Theory and Algorithms for Penalization, Graphical Models, and Surrogate Marker Evaluation

Rui Zhuang

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

Doctor of Philosophy

University of Washington

2019

Reading Committee:

Johannes Lederer, Chair

Noah Simon, Chair

Ying Qing Chen

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

Theory and Algorithms for Penalization, Graphical Models, and Surrogate Marker Evaluation

Rui Zhuang

Co-Chairs of the Supervisory Committee:
Adjunct Assistant Professor Johannes Lederer
Department of Biostatistics

Associate Professor Noah Simon
Department of Biostatistics

In this dissertation, we study three problems: oracle inequality in high-dimensional statistics theory, graphical models, and surrogate measures in clinical trials. First, we introduce a general slow rate bound for maximum regularized likelihood estimators in Kullback-Leibler divergence. The result applies to a wide variety of models and estimators where the densities have a convex parametrization, and the regularization is definite and positively homogenous. Next, we introduce a general framework, the so-called exponential trace models, for undirected graphical models. We employ a sampling-based approximation algorithm to compute the maximum likelihood estimator. The models apply to a wide range of data, such as continuous, discrete, and different combinations of those. Finally, we review the primary frameworks of surrogate measures and propose two new ones, the population surrogacy fraction of treatment effect and time-varying F-measure. The new measures complement the existing statistical framework and apply to the HIV Prevention Trial Network 052 Study.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I want to thank my dissertation reading committee, Johannes Lederer, Noah Simon, and Ying Qing Chen, for their guidance and support in my doctoral study. They are role models of my statistician career. I also want to thank Susanne May and Scott Emerson. They were my advisors when I made my very first attempts in biostatistics research. They inspire me and enlighten my passion for the endeavor. I am also grateful for Gitana Garofalo and Lurdes Inoue for making the program such a big warm family.

Last and most, I want to thank my parents and husband. Their continued trust and support equip me the courage and freedom to pursue my dreams.

DEDICATION

to my family

Chapter 1
INTRODUCTION

In this chapter, we introduce the background and motivation of the three problems studied in the dissertation: oracle inequality for high-dimensional statistics theory, graphical models, and surrogate measure in clinical trials.

1.1 Oracle Inequality in High-dimensional Statistics Theory

Advances in information technology have led us into the era of high-dimensional data. A distinguishing feature of high-dimensional data is that the number of features far exceeds the number of observations. This type of data brings new opportunities and challenges to the field of statistics. In particular, the massive number of features are filled with noise and low-quality signal. Distinguishing between sufficient information and redundant information challenges traditional low-dimensional methods. Applying prior scientific belief facilitates these tasks. One particular way is to add a penalty (for example, a sparsity penalty) to usual (goodness-of-fit) loss functions employed in low dimensional statistical estimation. The penalty usually comes with a tuning parameter (also known as regularization parameter) that balances the two parts: Goodness-of-fit and penalization. Maximum regularized likelihood estimators (MRLEs) are arguably the most established class of estimators in high-dimensional statistics. They have a penalized negative log-likelihood as the loss function. They are prevalent in generalized linear regression, tensor response regression, and graphical modeling with high-dimensional data.

To understand the finite sample performance for high-dimensional methods, statisticians formulate oracle inequalities to bound the distance of interest between the estimator and the truth. There are two main types of oracle inequality bounds: fast rate bounds and

slow rate bounds. There are two main differences between these bounds. First, fast rate bounds are proportional to the square of the tuning parameter, while slow rate bounds are proportional to the tuning parameter. Second, fast rate bounds invoke restricted eigenvalue-type conditions while slow rate bounds hold without questionable assumptions. Fast rate bounds are popular in the community of high-dimensional statistics. They are optimal in the sense of matching minimax rates (Verzelen, 2012). However, the assumptions that fast rate bounds rely on involve the true value of the unknown parameter. The assumptions are not only stringent but also unverifiable in practice (Bunea et al., 2007a; Dalalyan and Tsybakov, 2007; Rigollet and Tsybakov, 2011; Dalalyan and Tsybakov, 2012a,b). In addition, the formulation of the assumptions is motivated by the process of proof. The assumptions are not always necessary for prediction performance guarantees. For example, restricted eigenvalue-type conditions in regression basically require that covariates are not too correlated. However, even perfectly colinear covariates do not necessarily hurt prediction (Hebiri and Lederer, 2013; Dalalyan et al., 2017). On the other hand, slow rate bound does not involve questionable assumptions and is optimal in the assumptionless setting (Foygel and Srebro, 2011; Zhang et al., 2017; Dalalyan et al., 2017).

