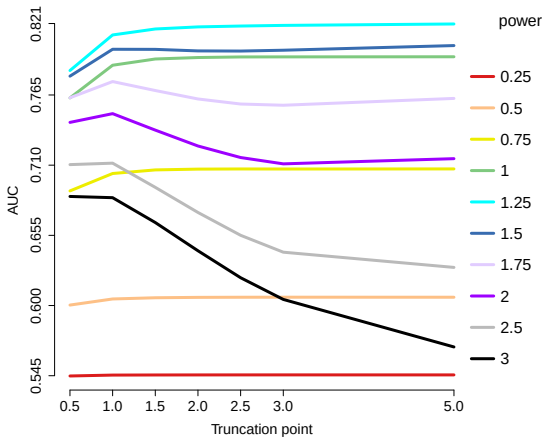
(a) $n = 80$, $\boldsymbol{\eta} = -0.5\mathbf{1}_{100}$, profiled estimator



(b) $n = 1000$, $\boldsymbol{\eta} = -0.5\mathbf{1}_{100}$, profiled estimator



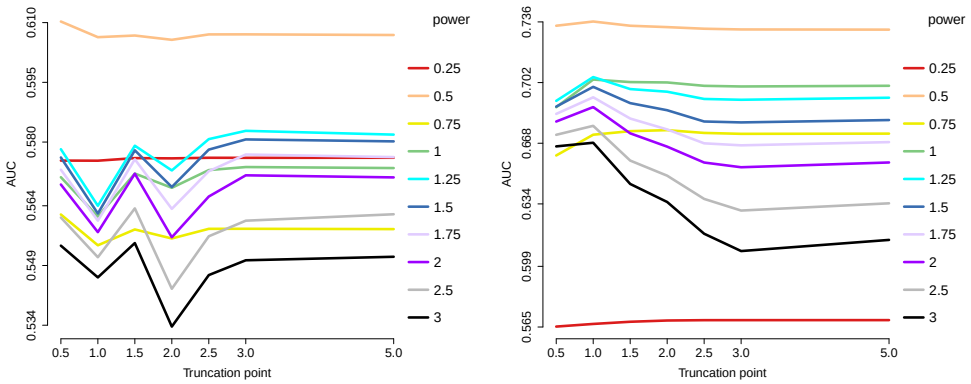
(c) $n = 80$, $\boldsymbol{\eta} = 0.5\mathbf{1}_{100}$, profiled estimator



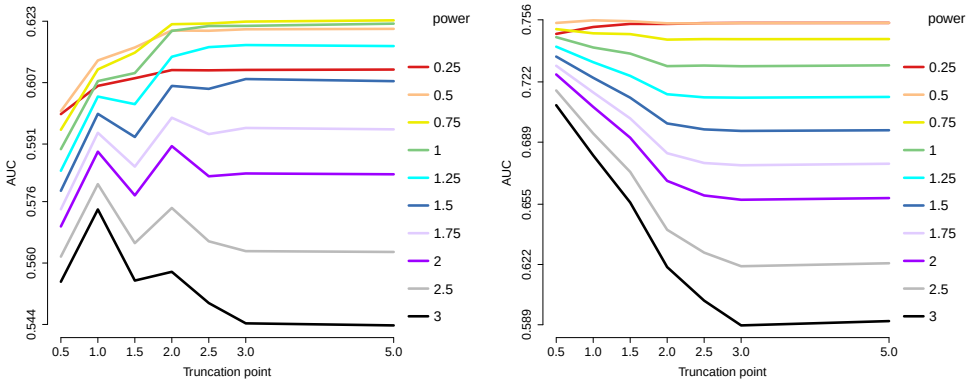
(d) $n = 1000$, $\boldsymbol{\eta} = 0.5\mathbf{1}_{100}$, profiled estimator

Figure A.6: AUCs for edge recovery using generalized score matching for the gamma models.

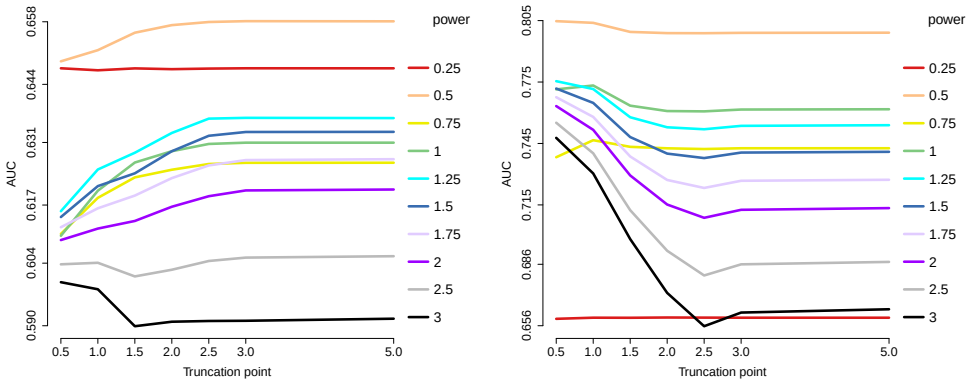
Figures A.7 and A.8 demonstrate the results for $a = 3/2$ and $b = 1/2$ or $b = 0$, respectively.



(a) $n = 80$, $\boldsymbol{\eta} = -0.5\mathbf{1}_{100}$, profiled est (b) $n = 1000$, $\boldsymbol{\eta} = -0.5\mathbf{1}_{100}$, profiled est

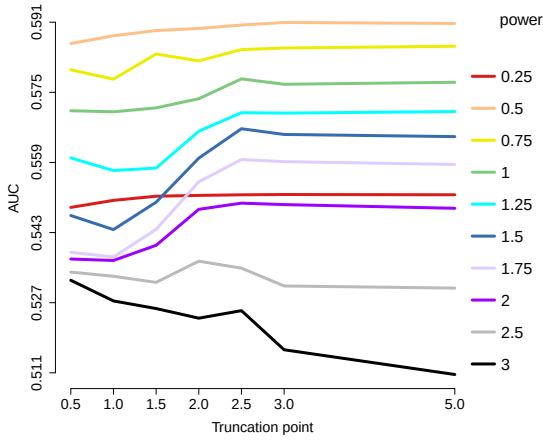


(c) $n = 80$, $\boldsymbol{\eta} = \mathbf{0}$, centered est (d) $n = 1000$, $\boldsymbol{\eta} = \mathbf{0}$, centered est

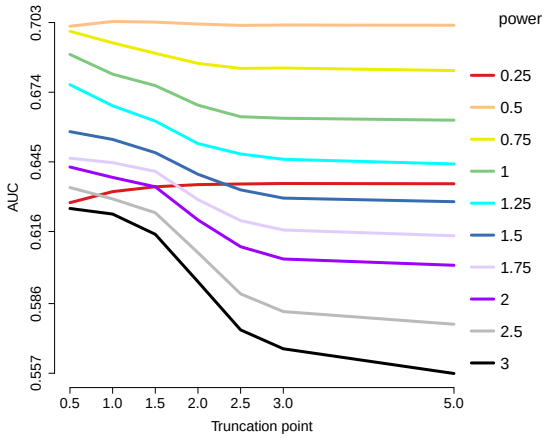


(e) $n = 80$, $\boldsymbol{\eta} = 0.5\mathbf{1}_{100}$, profiled est (f) $n = 1000$, $\boldsymbol{\eta} = 0.5\mathbf{1}_{100}$, profiled est

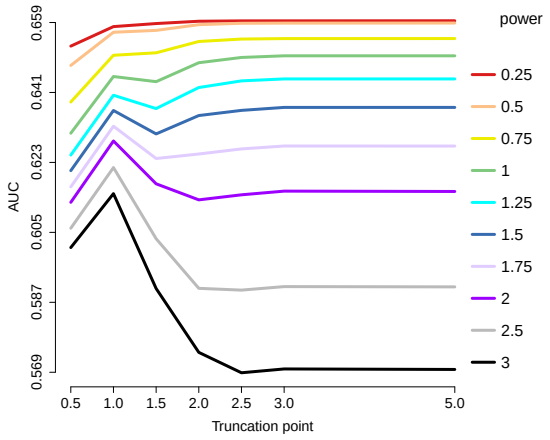
Figure A.7: AUCs for edge recovery using generalized score matching for $a = 3/2$, $b = 1/2$.



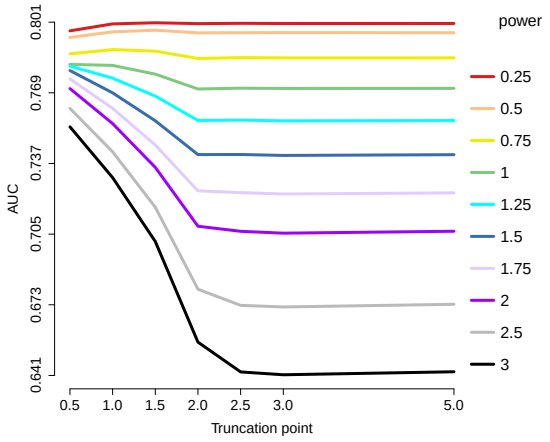
(a) $n = 80$, $\boldsymbol{\eta} = -0.5\mathbf{1}_{100}$, profiled estimator



(b) $n = 1000$, $\boldsymbol{\eta} = -0.5\mathbf{1}_{100}$, profiled estimator



(c) $n = 80$, $\boldsymbol{\eta} = 0.5\mathbf{1}_{100}$, profiled estimator



(d) $n = 1000$, $\boldsymbol{\eta} = 0.5\mathbf{1}_{100}$, profiled estimator

Figure A.8: AUCs for edge recovery using generalized score matching for $a = 3/2$, $b = 0$.

Appendix B

APPENDIX TO CHAPTER 3

Before proving Lemma 13 we first prove the following lemma.

Lemma 38. *Suppose $h_1, \ldots, h_m$ are absolutely continuous in any bounded sub-interval of $\mathbb{R}_+$. Then for any $j = 1, \ldots, m$ and any $\boldsymbol{x}_{-j} \in \mathcal{S}_{-j,\mathcal{D}}$, $(h_j \circ \varphi_j)$ is absolutely continuous in x_j in any bounded sub-interval of $\mathcal{C}_{j,\mathcal{D}}(\boldsymbol{x}_{-j})$.*

Proof of Lemma 38. In the proof we drop the dependency on $\boldsymbol{x}_{-j}$ in notation. By assumption, under Equation 3.5 any bounded sub-interval $[a, b]$ of $\mathcal{C}_{j,\mathcal{D}}(\boldsymbol{x}_{-j})$ must be a sub-interval of $[a_{k,j}, b_{k,j}]$ for some k (for simplicity we do not differentiate among $[a, b]$, $(a, b]$, $[a, b)$ and (a, b) here).

(1) If $a_{k,j} > -\infty$ and $b_{k,j} < +\infty$, denote $C_0 \equiv \min\{C_j, (b_{k,j} - a_{k,j})/2\}$ and rewrite

$$\begin{aligned}
&(h_j \circ \varphi_j)(\boldsymbol{x}) \\
=& h_j\left(\min(C_j, x_j - a_{k,j}, b_{k,j} - x_j)\right) \\
=& h_j(x_j - a_{k,j})\mathbb{1}_{x_j \in [a_{k,j}, a_{k,j}+C_0]} + h_j(C_j)\mathbb{1}_{x_j \in [a_{k,j}+C_0, b_{k,j}-C_0]} + h_j(b_{k,j} - x_j)\mathbb{1}_{x_j \in [b_{k,j}-C_0, b_{k,j}]}.
\end{aligned}$$

Then by absolute continuity of h_j in $[a_{k,j}, b_{k,j}]$ it is apparent that $(h_j \circ \varphi_j)$ is differentiable in x_j a.e. with derivative

$$\partial_j(h_j \circ \varphi_j)(\boldsymbol{x}) = h_j'(x_j - a_{k,j})\mathbb{1}_{x_j \in [a_{k,j}, a_{k,j}+C_0]} - h_j'(b_{k,j} - x_j)\mathbb{1}_{x_j \in [b_{k,j}-C_0, b_{k,j}]}.$$

Then by the absolute continuity of h_j again, for $x_j \in [a_{k,j}, b_{k,j}]$,

$$\begin{aligned}
&\int_{a_{k,j}}^{x_j} \partial_j(h_j \circ \varphi_j)(t_j; \boldsymbol{x}_{-j})\, \mathrm{d}t_j \\
=& h_j(x_j - a_{k,j})\mathbb{1}_{x_j \in [a_{k,j}, a_{k,j}+C_0]} + h_j(C_j)\mathbb{1}_{x_j \in [a_{k,j}+C_0, b_{k,j}-C_0]} + h_j(b_{k,j} - x_j)\mathbb{1}_{x_j \in [b_{k,j}-C_0, b_{k,j}]} \\
=& (h_j \circ \varphi_j)(\boldsymbol{x}),
\end{aligned}$$

which proves that $(h_j \circ \varphi_j)(\boldsymbol{x})$ is absolutely continuous in x_j in $[a_{k,j}, b_{k,j}]$, and hence in $[a, b] \subset [a_{k,j}, b_{k,j}]$.

(2) If $a_{k,j} > -\infty$ and $b_{k,j} = +\infty$, on $[a, b]$ $(h_j \circ \varphi_j)(\boldsymbol{x}) = h_j(\min(C_j, x_j - a_{k,j}))$ is an absolutely continuous function in a linear function of x_j truncated above by C_j, and is thus trivially absolutely continuous in $[a, b]$.

(3) If $a_{k,j} = -\infty$ and $b_{k,j} < +\infty$, on $[a,b]$ $(h_j \circ \varphi_j)(\boldsymbol{x}) = h_j(\min(C_j, b_{k,j} - x_j))$ is an absolutely continuous function in a linear function of x_j truncated above by C_j, and is thus trivially absolutely continuous in $[a,b]$.

(4) If $a_{k,j} = -\infty$ and $b_{k,j} = +\infty$, $(h_j \circ \varphi_j)(\boldsymbol{x}) = h_j(C_j)$ is constant and hence trivially absolutely continuous in $[a,b]$.

□

Proof of Lemma 13. By simple manipulation

$$\begin{aligned}
J_{\boldsymbol{h},\boldsymbol{C},\mathcal{D}}(p) &\equiv \frac{1}{2}\int_{\mathcal{D}} p_0(\boldsymbol{x}) \left\| \nabla \log p(\boldsymbol{x}) \odot (\boldsymbol{h} \circ \boldsymbol{\varphi})^{1/2}(\boldsymbol{x}) - \nabla \log p_0(\boldsymbol{x}) \odot (\boldsymbol{h} \circ \boldsymbol{\varphi})^{1/2}(\boldsymbol{x}) \right\|_2^2 \mathrm{d}\boldsymbol{x} \\
&= \frac{1}{2}\sum_{j=1}^{m} \int_{\mathcal{D}} p_0(\boldsymbol{x})(h_j \circ \varphi_j)(\boldsymbol{x}) \left(\partial_j \log p_0(\boldsymbol{x}) - \partial_j \log p(\boldsymbol{x})\right)^2 \mathrm{d}\boldsymbol{x} \\
&= \frac{1}{2}\sum_{j=1}^{m} \int_{\mathcal{D}} p_0(\boldsymbol{x})(h_j \circ \varphi_j)(\boldsymbol{x}) \left(\partial_j \log p(\boldsymbol{x})\right)^2 \mathrm{d}\boldsymbol{x} \\
&\quad - \sum_{j=1}^{m} \int_{\mathcal{D}} p_0(\boldsymbol{x})(h_j \circ \varphi_j)(\boldsymbol{x}) \partial_j \log p_0(\boldsymbol{x}) \partial_j \log p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + \text{const.}
\end{aligned} \tag{B.1}$$

By (B.1) it suffices to prove for all $j = 1, \ldots, m$ that

$$\int_{\mathcal{D}} p_0(\boldsymbol{x})(h_j \circ \varphi_j)(\boldsymbol{x}) \partial_j \log p_0(\boldsymbol{x}) \partial_j \log p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = -\int_{\mathcal{D}} p_0(\boldsymbol{x}) \partial_j \left[(h_j \circ \varphi_j)(\boldsymbol{x}) \partial_j \log p(\boldsymbol{x})\right] \mathrm{d}\boldsymbol{x}. \tag{B.2}$$

$\int_{\mathcal{D}} p_0(\boldsymbol{x}) \left\| \nabla \log p(\boldsymbol{x}) \odot (\boldsymbol{h} \circ \boldsymbol{\varphi})^{1/2}(\boldsymbol{x}) \right\|_2^2 \mathrm{d}\boldsymbol{x}$ and $\int_{\mathcal{D}} p_0(\boldsymbol{x}) \left\| \nabla \log p_0(\boldsymbol{x}) \odot (\boldsymbol{h} \circ \boldsymbol{\varphi})^{1/2}(\boldsymbol{x}) \right\|_2^2 \mathrm{d}\boldsymbol{x}$ are both finite by assumption, so by $|2ab| \leq a^2 + b^2$ the integrand in the left-hand side of (B.2) is integrable. Then by Fubini-Tonelli

$$\begin{aligned}
&\int_{\mathcal{D}} p_0(\boldsymbol{x})(h_j \circ \varphi_j)(\boldsymbol{x}) \partial_j \log p_0(\boldsymbol{x}) \partial_j \log p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\
&= \int_{\mathcal{S}_{-j}} \int_{\mathcal{C}_j(\boldsymbol{x}_{-j})} \underbrace{(h_j \circ \varphi_j)(\boldsymbol{x}) \partial_j p_0(\boldsymbol{x}) \partial_j \log p(\boldsymbol{x})}_{\equiv f(\boldsymbol{x})} \mathrm{d}x_j \, \mathrm{d}\boldsymbol{x}_{-j} \\
&= \int_{\mathcal{S}_{-j}} \int_{\mathbb{R}} \mathbb{1}_{\mathcal{C}_j(\boldsymbol{x}_{-j})}(x_j) f(\boldsymbol{x}) \mathrm{d}x_j \, \mathrm{d}\boldsymbol{x}_{-j}
\end{aligned}$$

$$= \int_{\mathcal{S}_{-j}} \int_{\mathbb{R}} \left[\sum_{k=1}^{K_j(\boldsymbol{x}_{-j})} \mathbb{1}_{[a_{k,j}(\boldsymbol{x}_{-j}),\, b_{k,j}(\boldsymbol{x}_{-j})]}(x_j) \right] f(x_j; \boldsymbol{x}_{-j}) \, \mathrm{d}x_j \, \mathrm{d}\boldsymbol{x}_{-j}$$

$$= \int_{\mathcal{S}_{-j}} \left[\sum_{k=1}^{K_j(\boldsymbol{x}_{-j})} \int_{a_{k,j}(\boldsymbol{x}_{-j})}^{b_{k,j}(\boldsymbol{x}_{-j})} f(x_j; \boldsymbol{x}_{-j}) \, \mathrm{d}x_j \right] \mathrm{d}\boldsymbol{x}_{-j} \tag{B.3}$$

where the interchangeability of integration and (potentially infinite) summation is justified by Fubini-Tonelli again. Then using the decomposition of the domain in (3.5) while omitting the dependency of $a_{k,j}$ and $b_{k,j}$ on $\boldsymbol{x}_{-j}$ in notation, for a.e. $\boldsymbol{x}_{-j} \in \mathcal{S}_{-j}$ and any $k = 1, \ldots, K_j(\boldsymbol{x}_{-j})$ we have

$$\begin{aligned}
\int_{a_{k,j}}^{b_{k,j}} f(\boldsymbol{x}) \, \mathrm{d}x_j &= \int_{a_{k,j}}^{b_{k,j}} (h_j \circ \varphi_j)(\boldsymbol{x}) \partial_j p_0(\boldsymbol{x}) \partial_j \log p(\boldsymbol{x}) \, \mathrm{d}x_j \\
&= \lim_{x_j \nearrow b_{k_j}^-} (h_j \circ \varphi_j)(\boldsymbol{x}) p_0(\boldsymbol{x}) \partial_j \log p(\boldsymbol{x}) - \lim_{x_j \searrow a_{k_j}^+} (h_j \circ \varphi_j)(\boldsymbol{x}) p_0(\boldsymbol{x}) \partial_j \log p(\boldsymbol{x}) \\
&\qquad - \int_{a_{k,j}}^{b_{k,j}} p_0(\boldsymbol{x}) \partial_j \left[(h_j \circ \varphi_j)(\boldsymbol{x}) \partial_j \log p(\boldsymbol{x}) \right] \mathrm{d}x_j \\
&= - \int_{a_{k,j}}^{b_{k,j}} p_0(\boldsymbol{x}) \partial_j \left[(h_j \circ \varphi_j)(\boldsymbol{x}) \partial_j \log p(\boldsymbol{x}) \right] \mathrm{d}x_j,
\end{aligned}$$

by integration by parts and by Assumption (A1) on the limits going to 0. The integration by parts is justified by the fundamental theorem of calculus for absolutely continuous functions (Lemma 38) as well as the product rule (cf. proof of Lemma 34). Thus, by going backwards using Fubini-Tonelli twice again, (B.3) becomes

$$\begin{aligned}
&\int_{\mathcal{S}_{-j}} \left\{ - \sum_{k=1}^{K_j(\boldsymbol{x}_{-j})} \int_{a_{k,j}(\boldsymbol{x}_{-j})}^{b_{k,j}(\boldsymbol{x}_{-j})} p_0(\boldsymbol{x}) \partial_j \left[(h_j \circ \varphi_j)(\boldsymbol{x}) \partial_j \log p(\boldsymbol{x}) \right] \mathrm{d}x_j \right\} \mathrm{d}\boldsymbol{x}_{-j} \\
= & - \int_{\mathcal{S}_{-j}} \int_{\mathcal{C}_j(\boldsymbol{x}_{-j})} p_0(\boldsymbol{x}) \partial_j \left[(h_j \circ \varphi_j)(\boldsymbol{x}) \partial_j \log p(\boldsymbol{x}) \right] \mathrm{d}x_j \, \mathrm{d}\boldsymbol{x}_{-j} \\
= & - \int_{\mathcal{D}} p_0(\boldsymbol{x}) \partial_j \left[(h_j \circ \varphi_j)(\boldsymbol{x}) \partial_j \log p(\boldsymbol{x}) \right] \mathrm{d}\boldsymbol{x},
\end{aligned}$$

proving (B.2). □

Proof of Theorem 16. Note that the condition that ${\boldsymbol{v}^a}^\top \mathbf{K} \boldsymbol{v}^a > 0 \ \forall \boldsymbol{v} \in \mathcal{D} \backslash \{\mathbf{0}\}$ implies that ${\boldsymbol{v}^a}^\top \mathbf{K} \boldsymbol{v}^a > 0 \ \forall \boldsymbol{v} \in \mathcal{D}_+ \equiv \{\boldsymbol{v}/\|\boldsymbol{v}\|_2 : \boldsymbol{v} \in \mathcal{D} \backslash \{\mathbf{0}\}\} \subseteq \{\boldsymbol{v} \in \mathbb{R}^m : \|\boldsymbol{v}\|_2 = 1\} \equiv \mathbb{S}^{m-1}$ with $\mathbb{S}^{m-1}$

compact, so

$$N_{\mathbf{K}} \equiv \inf_{\boldsymbol{v}\in\mathcal{D}\backslash\{\mathbf{0}\}} \boldsymbol{v}^{a\top}\mathbf{K}\boldsymbol{v}^a/\boldsymbol{v}^{a\top}\boldsymbol{v}^a = \inf_{\boldsymbol{v}\in\mathcal{D}_+} \boldsymbol{v}^{a\top}\mathbf{K}\boldsymbol{v}^a/\boldsymbol{v}^{a\top}\boldsymbol{v}^a \geq \inf_{\boldsymbol{v}\in\mathbb{S}^{m-1}} \boldsymbol{v}^{a\top}\mathbf{K}\boldsymbol{v}^a/\boldsymbol{v}^{a\top}\boldsymbol{v}^a > 0.$$

(1) *Case $a > 0$ and $b > 0$ (CC1, CC2):* Since p is bounded everywhere, it is integrable over a bounded $\mathcal{D}$ (proving (CC1)). Otherwise, assume $\mathcal{D}$ is unbounded. If either a or b is non-integer, then $\mathcal{D} \subset \mathbb{R}^m_+$ and a sufficient condition is $\boldsymbol{v}^a\mathbf{K}\boldsymbol{v}^a > 0\ \forall \boldsymbol{v} \in \mathcal{D}\backslash\{\mathbf{0}\}$ and either $\boldsymbol{\eta}^\top\boldsymbol{v}^b \leq 0\ \forall \boldsymbol{v} \in \mathcal{D}$ or $2a > b > 0$, corresponding to (i) and (ii) in the Proof of Theorem 5 in Section A.1.3, respectively. If a and b are both integers, $\mathcal{D}$ may be any subset of $\mathbb{R}^m$ and the same sufficient condition can be implied following the same proof in Section A.1.3, with integration over $(-\infty, +\infty)$ instead of $(0, +\infty)$. This proves (CC2).

(2) *Case $a > 0$ and $b = 0$ (CC3):* By definition $\mathcal{D} \subseteq \mathbb{R}^m_+$. If $\mathcal{D}$ is bounded, $-\frac{1}{2a}\boldsymbol{x}^{a\top}\mathbf{K}\boldsymbol{x}^a$ as a continuous function is bounded, and it thus suffices to bound $\int_{\mathcal{D}} \exp\left(\boldsymbol{\eta}^\top \log(\boldsymbol{x})\right) \mathrm{d}\boldsymbol{x} = \int_{\mathcal{D}} \prod_{j=1}^m x_j^{\eta_j}\, \mathrm{d}\boldsymbol{x} \leq \prod_{j=1}^m \int_{\rho_j(\mathcal{D})} x_j^{\eta_j}\, \mathrm{d}x_j < +\infty$ if $\eta_j > -1$ for all j such that $0 \in \rho_j(\mathcal{D})$, where for the $\leq$ step we used the fact that $x_j > 0$. This proves (CC3) (i).

If $\mathcal{D}$ is unbounded and $\boldsymbol{v}^{a\top}\mathbf{K}\boldsymbol{v}^a > 0$ for all $\boldsymbol{v} \in \mathcal{D}\backslash\{\mathbf{0}\}$, using the fact that $\exp(\cdots) > 0$,

$$\begin{aligned}\int_{\mathcal{D}} p_{\boldsymbol{\eta},\mathbf{K}}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} &= \int_{\mathcal{D}} \exp\left(-\boldsymbol{x}^{a\top}\mathbf{K}\boldsymbol{x}^a/(2a) + \boldsymbol{\eta}^\top \log(\boldsymbol{x})\right) \mathrm{d}\boldsymbol{x} \\ &\leq \prod_{j=1}^m \int_{\rho_j(\mathcal{D})} \exp\left(-N_{\mathbf{K}} x_j^{2a}/(2a) + \eta_j \log(x_j)\right) \mathrm{d}x_j.\end{aligned}$$

Note that the indefinite integral of the last display is

$$-\frac{1}{2a} x^{1+\eta_j} \left(\frac{N_{\mathbf{K}}}{2a} x^{2a}\right)^{-(1+\eta_j)/(2a)} \Gamma\left[\frac{1+\eta_j}{2a}, \frac{N_{\mathbf{K}} x^{2a}}{2a}\right]$$

so the definite integral is finite if and only if $\eta_j > -1$ for all j s.t. $0 \in \rho_j(\mathcal{D})$. This proves (CC3) (ii).

If $\mathcal{D}$ is unbounded and $\boldsymbol{v}^{a\top}\mathbf{K}\boldsymbol{v}^a \geq 0$ for all $\boldsymbol{v} \in \mathcal{D}$, $\int_{\mathcal{D}} p_{\boldsymbol{\eta},\mathbf{K}}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \leq \prod_{j=1}^m \int_{\rho_j(\mathcal{D})} x_j^{\eta_j}\, \mathrm{d}x_j < \infty$ if $\eta_j > -1$ for all j s.t. $0 \in \rho_j(\mathcal{D})$ and $\eta_j < -1$ for all j s.t. $\rho_j(\mathcal{D})$ is unbounded. This proves (CC3) (iii).

(3) *Case $a = 0$, $\mathcal{D}$ is bounded and $0 \notin \rho_j(\mathcal{D})$ for all j (CC4):* If $\mathcal{D}$ is bounded and $0 \notin \rho_j(\mathcal{D})$ for all j, then $\log(\mathcal{D})$ is bounded, and since the integrand is continuous and bounded, the integral is finite without any further requirements.

(4) *Case $a = 0$ and $b = 0$ (CC5):* Assume $\log(\boldsymbol{x})^\top \mathbf{K} \log(\boldsymbol{x}) > 0$ for all $\boldsymbol{x} \in \mathcal{D}$, then

$$\begin{aligned}
\int_{\mathcal{D}} p_{\boldsymbol{\eta},\mathbf{K}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} &= \int_{\mathcal{D}} \exp\left(-\frac{1}{2}\log(\boldsymbol{x})^\top \mathbf{K} \log(\boldsymbol{x}) + \boldsymbol{\eta}^\top \log(\boldsymbol{x})\right) \mathrm{d}\boldsymbol{x} \\
&= \int_{\log(\mathcal{D})} \exp\left(-\frac{1}{2}\boldsymbol{x}^\top \mathbf{K}\boldsymbol{x} + (\boldsymbol{\eta} + \mathbf{1}_m)^\top \boldsymbol{x}\right) \mathrm{d}\boldsymbol{x} \\
&< \prod_{j=1}^{m} \int_{\log(\rho_j(\mathcal{D}))} \exp\left(-N_{\mathbf{K}} x_j^2/2 + (\eta_j + 1)x_j\right) \mathrm{d}x_j \\
&< \prod_{j=1}^{m} \int_{-\infty}^{\infty} \exp\left(-N_{\mathbf{K}} x_j^2/2 + (\eta_j + 1)x_j\right) \mathrm{d}x_j < +\infty
\end{aligned}$$

since the integrand is proportional to a univariate Gaussian density.

(5) *Case $a = 0$ and $b > 0$ (CC6, CC7):* Assume $\log(\boldsymbol{x})^\top \mathbf{K} \log(\boldsymbol{x}) > 0$ for all $\boldsymbol{x} \in \mathcal{D}$ and $\eta_j \leq 0$ for all j s.t. $\rho_j(\mathcal{D})$ is unbounded (from above). Then

$$\begin{aligned}
\int_{\mathcal{D}} p_{\boldsymbol{\eta},\mathbf{K}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} &= \int_{\mathcal{D}} \exp\left(-\frac{1}{2}\log(\boldsymbol{x})^\top \mathbf{K} \log(\boldsymbol{x}) + \boldsymbol{\eta}^\top \boldsymbol{x}^b\right) \mathrm{d}\boldsymbol{x} \\
&= \int_{\log(\mathcal{D})} \exp\left(-\frac{1}{2}\boldsymbol{x}^\top \mathbf{K}\boldsymbol{x} + \mathbf{1}_m^\top \boldsymbol{x} + \boldsymbol{\eta}^\top \exp(b\boldsymbol{x})\right) \mathrm{d}\boldsymbol{x} \\
&< \prod_{j=1}^{m} \int_{\log(\rho_j(\mathcal{D}))} \exp\left(-N_{\mathbf{K}} x_j^2/2 + x_j + \eta_j \exp(bx_j)\right) \mathrm{d}x_j \\
&\leq \prod_{j=1}^{m} \int_{-\infty}^{\infty} c_j \exp\left(-N_{\mathbf{K}} x_j^2/2 + x_j\right) \mathrm{d}x_j < +\infty,
\end{aligned}$$

where $c_j \equiv 1$ if $\eta_j \leq 0$ or $c_j \equiv \exp\left(\eta_j \left(\sup \rho_j(\mathcal{D})\right)^b\right) > +\infty$ otherwise. This proves (CC6).

Finally, if $\log(\mathcal{D})$ is unbounded and $\log(\boldsymbol{x})^\top \mathbf{K} \log(\boldsymbol{x}) \geq 0$ for all $\boldsymbol{x} \in \mathcal{D}$, the integral is bounded by

$$\prod_{j=1}^{m} \int_{\log(\rho_j(\mathcal{D}))} \exp\left(x_j + \eta_j \exp(bx_j)\right) \mathrm{d}x_j$$

which is finite if and only if $\eta_j < 0$ for all j s.t. $\rho_j(\mathcal{D})$ is unbounded (from above). This proves (CC7).

□

Proof of Theorem 18. It suffices to consider the case $\mathcal{D} = \mathbb{R}^m_+$ for general a and b as well as $\mathcal{D} = \mathbb{R}^m$ for integer $a > 0$ and $b > 0$ (so that (3.17) is well defined on $\mathbb{R}^m$): For (A.1), the irregularities only occur at the boundary points, but with the composition $(h_j \circ \varphi_j)(\boldsymbol{x})$ with x_j approaching any finite boundary point behaves like $h_j(x_j)$ with $x_j \searrow 0^+$ in $\mathcal{D} = \mathbb{R}^m_+$, and $(h_j \circ \varphi_j)(\boldsymbol{x})$ with $x_j \to \infty$ behaves like $h_j(x_j)$ with $x_j \to \infty$ in $\mathcal{D} = \mathbb{R}^m_+$ (or $\mathbb{R}^m$ if applicable). For (A.2), obviously integrability over $\mathcal{D}$ follows from that over $\mathcal{D} = \mathbb{R}^m_+$ or $\mathbb{R}^m$. (A.3) is trivially satisfied by a power function h_j.

As in the proof of Theorem 16, $N_{\mathbf{K}} \equiv \inf_{\boldsymbol{v} \in \mathcal{D}} \boldsymbol{v}^{a\top} \mathbf{K} \boldsymbol{v}^a / \boldsymbol{v}^{a\top} \boldsymbol{v}^a > 0$.

(1) The case for $a > 0$ and $b \geq 0$ and $\mathcal{D} = \mathbb{R}^m_+$ is covered in Chapter 2. The proof for the case for $a > 0$ and $b > 0$ and $\mathcal{D} = \mathbb{R}^m$ is analogous and omitted.

(2) *Case $a = 0$ and $b = 0$:*

$$
\begin{aligned}
&|p_0(\boldsymbol{x}) \partial_j \log p(\boldsymbol{x})| \\
&\propto \exp\left(-\frac{1}{2}\log(\boldsymbol{x})^\top \mathbf{K}_0 \log(\boldsymbol{x}) + \boldsymbol{\eta}_0^\top \log(\boldsymbol{x})\right) \left| x_j^{-1}\left(\eta_j - \boldsymbol{\kappa}_{j,-j}^\top \log(\boldsymbol{x}_{-j})\right) - \kappa_{jj} x_j^{-1} \log x_j \right| \\
&\leq \left| \left(\eta_j - \boldsymbol{\kappa}_{j,-j}^\top \log \boldsymbol{x}_{-j}\right) \exp\left[-N_{\mathbf{K}_0}(\log x_j)^2/2 + (\eta_j - 1)\log x_j\right] \right. \\
&\left. -\kappa_{jj} \exp\left[-N_{\mathbf{K}_0}(\log x_j)^2/2 + (\eta_j - 1)\log x_j\right] \log x_j \right| \times \prod_{k \neq m} \exp\left(-N_{\mathbf{K}_0}(\log x_k)^2/2 + \eta_j \log x_k\right) \\
&\propto \mathcal{O}\left[\exp\left(-N_{\mathbf{K}_0} y_j^2/2 + (\eta_j - 1) y_j\right)\right] + \mathcal{O}\left[\exp\left(-N_{\mathbf{K}_0} y_j^2/2 + (\eta_j - 1) y_j\right) y_j\right]
\end{aligned}
$$

which apparently vanishes as $x_j \searrow 0^+$ and $x_j \nearrow +\infty$ with $y_j \equiv \log(x_j)$ since it is dominated by a constant times a Gaussian density in y_j. Thus, by Proposition 14, (A.1) is satisfied with any $\alpha_j \geq 0$. Likewise, for (A.2),

$$
\begin{aligned}
&\int_{\mathbb{R}^m_+} p_0(\boldsymbol{x}) \left\| \nabla \log p(\boldsymbol{x}) \odot (\boldsymbol{h} \circ \boldsymbol{\varphi})^{1/2}(\boldsymbol{x}) \right\|_2^2 \mathrm{d}\boldsymbol{x} \\
\leq &\,\mathrm{const} \cdot \sum_{j=1}^m \int_{\mathbb{R}^m_+} \prod_{k=1}^m \exp\left[-N_{\mathbf{K}_0}(\log x_k)^2/2 + \eta_k \log(x_k)\right] \times \\
&\qquad h_j(x_j) \left[x_j^{-1}\left(\eta_j - \boldsymbol{\kappa}_{j,-j}^\top \log(\boldsymbol{x}_{-j})\right) - \kappa_{jj} x_j^{-1} \log x_j\right]^2 \mathrm{d}\boldsymbol{x},
\end{aligned}
$$

which can be decomposed into a sum of products of univariate integrals of the form

$$\text{const} \cdot \exp\left(-N_{\mathbf{K}_0}(\log x_j)^2/2 + A\log(x_j)\right)(\log x_j)^B (h_j(x_j))^C$$

with $B = 0, 1, 2$, $C = 0, 1$, and constants A. With $h_j(x_j) = x_j^{\alpha_j}$ for any $\alpha_j \geq 0$ this is bounded by some Gaussian density in $\log x_j$, and thus $\int_{\mathbb{R}_+^m} p_0(\boldsymbol{x}) \left\|\nabla \log p(\boldsymbol{x}) \odot (\boldsymbol{h} \circ \boldsymbol{\varphi})^{1/2}(\boldsymbol{x})\right\|_2^2 \mathrm{d}\boldsymbol{x} < +\infty$. Proof for $\int_{\mathbb{R}_+^m} p_0(\boldsymbol{x}) \left\|[\nabla \log p(\boldsymbol{x}) \odot (\boldsymbol{h} \circ \boldsymbol{\varphi})(\boldsymbol{x})]'\right\|_1 \mathrm{d}\boldsymbol{x} < +\infty$ is similar and is omitted.

(3) *Case $a = 0$ and $b > 0$:* Recall $\rho_j(\mathcal{D}) \equiv \overline{\{x_j : \boldsymbol{x} \in \mathcal{D}\}}$. Let $\rho_j^*(\mathcal{D}) \equiv \sup \rho_j(\mathcal{D})$. Since we assume that $\eta_j \leq 0$ for any j such that $\rho_j^*(\mathcal{D}) < +\infty$,

$$\begin{aligned}
&p_0(\boldsymbol{x})\partial_j \log p(\boldsymbol{x}) \\
&\propto \exp\left(-\frac{1}{2}\log(\boldsymbol{x})^\top \mathbf{K}_0 \log(\boldsymbol{x}) + \frac{1}{b}\boldsymbol{\eta}_0^\top \boldsymbol{x}^b\right)\left[\eta_j x_j^{b-1} - x_j^{-1}\boldsymbol{\kappa}_{j,-j}^\top \log(\boldsymbol{x}_{-j}) - \kappa_{jj} x_j^{-1}\log x_j\right] \\
&\leq \exp\left(-\frac{1}{2}\log(\boldsymbol{x})^\top \mathbf{K}_0 \log(\boldsymbol{x}) + \frac{1}{b}\sum_{j:\rho_j^*(\mathcal{D})<+\infty} \eta_{0j}(\rho_j^*(\mathcal{D}))^b\right) \times \\
&\qquad \left[-x_j^{-1}\boldsymbol{\kappa}_{j,-j}^\top \log(\boldsymbol{x}_{-j}) - \kappa_{jj} x_j^{-1}\log x_j\right] \\
&\propto \exp\left(-\frac{1}{2}\log(\boldsymbol{x})^\top \mathbf{K}_0 \log(\boldsymbol{x})\right)\left[-x_j^{-1}\boldsymbol{\kappa}_{j,-j}^\top \log(\boldsymbol{x}_{-j}) - \kappa_{jj} x_j^{-1}\log x_j\right]
\end{aligned}$$

is bounded by the corresponding quantity in the $a = b = 0$ case with $\boldsymbol{\eta} = \mathbf{0}_m$, and (A.1) is thus satisfied. Similarly, the two quantities for (A.2) are bounded by a constant times those in the $a = b = 0$ case with $\boldsymbol{\eta} = \mathbf{0}_m$ and (A.2) is thus also satisfied. □

Proof of Theorem 19. Fix $j = 1, \ldots, m-1$ and $\boldsymbol{x}_{-j,-m} \in \mathcal{S}_{-j,\mathcal{D}_{-m}}$, i.e. $\boldsymbol{x}_{-j,-m} \in \mathbb{R}_+^{m-2}$ such that $\mathbf{1}_{m-2}^\top \boldsymbol{x}_{-j,-m} < 1$. In our discussion, for ease of notation, given x_j and $\boldsymbol{x}_{-j,-m}$, we may still write $x_m \equiv 1 - \mathbf{1}_{m-1}^\top \boldsymbol{x}_{-m}$, a function in x_j and $\boldsymbol{x}_{-j,-m}$, and for simplicity we may drop its dependence on x_j and $\boldsymbol{x}_{-j,-m}$. Note that $\mathcal{C}_j(\boldsymbol{x}_{-j,-m}) = \left(0, 1 - \mathbf{1}_{m-2}^\top \boldsymbol{x}_{-j,-m}\right)$.

(1) *Case $a > 0$ and $b \geq 0$:* For (A.1), we need $p_0(\boldsymbol{x}_{-m})\partial_j \log p(\boldsymbol{x}_{-m})(h_j \circ \varphi_j)(\boldsymbol{x}_{-m}) \to 0$ as $x_j \searrow 0^+$ and $x_j \nearrow 1 - \mathbf{1}_{m-2}^\top \boldsymbol{x}_{-j,-m}$. As x_j goes to any finite constant, by (3.25) $p_0(\boldsymbol{x}_{-m})$ converges to a non-zero constant when $b > 0$, or a finite constant times the limit of $x_j^{\eta_j} x_m (x_j)^{\eta_m}$ when $b = 0$. Note that

$$\partial_j \log p(\boldsymbol{x}_{-m}) = -\left(\boldsymbol{\kappa}_{-m,j}^\top \boldsymbol{x}_{-m}^a\right) x_j^{a-1} + \left(\boldsymbol{\kappa}_{-m,m}^\top \boldsymbol{x}_{-m}^a\right) x_m^{a-1}(x_j) - x_m^a(x_j)\kappa_{jm} x_j^{a-1} + x_m^{2a-1}(x_j)\kappa_{mm} + \eta_j x_j^{b-1} - \eta_m x_m^{b-1}(x_j). \quad \text{(B.4)}$$

i) If $b > 0$, by arguments above we only consider $(h_j \circ \varphi_j)(\boldsymbol{x}_{-m})\partial_j \log p(\boldsymbol{x}_{-m})$.

   a) As $x_j \searrow 0^+$, $x_m \nearrow 1 - \mathbf{1}_{m-2}^\top \boldsymbol{x}_{-j,-m} > 0$, so $\partial_j \log p(\boldsymbol{x}_{-m}) = \mathcal{O}\left(x_j^{a-1}\right) + \mathcal{O}\left(x_j^{b-1}\right) + \mathcal{O}(1)$. Thus we need $\alpha_j > \max\{0, 1-a, 1-b\}$ so that $(h_j \circ \varphi_j)(\boldsymbol{x}_{-m})\partial_j \log p(\boldsymbol{x}_{-m}) \to 0$.

   b) The case where $x_j \nearrow 1 - \mathbf{1}_{m-2}^\top \boldsymbol{x}_{-j,-m}$ and $x_m \searrow 0^+$ is an analog of a) by noting that φ_j is symmetric in x_j about the midpoint of its domain $(1 - \mathbf{1}_{m-2}^\top \boldsymbol{x}_{-j,-m})/2$.

ii) If $b = 0$, we need $x_j^{\eta_j} x_m^{\eta_m}(x_j)(h_j \circ \varphi_j)(\boldsymbol{x}_{-m})\partial_j \log p(\boldsymbol{x}_{-m}) \to 0$. Note that this quantity has the same form as in a) just with η_j or η_m added to the a and b $(= 0)$ in the exponents, we thus require $\alpha_j > \max\{0, 1-a-\eta_j, 1-a-\eta_m, 1-\eta_j, 1-\eta_m\} = \max\{0, 1-\eta_j, 1-\eta_m\}$.