We are particularly interested in slow rate bounds for prediction in high-dimensional statistics. Slow rate bounds have been derived for lasso-type estimators (Greenshtein and Ritov, 2004; Rigollet and Tsybakov, 2011; Massart and Meynet, 2011; Koltchinskii et al., 2011; Huang and Zhang, 2012; Chatterjee, 2013; Bühlmann, 2013; Chatterjee, 2014; Lederer et al., 2016; Dalalyan et al., 2017). However, many other examples still lack such guarantees, and a broadly applicable slow rate bound is still needed. We are thus motivated to work on general theory and provide support for the encompassed examples.

1.2 Graphical Models

Dependencies in multivariate measurements are vital for uncovering relationships among the underlying processes. Microbiologists, for example, collect data about the number of bacteria in stool samples. The data for each stool sample can be summarized in a measurement

vector, where each coordinate contains the number of a given type of bacteria in the sample. It is expected that the dependencies among the coordinates can provide insights into the microbiological environment of the human gut. Graphical modeling is a popular approach to model and visualize dependencies in multivariate data (Drton and Maathuis, 2017; Lauritzen, 1996a; Wainwright and Jordan, 2008).

Graphical models combine graph theory and probability theory. They use graphs to represent the dependence structure among random variables (Lauritzen, 1996a). There are two main types: directed and undirected graphical models. Directed graphical models emphasize the directional parent-child relationship, while undirected models are concerned about the symmetric local relationship. In this dissertation, we only cover undirected graphical models (which are also called Markov random fields). Consider an undirected graph $G(V, E)$, where V is the set of nodes, and E is the set of undirected edges. Specifically, suppose there is a one-to-one bijective mapping between a set of random variables and the nodes in set V. There is an edge between node i and j if and only if $(i, j) \in E$ and $(j, i) \in E$. The graph represents a family of joint probability distributions sharing the same (conditional) (in)dependence structure. In the following, we discuss the conditional independence property of undirected graphical models using a toy example in Figure 1.1. In the graph, X_A, X_B, X_C are node sets for index subsets $A, B,$ and C. If blocking all the nodes of X_B disconnects all the paths connecting a node in X_A and a node in X_C, then we have conditional independence: $X_A \perp X_C \mid X_B$. Otherwise, conditional independence does not hold.

The exponential trace framework (Zhuang et al., 2016) unifies a variety of known classes of graphical models, such as Gaussian graphical models, Ising models (Brush, 1967) and multivariate extensions of it (Loh and Wainwright, 2012). It also allows for very different kinds of other discrete and continuous data — and even combinations of different types of data. Exponential trace models consider arbitrary (non-empty) finite or continuous domains $\mathcal{D} \subset \mathbb{R}^p$ and random vectors $\mathbf{x} \in \mathcal{D}$ that have densities of the form

$$f_{\mathrm{M}}(\mathbf{x}) = \exp\big(- \langle \mathrm{M},\, T(\mathbf{x}) \rangle_{\mathrm{tr}} + \xi(\mathbf{x}) - a(\mathrm{M}) \big) \qquad \big(\mathrm{M} \in \mathfrak{M} \big)$$