In conclusion, (A.1) requires $\alpha_j \geq \max\{0, 1-a, 1-b\}$ for $b > 0$ or $\alpha_j > \max\{0, 1-\eta_j\}$. For (A.2), we only prove the first integrability condition, since the second integrability condition is similar. For the first, we need to show that

$$\int_{\boldsymbol{x}_{-m} \succ \mathbf{0},\, \mathbf{1}_{m-1}^\top \boldsymbol{x}_{-m} < 1} p_0(\boldsymbol{x}_{-m})(h_j \circ \varphi_j)(\boldsymbol{x}_{-m}) \left(\partial_j \log p(\boldsymbol{x}_{-m})\right)^2 \mathrm{d}\boldsymbol{x}_{-m} < +\infty.$$

Using the fact that $\mathbf{0} \prec \boldsymbol{x}_{-m} \prec \mathbf{1}$ and $0 < x_j^a < 1$, $0 < x_m^a < 1$ with the triangle inequality multiple times, we have

$$\begin{aligned} |\partial_j \log p(\boldsymbol{x}_{-m})| &\leq \sum_{i=1}^{m-1} |\kappa_{ij}| x_j^{a-1} + \sum_{i=1}^{m-1} |\kappa_{im}| x_m^{a-1} + |\kappa_{jm}| x_j^{a-1} + x_m^{a-1}|\kappa_{mm}| + |\eta_j| x_j^{b-1} + |\eta_m| x_m^{b-1} \\ &\leq \left\|\!\left|\mathbf{K}\right|\!\right\|_1 x_j^{a-1} + \left\|\!\left|\mathbf{K}\right|\!\right\|_1 x_m^{a-1} + |\eta_j| x_j^{b-1} + |\eta_m| x_m^{b-1}, \end{aligned}$$

where $\left\|\!\left|\mathbf{K}\right|\!\right\|_1 \equiv \max_{j=1,\ldots,m} \sum_{i=1}^m |\kappa_{ij}|$. We again consider the following two cases.

i) If $b > 0$, $p_0(\boldsymbol{x}_{-m})$ is bounded by an absolute constant, which we therefore ignore. We first fix $\boldsymbol{x}_{-j,-m}$ and denote $y_j(\boldsymbol{x}_{-j,-m}) \equiv 1 - \mathbf{1}_{m-2}^\top \boldsymbol{x}_{-j,-m} = x_j + x_m$. Then

$$\int_0^{y_j} (h_j \circ \varphi_j)(\boldsymbol{x}_{-m})(\partial_j \log p(\boldsymbol{x}_{-m}))^2 \,\mathrm{d}x_j$$

$$
\begin{aligned}
&\leq \int_0^{y_j} (h_j \circ \varphi_j)(\boldsymbol{x}_{-m}) \left(|||\mathbf{K}|||_1 x_j^{a-1} + |||\mathbf{K}|||_1 x_m^{a-1}(x_j) + |\eta_j| x_j^{b-1} + |\eta_m| x_m^{b-1}(x_j) \right)^2 \mathrm{d}x_j \\
&\leq \int_0^{y_j/2} (h_j \circ \varphi_j)(\boldsymbol{x}_{-m}) \left(|||\mathbf{K}|||_1 x_j^{a-1} + |\eta_j| x_j^{b-1} + c_{1,m}(\boldsymbol{x}_{-j,-m}) \right)^2 \mathrm{d}x_j \\
&\qquad + \int_{y_j/2}^{y_j} (h_j \circ \varphi_j)(\boldsymbol{x}_{-m}) \left(|||\mathbf{K}|||_1 x_m^{a-1}(x_j) + |\eta_m| x_m^{b-1}(x_j) + c_{1,j}(\boldsymbol{x}_{-j,-m}) \right)^2 \mathrm{d}x_j \\
&= \int_0^{y_j/2} h_j(x_j) \left(|||\mathbf{K}|||_1 x_j^{a-1} + |\eta_j| x_j^{b-1} + c_{1,m}(\boldsymbol{x}_{-j,-m}) \right)^2 \mathrm{d}x_j \\
&\qquad + \int_0^{y_j/2} h_j(x_j) \left(|||\mathbf{K}|||_1 x_j^{a-1} + |\eta_m| x_j^{b-1} + c_{1,j}(\boldsymbol{x}_{-j,-m}) \right)^2 \mathrm{d}x_j
\end{aligned}
$$

where in the last step we used change of variable $x_j \leftarrow x_m(x_j) = y_j - x_j$ for the second term, and where

$$
0 < c_{1,j} \equiv \max_{y_j/2 \leq x_j \leq y_j} \left(|||\mathbf{K}|||_1 x_j^{a-1} + |\eta_j| x_j^{b-1} \right) = \mathcal{O}\left(y_j^{a-1}\right) + \mathcal{O}\left(y_j^{b-1}\right) + \mathcal{O}(1) < +\infty,
$$

where the $\mathcal{O}$ depends on $\mathbf{K}$ and $\boldsymbol{\eta}$. We thus have

$$
\begin{aligned}
&\int_{\substack{\boldsymbol{x}_{-m} \succ \mathbf{0}, \\ \mathbf{1}_{m-1}^\top \boldsymbol{x}_{-m} < 1}} (h_j \circ \varphi_j)(\boldsymbol{x}_{-m}) \left(\partial_j \log p(\boldsymbol{x}_{-m}) \right)^2 \mathrm{d}\boldsymbol{x}_{-m} \\
&= \int_{\substack{\boldsymbol{x}_{-j,-m} \succ \mathbf{0}, \\ y_j(\boldsymbol{x}_{-j,-m}) > 0}} \int_0^{y_j(\boldsymbol{x}_{-j,-m})} (h_j \circ \varphi_j)(\boldsymbol{x}_{-m}) (\partial_j \log p(\boldsymbol{x}_{-m}))^2 \, \mathrm{d}x_j \, \mathrm{d}\boldsymbol{x}_{-j,-m} \\
&\leq \int_{\substack{\boldsymbol{x}_{-j,-m} \succ \mathbf{0}, \\ y_j(\boldsymbol{x}_{-j,-m}) > 0}} \int_0^{y_j(\boldsymbol{x}_{-j,-m})/2} \\
&\qquad\qquad h_j(x_j) \left(\mathcal{O}\left(x_j^{a-1}\right) + \mathcal{O}\left(x_j^{b-1}\right) + \mathcal{O}\left(y_j^{a-1}\right) + \mathcal{O}\left(y_j^{b-1}\right) + \mathcal{O}(1) \right)^2 \mathrm{d}x_j \, \mathrm{d}\boldsymbol{x}_{-j,-m} \\
&\leq \int_{\substack{\boldsymbol{x}_{-j,-m} \succ \mathbf{0}, \\ y_j(\boldsymbol{x}_{-j,-m}) > 0}} \int_0^{y_j(\boldsymbol{x}_{-j,-m})/2} h_j(x_j) \left(\mathcal{O}\left(x_j^{a-1}\right) + \mathcal{O}\left(x_j^{b-1}\right) + \mathcal{O}(1) \right)^2 \mathrm{d}x_j \, \mathrm{d}\boldsymbol{x}_{-j,-m} \\
&\quad + \int_{\substack{\boldsymbol{x}_{-j,-m} \succ \mathbf{0}, \\ y_j(\boldsymbol{x}_{-j,-m}) > 0}} \int_0^{y_j(\boldsymbol{x}_{-j,-m})/2} h_j(x_j) \left(\mathcal{O}\left(y_j^{a-1}\right) + \mathcal{O}\left(y_j^{b-1}\right) + \mathcal{O}(1) \right)^2 \mathrm{d}x_j \, \mathrm{d}\boldsymbol{x}_{-j,-m} \\
&\leq \int_{\mathbf{1} \succ \boldsymbol{x}_{-j,-m} \succ \mathbf{0}} \int_0^1 x_j^{\alpha_j} \left(\mathcal{O}\left(x_j^{a-1}\right) + \mathcal{O}\left(x_j^{b-1}\right) + \mathcal{O}(1) \right)^2 \mathrm{d}x_j \, \mathrm{d}\boldsymbol{x}_{-j,-m} \\
&\qquad\qquad + \int_{\substack{\boldsymbol{x}_{-j,-m} \succ \mathbf{0}, \\ y_j(\boldsymbol{x}_{-j,-m}) > 0}} \frac{(y_j/2)^{\alpha_j+1}}{\alpha_j + 1} \left(\mathcal{O}\left(y_j^{a-1}\right) + \mathcal{O}\left(y_j^{b-1}\right) + \mathcal{O}(1) \right)^2 \mathrm{d}\boldsymbol{x}_{-j,-m} \\
&\quad \leq \int_0^1 \mathcal{O}\left(x_j^{2a-2+\alpha_j}\right) + \mathcal{O}\left(x_j^{2b-2+\alpha_j}\right) + \mathcal{O}\left(x_j^{\alpha_j}\right) \mathrm{d}x_j
\end{aligned}
$$

$$+ \sum_{p\in\{2a-1+\alpha_j,\, 2b-1+\alpha_j,\, \alpha_j+1\}} \int_{\substack{\boldsymbol{x}_{-j,-m}\succ\mathbf{0},\\ \mathbf{1}_{m-2}^\top \boldsymbol{x}_{-j,-m}<1}} \mathcal{O}\left(\left(1-\mathbf{1}_{m-2}^\top \boldsymbol{x}_{-j,-m}\right)^p\right) \mathrm{d}\boldsymbol{x}_{-j,-m}$$

$$= \int_0^1 o\left(x_j^{a-1}\right) + o\left(x_j^{b-1}\right) + o\left(x^0\right) \mathrm{d}x_j + \sum_{p\in\{a,b,1\}} \mathcal{O}\left(\Gamma(p+1)/\Gamma(p+m-1)\right) < +\infty$$

for $\alpha_j > \max\{0, 1-a, 1-b\}$, where the second term of the last quantity follows from the inverse normalizing constant of the Dirichlet distribution with parameters $(\mathbf{1}_{m-2}, p+1)$.

ii) If $b = 0$, $p_0(\boldsymbol{x}_{-m})$ is now bounded by $C_2 \prod_{j=1}^m x_j^{\eta_{0j}}$ where C_2 is the inverse normalizing constant of $p_0(\boldsymbol{x}_{-m})$ times $\sup_{\boldsymbol{x}\succ\mathbf{0}, \mathbf{1}^\top\boldsymbol{x}=1} \exp\left(-\boldsymbol{x}^{a\top}\mathbf{K}_0\boldsymbol{x}^a/(2a)\right)$, a positive and finite constant. Then by the same reasoning as in i), with $y_j(\boldsymbol{x}_{-j,-m}) \equiv 1-\mathbf{1}_{m-2}^\top \boldsymbol{x}_{-j.-m}$ and noting that $\boldsymbol{\eta} \succ -\mathbf{1}_m$,

$$\int_{\boldsymbol{x}_{-m}\succ\mathbf{0},\, \mathbf{1}_{m-1}^\top\boldsymbol{x}_{-m}<1} p_0(\boldsymbol{x}_{-m})(h_j\circ\varphi_j)(\boldsymbol{x}_{-m}) \left(\partial_j \log p(\boldsymbol{x}_{-m})\right)^2 \mathrm{d}\boldsymbol{x}_{-m}$$

$$\leq \int_{\boldsymbol{x}_{-j,-m}\succ\mathbf{0},\, y_j>0} \int_0^{y_j/2} C_2 \prod_{k=1}^m x_k^{\eta_{0k}} h_j(x_j) \left(\mathcal{O}\left(x_j^{a-1}\right) + \mathcal{O}\left(x_j^{-1}\right) + \mathcal{O}(1)\right)^2 \mathrm{d}x_j \,\mathrm{d}\boldsymbol{x}_{-j,-m}$$

$$+ \int_{\boldsymbol{x}_{-j,-m}\succ\mathbf{0},\, y_j>0} \int_0^{y_j/2} C_2 \prod_{k=1}^m x_k^{\eta_{0k}} h_j(x_j) \left(\mathcal{O}\left(y_j^{a-1}\right) + \mathcal{O}\left(y_j^{-1}\right) + \mathcal{O}(1)\right)^2 \mathrm{d}x_j \,\mathrm{d}\boldsymbol{x}_{-j,-m}$$

$$\leq C_2 \prod_{k\neq j,m} \int_0^1 x_k^{\eta_{0k}} \mathrm{d}x_k \int_0^1 x_j^{\eta_{0k}+\alpha_j} \left(\mathcal{O}\left(x_j^{-1}\right) + \mathcal{O}(1)\right)^2 \mathrm{d}x_j$$

$$+ C_2 \int_{\boldsymbol{x}_{-j,-m}\succ\mathbf{0},\, y_j>0} \frac{(y_j/2)^{\alpha_j+1+\eta_{0j}}}{\alpha_j+1+\eta_{0j}} \prod_{k\neq j,m} x_k^{\eta_{0k}} \left(\mathcal{O}\left(y_j^{-1}\right) + \mathcal{O}(1)\right)^2 \mathrm{d}\boldsymbol{x}_{-j,-m}$$

$$\leq C_2 \prod_{k\neq j,m} \frac{1}{\eta_{0k}+1} \int_0^1 \mathcal{O}\left(x_j^{-1}\right) + \mathcal{O}\left(x_j\right) \mathrm{d}x_j$$

$$+ \sum_{p\in\{0,2\}} C_2 \int_{\substack{\boldsymbol{x}_{-j,-m}\succ\mathbf{0},\\ y_j>0}} \prod_{k\neq j,m} x_k^{\eta_{0k}} \mathcal{O}\left(y_j^p\right) \mathrm{d}\boldsymbol{x}_{-j,-m}$$

$$< +\infty$$

since the integral in the second term is the inverse normalizing constant of the Dirichlet distribution with parameters $(\boldsymbol{\eta}_{0,-j,-m} + \mathbf{1}_{m-2}, p+1)$, i.e. $\frac{\Gamma(p+1)\prod_{k\neq j,m}\Gamma(\eta_{0k}+1)}{\Gamma\left(\mathbf{1}_{m-2}^\top\boldsymbol{\eta}_{0,-j,-m}+p+m-1\right)} < +\infty$.

This ends the proof for the first integrability condition for (A.1) for $a > 0$. For the second half, the integrand we consider is $p_0(\boldsymbol{x}_{-m}) \left|\partial_j \left(\partial_j \log p(\boldsymbol{x}_{-m})(h_j\circ\varphi)(\boldsymbol{x}_{-m})\right)\right|$. The arguments

are similar to those for the first condition, where we first bound $\partial_{jj} \log p(\boldsymbol{x})$ using sums of products of powers of $\boldsymbol{x}$. Then for each fixed $\boldsymbol{x}_{-j,-m}$ we split the domain of x_j into two halves and deal with the potential singularity at $x_j \searrow 0^+$, where one can show that the requirement on α_j is just enough for the integrand to be $o\left(x_j^{-1}\right)$ and thus the integral is finite. The detailed proof is tedious and is omitted.

(2) *Case $a = 0$ and $b \geq 0$:* First consider $b = 0$. We again write $y_j \equiv 1 - \mathbf{1}_{m-2}^\top \boldsymbol{x}_{-j,-m}$. Fixing $\boldsymbol{x}_{-j,-m} \in \mathcal{S}_{-j,\mathcal{D}_{-m}}$,

$$
\begin{aligned}
& p_0(\boldsymbol{x}_{-m}) \left|\partial_j \log p(\boldsymbol{x}_{-m})\right| && \text{(B.5)} \\
\propto \exp & \left[-\frac{1}{2}\log(\boldsymbol{x}_{-m})^\top \mathbf{K}_{0,-m,-m} \log(\boldsymbol{x}_{-m}) - \log(\boldsymbol{x}_{-m})^\top \boldsymbol{\kappa}_{0,-m,m} \log x_m \right. \\
& \qquad \left. -\frac{1}{2}\kappa_{0,m,m} \left(\log x_m\right)^2 + \boldsymbol{\eta}_{0,-m}^\top \log(\boldsymbol{x}_{-m}) + \eta_{0,m} \log x_m \right] \times \\
& \left|-\boldsymbol{\kappa}_{-m,j}^\top \log(\boldsymbol{x}_{-m})/x_j + \boldsymbol{\kappa}_{-m,m}^\top \log(\boldsymbol{x}_{-m})/x_m - \kappa_{jm} \log x_m/x_j \right. \\
& \qquad \left. +\kappa_{mm} \log x_m/x_m + \eta_j/x_j - \eta_m/x_m \right|. && \text{(B.6)} \\
\leq \prod_{k \neq j,m} & \exp\left[-\frac{N_{\mathbf{K}_{0,-m,-m}}}{2}(\log x_k)^2 + \eta_{0,k} \log x_k\right] \exp\left[-\frac{N_{\mathbf{K}_{0,-m,-m}}}{2}(\log x_j)^2 + \eta_{0,j} \log x_j \right. \\
& \qquad \left. -\log(\boldsymbol{x}_{-m})^\top \boldsymbol{\kappa}_{0,-m,m} \log x_m - \frac{1}{2}\kappa_{0,m,m} \left(\log x_m\right)^2 + \eta_{0,m} \log x_m \right] \times \\
& \left|-\boldsymbol{\kappa}_{-m,j}^\top \log(\boldsymbol{x}_{-m})/x_j + \boldsymbol{\kappa}_{-m,m}^\top \log(\boldsymbol{x}_{-m})/x_m - \kappa_{jm} \log x_m/x_j \right. \\
& \qquad \left. +\kappa_{mm} \log x_m/x_m + \eta_j/x_j - \eta_m/x_m \right|. && \text{(B.7)}
\end{aligned}
$$

which is $\mathcal{O}\left(\exp\left(\mathcal{O}\left((\log x_j)^2\right) + \mathcal{O}\left(\log x_j\right) + \mathcal{O}\left(\log\log x_j\right)\right)\right)$ as $x_j \searrow 0^+$. Since the coefficient on the leading term is negative the entire term goes to 0. By symmetry the quantity goes to zero also when $x_j \nearrow y_j^-$. Thus, (A.1) holds for any $\alpha_j \geq 0$.

Similarly, both $p_0(\boldsymbol{x}_{-m})(h_j \circ \varphi_j)(\boldsymbol{x}_{-m})(\partial_j \log p(\boldsymbol{x}_{-m}))^2$ and $p_0(\boldsymbol{x}_{-m})|\partial(\partial_j \log p(\boldsymbol{x}_{-m})(h_j \circ \varphi)(\boldsymbol{x}_{-m}))|$ are products of $\prod_{j=1}^m \mathcal{O}\left(\exp\left(-(\log x_j)^2\right)\right)$ times a polynomial, and are thus bounded and go to 0 at the boundaries of $\mathcal{D}_{-m}$. Thus extending the

integrands to 0 at the boundaries, they are continuous and bounded in the compact $\overline{\mathcal{D}_{-m}}$, so integrals $\int_{\mathcal{D}_{-m}} p_0(\boldsymbol{x}_{-m})(h_j \circ \varphi_j)(\boldsymbol{x}_{-m})(\partial_j \log p(\boldsymbol{x}_{-m}))^2 \,\mathrm{d}\boldsymbol{x}_{-m}$ and $\int_{\mathcal{D}_{-m}} p_0(\boldsymbol{x}_{-m}) \left|\partial \left(\partial_j \log p(\boldsymbol{x}_{-m})(h_j \circ \varphi)(\boldsymbol{x}_{-m})\right)\right| \mathrm{d}\boldsymbol{x}_{-m}$ are finite, thus proving (A.2).

For $a = 0$ and $b > 0$, $\boldsymbol{\eta} \preceq \mathbf{0}$, and the proof is similar and is omitted. In particular, $p_0(\boldsymbol{x}_{-m})$ is bounded by that with $a = 0$, $b = 0$, $\boldsymbol{\eta} \equiv \mathbf{0}_m$, and thus its product with any polynomial is bounded and goes to 0 at the boundary of $\overline{\mathcal{D}_{-m}}$. □

Proof of Theorem 20. For notational simplicity, denote $\widetilde{\mathbf{K}} \equiv \mathbf{K}_1 - \mathbf{K}_2$ with columns $\widetilde{\boldsymbol{\kappa}}_{,1} \ldots, \widetilde{\boldsymbol{\kappa}}_{,m}$, and denote $\widetilde{\boldsymbol{\eta}} \equiv \boldsymbol{\eta}_1 - \boldsymbol{\eta}_2$. Assume that either $\widetilde{\mathbf{K}} \neq \mathbf{0}_{m\times m}$ or $\widetilde{\boldsymbol{\eta}} \neq \mathbf{0}_m$, otherwise there is nothing to prove. By (3.26), writing $x_m \equiv 1 - \mathbf{1}_{m-1}^\top \boldsymbol{x}_{-m}$ and $\boldsymbol{x} = (\boldsymbol{x}_{-m}; x_m)$ and taking the gradient of the log of both sides of the equation with respect to x_j, $j = 1, \ldots, m-1$, we have

$$\left(x_j^{a-1}\widetilde{\boldsymbol{\kappa}}_{,j} - x_m^{a-1}\widetilde{\boldsymbol{\kappa}}_{,m}\right)^\top \boldsymbol{x}^a = \widetilde{\eta}_j x_j^{b-1} - \widetilde{\eta}_m x_m^{b-1} \tag{B.8}$$

for all $\boldsymbol{x}_{-m} \in \mathcal{D}_{-m} \equiv \left\{\boldsymbol{x}_{-m} \in \mathbb{R}_+^{m-1} \middle|\, \boldsymbol{x}_{-m} \succ \mathbf{0}, \mathbf{1}_{m-1}^\top \boldsymbol{x}_{-m} < 1\right\}$. In the following when $a = 0$ by x^a we mean $\log(x)$, and by x^{a-1} we mean $1/x$ and we do not treat this case differently as the same expressions still hold.

(1) Suppose $\left(x_j^{a-1}\widetilde{\boldsymbol{\kappa}}_{,j} - x_m^{a-1}\widetilde{\boldsymbol{\kappa}}_{,m}\right)_{-j,-m} = \mathbf{0}_{m-2}$ for all $\boldsymbol{x}_{-m} \in \mathcal{D}_{-m}$ and $x_m = 1 - \mathbf{1}_{-m}^\top \boldsymbol{x}_{-m}$ or $\tilde{\eta}_j$.

   (i) Suppose $a = 1$, then (B.8) becomes $\left(\widetilde{\kappa}_{j,j} - \widetilde{\kappa}_{j,m}\right) x_j + \left(\widetilde{\kappa}_{m,j} - \widetilde{\kappa}_{m,m}\right) x_m = \widetilde{\eta}_j x_j^{b-1} - \widetilde{\eta}_m x_m^{b-1}$, and we must have $b = 2$ or $b = 1$ or $\widetilde{\eta}_j = \widetilde{\eta}_m = 0$, i.e. $\boldsymbol{\eta}_1 = \boldsymbol{\eta}_2$.

   (ii) Suppose $a \neq 1$. The assumption implies that $\left(\widetilde{\boldsymbol{\kappa}}_j\right)_{-j,-m} = \left(\widetilde{\boldsymbol{\kappa}}_m\right)_{-j,-m} = \mathbf{0}_{m-2}$. Then (B.8) becomes $\left(x_j^{a-1}\widetilde{\kappa}_{j,j} - x_m^{a-1}\widetilde{\kappa}_{j,m}\right) x_j^a + \left(x_j^{a-1}\widetilde{\kappa}_{m,j} - x_m^{a-1}\widetilde{\kappa}_{m,m}\right) x_m^a = \widetilde{\eta}_j x_j^{b-1} - \widetilde{\eta}_m x_m^{b-1}$. Since $x_j > 0$ and $x_m > 0$ are arbitrary (as $\mathbf{1}_{-m}^\top \boldsymbol{x}_{-m}$ can vary) as long as $x_j + x_m < 1$, the cross terms must not exist, and so $\widetilde{\kappa}_{j,m} = \widetilde{\kappa}_{m,j} = 0$. It thus follows that $\tilde{\boldsymbol{\kappa}}_{-j,j} = \tilde{\boldsymbol{\kappa}}_{-m,m} = \mathbf{0}_{m-1}$ and hence $\widetilde{\mathbf{K}}$ is diagonal, and the original equality becomes $-\frac{1}{2}\mathrm{diag}(\widetilde{\mathbf{K}})^\top (\boldsymbol{x}^a)^2 + \widetilde{\boldsymbol{\eta}}^\top \boldsymbol{x}^b = 0$, in which by $\boldsymbol{x}^0$ we mean $\log(\boldsymbol{x})$. Thus we must have $2a = b \neq 0$ and $\mathbf{K}_1 - \mathbf{K}_2 = 2\boldsymbol{\eta}_1 - 2\boldsymbol{\eta}_2$.

(2) Now fix x_j and x_m such that $\left(x_j^{a-1}\widetilde{\boldsymbol{\kappa}}_{\cdot,j} - x_m^{a-1}\widetilde{\boldsymbol{\kappa}}_{\cdot,m}\right)_{-j,-m} \neq \mathbf{0}_{m-2}$. Note that $\mathbf{1}_{m-2}^\top \boldsymbol{x}_{-j,-m} = 1 - x_j - x_m$ is also fixed. Now the right-hand side and the first vector on the left-hand side of (B.8) are both constant, while $\boldsymbol{x}_{-j,-m}$ is allowed to vary freely as long as their sum is fixed. A necessary condition of (B.8) is thus

$$\left(x_j^{a-1}\widetilde{\boldsymbol{\kappa}}_{\cdot,j} - x_m^{a-1}\widetilde{\boldsymbol{\kappa}}_{\cdot,m}\right)_{-j,-m}^\top \boldsymbol{x}_{-j,-m}^a = \text{const depending on } x_j \text{ and } x_m \text{ only} \tag{B.9}$$

for all $\boldsymbol{x}_{-j,-m}^a \in \mathcal{U}_{x_j,x_m} \equiv \{\boldsymbol{y}^a : \boldsymbol{y} \succ \mathbf{0}_{m-2}, \mathbf{1}_{m-2}\boldsymbol{y} = \mathbf{1}_{m-2}^\top \boldsymbol{x}_{-j,-m} = 1 - x_j - x_m\}$.

Suppose $a \neq 1$. Then $\mathcal{U}_{x_j,x_m}$ is not entirely on a hyperplane, and by assumption $\left(x_j^{a-1}\widetilde{\boldsymbol{\kappa}}_{\cdot,j} - x_m^{a-1}\widetilde{\boldsymbol{\kappa}}_{\cdot,m}\right)_{-j,-m}$ is not a zero vector, so the equality cannot hold. We thus have $a = 1$, so that $\mathcal{U}_{x_j,x_m}$ lies on the hyperplane $\mathcal{H}_{x_j,x_m} \equiv \{\boldsymbol{y} : \mathbf{1}_{m-2}^\top \boldsymbol{y} = 1 - x_j - x_m\}$. Since $\mathcal{H}_{x_j,x_m} \equiv \{c\boldsymbol{x}_{-j,-m} : c \in \mathbb{R}, \boldsymbol{x} \in \mathcal{U}_{x_j,x_m}\}$, (B.9) must hold for all $\boldsymbol{x}_{-j,-m}$ in the hyperplane $\mathcal{H}_{x_j,x_m}$, and by the assumption that $\left(\widetilde{\boldsymbol{\kappa}}_{\cdot,j} - \widetilde{\boldsymbol{\kappa}}_{\cdot,m}\right)_{-j,-m}$ is nonzero it must be a constant multiple of $\mathbf{1}_m$, and the right-hand side of (B.9) is hence $c_0(1 - x_j - x_m)$ for some absolute constant $c_0 \neq 0$ assuming $\left(\widetilde{\boldsymbol{\kappa}}_{\cdot,j} - \widetilde{\boldsymbol{\kappa}}_{\cdot,m}\right)_{-j,-m} = c_0\mathbf{1}_m$. Plugging this back in (B.8) we get

$$c_0(1 - x_j - x_m) + \left(\widetilde{\kappa}_{j,j} - \widetilde{\kappa}_{j,m}\right) x_j + \left(\widetilde{\kappa}_{m,j} - \widetilde{\kappa}_{m,m}\right) x_m = \widetilde{\eta}_j x_j^{b-1} - \widetilde{\eta}_m x_m^{b-1},$$

and hence as in (1) (i) we have $b = 2$ or $b = 1$ or $\boldsymbol{\eta}_1 = \boldsymbol{\eta}_2$.

□

Proof of Theorem 21. For any $\boldsymbol{x}_{-m} \in \mathcal{D}_{-m}$, write $\iota_{+m}(\boldsymbol{x}_{-m}) = \left(\boldsymbol{x}_{-m}, 1 - \mathbf{1}_{m-1}^\top \boldsymbol{x}_{-m}\right)$. We first prove the finiteness of the normalizing constant. If $\mathbf{K}$ is positive definite, the inverse normalizing constant is

$$\begin{aligned}
&\int_{\mathcal{D}_{-m}} \exp\left(-\frac{1}{2}\log \iota_{+m}(\boldsymbol{x}_{-m})^\top \mathbf{K} \log \iota_{+m}(\boldsymbol{x}_{-m}) + \boldsymbol{\eta}^\top \log \iota_{+m}(\boldsymbol{x}_{-m})\right) \mathrm{d}\boldsymbol{x}_{-m} \\
\leq &\int_{\mathcal{D}_{-m}} \exp\left(\sum_{j=1}^m \left(-\lambda_{\min}(\mathbf{K}) \left(\log \iota_{+m}\left(\boldsymbol{x}_{-m}\right)\right)_j^2 + \eta_j \left(\log \iota_{+m}\left(\boldsymbol{x}_{-m}\right)\right)_j\right)\right) \mathrm{d}\boldsymbol{x}_{-m} \\
\leq &\int_{\mathcal{D}_{-m}} \exp\left(\sum_{j=1}^m \frac{\eta_j^2}{4\lambda_{\min}(\mathbf{K})}\right) \mathrm{d}\boldsymbol{x}_{-m} < +\infty,
\end{aligned}$$

proving (1). Now assume $\mathbf{K}$ is no longer positive definite. If $\mathbf{K}\mathbf{1}_m = \mathbf{0}_m$, for any $\boldsymbol{a} \in \mathbb{R}^m$,

$$\begin{aligned}
\boldsymbol{a}^\top \mathbf{K} \boldsymbol{a} &= (\boldsymbol{a}_{-j}, a_j)^\top \begin{bmatrix} \mathbf{K}_{-j,-j} & \boldsymbol{\kappa}_{-j,j} \\ \boldsymbol{\kappa}_{-j,j}^\top & \kappa_{jj} \end{bmatrix} (\boldsymbol{a}_{-j}, a_j) \\
&= (\boldsymbol{a}_{-j}, a_j)^\top \begin{bmatrix} \mathbf{K}_{-j,-j} & -\mathbf{K}_{-j,-j}\mathbf{1}_{m-1} \\ -\mathbf{1}_{m-1}^\top \mathbf{K}_{-j,-j} & \mathbf{1}_{m-1}^\top \mathbf{K}_{-j,-j} \mathbf{1}_{m-1} \end{bmatrix} (\boldsymbol{a}_{-j}, a_j) \\
&= (\boldsymbol{a}_{-j} - a_j \mathbf{1}_m)^\top \mathbf{K}_{-j,-j} (\boldsymbol{a}_{-j} - a_j \mathbf{1}_{m-1}),
\end{aligned}$$

which is zero if and only if $\boldsymbol{a}_{-j} = a_j \mathbf{1}_{m-1}$, i.e. $a_1 = \cdots = a_m$, and is positive otherwise. Thus, if $\mathbf{K}\mathbf{1}_m = \mathbf{0}_m$, the condition that $\mathbf{K}_{-j,-j}$ is positive definite for some $j = 1, \ldots, m$ is equivalent to that $\mathbf{K}_{-j,-j}$ is positive definite for all $j = 1, \ldots, m$, and implies that $\mathbf{K}$ is positive semi-definite.

If $\mathbf{K}$ is positive semi-definite and $\boldsymbol{\eta} \succ -\mathbf{1}_m$, the inverse normalizing constant is

$$\begin{aligned}
&\int_{\mathcal{D}_{-m}} \exp\left(-\frac{1}{2} \log \iota_{+m}(\boldsymbol{x}_{-m})^\top \mathbf{K} \log \iota_{+m}(\boldsymbol{x}_{-m}) + \boldsymbol{\eta}^\top \log \iota_{+m}(\boldsymbol{x}_{-m}) \right) \mathrm{d}\boldsymbol{x}_{-m} \\
\leq &\int_{\mathcal{D}_{-m}} \exp\left(\boldsymbol{\eta}^\top \log \iota_{+m}(\boldsymbol{x}_{-m}) \right) \mathrm{d}\boldsymbol{x}_{-m} = \frac{\prod_{j=1}^m \Gamma(\eta_j + 1)}{\Gamma\left(\mathbf{1}_m^\top \boldsymbol{\eta} + m\right)} < +\infty
\end{aligned}$$

since the last quantity is the inverse normalizing constant of the Dirichlet distribution with parameters $(\boldsymbol{\eta} + \mathbf{1}_m)$, proving (3).

On the other hand, suppose $\mathbf{K}\mathbf{1}_m = \mathbf{0}_m$ and $\mathbf{K}_{-j,-j}$ is positive definite for some/all $j = 1, \ldots, m$. Again letting $\boldsymbol{x}_{-m}$ be the free variables and letting $x_m = 1 - \mathbf{1}_{m-1}^\top \boldsymbol{x}_{-m}$, define the *additive log-ratio transformation* applied to $\boldsymbol{x}$: $\boldsymbol{y}_{-m} \equiv \mathrm{alt}(\boldsymbol{x}) \equiv \log \boldsymbol{x}_{-m} - (\log x_m)\mathbf{1}_{m-1}$, a random vector supported on $\mathbb{R}^{m-1}$. Append an extra $y_m = 0$ for ease of notation. The transformation is thus bijective and the inverse transformation, the *additive logistic transformation* $\boldsymbol{x} = \exp(\boldsymbol{y})/\mathbf{1}_m^\top \exp(\boldsymbol{y})$. Since

$$\partial x_k / \partial y_j = -x_k x_j, \quad \partial x_j / \partial y_j = x_j(1 - x_j)$$

for $j \neq k$, $j, k = l, \ldots, m-1$, we have

$$\left|\frac{\partial \boldsymbol{x}_{-m}}{\partial \boldsymbol{y}_{-m}}\right| = \begin{vmatrix} x_1(1-x_1) & -x_1x_2 & \cdots & -x_1x_{m-1} \\ -x_1x_2 & x_2(1-x_2) & \cdots & -x_2x_{m-1} \\ \vdots & \vdots & \ddots & \vdots \\ -x_1x_{m-1} & -x_2x_{m-1} & \cdots & x_{m-1}(1-x_{m-1}) \end{vmatrix} = \prod_{j=1}^{m} x_j = \exp\left(\mathbf{1}_m^\top \log \boldsymbol{x}\right).$$

Then by $\mathbf{1}_m^\top \mathbf{K} = \mathbf{K}\mathbf{1}_m = \mathbf{0}_m$, $\boldsymbol{y}_{-m}$ has density proportional to

$$\begin{aligned} & p(\boldsymbol{x}_{-m})\,|\partial \boldsymbol{x}_{-m}/\partial \boldsymbol{y}_{-m}| \\ \propto\ & \exp\Bigg(-\frac{1}{2}\left(\log \boldsymbol{x} - (\log x_m)\mathbf{1}_m\right)^\top \mathbf{K}\left(\log \boldsymbol{x} - (\log x_m)\mathbf{1}_m\right) \\ & \qquad + (\boldsymbol{\eta} + \mathbf{1}_m)^\top \left(\log \boldsymbol{x} - (\log x_m)\mathbf{1}_m\right) + (\log x_m)\mathbf{1}_m^\top(\boldsymbol{\eta} + \mathbf{1}_m)\Bigg) \\ =\ & \exp\left(-\frac{1}{2}\boldsymbol{y}^\top \mathbf{K}\boldsymbol{y} + (\boldsymbol{\eta} + \mathbf{1}_m)^\top \boldsymbol{y} + \mathbf{1}_m^\top(\boldsymbol{\eta} + \mathbf{1}_m)\log x_m\right) \\ =\ & \exp\left(-\frac{1}{2}\boldsymbol{y}_{-m}^\top \mathbf{K}_{-m,-m}\boldsymbol{y}_{-m} + (\boldsymbol{\eta}_{-m} + \mathbf{1}_{m-1})^\top \boldsymbol{y}_{-m} - \left(\mathbf{1}_m^\top \boldsymbol{\eta} + m\right)\log\left(1 + \mathbf{1}_{m-1}^\top \exp(\boldsymbol{y}_{-m})\right)\right). \end{aligned}$$

Note that the $\log x_m = -\log\left(1 + \mathbf{1}_{m-1}^\top \exp(\boldsymbol{y}_{-m})\right) < 0$, so for $\mathbf{1}_m^\top \boldsymbol{\eta} + m \geq 0$ the last display is always upper-bounded by a constant times a normal density with a positive definite inverse covariance matrix $\mathbf{K}_{-m,-m}$, and thus the normalizing constant is finite, thus proving (2).

As for (A.1), fix $j = 1, \ldots, m-1$ and any $\ell \in \{1, \ldots, m-1\}\backslash\{j\}$, and write $\boldsymbol{z} \equiv \log \boldsymbol{x} - (\log x_\ell)\mathbf{1}_m$. Fix any $\boldsymbol{x}_{-j,-m} \in \mathbb{R}^{m-2}$ with $\boldsymbol{x}_{-j,-m} \succ \mathbf{0}$, $\mathbf{1}_{m-2}^\top \boldsymbol{x}_{-j,-m} < 1$. Then if (1) $\mathbf{K}_0$ is positive definite or (2) $\mathbf{K}_0\mathbf{1}_m = \mathbf{0}_m$ and $\mathbf{K}_{0,-\ell,-\ell}$ is positive definite, by the proof above, $p_0(\boldsymbol{x}_{-m})x_i^t$ is upper bounded by a finite constant depending on $\mathbf{K}_0$ and $\boldsymbol{\eta}_0$ for any $t \in \mathbb{R}$ and $i = j, m$, since it is a constant times the density with parameters $\mathbf{K}_0$ and $\boldsymbol{\eta}_0 + t\boldsymbol{e}_i$, and since we did not impose any restriction on the $\boldsymbol{\eta}$ parameter. On the other hand, for (3) $\mathbf{K}_0\mathbf{1}_m = \mathbf{0}_m$, $\mathbf{K}_{0,-\ell,-\ell}$ is positive semi-definite and $\boldsymbol{\eta}_0 \succ -\mathbf{1}_m$,

$$\begin{aligned} p_0(\boldsymbol{x}_{-m}) &\propto \exp\left(-\frac{1}{2}\log \boldsymbol{x}^\top \mathbf{K}_0 \log \boldsymbol{x} + \boldsymbol{\eta}_0^\top \log \boldsymbol{x}\right) \\ &= \exp\left(-\frac{1}{2}\boldsymbol{z}_{-\ell}^\top \mathbf{K}_{0,-\ell,-\ell}\boldsymbol{z}_{-\ell} + \boldsymbol{\eta}_{0,-\ell}^\top \boldsymbol{z}_{-\ell} + \left(\mathbf{1}_m^\top \boldsymbol{\eta}_0\right)\log x_\ell\right) \end{aligned}$$

$$\leq \exp\left(\boldsymbol{\eta}_{0,-\ell}^\top \boldsymbol{z}_{-\ell} + \left(\mathbf{1}_m^\top \boldsymbol{\eta}_0\right) \log x_\ell\right)$$
$$\propto \exp\left(\eta_{0,j} z_j + \eta_{0,m} z_m\right). \tag{B.10}$$

On the other hand,

$$\begin{aligned}
&|\partial_j \log p(\boldsymbol{x}_{-m})| \min\{x_j, x_m\}^{\alpha_j} \\
&= \left|-\boldsymbol{\kappa}_{\cdot,j}^\top \log \boldsymbol{x}/x_j + \boldsymbol{\kappa}_{\cdot,m}^\top \log \boldsymbol{x}/x_m + \eta_j/x_j - \eta_m/x_m\right| \min\{x_j, x_m\}^{\alpha_j} \\
&= \left(\left|\boldsymbol{\kappa}_{\cdot,j}^\top \log \boldsymbol{x}/x_j\right| + \left|\boldsymbol{\kappa}_{\cdot,m}^\top \log \boldsymbol{x}/x_m\right| + |\eta_j/x_j| + |\eta_m/x_m|\right) \min\{x_j, x_m\}^{\alpha_j} \\
&\leq \left(\left|\boldsymbol{\kappa}_{\cdot,j}^\top \log \boldsymbol{x}\right| + \left|\boldsymbol{\kappa}_{\cdot,m}^\top \log \boldsymbol{x}\right| + |\eta_j| + |\eta_m|\right) \min\{x_j, x_m\}^{\alpha_j - 1}.
\end{aligned}$$

Thus, as $x_j \searrow 0^+$ or $x_m \searrow 0^+$ (i.e. $x_j \nearrow 1 - \mathbf{1}_{m-2}^\top \boldsymbol{x}_{-j,-m}$), by multiplying the two bounds we have $p_0(\boldsymbol{x}_{-m}) \left|\partial_j \log p(\boldsymbol{x}_{-m})\right| (h_j \circ \varphi_j)(\boldsymbol{x}) \searrow 0^+$ for any α_j for (1) and (2) (by letting t to e.g. $\alpha_j - 2$ in the discussion above), or for (3) by a constant times

$$\left(\left|\boldsymbol{\kappa}_{\cdot,j}^\top \log \boldsymbol{x}\right| + \left|\boldsymbol{\kappa}_{\cdot,m}^\top \log \boldsymbol{x}\right| + |\eta_j| + |\eta_m|\right) \min\{x_j, x_m\}^{\alpha_j - 1} x_j^{\eta_{0,j}} x_m^{\eta_{0,m}} \searrow 0^+$$

if $\alpha_j > \max\{1 - \eta_{0,j}, 1 - \eta_{0,m}\}$.

As for (A.2), the results follow by a similar discussion for the Gamma model (a-b model with $b = 0$) on the standard simplex in Section 3.6. □

Proof of Theorem 24. It suffices to bound $\boldsymbol{\Gamma}$ and $\boldsymbol{g}$ using their forms in Section 3.5.2 and apply Theorem 1 in Lin et al. (2016). Thus, we first find the bounds of $(h_j \circ \varphi_j)(\boldsymbol{x}) x_j^{p_j} x_k^{p_k} x_\ell^{p_\ell}$ with $h_j(x) = x^{\alpha_j}$, $\alpha_j \geq 0$, $\alpha_j \geq -p_j$, $p_j \in \mathbb{R}$, $p_k \geq 0$, $p_\ell \geq 0$ and $x_i \in [u_i, v_i]$ for $i = 1, \ldots, m$. Suppose without loss of generality that j, k, ℓ are all different, as

$$\max_{x_j} f_{j,1}(x_j) \max_{x_j} f_{j,2}(x_j) \max_{x_j} f_{j,3}(x_j) \geq \max_{x_j} \left(f_{j,1}(x_j) f_{j,2}(x_j) f_{j,3}(x_j)\right) \geq 0$$

for any nonnegative functions $f_{j,1}$, $f_{j,2}$, $f_{j,3}$.