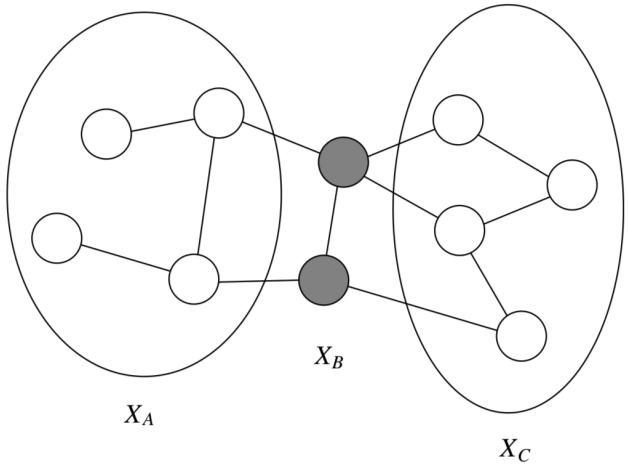

Figure 1.1: An example of undirected graphical model from Jordan et al. (1999). X_A, X_B, X_C represent node sets for index subsets $A, B,$ and C.

with respect to some σ-finite measure ν on $\mathbb{R}^p$. Here, the matrix-valued parameter M encodes the dependence structure of the random vector X, the matrix-valued function $T(\cdot)$ on $\mathcal{D}$ determines how the data enters the model, $a(\mathrm{M})$ is the partition or normalization function, and $\mathfrak{M}$ is a convex set of matrices M with finite normalization. Given i.i.d. observations $(X^1, \ldots, X^n)$, the maximum likelihood estimator is of the form

$$\widehat{\mathrm{M}} \in \operatorname*{argmin}_{\mathrm{M}} \Big\{ \langle \mathrm{M}, \overline{T} \rangle_{\mathrm{tr}} + a(\mathrm{M}) \Big\},$$

where $\overline{T} := \sum_{i=1}^n T(X^i)/n$, $a(\mathrm{M}) = \log \int \exp\Big(- \langle \mathrm{M}, T(\mathbf{x}) \rangle_{\mathrm{tr}} + \xi(\mathbf{x}) \Big) d\mathbf{x}$. In general, the normalization term lacks a closed-form expression.

The framework is amenable to theoretically efficient estimation and inference based on maximum likelihood. However, the normalization term of the likelihood function is not tractable, and hence, the computation of the maximum likelihood estimators is challenging. In Chapter 3, we use a sampling-based approximation method to tackle this problem and uncover the corresponding graphs. This technique is also applicable to other problems where there is difficulty in computing the exact objective function.

1.3 Surrogate Measure in Medical Research

Large, long-term randomized clinical trials can be very resource-demanding and time-consuming. Investigators are hence often motivated to find appropriate surrogate endpoints to substitute for rare or distal clinically meaningful endpoints to compare specific interventions or treatments in trials (Ellenberg and Hamilton, 1989; Temple, 1995; Fleming and DeMets, 1996). The term "surrogate endpoint" generally refers to a biomarker that is expected to predict clinical benefit or harm reliably (Colburn et al., 2001). Hence, the terms "surrogate marker" and "surrogate endpoint" are used interchangeably in the literature. When measured early and conveniently, surrogate markers are expected to reduce the sample size and shorten the overall trial duration, which improves the trials' financial efficiency and feasibility (Wittes et al., 1989; Benjamin et al., 2016).

The use of surrogate markers is controversial. On the one hand, it is recommended to be conservative when substituting surrogate markers for clinically meaningful endpoints (Grimes and Schulz, 2005; Fleming and Powers, 2012). The relationship among surrogate markers, clinical endpoints, and interventions is usually so complicated that the surrogates may fail to provide reliable evidence about the benefit-to-risk profile of interventions (Fleming and DeMets, 1996; Temple, 1999; Fleming and Powers, 2012). In addition, the reliability and validity of surrogate markers are usually context-specific (Fleming and Powers, 2012). There are many examples where results based on inappropriate surrogate markers may provide misleading information (Landau, 1990; Gallin et al., 1991; Echt et al., 1991). On the other hand, realistic and ethical issues, for example, the difficulty of studying the clinically meaningful endpoints and seriousness of various medical conditions, oblige people to use surrogate markers at times (Chakravarty, 2005; Temple, 1999). With limited resources, the use of surrogate makers allows for more studies to be conducted: This accelerates the early screening of new interventions (Cook and DeMets, 2007). A recent review showed that pivotal trials using surrogate end points as their primary endpoint formed the exclusive basis of FDA approvals for 91 of 206 (44%) indications for novel therapeutic agents between 2005 and 2012

(Downing et al., 2014). Beyond substituting for clinically meaningful endpoints, surrogate makers play an indispensable role in secondary analyses that facilitate our understanding of pathogenic processes and pharmacological responses to therapeutic intervention (Lonn, 2001; Buhr, 2012; Colburn et al., 2001). This information is especially valuable for public health research and policy-making for disease prevention. Consequently, it is crucial to identify and validate appropriate surrogate markers. Surrogate markers are usually proposed by their "biological relevance" (Ellenberg and Hamilton, 1989). In particular, markers on the intermediate pathway between the interventions and the clinically meaningful endpoints are good candidates (Ellenberg and Hamilton, 1989; Hillis and Seigel, 1989; Wittes et al., 1989). However, due to our often limited understand of the mechanisms behind diseases and treatments, this biological rationale does not always guarantee the validity and reliability of surrogate markers (Schatzkin, 2000). It is essential to develop statistical methods as auxiliary tools for validating surrogate markers.

Although there is a long history of development of statistical methods for validating surrogate endpoints, there are still many open challenges. In particular, there is no unified robust statistical validation or consensus on processes for pragmatic evaluation and adoption of surrogate endpoints (Buyse et al., 2010).

1.4 Summary of Our Contributions

In Chapter 2, we derive guarantees for MRLEs in Kullback-Leibler divergence, a general measure of prediction accuracy. We assume only that the densities have a convex parametrization and that the regularization is definite and positively homogenous. The results thus apply to a wide variety of models and estimators, such as tensor regression and graphical models with convex and non-convex regularized methods. The main results show that MRLEs are broadly consistent in prediction — regardless of whether the restricted eigenvalue condition holds (or other similar conditions). The contribution of this work is two-fold. First, the main result provides a general prediction guarantee for MRLEs in terms of the Kullback-Leibler loss. In addition to being the first assumptionless bound in such a broad setting, the

result specializes correctly to known results, such as for lasso, where the corresponding rates are optimal up to log-factors. Second, we show that applications of the general theorem to specific examples lead to new guarantees in tensor response regression, generalized linear tensor regression, and graphical modeling. The theory thus also establishes new insights into individual cases of MRLEs.

In Chapter 3, we introduce a general framework for characterizing undirected graphical models. This framework generalizes Gaussian graphical models to a wide range of continuous, discrete, and combinations of different types of data. The models in the framework, called exponential trace models, are amenable to estimation based on maximum likelihood. We introduce a sampling-based approximation algorithm for computing the maximum likelihood estimator, and we apply this pipeline to learn simultaneous neural activity from spike data.

In Chapter 4, we consider the statistical challenges involved in measuring and ranking surrogate markers quantitatively. We review the main methodology frameworks for validating surrogate markers in clinical trails and link them to the problem of mediation analysis in Public Health. Motivated by the concept of population attributable fraction, we propose a new measure, the so-called treatment surrogacy fraction, in the setting of clinical trials. The new measure carries an appealing population impact interpretation and supplements the existing statistical surrogate measures by providing absolute information. In addition, we define the time-varying F-measure, a model-free surrogate measure for time-varying internal markers and time-to-event outcomes in survival settings. The measures are illustrated using the HIV Prevention Trial Network 052 Study, a landmark HIV/AIDS prevention trial.

Chapter 2

MAXIMUM REGULARIZED LIKELIHOOD ESTIMATORS: A GENERAL PREDICTION THEORY AND APPLICATIONS

This chapter is a slightly revised version of my work published in Stat (Zhuang and Lederer, 2018).

2.1 Introduction

2.1.1 Overview

Maximum regularized likelihood estimators (MRLEs) are widely used in generalized linear regression, tensor response regression, and graphical modeling with high-dimensional data. It is thus of major interest to develop theory for this class of estimators.