As x_j approaches its boundary, $\varphi_j(\boldsymbol{x}) \searrow 0^+$ and hence $(h_j \circ \varphi_j)(\boldsymbol{x}) x_j^{p_j} \searrow 0^+$ if $\alpha_j > -p_j$. The lower bound 0 for $(h_j \circ \varphi_j)(\boldsymbol{x}) x_j^{p_j} x_k^{p_k} x_\ell^{p_\ell}$ is thus tight enough.

As for the upper bound, the only way for the quantity to be unbounded from above is when $x_j \searrow 0^+$ and $p_j < 0$, but as $x_j \searrow 0^+$, $(h_j \circ \varphi_j)(\boldsymbol{x}) = x_j^{\alpha_j}$ so this cannot happen with

the choice of $\alpha_j \geq -p_j$. Noting that h_j is monotonically increasing, we consider the following cases:

(1) Suppose $x_j \geq (u_j + v_j)/2$. Then $(h_j \circ \varphi_j)(\boldsymbol{x}) \leq h_j(\min\{C_j, v_j - x_j\}) \leq h_j(\min\{C_j, (v_j - u_j)/2\}) \leq \min\{C_j^{\alpha_j}, (v_j - u_j)^{\alpha_j}/2^{\alpha_j}\}$, $x_j^{p_j} \leq (u_j + v_j)^{p_j}/2^{p_j}$ if $p_j < 0$ or $x_j^{p_j} \leq v_j^{p_j}$ if $p_j \geq 0$.

(2) Suppose $x_j \leq (u_j + v_j)/2$. Then $(h_j \circ \varphi_j)(\boldsymbol{x}) x_j^{p_j} \leq h_j(\min\{C_j, x_j - u_j\}) x_j^{p_j} = \min\{C_j^{\alpha_j}, (x_j - u_j)^{\alpha_j}\} x_j^{p_j}$. Now let $f(x) = (\min\{C_j, x - u_j\})^{\alpha_j} x^{p_j}$. Then $(\log f(x))' = \alpha_j/(x - u_j)\mathbb{1}_{x<u_j+C_j} + p_j/x$. For $x \geq u_j + C_j$ this has the same sign as p_j, otherwise it is equal to $((\alpha_j + p_j)x - u_j p_j)/(x(x - u_j)) \geq 0$ on (u_j, v_j) since $x > u_j$, $\alpha_j \geq -p_j$ and $\alpha_j \geq 0$. This implies that if $p_j \geq 0$ or $v_j - u_j \leq 2C_j$, f is increasing on $(u_j, (u_j + v_j)/2)$, and so $(h_j \circ \varphi_j)(\boldsymbol{x}) x_j^{p_j} \leq \min\{C_j, (v_j - u_j)/2\}^{\alpha_j} (u_j + v_j)^{p_j}/2^{p_j}$; otherwise, f is increasing on $(u_j, u_j + C_j)$ and decreasing on $(u_j + C_j, (u_j + v_j)/2)$, so $(h_j \circ \varphi_j)(\boldsymbol{x}) x_j^{p_j} \leq C_j^{\alpha_j}(u_j + C_j)^{p_j}$.

Thus, defining

$$\zeta(C_j, u_j, v_j, \alpha_j, p_j) \equiv \begin{cases} \min\{C_j, (v_j - u_j)/2\}^{\alpha_j} (u_j + v_j)^{p_j}/2^{p_j}, & p_j < 0, v_j - u_j \leq 2C_j, \\ \min\{C_j, (v_j - u_j)/2\}^{\alpha_j} (u_j + C_j)^{p_j}, & p_j < 0, v_j - u_j > 2C_j, \\ \min\{C_j, (v_j - u_j)/2\}^{\alpha_j} v_j^{p_j}, & p_j \geq 0, \end{cases}$$

we have $0 \leq (h_j \circ \varphi_j)(\boldsymbol{x}) x_j^{p_j} x_k^{p_k} x_\ell^{p_\ell} \leq \zeta(C_j, u_j, v_j, \alpha_j, p_j) v_k^{p_k} v_\ell^{p_\ell}$. Now assume additionally that $\alpha_j \geq \max\{1, 1 - p_j\}$, then by $h_j'(x) = \alpha_j x_j^{\alpha_j - 1}$, $0 \leq \partial_j (h_j \circ \varphi_j)(\boldsymbol{x}) x_j^{p_j} x_k^{p_k} \leq \alpha_j \zeta(C_j, u_j, v_j, \alpha_j - 1, p_j) v_k^{p_k}$.

First assume $a > 0$. Then assuming $\alpha_1, \ldots, \alpha_m \geq \max\{1, 2-2a, 2-2b, 1-a, 2-a, 2-b\} = \max\{1, 2 - a, 2 - b\}$, using the form of $\boldsymbol{\Gamma}$ and $\boldsymbol{g}$ in Section 3.5.2, for all j, k, ℓ we have $0 \leq \gamma_{j,k,\ell}(\mathbf{x}) \leq \varsigma_{\boldsymbol{\Gamma}} \equiv \max_{j,k=1,\ldots,m} \max\{\zeta(C_j, u_j, v_j, \alpha_j, 2a - 2) v_k^{2a}, \zeta(C_j, u_j, v_j, \alpha_j, 2b - 2)\}$ and $0 \leq g_{j,k}(\mathbf{x}) \leq \varsigma_{\boldsymbol{g}} \equiv \max_{j,k=1,\ldots,m} \max\{\alpha_j \zeta(C_j, u_j, v_j, \alpha_j - 1, a - 1) v_k^a + |a - 1| \zeta(C_j, u_j, v_j, \alpha_j, a - 2) v_k^a + a\zeta(C_j, u_j, v_j, \alpha_j, 2a - 2), \alpha_j \zeta(C_j, u_j, v_j, \alpha_j - 1, b - 1) + |b - 1| \zeta(C_j, u_j, v_j, \alpha_j, b - 2)\}$.

Then by Hoeffding's inequality,

$$\mathbb{P}\left(\max_{j,k,\ell}|\gamma_{j,k,\ell}-\mathbb{E}_0\gamma_{j,k,\ell}|\geq\epsilon_1/2\right)\leq 2\exp\left(-n\epsilon_1^2/(2\varsigma_{\mathbf{\Gamma}}^2)\right), \tag{B.11}$$

$$\mathbb{P}\left(\max_{j,k}|g_{j,k}-\mathbb{E}_0 g_{j,k}|\geq\epsilon_2\right)\leq 2\exp\left(-2n\epsilon_2^2/\varsigma_{\boldsymbol{g}}^2\right). \tag{B.12}$$

Let $\epsilon_1 \equiv \varsigma_{\mathbf{\Gamma}}\sqrt{2(\log m^\tau+\log 4)/n}$ and $\epsilon_2 \equiv \varsigma_{\boldsymbol{g}}\sqrt{(\log m^\tau+\log 4)/(2n)}$. With the choice of $\delta \leq 1+\sqrt{(\log m^\tau+\log 4)/(2n)}$ and by the fact that $0\leq\max_{j,k,\ell}\gamma_{j,k,\ell}\leq\varsigma_{\mathbf{\Gamma}}=\epsilon_1/(2\delta-2)$, (B.11) and (B.12) imply that

$$\begin{aligned}\mathbb{P}\left(\max_{j,k,\ell}|\delta\gamma_{j,k,\ell}-\mathbb{E}_0\gamma_{j,k,\ell}|\geq\epsilon_1\right)&\leq\mathbb{P}\left(\max_{j,k,\ell}|\gamma_{j,k,\ell}-\mathbb{E}_0\gamma_{j,k,\ell}|+(\delta-1)\max_{j,k,\ell}\gamma_{j,k,\ell}\geq\epsilon_1\right)\\&\leq\mathbb{P}\left(\max_{j,k,\ell}|\gamma_{j,k,\ell}-\mathbb{E}_0\gamma_{j,k,\ell}|\geq\epsilon_1/2\right)\leq m^{-\tau}/2, \qquad \text{(B.13)}\\ \mathbb{P}\left(\max_{j,k}|g_{j,k}-\mathbb{E}_0 g_{j,k}|\geq\epsilon_2\right)&\leq m^{-\tau}/2. \qquad \text{(B.14)}\end{aligned}$$

The results then follow by applying Theorem 1 in Lin et al. (2016).

In the case where $a=0$, and $u_k>0$ for all k,

$$|(h_j\circ\varphi_j)(\boldsymbol{x})x_j^{p_j}\log(x_k)\log(x_\ell)|\leq\zeta(C_j,u_j,v_j,\alpha_j,p_j)\cdot\max\{|\log(u_k)\log(u_\ell)|,|\log(v_k)\log(v_\ell)|\}$$

and everything else follows similarly as for $a>0$. □

Proof of Lemma 25. We show that X_j^{2a} for $a>0$ or $\log X_j$ for $a=0$ is sub-exponential by showing its moment-generating function is finite. Then the sub-exponentiality follows from Theorem 2.13 of Wainwright (2019).

First consider the case where $a=0$. In Corollary 17, we only require $\mathbf{K}$ to be positive definite without any restrictions on $\boldsymbol{\eta}$, and thus for any $t\in\mathbb{R}$, $\mathbb{E}_0\exp(t\log X_j)$ is the inverse normalizing constant for the model with parameters $\mathbf{K}_0$ and $\boldsymbol{\eta}_0+t\boldsymbol{e}_j$, where $\boldsymbol{e}_j$ is the vector with the j-th coordinate equal to 1 and the rest equal to 0, and is thus finite.

Next, consider $a>0$. Corollary 17 requires $\mathbf{K}_0$ to be positive definite, and in addition $\boldsymbol{\eta}_0\succ-\mathbf{1}_m$ if $b=0$. Then, again writing $\boldsymbol{x}^0/0=\log\boldsymbol{x}$ for the b part,

$$p_0(\boldsymbol{x})\exp\left(tx_j^{2a}\right)\propto\exp\left(-\frac{1}{2a}\boldsymbol{x}^{a\top}\mathbf{K}_0\boldsymbol{x}^a+\frac{1}{b}\boldsymbol{\eta}_0^\top\boldsymbol{x}^b+tx_j^{2a}\right)$$

$$\leq \exp\left(\sum_{k=1}^{m}\left(\left(-\lambda_{\min}\left(\mathbf{K}_0\right)+2at\mathbb{1}_{k=j}\right)x_k^{2a}/(2a)+\eta_{0,k}x_k^b/b\right)\right),$$

a constant times the density for the model with parameters $\operatorname{diag}\left(\lambda_{\min}\left(\mathbf{K}_0\right)\mathbf{1}_m-2at\boldsymbol{e}_j\right)$ and $\boldsymbol{\eta}_0$. Thus, $\mathbb{E}_0\exp\left(tX_j^{2a}\right)$ is finite for $t\in\left(-\infty,\lambda_{\min}\left(\mathbf{K}_0\right)/(2a)\right)\ni 0$. □

Proof of Corollary 26. Let the sub-exponential norm of X_j^{2a} be $\left\|X_j^{2a}\right\|_{\psi_1}\equiv\sup_{q\geq 1}(\mathbb{E}_0|X_j|^{2aq})^{1/q}/q$, then by Lemma 35.6)

$$\mathbb{P}\left(\left|X_j^{2a}-\mathbb{E}_0X_j^{2a}\right|\geq\epsilon_{3,j}\right)\leq\exp\left(-\min\left(\frac{\epsilon_3^2}{8e^2\left\|X_j^{2a}\right\|_{\psi_1}^2},\frac{\epsilon_3}{4e\left\|X_j^{2a}\right\|_{\psi_1}}\right)\right).$$

Letting

$$\epsilon_{3,j}\equiv\max\Big\{2\sqrt{2}e\left\|X_j^{2a}\right\|_{\psi_1}\sqrt{\log 3+\log n+\tau\log m+\log\left(m-|\rho_{\mathcal{D}}^*|\right)},\\ 4e\left\|X_j^{2a}\right\|_{\psi_1}\left(\log 3+\log n+\tau\log m+\log\left(m-|\rho_{\mathcal{D}}^*|\right)\right)\Big\},$$

then $\max\left\{\mathbb{E}_0X_j^{2a}-\epsilon_{3,j},0\right\}^{1/(2a)}\leq X_j^{(i)}\leq\left(\mathbb{E}_0X_j^{2a}+\epsilon_{3,j}\right)^{1/(2a)}$ for all $j\notin\rho_{\mathcal{D}}^*$ and $i=1,\ldots,n$ with probability at least $1-1/(3m^\tau)$. The rest follows as in the proof of Theorem 24.

□

Proof of Corollary 27. It suffices to bound $\boldsymbol{\Gamma}$ and $\boldsymbol{g}$ using their forms in Section 3.6.1 and apply Theorem 1 in Lin et al. (2016). We first bound $(h_j\circ\varphi_j)(\boldsymbol{x})x_j^{p_j}x_m^{p_m}$ with $h_j(x)=x^{\alpha_j}$, $\alpha_j\geq\max\{0,-p_j,-p_m,-p_j-p_m\}$, $p_j,p_m\in\mathbb{R}$, and $0<x_j+x_m<1$. By the definition of $\boldsymbol{\varphi}$ on simplices, $\varphi_j(\boldsymbol{x})=\min\{C_j,x_j,x_m\}$, so $(h_j\circ\varphi_j)(\boldsymbol{x})x_j^{p_j}x_m^{p_m}=\min\{C_j,x_j,x_m\}^{\alpha_j}x_j^{p_j}x_m^{p_m}$ and is tightly lower bounded by 0. Noting that $\min\{x_j,x_m\}<1/2$, we consider the following cases.

(1) If $C_j<\min\{x_j,x_m\}$, then $C_j<1/2$ and the quantity is $C_j^{\alpha_j}x_j^{p_j}x_m^{p_m}\leq C_j^{\alpha_j+(p_j)_-+(p_m)_-}<2^{-\alpha_j-(p_j)_--(p_m)_-}$ where $(y)_-=y$ if $y<0$ and 0 otherwise.

(2) Otherwise suppose $x_j\leq x_m$ and $x_j\leq C_j$, then the quantity is equal to $x_j^{\alpha_j+p_j}x_m^{p_m}$, which if $p_m\leq 0$ is upper bounded by $x_j^{\alpha_j+p_j+p_m}<2^{-\alpha_j-p_j-p_m}<1$; if $p_m>0$ it is upper

bounded by $((\alpha_j+p_j)/(\alpha_j+p_j+p_m))^{\alpha_j+p_j}(p_m/(\alpha_j+p_j+p_m))^{p_m}$ if $(\alpha_j+p_j)/(\alpha_j+p_j+p_m) \le 1/2$ or by $2^{-\alpha_j-p_j-p_m}$ otherwise. Note that the statement for $p_m > 0$ covers the one for $p_m \le 0$. The conclusion for $x_m \le x_j$ and $x_m \le C_j$ follows by symmetry, and note that at most one of $(\alpha_j+p_j)/(\alpha_j+p_j+p_m) \le 1/2$ and $(\alpha_j+p_m)/(\alpha_j+p_j+p_m) \le 1/2$ can hold.

In conclusion, defining $\zeta_2(\alpha_j, p_j, p_m) = \begin{cases} \left(\frac{\alpha_j+p_j}{\alpha_j+p_j+p_m}\right)^{\alpha_j+p_j}\left(\frac{p_m}{\alpha_j+p_j+p_m}\right)^{p_m}, & \text{if } p_m \ge \alpha_j+p_j, \\ \left(\frac{\alpha_j+p_m}{\alpha_j+p_j+p_m}\right)^{\alpha_j+p_m}\left(\frac{p_j}{\alpha_j+p_j+p_m}\right)^{p_j}, & \text{if } p_j \ge \alpha_j+p_m, \\ 2^{-\alpha_j-p_j-p_m}, & \text{otherwise,} \end{cases}$ we have $(h_j \circ \varphi_j)(\boldsymbol{x}) x_j^{p_j} x_m^{p_m} \le \zeta_2(\alpha_j, p_j, p_m) < 1$. Similarly $\partial_j(h_j \circ \varphi_j)(\boldsymbol{x}) x_j^{p_j} x_m^{p_m} \le \alpha_j \zeta_2(\alpha_j - 1, p_j, p_m) < \alpha_j$ if $\alpha_j - 1 \ge \max\{0, -p_j, -p_m, -p_j-p_m\}$.

Then for all j, k, ℓ, as long as $\alpha_j \ge \max\{1, 2-a, 2-b\}$, we have $0 \le \gamma_{j,k,\ell} < 1$, and similarly $0 \le g_{j,k} < \max_{j=1,\dots,m} \alpha_j + \max\{|a-1|+2a, |b-1|\}$. The rest follows from the same proof as Theorem 24.

Note that using the form of $\boldsymbol{\Gamma}$ in Section 3.6.1, a tighter bound for $\gamma_{j,k,\ell}$ is $\max_{j,k=1,\dots,m} \max\{\zeta_2(\alpha_j, 2a-2, 0), \zeta_2(\alpha_j, 4a-2, 0), \zeta_2(\alpha_j, 2a-2, 2a), \zeta_2(\alpha_j, 2b-2, 0), \zeta_2(\alpha_j, 0, 2a-2), \zeta_2(\alpha_j, 2a, 2a-2), \zeta_2(\alpha_j, 0, 4a-2), \zeta_2(\alpha_j, 0, 2b-2), \zeta_2(\alpha_j, a-1, a-1), \zeta_2(\alpha_j, 2a-1, 2a-1), \zeta_2(\alpha_j, a-1, 3a-1), \zeta_2(\alpha_j, 3a-1, a-1), \zeta_2(\alpha_j, b-1, b-1)\}$, and the one for $g_{j,k}$ can be similarly written in terms of $\zeta_2(\alpha_j, \cdot, \cdot)$ and $\alpha_j \zeta_2(\alpha_j - 1, \cdot, \cdot)$. □

Proof of Lemma 28. For $a > 0$ or $b > 0$, the proof of Lemma 25 works even for the simplex domain. We thus only consider the case where $a = b = 0$. Similar to the proof there, we prove by showing that the moment-generating function of $\log X_j$ is finite and invoking Theorem 2.13 of Wainwright (2019). According to Theorem 21, assume

(1) $\mathbf{K}_0$ is positive definite, or

(2) $\mathbf{K}_0 \mathbf{1}_m = \mathbf{0}$, $\mathbf{K}_{0,-k,-k}$ is positive definite for some $k = 1, \dots, m$, and $\mathbf{1}_m^\top \boldsymbol{\eta} + m \ge 0$, or

(3) $\mathbf{K}_0 \mathbf{1}_m = \mathbf{0}$, $\mathbf{K}_0$ is positive semi-definite, and $\boldsymbol{\eta} \succ -\mathbf{1}_m$.

For any suitable t, $\mathbb{E}_0 \exp(t \log X_j)$ is the inverse normalizing constant for the model with parameters $\mathbf{K}_0$ and $\boldsymbol{\eta}_0 + t\boldsymbol{e}_j$, and is thus finite for (1) with $t \in \mathbb{R}$ and (3) with $t \in (-1 - \eta_{0,j}, +\infty) \ni 0$. For (2), recall that in the proof of Theorem 21 we have shown that for any $k = 1, \ldots, m$, the density of $\log \boldsymbol{X}_{-k} - (\log X_k)\, \mathbf{1}_{m-1}$ is bounded by a constant times a Gaussian density, and thus $\mathbb{E}_0 \left[X_j^t / X_k^t\right] = \mathbb{E}_0 \exp\left(t(\log X_j - \log X_k)\right) < +\infty$ for any $k = 1, \ldots, m$ and $t \in \mathbb{R}$. So for any $t < 0$

$$\begin{aligned}
\mathbb{E}_0 X_j^t &\leq \mathbb{E}_0 \left[X_j^t | X_j \geq 1/m\right] \mathbb{P}\left(X_j \geq 1/m\right) + \sum_{k \neq j} \mathbb{E}_0 \left[X_j^t \middle| X_k \geq 1/m\right] \mathbb{P}\left(X_k \geq 1/m\right) \\
&\leq m^{-t} \mathbb{P}\left(X_j \geq 1/m\right) + \sum_{k \neq j} m^{-t} \mathbb{E}_0 \left[X_j^t / X_k^t \middle| X_k \geq 1/m\right] \mathbb{P}\left(X_k \geq 1/m\right) \\
&\leq m^{-t} + \sum_{k \neq j} m^{-t} \mathbb{E}_0 \left[X_j^t / X_k^t\right] < +\infty.
\end{aligned}$$

On the other hand, $\mathbb{E}_0 X_j^t \leq 1$ for $t \geq 0$. Thus, $\mathbb{E}_0 \exp(t \log X_j) < +\infty$ for any $t \in \mathbb{R}$ for (2). Hence, for all of (1)–(3) we have $\mathbb{E}_0 \exp(t \log X_j) < +\infty$ for t in a neighborhood around 0. □

Proof of Corollary 29. Let the sub-exponential norm of $\log X_j$ be $\| \log X_j \|_{\psi_1} \equiv \sup_{q \geq 1} (\mathbb{E}_0 |\log X_j|^q)^{1/q} / q$, then by Lemma 35.6),

$$\mathbb{P}\left(-\log X_j + \mathbb{E}_0 \log X_j \geq \epsilon_3\right) \leq \exp\left(-\min\left(\frac{\epsilon_3^2}{8e^2 \| \log X_j \|_{\psi_1}^2}, \frac{\epsilon_3}{4e \| \log X_j \|_{\psi_1}}\right)\right).$$

Letting

$$\begin{aligned}
\epsilon_3 \equiv \max\Big\{ & 2\sqrt{2}e \max_j \| \log X_j \|_{\psi_1} \sqrt{\log 3 + \log n + (\tau + 1) \log m}, \\
& 4e \max_j \| \log X_j \|_{\psi_1} (\log 3 + \log n + (\tau + 1) \log m) \Big\},
\end{aligned}$$

then $0 \leq -\log X_j^{(i)} \leq \max_k \mathbb{E}_0 \log X_k + \epsilon_3$ for all $j = 1, \ldots, m$ and $i = 1, \ldots, n$ with probability at least $1 - 1/(3m^\tau)$. The rest follows as in the proof of Theorem 24 and Corollary 27.

□

Appendix C

APPENDIX TO CHAPTER 4

In this appendix we present proofs for the theorems and corollaries in Chapter 4.

We first prove the following lemma that states that if two sums of distinct (ignoring the multiplicative constant) exponentials of polynomials in $\boldsymbol{y} \in \mathbb{R}^m$ agree almost everywhere in $\mathbb{R}^m$, then they must have the same number of terms and there must be a 1-1 correspondence between the terms.

Lemma 39. *Let the number of variable be $m \geq 1$ and the degree be $p \geq 1$. Let $\mathcal{D} \equiv \{\boldsymbol{d} \in \mathbb{Z}^m_{\geq 0} : 1 \leq \sum_{j=1}^m d_j \leq p\}$ be the set of nonnegative integer-valued m-vectors with ℓ_1 norm $\in [1, p]$. Given a vector $\boldsymbol{a} \in \mathbb{R}^{|\mathcal{D}|}$ indexed by $\boldsymbol{d} \in \mathcal{D}$ (i.e. $a_{\boldsymbol{d}} \in \mathbb{R}$ for all $\boldsymbol{d} \in \mathcal{D}$), define*

$$f^{(m)}(\boldsymbol{y}; \boldsymbol{a}) \equiv \exp\left(\sum_{\boldsymbol{d} \in \mathcal{D}} a_{\boldsymbol{d}} \prod_{j=1} y_i^{d_j}\right),$$

the exponential of the corresponding polynomial of degree $\leq p$ in $\boldsymbol{y} \in \mathbb{R}^m$. Note that $f^{(m)}$ does not have a constant term, and has degrees $\boldsymbol{d} \in \mathcal{D}$ and coefficients $\boldsymbol{a}$.

Suppose we have

$$\sum_{i=1}^{N_a} a_0^i f^{(m)}(\boldsymbol{y}; \boldsymbol{a}^i) = \sum_{i=1}^{N_b} b_0^i f^{(m)}(\boldsymbol{y}; \boldsymbol{b}^i) \tag{C.1}$$

for almost every $\boldsymbol{y} \equiv (y_1, \ldots, y_m) \in \mathbb{R}^m$ with respect to the Lebesgue measure, where $N_a \geq 0$, $N_b \geq 0$, $\{\boldsymbol{a}^i\}_{i=1}^{N_a}$ are N_a distinct vectors in $\mathbb{R}^{|\mathcal{D}|}$, $\{\boldsymbol{b}^i\}_{i=1}^{N_b}$ are N_b distinct vectors in $\mathbb{R}^{|\mathcal{D}|}$ (otherwise just combine the coefficients), and $a_0^i, b_0^i \in \mathbb{R}\backslash\{0\}$ for all i. In other words, both sides of (C.1) are a sum of distinct exponentials of polynomials.

Then we must have $N_a = N_b$ and there is a permutation π of $\{1, \ldots, N_a\}$ such that $\boldsymbol{a}^i = \boldsymbol{b}^{\pi(i)}$ and $a_0^i = b_0^{\pi(i)}$, i.e. there is a 1-1 correspondence between the summands on both sides of (C.1).

Proof of Lemma 39. First note that both sides of (C.1) are continuous functions, and so is their difference, which is 0 almost everywhere by assumption. Thus, the inverse image of the open set $\mathbb{R}\backslash\{0\}$ under the difference is also open, and must be the empty set since it has measure 0. (C.1) thus holds for all $\boldsymbol{y} \in \mathbb{R}^m$.

We prove by induction on m, and first show the result for $m = 1$. In this case, $f^{(1)}(y_1; \boldsymbol{a}) \equiv \exp(a_1 y_1 + \cdots + a_p y_1^p)$, and $\boldsymbol{a}$ is just a p-vector.

First suppose $N_a \neq 0$ and $N_b \neq 0$. Observe that as $x \nearrow +\infty$, if $a_0 \neq 0$, the function $a_0 \exp(a_1 x + \cdots + a_p x^p)$ goes to

(i) $a_0 \neq 0$ if $a_1 = \cdots = a_p = 0$, or

(ii) 0 if $a_{d_{\max\neq 0}(\boldsymbol{a})} < 0$ where $d_{\max\neq 0}(\boldsymbol{a})$ is the largest $d \in \{1, \ldots, p\}$ such that $a_d \neq 0$, or

(iii) $+\infty$ if $a_{d_{\max\neq 0}(\boldsymbol{a})} > 0$.

Rearrange the terms on the left of (C.1) so that for each $1 \leq i < j \leq N_a$ we have $(\boldsymbol{a}^i - \boldsymbol{a}^j)_{d_{\max\neq 0}(\boldsymbol{a}^i - \boldsymbol{a}^j)} > 0$, and denote this total order as $\boldsymbol{a}^i > \boldsymbol{a}^j$. Rearrange the right-hand side similarly. By the assumption that $\{\boldsymbol{a}^i\}_{i=1}^{N_a}$ are distinct, $\boldsymbol{a}^i - \boldsymbol{a}^j \neq 0$, so $d_{\max\neq 0}(\boldsymbol{a}^i - \boldsymbol{a}^j)$ exists and this rearrangement is possible. Now dividing both sides of (C.1) by $f^{(1)}(y_1; \boldsymbol{a}^1) = \exp(a_1^1 y_1 + \cdots + a_p^1 y_1^p)$ we have

$$a_0^1 + \sum_{i=2}^{N_a} a_0^i f^{(1)}(y_1; \boldsymbol{a}^i - \boldsymbol{a}^1) = \sum_{i=1}^{N_b} b_0^i f^{(1)}(y_1; \boldsymbol{b}^i - \boldsymbol{a}^1). \tag{C.2}$$

Since $a_0^1 \neq 0$, and by the unique maximality of $\boldsymbol{a}^1$, as $y_1 \nearrow +\infty$, all terms in the summation on the left go to 0 (case (ii)). Thus, the right-hand side necessarily also goes to $a_0^1 \neq 0$, landing us in case (i) for at least one (and only one because $\boldsymbol{b}^i$ are unique) term on the right, i.e. $\boldsymbol{b}^i - \boldsymbol{a}^1 = \boldsymbol{0}$. (A nonzero finite limit cannot come from a sum of terms that go to $+\infty$ with positive and negative weights, since they must grow at different rates by uniqueness of $\boldsymbol{b}^i - \boldsymbol{a}^1$.) Since summands on both sides are sorted, we must have $\boldsymbol{b}^1 = \boldsymbol{a}^1$.

Then (C.2) becomes $a_0^1 - b_0^1 + \sum_{i=2}^{N_a} a_0^i f^{(1)}(y_1; \boldsymbol{a}^i - \boldsymbol{a}^1) = \sum_{i=2}^{N_b} b_0^i f^{(1)}(y_1; \boldsymbol{b}^i - \boldsymbol{a}^1)$. If $a_0^1 \neq b_0^1$, by the same reasoning there exists another $i \in \{2, \ldots, N_b\}$ such that $\boldsymbol{b}^i - \boldsymbol{a}^1 = \boldsymbol{0}$, violating uniqueness of $\{\boldsymbol{b}^i\}_{i=1}^{N_b}$. Thus, $a_0^1 = b_0^1$ and $\boldsymbol{a}^1 = \boldsymbol{b}^1$, and we have reduced the number of summands on both sides of (C.2) by 1 to

$$\sum_{i=2}^{N_a} a_0^i f^{(1)}(y_1; \boldsymbol{a}^i - \boldsymbol{a}^1) = \sum_{i=2}^{N_b} b_0^i f^{(1)}(y_1; \boldsymbol{b}^i - \boldsymbol{a}^1).$$

Continuing this process by each time dividing both sides by $f^{(1)}(y_1; \boldsymbol{a}^j - \boldsymbol{a}^{j-1})$, we would have matched $\min\{N_a, N_b\}$ pairs of coefficients between the a and the b groups. If $N_a \neq N_b$, assume $N_a > N_b$ without loss of generality, then

$$\sum_{i=N_b+1}^{N_1} a_0^i f^{(1)}(y_1; \boldsymbol{a}^i - \boldsymbol{a}^{N_b}) = \text{const.}$$

Here the right-hand side is a constant that could be nonzero, because the argument for $a_0^1 = b_0^1$ in our first elimination step does not apply here. Dividing both sides by $f^{(1)}(y_1; \boldsymbol{a}^{N_b+1} - \boldsymbol{a}^{N_b})$, we have $a_0^{N_b+1} + \sum_{i=N_b+2}^{N_1} a_0^i f^{(1)}(y_1; \boldsymbol{a}^i - \boldsymbol{a}^{N_b+1}) = f^{(1)}(y_1; \boldsymbol{a}^{N_b} - \boldsymbol{a}^{N_b+1})$. By maximality of $\boldsymbol{a}^{N_b+1}$ among $\boldsymbol{a}^{N_b+1}, \ldots, \boldsymbol{a}^{N_a}$, the left-hand side goes to $a_0^{N_b+1} \neq 0$ as $y_1 \nearrow +\infty$, while since $\boldsymbol{a}^{N_b} > \boldsymbol{a}^{N_b+1}$, the right-hand side goes to $+\infty$, a contradiction. Thus, $N_a = N_b$, $a_0^i = b_0^i$ and $\boldsymbol{a}^i = \boldsymbol{b}^i$ for $i = 1, \ldots, N_a$, proving the $m = 1$ case when $N_a \neq 0$ and $N_b \neq 0$.

Now consider the case where one of N_a and N_b is 0; assume without loss of generality that $N_b = 0$, then by division by $f^{(1)}(y_1; \boldsymbol{a}^1)$, the right-hand side is constant 0, while the left-hand side goes to $a_0^1 \neq 0$ unless $N_a = 0$, so $N_a = N_b = 0$.

Now suppose the result holds for some $m - 1 \geq 1$, and suppose either $N_a \neq 0$ or $N_b \neq 0$, otherwise there is nothing to prove. We denote $\boldsymbol{a}_1$ as the subvector of $\boldsymbol{a}$ corresponding to $\boldsymbol{d}$ with $d_1 \geq 1$, i.e. $\{a_{\boldsymbol{d}}\}_{\boldsymbol{d} \in \mathcal{D},\, d_1 \geq 1}$, and $\boldsymbol{a}_{-1}$ as that of $\boldsymbol{a}$ with $d_1 = 0$. Separating out the terms involving y_1,

$$\begin{aligned} f^{(m)}(\boldsymbol{y}; \boldsymbol{a}^i) &= \exp\left\{ \sum_{d=1}^{p} \left(\sum_{\boldsymbol{d} \in \mathcal{D},\, d_1 = d} a_{\boldsymbol{d}}^i \prod_{j=2}^{m} y_j^{d_j} \right) y_1^d \right\} \exp\left(\sum_{\boldsymbol{d} \in \mathcal{D},\, d_1 = 0} a_{\boldsymbol{d}}^i \prod_{j=2}^{m} y_j^{d_j} \right) \\ &= f^{(1)}\left(y_1; \boldsymbol{a}_{1*}^i(\boldsymbol{y}_{-1})\right) f^{(m-1)}(\boldsymbol{y}_{-1}; \boldsymbol{a}_{-1}^i), \end{aligned}$$

where $\boldsymbol{a}_{1*}^i(\boldsymbol{y}_{-1}) : \mathbb{R}^{m-1} \to \mathbb{R}^p$ is a vector-valued function in $\boldsymbol{y}_{-1}$, with d-th coordinate a polynomial $\sum_{\boldsymbol{d} \in \mathcal{D},\, d_1 = d} a_{\boldsymbol{d}}^i \prod_{j=2}^{m} y_j^{d_j}$, and coefficients corresponding to $\boldsymbol{a}_1^i$. Note that there is a one-to-one correspondence between such a function $\boldsymbol{a}_{1*}^i$ and vector $\boldsymbol{a}_1^i$. So we can rewrite (C.1) as

$$\sum_{i=1}^{N_a} a_0^i f^{(1)}(y_1; \boldsymbol{a}_{1*}^i(\boldsymbol{y}_{-1})) f^{(m-1)}(\boldsymbol{y}_{-1}; \boldsymbol{a}_{-1}^i) = \sum_{i=1}^{N_b} b_0^i f^{(1)}(y_1; \boldsymbol{b}_{1*}^i(\boldsymbol{y}_{-1})) f^{(m-1)}(\boldsymbol{y}_{-1}; \boldsymbol{b}_{-1}^i)$$

for all $\boldsymbol{y} \in \mathbb{R}^m$. Then collecting terms with the same $f^{(1)}$ (same $\boldsymbol{a}_1^i$ ($\boldsymbol{a}_{1*}^i$) or $\boldsymbol{b}_1^i$ ($\boldsymbol{b}_{1*}^i$)),

$$\sum_{\ell=1}^{C} f^{(1)}(y_1; \boldsymbol{c}_{1*}^{\ell}(\boldsymbol{y}_{-1})) \left\{ \sum_{j=1}^{n_\ell^a} a_0^{k_{\ell j}^a} f^{(m-1)}(\boldsymbol{y}_{-1}; \boldsymbol{a}_{-1}^{k_{\ell j}^a}) + \sum_{j=1}^{n_\ell^b} b_0^{k_{\ell j}^b} f^{(m-1)}(\boldsymbol{y}_{-1}; \boldsymbol{b}_{-1}^{k_{\ell j}^b}) \right\} = 0, \quad \text{(C.3)}$$

where $C > 0$, each $\boldsymbol{c}_1^\ell$ (coefficients for $\boldsymbol{c}_{1*}^\ell$) is some $\boldsymbol{a}_1^i$ or $\boldsymbol{b}_1^i$, and $\{\boldsymbol{c}_1^\ell\}_{\ell=1}^C$ are distinct. Here, let $\{k_{11}^a, \ldots, k_{1,n_1^a}^a, \ldots, k_{C1}^a, \ldots, k_{Cn_C^a}^a\}$ be a permutation of $\{1, \ldots, N_a\}$, and $\{k_{11}^b, \ldots, k_{1,n_1^b}^b, \ldots, k_{C1}^b, \ldots, k_{Cn_C^b}^b\}$ a permutation of $\{1, \ldots, N_b\}$.

Since $\{\boldsymbol{c}_1^\ell\}_{\ell=1}^C$ are distinct, $\{\boldsymbol{c}_{1*}^\ell\}_{\ell=1}^C$ are distinct finite polynomials in $\boldsymbol{y}_{-1} \in \mathbb{R}^{m-1}$. For each pair of such distinct polynomials, the lemma of Okamoto (1973) implies that they only agree at a Lebesgue-null subset of $\mathbb{R}^{n-1}$, so all polynomials are distinct except on a null set. Thus, for almost every fixed $\boldsymbol{y}_{-1} \in \mathbb{R}^{m-1}$, the left-hand side of (C.3) is a sum of $C > 0$ distinct $f^{(1)}$'s in y_1 multiplied by constant weights depending on $\boldsymbol{y}_{-1}$. But the right-hand side is a sum of 0 terms, so by the result for $m = 1$ we necessarily have

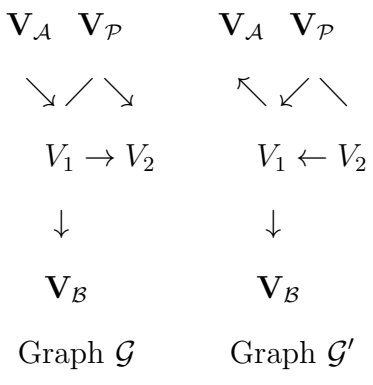
$$\sum_{j=1}^{n_\ell^a} a_0^{k_{\ell j}^a} f^{(m-1)}(\boldsymbol{y}_{-1}; \boldsymbol{a}_{-1}^{k_{\ell j}^a}) = \sum_{j=1}^{n_\ell^b} -b_0^{k_{\ell j}^b} f^{(m-1)}(\boldsymbol{y}_{-1}; \boldsymbol{b}_{-1}^{k_{\ell j}^b}) \quad \text{(C.4)}$$

for all $\ell = 1, \ldots, C$ for almost every $\boldsymbol{y}_{-1}$. Fixing $\ell \in \{1, \ldots, C\}$, for any $1 \leq j_1 < j_2 \leq n_\ell^a$, $\boldsymbol{a}^{k_{\ell j_1}^a} \neq \boldsymbol{a}^{k_{\ell j_2}^a}$ and $\boldsymbol{a}_1^{k_{\ell j_1}^a} = \boldsymbol{a}_1^{k_{\ell j_2}^a}$ implies $\boldsymbol{a}_{-1}^{k_{\ell j_1}^a} \neq \boldsymbol{a}_{-1}^{k_{\ell j_2}^a}$, and similarly for $\boldsymbol{b}$. Thus, each term on the left-hand side of (C.4) has its unique coefficients, and similarly for the right-hand side. Since (C.4) holds for almost every $\boldsymbol{y}_{-1}$, by the result for $m-1$ variables, we must have $n_\ell^a = n_\ell^b$ and each $a_0^{k_{\ell j}^a} = b_0^{k_{\ell \pi(j)}^b}$ and $\boldsymbol{a}_{-1}^{k_{\ell j}^a} = \boldsymbol{b}_{-1}^{k_{\ell \pi(j)}^b}$ for some permutation π of $\{1, \ldots, n_\ell^a\}$, which in turn implies $\boldsymbol{a}^{k_{\ell j}^a} = \boldsymbol{b}^{k_{\ell \pi(j)}^b}$ for all $j = 1, \ldots, n_\ell^a$ by construction of the groups $\ell = 1, \ldots, C$. Since this holds for all ℓ, $N_a = \sum_{\ell=1}^C n_\ell^a = \sum_{\ell=1}^C n_\ell^b = N_b$, and we have thus again matched each $\boldsymbol{a}^\ell$ with a $\boldsymbol{b}^\ell$ as well as the corresponding a_0's with b_0's. This ends the proof for m, and the entire proof. □

Proof of Theorem 30. Suppose $\mathcal{G}$ and $\mathcal{G}'$ have the same node set $\mathcal{V}$ and are Markov equivalent, otherwise the distributions represented by them are trivially not identical.

Now suppose $p(\boldsymbol{Y})$ is Markov and faithful with respect to $\mathcal{G}$ and $\mathcal{G}'$, and factorize w.r.t. both graphs with *strong Hurdle polynomial* parameters. Then by Proposition 31,

there exist V_1 and V_2 such that $V_1 \to V_2$ in $\mathcal{G}$, $V_2 \to V_1$ in $\mathcal{G}'$ and $\mathcal{P} \equiv \text{pa}_{\mathcal{G}}(V_2)\backslash\{V_1\} = \text{pa}_{\mathcal{G}'}(V_1)\backslash\{V_2\}$. Following the arguments in the proof of Proposition 31 in Peters et al. (2014), recursively marginalizing out nodes without children but having the same parents in both graphs, we eventually obtain structures as follows, where $\mathcal{A}$ and $\mathcal{B}$ are some unknown node sets and V_2 does not have any children in Graph $\mathcal{G}$.



We consider the (α, β, k)-parametrization only, since the result for the (p, μ, σ^2) naturally follows from their relationship (4.6). For notational simplicity write V_1 and V_2 as nodes 1 and 2. Suppose after marginalization above we are left with nodes $\mathcal{V}_0 \subseteq \mathcal{V}$ which include 1, 2, $\mathbf{V}_\mathcal{A}$, $\mathbf{V}_\mathcal{B}$ and $\mathbf{V}_\mathcal{P}$ illustrated above. Now let $Y_U = 0$ for all $U \in \mathcal{V}_0\backslash\{2\}$, and let $Y_2 \neq 0$. Then the joint distribution $p(Y_2 = y_2 \neq 0, \boldsymbol{y}_{\mathcal{V}_0} = \mathbf{0})$ using $\mathcal{G}$ is proportional to

$$\prod_{V \in \mathcal{V}_0} \frac{\exp\{\alpha_V(\boldsymbol{y}_{\text{pa}_\mathcal{G}(V)})\mathbb{1}_{y_v} + \beta_v(\boldsymbol{y}_{\text{pa}_\mathcal{G}(V)})y_V - k_V y_V^2/2\}}{\sqrt{2\pi/k_V}\exp\{\alpha_V(\boldsymbol{y}_{\text{pa}_\mathcal{G}(V)}) + \beta_V(\boldsymbol{y}_{\text{pa}_\mathcal{G}(V)})^2/(2k_V)\} + 1}\Bigg|_{y_2 \neq 0, \boldsymbol{y}_{\mathcal{V}_0\backslash\{2\}} = \mathbf{0}}$$
$$\propto \exp\{\beta_2(\mathbf{0})y_2 - k_2 y_2^2/2\}$$

since 2 does not have any child in $\mathcal{G}$. But using $\mathcal{G}'$, the same joint distribution is proportional to

$$\prod_{V \in \mathcal{V}_0} \frac{\exp\{\alpha'_V(\boldsymbol{y}_{\text{pa}_{\mathcal{G}'}(V)})\mathbb{1}_{y_V} + \beta'_V(\boldsymbol{y}_{\text{pa}_{\mathcal{G}'}(V)})y_V - k'_V y_V^2/2\}}{\sqrt{2\pi/k'_V}\exp\{\alpha'_V(\boldsymbol{y}_{\text{pa}_{\mathcal{G}'}(V)}) + \beta'_V(\boldsymbol{y}_{\text{pa}_{\mathcal{G}'}(V)})^2/(2k'_V)\} + 1}\Bigg|_{y_2 \neq 0, \boldsymbol{y}_{\mathcal{V}_0\backslash\{2\}} = \mathbf{0}}$$
$$\propto \exp\{\beta'_2(\mathbf{0})y_2 - k'_2 y_2^2/2\}\times$$

$$\prod_{U \in \mathcal{P} \cup \{1\},\, 2 \in \mathrm{pa}_{\mathcal{G}'}(U)} \frac{1}{\sqrt{2\pi/k'_U}\exp\{\alpha'_U(y_2, \mathbf{0}) + \beta'_U(y_2, \mathbf{0})^2/(2k'_U)\} + 1}$$

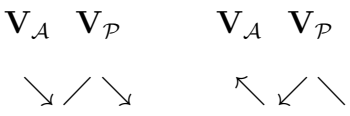
where in the case where $\mathrm{pa}_{\mathcal{G}'}(2) = \varnothing$ replace $\alpha'_2(\mathbf{0})$ and $\beta'_2(\mathbf{0})$ by constants α'_2 and β'_2, and $\alpha'_U(y_2, \mathbf{0})$ and $\beta'_U(y_2, \mathbf{0})$ denote setting all parents other than 2 in the Hurdle polynomials α'_U and β'_U to $\mathbf{0}$. Since the two joint distributions derived from both graphs must be proportional to each other, we get for $y_2 \neq 0$

$$\begin{aligned}
&\exp\left[y_2\{\beta'_2(\mathbf{0}) - \beta_2(\mathbf{0})\} - (k'_2 - k_2)y_2^2/2\right] \\
&\qquad \propto \prod_{U \in \mathcal{P} \cup \{1\},\, 2 \in \mathrm{pa}_{\mathcal{G}'}(U)} \left[\sqrt{2\pi/k'_U}\exp\left\{\alpha'_U(y_2, \mathbf{0}) + \beta'_U(y_2, \mathbf{0})^2/(2k'_U)\right\} + 1\right]. \quad \text{(C.5)}
\end{aligned}$$

Note that $2 \in \mathrm{pa}_{\mathcal{G}'}(1)$ and thus the product on the right of (C.5) has at least one term. Thus, supposing that for at least one of $U \in \mathcal{P} \cup \{1\}$ such that $2 \in \mathrm{pa}_{\mathcal{G}'}(U)$, $\alpha'_U(Y_2, \mathbf{0}) + \beta'_U(Y_2, \mathbf{0})^2/(2k'_U)$ is nonconstant in $Y_2 \neq 0$, then the right-hand side of (C.5) can be expanded into a sum of at least two exponentials of polynomials in y_2 (including the constant 1 as a degenerated exponential polynomial), while the left-hand side is a single polynomial in y_2. This is a contradiction according to Lemma 39, and thus the assumption of having *strong Hurdle polynomials* as the parameters in the Hurdle conditionals implies that $p(\boldsymbol{Y})$ cannot be represented by both $\mathcal{G}$ and $\mathcal{G}'$, which ends the proof.

□

Proof of Theorem 32. As in the proof of Theorem 30 using Proposition 31, under the assumptions there exist V_1 and V_2 such that $\mathcal{P} \equiv \mathrm{pa}_{\mathcal{G}}(V_2)\backslash\{V_1\} = \mathrm{pa}_{\mathcal{G}'}(V_1)\backslash\{V_2\}$ with $V_1 \to V_2$ in $\mathcal{G}$ and $V_2 \to V_1$ in $\mathcal{G}'$. Following the arguments in the proof of Proposition 31 in Peters et al. (2014), recursively marginalizing out nodes without children but having the same parents in both graphs, we again obtain structures as follows.



$$\begin{array}{cc} V_1 \to V_2 & V_1 \leftarrow V_2 \\ \downarrow & \downarrow \\ \mathbf{V}_{\mathcal{B}} & \mathbf{V}_{\mathcal{B}} \\ \text{Graph } \mathcal{G} & \text{Graph } \mathcal{G}' \end{array}$$

To ease the notation assume we again write $V_1 = 1$ and $V_2 = 2$.Note that the distribution of each node conditional on some other nodes is the sum of a point mass at 0 and a continuous distribution over $\mathbb{R}$, which follows by induction and the fact that the indefinite integral of a continuous density is continuous and that the sum of continuous densities is continuous. We focus on the continuous components, and wish to reach the conclusion using the factorization

$$\begin{aligned} P(y_1, y_2 | \boldsymbol{Y}_{\mathcal{P}} = \boldsymbol{y}_{\mathcal{P}}) &= P(y_1 | \boldsymbol{Y}_{\mathcal{P}} = \boldsymbol{y}_P) P(y_2 | y_1, \boldsymbol{Y}_{\mathcal{P}} = \boldsymbol{y}_{\mathcal{P}}) \\ &= P(y_2 | \boldsymbol{Y}_{\mathcal{P}} = \boldsymbol{y}_{\mathcal{P}}) P(y_1 | y_2, \boldsymbol{Y}_{\mathcal{P}} = \boldsymbol{y}_{\mathcal{P}}), \end{aligned}$$

where the second terms in both decompositions are a regular Hurdle conditional w.r.t. $\mathcal{G}$ and $\mathcal{G}'$, respectively, and we write the first terms as

$$P(y_1 | \boldsymbol{Y}_{\mathcal{P}} = \boldsymbol{y}_{\mathcal{P}}) \propto \exp\{\mathbb{1}_{y_1} \delta_1 + f_1(y_1)\}$$

and

$$P(y_2 | \boldsymbol{Y}_{\mathcal{P}} = \boldsymbol{y}_{\mathcal{P}}) \propto \exp\{\mathbb{1}_{y_2} \delta_2' + f_2'(y_1)\}$$

in terms of the conditional densities w.r.t. λ. Here f_1 and f_2' are continuous functions in $\mathbb{R}$ with no additive constant term, and δ_1 and δ_2' are constants.

We prove the results in the (α, β, k)-parameterization only, since results for the (p, μ, σ^2)-parameterization would follow from their relationship (4.6). In our model, we assumed the α and β parameters for each node to be polynomial in the parents and their indicators. We also assumed that for each node, either the β function is nonconstant in any of the parents, or α depends on the value of all of its parents.

Consider a generic β function associated with some generic parent set $\mathcal{P} \equiv \mathcal{P}_1 \sqcup \{p_0\}$ with $p_0 \notin \mathcal{P}_1 \neq \varnothing$ and suppose that β is nonconstant in any of $\mathcal{P}$, and write $\beta(\boldsymbol{y}_{\mathcal{P}})$ equivalently

as $\beta(y_{p_0}, \boldsymbol{y}_{\mathcal{P}_1})$. Then $\beta(\boldsymbol{y}_{\mathcal{P}})$ has the form $\beta_{-1}(\boldsymbol{y}_{\mathcal{P}_1}) + \beta_0(\boldsymbol{y}_{\mathcal{P}_1})\mathbb{1}_{y_1} + \sum_{i=1}^{k} \beta_i(\boldsymbol{y}_{\mathcal{P}_1}) y_1^i$, where by construction β_{-1} through β_k are (potentially constant or even zero) Hurdle polynomials in $\boldsymbol{y}_{\mathcal{P}_1}$, but there must exist some $j = 0, \ldots, k$ such that β_j is nonzero. By the lemma of Okamoto (1973), $\beta_j(\boldsymbol{y}_{\mathcal{P}_1}) \neq 0$ for (Lebesgue) almost every $\boldsymbol{y}_{\mathcal{P}_1} \in \mathbb{R}^{|\mathcal{P}_1|}$. Thus, $\beta(y_{p_0}, \boldsymbol{y}_{\mathcal{P}_1})$ is nonconstant in y_{p_0} for almost every $\boldsymbol{y}_{\mathcal{P}_1} \in \mathbb{R}^{|P_1|}$. Formally, define

$$\mathcal{Y}_{\beta,p_0,\mathcal{P}_1} \equiv \left\{ \boldsymbol{y}_{\mathcal{P}_1} \in \mathbb{R}^{|\mathcal{P}_1|} : \beta\left(y_{p_0}, \boldsymbol{y}_{\mathcal{P}_1}\right) \text{ nonconstant function in } y_{p_0} \right\}.$$

Thus $\mathbb{R}^{|\mathcal{P}_1|} \backslash \mathcal{Y}_{\beta,p_0,\mathcal{P}_1}$ has zero Lebesgue measure assuming β is nonconstant in its any of $\mathcal{P}$. Hence, by a similar argument, under the assumptions of the theorem, letting

$$\mathcal{Y}_{\alpha,\beta,p_0,\mathcal{P}_1} \equiv \left\{ \boldsymbol{y}_{\mathcal{P}_1} \in \mathbb{R}^{|\mathcal{P}_1|} : \beta\left(y_{p_0}, \boldsymbol{y}_{\mathcal{P}_1}\right) \text{ nonconstant function in } y_{p_0} \text{ or } \alpha\left(y_{p_0}, \boldsymbol{y}_{\mathcal{P}_1}\right) \text{ depends on the value of } y_{p_0} \right\},$$

$\mathbb{R}^{|\mathcal{P}_1|} \backslash \mathcal{Y}_{\alpha,\beta,p_0,\mathcal{P}_1}$ has zero Lebesgue measure.

Now we go back to $\mathcal{G}$ and $\mathcal{G}'$. Suppose $\mathcal{P} \neq \varnothing$ and that the Hurdle density of node 2 conditional on $\{1\} \sqcup \mathcal{P}$ in $\mathcal{G}$ have α and β parameters $\alpha_2(y_1, \boldsymbol{y}_{\mathcal{P}})$ and $\beta_2(y_1, \boldsymbol{y}_{\mathcal{P}})$, and let those for 1 conditional on $\{2\} \sqcup \mathcal{P}$ in $\mathcal{G}'$ be $\alpha_1'(y_2, \boldsymbol{y}_{\mathcal{P}})$ and $\beta_1'(y_2, \boldsymbol{y}_{\mathcal{P}})$. We also denote $\mathcal{Y}_* \equiv \mathcal{Y}_{\alpha_2,\beta_2,1,\mathcal{P}} \cap \mathcal{Y}_{\alpha_1',\beta_1',2,\mathcal{P}}$, which by discussion above contains almost every $\boldsymbol{y}_{\mathcal{P}} \subset \mathbb{R}^{|\mathcal{P}|}$.

From now on we thus fix $\boldsymbol{y}_{\mathcal{P}} \in \mathcal{Y}_*$ and condition on $\boldsymbol{Y}_{\mathcal{P}} = \boldsymbol{y}_{\mathcal{P}}$, and omit the dependency of the α and β functions on $\mathcal{P}$, and write them as scalar functions instead notation-wise. By discussion above, β_2 becomes a nonconstant function in y_1 and β_1' becomes a nonconstant function in y_2. Note that for $\mathcal{P} = \varnothing$, we do not fix or condition on any parent variables and α_1', α_2, β_1' and β_2 are automatically univariate functions, with β_1' and β_2 nonconstant by assumption.

The joint density of $P(y_1, y_2 | \boldsymbol{Y}_{\mathcal{P}} = \boldsymbol{y}_{\mathcal{P}})$ w.r.t. λ thus has two characterizations (up to normalizing constants)

$$\frac{\exp\{\mathbb{1}_{y_1}\delta_1 + f_1(y_1) + \mathbb{1}_{y_2}\alpha_2(y_1) + y_2\beta_2(y_1) - y_2^2 k_2/2\}}{\sqrt{2\pi/k_2}\exp\{\alpha_2(y_1) + \beta_2(y_1)^2/(2k_2)\} + 1}$$

$$\propto \frac{\exp\{\mathbb{1}_{y_2}\delta_2' + f_2'(y_2) + \mathbb{1}_{y_1}\alpha_1'(y_2) + y_1\beta_1'(y_2) - y_1^2k_1'/2\}}{\sqrt{2\pi/k_1'}\exp[\alpha_1'(y_2) + \{\beta_1'(y_2)\}^2/(2k_1')] + 1}, \quad \text{(C.6)}$$

where $\alpha_2(y_1)$ has the form $c_{\alpha_2,-1} + c_{\alpha_2,0}\mathbb{1}_{y_1} + c_{\alpha_2,1}y_1 + \cdots + c_{\alpha_2,k}y_1^k$ with coefficients being polynomials in $\boldsymbol{y}_{\mathcal{P}}$ and their indicators (or constants if $\mathcal{P} = \varnothing$), and similarly for $\beta_2(y_1)$, $\alpha_1'(y_2)$ and $\beta_1'(y_2)$. Note that if the values of $\mathbb{1}_{y_1}$ and $\mathbb{1}_{y_2}$ are given, these four functions are just polynomials in y_1 and y_2, respectively.

First condition on the event $\mathbb{1}_{y_1} = \mathbb{1}_{y_2} = 1$ that has a positive probability. Then (C.6) becomes

$$\frac{\exp\{f_1(y_1) + \alpha_2(y_1) + y_2\beta_2(y_1) - y_2^2k_2/2\}}{\sqrt{2\pi/k_2}\exp\{\alpha_2(y_1) + \beta_2(y_1)^2/(2k_2)\} + 1}\mathbb{1}_{y_1}\mathbb{1}_{y_2},$$
$$\propto \frac{\exp\{f_2'(y_2) + \alpha_1'(y_2) + y_1\beta_1'(y_2) - y_1^2k_1'/2\}}{\sqrt{2\pi/k_1'}\exp\{\alpha_1'(y_2) + (\beta_1'(y_2))^2/(2k_1')\} + 1}\mathbb{1}_{y_1}\mathbb{1}_{y_2}, \quad \text{(C.7)}$$

for all $(y_1, y_2) \in (\mathbb{R}\backslash\{0\})^2$. (C.7) has the form

$$\frac{\exp\{f_1(y_1) + P_1(y_1, y_2)\}}{\exp\{P_2(y_1)\} + 1} = \frac{\exp\{f_2'(y_2) + P_3(y_1, y_2)\}}{\exp\{P_4(y_2)\} + 1},$$

where P_1 and P_3 are polynomials in y_1 and y_2 simultaneously, possibly with interactions from the $y_2\beta_2(y_1)$ and $y_1\beta_1'(y_2)$ terms, and P_2 and P_4 are univariate polynomials in y_1, y_2, respectively. By cross-multiplication,

$$\exp\{f_1(y_1) + P_1(y_1, y_2) + P_4(y_2)\} + \exp\{f_1(y_1) + P_1(y_1, y_2)\}$$
$$= \exp\{f_2'(y_2) + P_3(y_1, y_2) + P_2(y_1)\} + \exp\{f_2'(y_2) + P_3(y_1, y_2)\}. \quad \text{(C.8)}$$

Differentiating both sides of (C.8) with respect to y_1,

$$\left[\frac{\partial}{\partial y_1}\{f_1(y_1) + P_1(y_1, y_2)\}\right]\exp\{f_1(y_1) + P_1(y_1, y_2) + P_4(y_2)\}$$
$$+ \exp\{f_1(y_1) + P_1(y_1, y_2))\}$$
$$= \left[\frac{\partial}{\partial y_1}\{P_3(y_1, y_2) + P_2(y_1)\}\right]\exp\{f_2'(y_2) + P_3(y_1, y_2) + P_2(y_1)\}$$
$$+ \left\{\frac{\partial}{\partial y_1}P_3(y_1, y_2)\right\}\exp\{f_2'(y_2) + P_3(y_1, y_2)\}. \quad \text{(C.9)}$$

Plugging (C.8) into the left-hand side of (C.9),

$$\begin{aligned}
&\left[\frac{\partial}{\partial y_1}\left\{f_1(y_1)+P_1(y_1,y_2)\right\}\right]\left[\exp\left\{f_2'(y_2)+P_3(y_1,y_2)+P_2(y_1)\right\}\right.\\
&\qquad\qquad\left.+\exp\left\{f_2'(y_2)+P_3(y_1,y_2)\right\}\right]\\
&=\left[\frac{\partial}{\partial y_1}\left\{P_3(y_1,y_2)+P_2(y_1)\right\}\right]\exp\left\{f_2'(y_2)+P_3(y_1,y_2)+P_2(y_1)\right\}\\
&\qquad\qquad+\left\{\frac{\partial}{\partial y_1}P_3(y_1,y_2)\right\}\exp\left\{f_2'(y_2)+P_3(y_1,y_2)\right\},
\end{aligned}$$

which simplifies to

$$\begin{aligned}
&\left[\frac{\partial}{\partial y_1}\left\{f_1(y_1)+P_1(y_1,y_2)-P_3(y_1,y_2)-P_2(y_1)\right\}\right]\\
&\qquad\qquad\times\exp\left\{f_2'(y_2)+P_3(y_1,y_2)+P_2(y_1)\right\}\\
&\qquad+\left[\frac{\partial}{\partial y_1}\left\{f_1(y_1)+P_1(y_1,y_2)-P_3(y_1,y_2)\right\}\right]\\
&\qquad\qquad\times\exp\left\{f_2'(y_2)+P_3(y_1,y_2)\right\}=0.
\end{aligned}$$

Since $\exp\left\{f_2'(y_2)+P_3(y_1,y_2)\right\}\neq 0$, this becomes

$$\begin{aligned}
&\left[\frac{\partial}{\partial y_1}\left\{f_1(y_1)+P_1(y_1,y_2)-P_3(y_1,y_2)-P_2(y_1)\right\}\right]\exp\left\{P_2(y_1)\right\}\\
&\qquad+\left[\frac{\partial}{\partial y_1}\left\{f_1(y_1)+P_1(y_1,y_2)-P_3(y_1,y_2)\right\}\right]=0. \quad \text{(C.10)}
\end{aligned}$$

Focusing on the components that involve y_2, we see that

$$\left[\frac{\partial}{\partial y_1}\left\{P_1(y_1,y_2)-P_3(y_1,y_2)\right\}\right]\left[\exp\left\{P_2(y_1)\right\}+1\right]$$

does not depend on y_2. Since $(\exp(P_2(y_1))+1)>0$, we have

$$\frac{\partial^2}{\partial y_1\partial y_2}\left\{P_1(y_1,y_2)-P_3(y_1,y_2)\right\}=0.$$

Recall that

$$P_1(y_1,y_2)-P_3(y_1,y_2)=\alpha_2(y_1)+y_2\beta_2(y_1)-y_2^2k_2/2-\alpha_1'(y_2)-y_1\beta_1'(y_2)+y_1^2k_1'/2. \quad \text{(C.11)}$$

So $0 = \frac{\partial^2}{\partial y_1 \partial y_2}\{P_1(y_1, y_2) - P_3(y_1, y_2)\} = \frac{\mathrm{d}\beta_2(y_1)}{\mathrm{d}y_1} - \frac{\mathrm{d}\beta_1'(y_2)}{\mathrm{d}y_2}$ implies that β_2 and β_1' are both linear with the same coefficient on the linear term. Now that β_2 has the form $\beta_2(y_1) = c_{\beta_2,-1} + c_{\beta_2,0}\mathbb{1}_{y_1} + c_{\beta_2,1}y_1$, write $\beta_{2;-1,0} \equiv c_{\beta_2,-1} + c_{\beta_2,0} = \beta_2(0) + c_{\beta_2,0}$ as a shorthand notation for β_2 with indicator set to 1 while y_1 set to 0. Similarly define $\beta'_{1;-1,0} \equiv c_{\beta_1',-1} + c_{\beta_1',0} = \beta_1'(0) + c_{\beta_1',0}$. Then for $y_1, y_2 \neq 0$ since $c_{\beta_2,1} = c_{\beta_1',1}$, we necessarily have

$$\begin{aligned} y_2\beta_2(y_1) - y_1\beta_1'(y_2) &= y_2(c_{\beta_2,-1} + c_{\beta_2,0} + c_{\beta_2,1}y_1) - y_1(c_{\beta_1',-1} + c_{\beta_1',0} + c_{\beta_1',1}y_2) \\ &= y_2\beta_{2;-1,0} - y_1\beta'_{1;-1,0}, \end{aligned}$$

and so by (C.11)

$$\begin{aligned} & P_1(y_1, y_2) - P_3(y_1, y_2) \\ = & \left(\alpha_2(y_1) - y_1\beta'_{1;-1,0} + y_1^2k_1'/2\right) - \left(\alpha_1'(y_2) - y_2\beta_{2;-1,0} + y_2^2k_2/2\right) \\ \equiv & P_{1,3}(y_1) - (\text{function in } y_2 \text{ only}). \end{aligned}$$

Plugging this into (C.10), we get

$$\left[\frac{\mathrm{d}}{\mathrm{d}y_1}\left\{f_1(y_1) + P_{1,3}(y_1) - P_2(y_1)\right\}\right]\exp\left\{P_2(y_1)\right\} + \left[\frac{\mathrm{d}}{\mathrm{d}y_1}\left\{f_1(y_1) + P_{1,3}(y_1)\right\}\right]$$

equals 0, or equivalently

$$\left[\frac{\mathrm{d}}{\mathrm{d}y_1}\left\{f_1(y_1) + P_{1,3}(y_1)\right\}\right]\left[\exp\{P_2(y_1)\} + 1\right] = \left\{\frac{\mathrm{d}}{\mathrm{d}y_1}P_2(y_1)\right\}\exp\left\{P_2(y_1)\right\}.$$

Then

$$\begin{aligned} f_1(y_1) &= \int \frac{\exp\{P_2(y_1)\}\{\mathrm{d}P_2(y_1)/\,\mathrm{d}y_1\}}{\exp\{P_2(y_1)\} + 1}\,\mathrm{d}y_1 - P_{1,3}(y_1) \\ &= \log\left[1 + \exp\{P_2(y_1)\}\right] - P_{1,3}(y_1) + \text{const.} \end{aligned}$$

So for $y_1 \neq 0$,

$$\begin{aligned} \exp(f_1(y_1)) &\propto \frac{1 + \exp\{P_2(y_1)\}}{\exp\{P_{1,3}(y_1)\}} \\ &= \frac{1 + \sqrt{2\pi/k_2}\exp\{\alpha_2(y_1) + \beta_2(y_1)^2/(2k_2)\}}{\exp\{\alpha_2(y_1) - \beta'_{1;-1,0}y_1 + y_1^2k_1'/2\}} \end{aligned}$$

$$= \exp\{-\alpha_2(y_1) + y_1\beta'_{1;-1,0} - y_1^2 k'_1/2\}$$
$$+\sqrt{2\pi/k_2}\exp\{y_1\beta'_{1;-1,0} + \beta_2(y_1)^2/(2k_2) - y_1^2 k'_1/2\}. \quad \text{(C.12)}$$

Now condition on the event $\mathbb{1}_{y_1} = 1$ and $\mathbb{1}_{y_2} = 0$. Then (C.6) becomes

$$\frac{\exp\{f_1(y_1)\}}{\sqrt{2\pi/k_2}\exp\{\alpha_2(y_1) + \beta_2(y_1)^2/(2k_2)\} + 1}\mathbb{1}_{y_1} \propto \exp\{y_1\beta'_1(0) - y_1^2 k'_1/2\}\mathbb{1}_{y_1},$$

which implies that for $y_1 \neq 0$,

$$\exp\{f_1(y_1)\} \propto \exp\{y_1\beta'_1(0) - y_1^2 k'_1/2\}$$
$$+ \sqrt{2\pi/k_2}\exp\left\{y_1\beta'_1(0) - y_1^2 k'_1/2 + \alpha_2(y_1) + \beta_2(y_1)^2/(2k_2)\right\}. \quad \text{(C.13)}$$

Applying Lemma 39 to (C.12) and (C.13), by matching the terms we have (conditional on $y_1 \neq 0$) either

$$-\alpha_2(y_1) + y_1\beta'_{1;-1,0} = y_1\beta'_1(0) + \text{const}; \quad \text{or} \quad \text{(C.14)}$$
$$-\alpha_2(y_1) + y_1\beta'_{1;-1,0} = y_1\beta'_1(0) + \alpha_2(y_1) + \beta_2(y_1)^2/(2k_2) + \text{const} \quad \text{and}$$
$$y_1\beta'_{1;-1,0} + \beta_2(y_1)^2/(2k_2) = y_1\beta'_1(0) + \text{const}. \quad \text{(C.15)}$$

Conditional on $y_1 \neq 0$, in the first case (C.14), $\alpha_2(y_1) = y_1 c_{\beta'_1,0} + \text{const}$; in the second case (C.15), $\alpha_2(y_1) + \beta_2(y_1)^2/(2k_2) = \text{const}$ and $\beta_2(y_1)^2/(2k_2) = -y_1 c_{\beta'_1,0} + \text{const}$, which implies $\beta_2(y_1) = \text{const}$ and $\alpha_2(y_1) = \text{const}$ for $y_1 \neq 0$, and $c_{\beta'_1,0} = 0$, which in turn implies (C.14). Thus, in either case, $\alpha_2(y_1) = c_{\alpha_2,0}\mathbb{1}_{y_1} + y_1 c_{\beta'_1,0} + \text{const}$, i.e. α_2 is linear (or constant) in $y_1 \neq 0$ with coefficient on y_1 equal to $c_{\beta'_1,0}$. By (C.14) for $y_1 \neq 0$,

$$\exp\{f_1(y_1)\} \propto \exp\{y_1\beta'_1(0) - y_1^2 k'_1/2\}$$
$$+ \sqrt{2\pi/k_2}\exp\{y_1\beta'_{1;-1,0} + \beta_2(y_1)^2/(2k_2) - y_1^2 k'_1/2\}, \quad \text{(C.16)}$$

clearly a single univariate Gaussian or a mixture of two univariate Gaussian distributions (since β_2 is at most linear in y_1). Similarly, we must have $\alpha'_1(y_2) = y_2\beta_{2;-1,0} - y_2\beta_2(0) + \text{const} = y_2 c_{\beta_2,0} + \text{const}$ for $y_2 \neq 0$, and for $y_2 \neq 0$

$$\exp\{f_2'(y_2)\} \propto \exp\{y_2\beta_2(0) - y_2^2 k_2/2\} + \sqrt{2\pi/k_1'}\exp\{y_2\beta_{2;-1,0} + \beta_1'(y_2)^2/(2k_1') - y_2^2 k_2/2\}. \quad \text{(C.17)}$$

Now suppose by contradiction that $\exp\{f_1(y_1)\}$ given $y_1 \neq 0$ has only one Gaussian component, instead of being a sum of two Gaussian densities. Then by (C.16), $\beta_1'(0) = \beta_{1;-1,0}'$ and $\beta_2(y_1)$ is a constant given $\mathbb{1}_{y_1}$, i.e. $\beta_2(y_1) = c_{\beta_2,-1} + c_{\beta_2,0} = \beta_{2;-1,0}$ for $y_1 \neq 0$. Plugging this into the left-hand side of (C.6) and integrating w.r.t. $\lambda(y_1)$, the continuous part ($y_2 \neq 0$) of the marginal distribution of y_2 given $\boldsymbol{Y}_{\mathcal{P}} \equiv \boldsymbol{y}_{\mathcal{P}}$ is

$$\exp\{f_2'(y_2)\} \propto \frac{\exp\{f_1(0) + \alpha_2(0) + y_2\beta_2(0) - y_2^2 k_2/2\}}{\sqrt{2\pi/k_2}\exp\{\alpha_2(0) + \beta_2(0)^2/(2k_2)\} + 1} + \exp\{y_2\beta_{2;-1,0} - y_2^2 k_2/2\}\int_{\mathbb{R}} \frac{\exp\{\delta_1 + f_1(y_1) + \alpha_2(y_1)\}}{\sqrt{2\pi/k_2}\exp\{\alpha_2(y_1) + \beta_2(y_1)^2/(2k_2)\} + 1}\,\mathrm{d}y_1,$$

which is a mixture between $\mathcal{N}(\beta_2(0)/k_2, 1/k_2)$ and $\mathcal{N}(\beta_{2;-1,0}/k_2, 1/k_2)$, i.e. the variance in both components are equal. Note that the integral in the second term is a Lebesgue integral. This together with (C.17) implies that $\beta_1'(y_2)$ cannot depend on the value of y_2 given $y_2 \neq 0$, i.e. $\beta_1'(y_2) = c_{\beta_1',-1} + c_{\beta_1',0} = \beta_{1;-1,0}'$. Since we already know that $\beta_1'(0) = \beta_{1;-1,0}'$ by discussion above, this implies that β_1' is an absolute constant in y_2 and $\mathbb{1}_{y_2}$, and also that α_2 may depend on y_1 only through $\mathbb{1}_{y_1}$, a contradiction to the assumption of the theorem.

Thus, (C.16) and (C.17) will both have to be mixtures of precisely two Gaussians, and so by definition the joint distribution $p(\boldsymbol{Y})$ of $\boldsymbol{Y}$ must be of 2-Gaussian type with respect to $\mathcal{G}$ and $\mathcal{G}'$. □

Proof of Corollary 33. When $|\mathcal{V}| = 2$, in Proposition 31 we always have $\mathcal{P} = \varnothing$ and V_1 does not have a parent in $\mathcal{G}$, so $P(Y_{V_1} = y | Y_{V_1} \neq 0)$ by definition is just a Gaussian, not a mixture two Gaussians, and hence $p(\boldsymbol{Y})$ cannot be of 2-Gaussian type with respect to any pairs of distinct Markov equivalent graphs.

Now consider $|\mathcal{V}| = 3$, and assume the two vertices with reversible edges in Proposition 31 are V_1 and V_2, and that $V_1 \to V_2$ in $\mathcal{G}$ and $V_1 \leftarrow V_2$ in $\mathcal{G}'$. If neither V_1 or V_2 has V_3 as its parent in both graphs, then we can marginalize V_3 out and it reduces to the 2-d case.

Suppose otherwise. Then we must have (1) $V_1 \to V_2 \leftarrow V_3$ in $\mathcal{G}$, or (2) $V_2 \to V_1 \leftarrow V_3$ in $\mathcal{G}'$, or (3) an additional edge between V_1 and V_3 added to (1), or (4) an additional edge between V_2 and V_3 added to (2).

For (1) and (2) both graphs are the only graph in their Markov equivalence class; for (3) the reversible edge becomes V_1—V_3 violating the assumption (and in fact one can marginalize out the common child V_2 and get back to the 2-d case), and similarly for (4). Thus, we have again ruled out the possibility of any pair of distinct Markov equivalent graphs with respect to which $p(\boldsymbol{Y})$ can be of 2-Gaussian type. □

Remark 6. *In the proof of Theorem 32, we proved that whenever $p(\boldsymbol{Y})$ factorizes with respect to two distinct graphs $\mathcal{G}$ and $\mathcal{G}'$ (whenever identifiability does not hold), everything up to (C.17) in the proof must hold. Specifically, conditioning on almost every $\boldsymbol{y}_{\mathcal{P}}$, α_2 and β_2 in $\mathcal{G}$ as well as α_1' and β_1' in $\mathcal{G}'$ can be at most linear in y_1 and y_2, respectively, namely*

$$\beta_1'(y_2) = c_{\beta_1',-1} + c_{\beta_1',0}\mathbb{1}_{y_2} + c_{\beta_1',1}y_2, \quad \beta_2(y_1) = c_{\beta_2,-1} + c_{\beta_2,0}\mathbb{1}_{y_1} + c_{\beta_2,1}y_1,$$
$$\alpha_1'(y_2) = c_{\alpha_1',-1} + c_{\alpha_1',0}\mathbb{1}_{y_2} + c_{\alpha_1',1}y_2, \quad \alpha_2(y_1) = c_{\alpha_2,-1} + c_{\alpha_2,0}\mathbb{1}_{y_1} + c_{\alpha_2,1}y_1,$$

with coefficients depending on $\boldsymbol{y}_{\mathcal{P}}$ where

$$c_{\alpha_1',1} = c_{\beta_2,0}, \quad c_{\alpha_2,1} = c_{\beta_1',0}, \quad c_{\beta_1',1} = c_{\beta_2,1}. \tag{C.18}$$

It is noted that, although not used in deriving our conclusion involving 2-Gaussian type distributions, we in addition also have the following results.

$$c_{\alpha_1',-1} = c_{\alpha_2,-1}, \quad c_{\alpha_1',0} = c_{\alpha_2,0}, \quad c_{\alpha_1',-1} + c_{\alpha_1',0} = c_{\alpha_2,-1} + c_{\alpha_2,0} = 0.$$

These might shed some light on how to show that distributions of 2-Gaussian type do not exist for a general $m \geq 4$.

Proof of Remark 6. By (C.6), (C.16), (C.17), the joint distribution of Y_1 and Y_2 conditional on $\boldsymbol{Y}_{\mathcal{P}}$ has two characterizations (up to normalizing constants)

$$\frac{\exp\{\mathbb{1}_{y_1}\delta_1 + y_1\beta_1'(0) - y_1^2k_1'/2 + \mathbb{1}_{y_2}\alpha_2(y_1) + y_2\beta_2(y_1) - y_2^2k_2/2\}}{\sqrt{2\pi/k_2}\exp\{\alpha_2(y_1) + \beta_2(y_1)^2/(2k_2)\} + 1}$$

$$+\frac{\sqrt{2\pi/k_2}\exp\{\mathbb{1}_{y_1}\delta_1+y_1\beta'_{1;-1,0}+\beta_2(y_1)^2/(2k_2)-y_1^2k'_1/2+\mathbb{1}_{y_2}\alpha_2(y_1)+y_2\beta_2(y_1)-y_2^2k_2/2\}}{\sqrt{2\pi/k_2}\exp\{\alpha_2(y_1)+\beta_2(y_1)^2/(2k_2)\}+1}$$
$$\propto\frac{\exp\{\mathbb{1}_{y_2}\delta'_2+y_2\beta_2(0)-y_2^2k_2/2+\mathbb{1}_{y_1}\alpha'_1(y_2)+y_1\beta'_1(y_2)-y_1^2k'_1/2\}}{\sqrt{2\pi/k'_1}\exp\{\alpha'_1(y_2)+\beta'_1(y_2)^2/(2k'_1)\}+1}$$
$$+\frac{\sqrt{2\pi/k'_1}\exp\{\mathbb{1}_{y_2}\delta'_2+y_2\beta_{2;-1,0}+\beta'_1(y_2)^2/(2k'_1)-y_2^2k_2/2+\mathbb{1}_{y_1}\alpha'_1(y_2)+y_1\beta'_1(y_2)-y_1^2k'_1/2\}}{\sqrt{2\pi/k'_1}\exp\{\alpha'_1(y_2)+\beta'_1(y_2)^2/(2k'_1)\}+1}. \tag{C.19}$$

Divide both sides by $\exp(y_1\beta'_1(0)+y_2\beta_2(0)-y_1^2k'_1/2-y_2^2k_2/2)$ and expanding $\beta'_1(y_2)$ and $\beta_2(y_1)$, this becomes

$$\frac{\exp\{\mathbb{1}_{y_1}\delta_1+\mathbb{1}_{y_2}\alpha_2(y_1)+y_2c_{\beta_2,0}\mathbb{1}_{y_1}+y_1y_2c_{\beta_2,1}\}}{\sqrt{2\pi/k_2}\exp\{\alpha_2(y_1)+\beta_2(y_1)^2/(2k_2)\}+1}$$
$$+\frac{\sqrt{2\pi/k_2}\exp\{\mathbb{1}_{y_1}\delta_1+y_1c_{\beta'_1,0}+\beta_2(y_1)^2/(2k_2)+\mathbb{1}_{y_2}\alpha_2(y_1)+y_2c_{\beta_2,0}\mathbb{1}_{y_1}+y_1y_2c_{\beta_2,1}\}}{\sqrt{2\pi/k_2}\exp\{\alpha_2(y_1)+\beta_2(y_1)^2/(2k_2)\}+1}$$

$$\propto\frac{\exp\{\mathbb{1}_{y_2}\delta'_2+\mathbb{1}_{y_1}\alpha'_1(y_2)+y_1c_{\beta'_1,0}\mathbb{1}_{y_2}+y_1y_2c_{\beta'_1,1}\}}{\sqrt{2\pi/k'_1}\exp\{\alpha'_1(y_2)+\beta'_1(y_2)^2/(2k'_1)\}+1}$$
$$+\frac{\sqrt{2\pi/k'_1}\exp\{\mathbb{1}_{y_2}\delta'_2+y_2c_{\beta_2,0}+\beta'_1(y_2)^2/(2k'_1)+\mathbb{1}_{y_1}\alpha'_1(y_2)+y_1c_{\beta'_1,0}\mathbb{1}_{y_2}+y_1y_2c_{\beta'_1,1}\}}{\sqrt{2\pi/k'_1}\exp\{\alpha'_1(y_2)+\beta'_1(y_2)^2/(2k'_1)\}+1}.$$

Now expanding $\alpha'_1(y_2)$ and $\alpha_2(y_1)$ and using the relationships in (C.18), we divide both sides by $\exp(y_1c_{\alpha_2,1}\mathbb{1}_{y_2}+y_2c_{\beta_2,0}\mathbb{1}_{y_1}+y_1y_2c_{\beta_2,1})=\exp(y_1c_{\beta'_1,0}\mathbb{1}_{y_2}+y_2c_{\alpha'_1,1}\mathbb{1}_{y_1}+y_1y_2c_{\beta_2,1})$ and get

$$\frac{\exp\{\mathbb{1}_{y_1}\delta_1+\mathbb{1}_{y_2}(c_{\alpha_2,-1}+c_{\alpha_2,0}\mathbb{1}_{y_1})\}}{\sqrt{2\pi/k_2}\exp\{\alpha_2(y_1)+\beta_2(y_1)^2/(2k_2)\}+1}$$
$$+\frac{\sqrt{2\pi/k_2}\exp\left\{\mathbb{1}_{y_1}\delta_1+y_1c_{\beta'_1,0}+\beta_2(y_1)^2/(2k_2)+\mathbb{1}_{y_2}(c_{\alpha_2,-1}+c_{\alpha_2,0}\mathbb{1}_{y_1})\right\}}{\sqrt{2\pi/k_2}\exp\{\alpha_2(y_1)+\beta_2(y_1)^2/(2k_2)\}+1}$$
$$=C_0\frac{\exp\{\mathbb{1}_{y_2}\delta'_2+\mathbb{1}_{y_1}(c_{\alpha'_1,-1}+c_{\alpha'_1,0}\mathbb{1}_{y_2})\}}{\sqrt{2\pi/k'_1}\exp\{\alpha'_1(y_2)+\beta'_1(y_2)^2/(2k'_1)\}+1}$$
$$+C_0\frac{\sqrt{2\pi/k'_1}\exp\{\mathbb{1}_{y_2}\delta'_2+y_2c_{\beta_2,0}+\beta'_1(y_2)^2/(2k'_1)+\mathbb{1}_{y_1}(c_{\alpha'_1,-1}+c_{\alpha'_1,0}\mathbb{1}_{y_2})\}}{\sqrt{2\pi/k'_1}\exp\{\alpha'_1(y_2)+\beta'_1(y_2)^2/(2k'_1)\}+1} \tag{C.20}$$

for some C_0. Setting $\mathbb{1}_{y_1}=\mathbb{1}_{y_2}=0$ (C.20) becomes

$$\frac{1+\sqrt{2\pi/k_2}\exp\{c_{\beta_2,-1}^2/(2k_2)\}}{\sqrt{2\pi/k_2}\exp\{c_{\alpha_2,-1}+c_{\beta_2,-1}^2/(2k_2)\}+1} \quad = \quad C_0\frac{1+\sqrt{2\pi/k_1'}\exp\{c_{\beta_1'1,-1}^2/(2k_1')\}}{\sqrt{2\pi/k_1'}\exp\{c_{\alpha_1',-1}+c_{\beta_1',-1}^2/(2k_1')\}+1}, \tag{C.21}$$

and with $\mathbb{1}_{y_1} \neq 0$, $\mathbb{1}_{y_2} = 0$ (C.20) becomes

$$\exp(\delta_1)\frac{1+\sqrt{2\pi/k_2}\exp\{y_1c_{\beta_1',0}+\beta_2(y_1)^2/(2k_2)\}}{\sqrt{2\pi/k_2}\exp\{c_{\alpha_2,-1}+c_{\alpha_2,0}+c_{\alpha_2,1}y_1+\beta_2(y_1)^2/(2k_2)\}+1} = C_0\exp(c_{\alpha_1',-1})\frac{1+\sqrt{2\pi/k_1'}\exp\{c_{\beta_1'1,-1}^2/(2k_1')\}}{\sqrt{2\pi/k_1'}\exp\{c_{\alpha_1',-1}+c_{\beta_1',-1}^2/(2k_1')\}+1}. \tag{C.22}$$

Since the right-hand side of (C.22) is a constant, by matching the numerator and the denominator of the left-hand side using Lemma 39, we must have either (i) $y_1c_{\beta_1',0}+\beta_2(y_1)^2/(2k_2) = c_{\alpha_2,-1}+c_{\alpha_2,0}+c_{\alpha_2,1}y_1+\beta_2(y_1)^2/(2k_2)$, or (ii) $y_1c_{\beta_1',0}+\beta_2(y_1)^2/(2k_2) = \text{const}$ for $y_1 \neq 0$. But (ii) implies that $c_{\beta_2,1} = c_{\beta_1',0} = 0$, which by $c_{\beta_1',1} = c_{\beta_2,1}$ implies that β_1' is an absolute constant in $y_2 \in \mathbb{R}$, a violation to the assumption. Thus (i) holds, and by $c_{\beta_1',0} = c_{\alpha_2,1}$ this implies that

$$\alpha_{2;-1,0} \equiv c_{\alpha_2,-1}+c_{\alpha_2,0} = 0, \quad \text{and by symmetry} \quad \alpha_{1;-1,0}' \equiv c_{\alpha_1',-1}+c_{\alpha_1',0} = 0. \tag{C.23}$$

Thus the left-hand side of (C.22) is just $\exp(\delta_1)$. Note that the right-hand side of (C.22) is $\exp(c_{\alpha_1',-1})$ times the right-hand side of (C.21). So by equating the left-hand side of (C.22) with $\exp(c_{\alpha_1',-1})$ times the left-hand side of (C.21) we have

$$\exp(\delta_1) = \exp\left(c_{\alpha_1,-1}'\right)\frac{1+\sqrt{2\pi/k_2}\exp\left\{c_{\beta_2,-1}^2/(2k_2)\right\}}{\sqrt{2\pi/k_2}\exp\left\{c_{\alpha_2,-1}+c_{\beta_2,-1}^2/(2k_2)\right\}+1)} \tag{C.24}$$

and similarly

$$\exp(\delta_2') = \exp\left(c_{\alpha_2,-1}\right)\frac{1+\sqrt{2\pi/k_1'}\exp\left\{c_{\beta_1'1,-1}^2/(2k_1')\right\}}{\sqrt{2\pi/k_1'}\exp\left\{c_{\alpha_1',-1}+c_{\beta_1',-1}^2/(2k_1')\right\}+1}. \tag{C.25}$$

Now by (C.23), with $\mathbb{1}_{y_1} = \mathbb{1}_{y_2} = 1$, (C.20) simplifies to $\exp(\delta_1) = C_0 \cdot \exp(\delta_2')$. Thus by

(C.21), (C.24) and (C.25), one get

$$C_0 = \frac{\exp(\delta_1)}{\exp(\delta_2')} = \frac{\exp(c'_{\alpha_1,-1})\frac{1+\sqrt{2\pi/k_2}\exp\left\{c^2_{\beta_2,-1}/(2k_2)\right\}}{\sqrt{2\pi/k_2}\exp\left\{c_{\alpha_2,-1}+c^2_{\beta_2,-1}/(2k_2))+1\right\}}}{\exp(c_{\alpha_2,-1})\frac{1+\sqrt{2\pi/k_1'}\exp\left\{c^2_{\beta_1' 1,-1}/(2k_1')\right\}}{\sqrt{2\pi/k_1'}\exp\left\{c_{\alpha_1',-1}+c^2_{\beta_1',-1}/(2k_1')\right\}+1}} = \frac{\exp(c'_{\alpha_1,-1})}{\exp(c_{\alpha_2,-1})}C_0$$

and thus $c'_{\alpha_1,-1} = c_{\alpha_2,-1}$. Combining with (C.23), we get

$$c_{\alpha_1',-1} = c_{\alpha_2,-1}, \quad c_{\alpha_1',0} = c_{\alpha_2,0}, \quad c_{\alpha_1',-1} + c_{\alpha_1',0} = c_{\alpha_2,-1} + c_{\alpha_2,0} = 0. \tag{C.26}$$

Note that this result holds as long as we assume identifiability does not hold.

□