Our specific goal is a general finite sample theory for prediction. Existing results are typically derived on a case-by-case basis. Moreover, many of these results also invoke restricted eigenvalues-type conditions (Bühlmann and van de Geer, 2011, Section 6). Such conditions are not only stringent and unverifiable in practice but also unsuitable for prediction. For example, restricted eigenvalue conditions in regression limit the correlations among the covariates. However, although correlations can affect the identifiability of the parameters, for prediction, even perfectly collinear covariates do not necessarily have a negative impact; in contrast, collinearity can even be beneficial (Hebiri and Lederer, 2013; Dalalyan et al., 2017). We are thus interested in a theory that does not involve additional assumptions and provides bounds for a general class of MRLEs. Besides its abstract value, such a general theory can also provide support for specific examples of MRLEs, such as the recently introduced approaches to tensor regression (Zhou et al., 2013; Li et al., 2013; Sun and Li, 2016), whose prediction properties have not been fully grasped.

In this work, we establish a general oracle inequality in terms of the Kullback-Leibler divergence. Oracle inequalities are a standard way to formulate finite sample bounds in high-dimensional statistics. The Kullback-Leibler divergence is a standard way to quantify prediction accuracies; it applies to any model and yet specializes to well-established and interpretable notions of prediction performance. Our proofs invoke only the convexity of the parametrization and the definiteness and positive homogeneity of the regularizers. This makes the result applicable to a variety of parametric and non-parametric models and allows for a broad class of convex and non-convex regularizers.

The remainder of this work is organized as follows. We introduce the framework and the general result in Section 2.2. We then provide examples in Section 2.3. We finally conclude with a brief discussion in Section 2.4. All proofs are deferred to the supplementary materials: proofs for the main result to Section 2.5.1, proofs for the examples to Section 2.5.2, and proofs for the bounds of the empirical process terms in Section 2.5.3. In addition, Section 2.5.4 contains notation and properties of tensors.

2.1.2 Related Literature

There are two types of oracle inequalities in the literature: so-called "fast rate bounds" and so-called "slow rate bounds." "Fast rate bounds" are proportional to the square of the regularization parameter. Many representatives of this type of bounds are found in the literature, such as Bunea et al. (2007b); Raskutti et al. (2015) for regression, Ravikumar et al. (2011) for graphical models, and more generally, Bühlmann and van de Geer (2011); van de Geer (2016) and references therein. For example, the corresponding bounds for lasso prediction are of the form $s \log p/(w^2 n)$, where s is the number of non-zero elements in the true regression vector, p is the number of parameters, w is the restricted eigenvalue, and n is the number of observations. These bounds are typically considered fast, because they can match minimax rates, see Verzelen (2012) and references therein. However, they rely on sparsity, and more importantly, they invoke restricted eigenvalue-type conditions or concern computationally challenging estimators instead (Bunea et al., 2007a; Dalalyan and

Tsybakov, 2007; Rigollet and Tsybakov, 2011; Dalalyan and Tsybakov, 2012a,b). Moreover, these eigenvalue-type assumptions are unverifiable and often unrealistic in practice, and even if they hold, the additional factors (such as s and $1/w^2$ for lasso) can be large.

On the other hand, oracle inequalities for prediction have be derived without sparsity or restricted eigenvalue conditions for lasso-type estimators (Greenshtein and Ritov, 2004; Rigollet and Tsybakov, 2011; Massart and Meynet, 2011; Koltchinskii et al., 2011; Huang and Zhang, 2012; Chatterjee, 2013; Bühlmann, 2013; Chatterjee, 2014; Lederer et al., 2016; Dalalyan et al., 2017). For example, the corresponding bounds for lasso prediction are of the form $\sqrt{\log p/n}\|\beta^*\|_1$, where β^* is the true regression vector. Such bounds are typically referred to as "slow rate bounds," because on a high level, the rates are $1/\sqrt{n}$ rather than $1/n$. However, there are no questionable assumptions involved, and for regression, it has even been shown that $1/\sqrt{n}$ is the optimal rate in the absence of further assumptions (Foygel and Srebro, 2011; Zhang et al., 2017; Dalalyan et al., 2017). Overall, this means that "fast rate bounds" are not necessarily fast and "slow rate bounds" are not necessarily slow. To correct the misleading nomenclature, Lederer et al. (2016) suggested replacing the term "fast rate bound" with "sparsity bound' and "slow rate bound" with "penalty bound."

Although some examples of MRLEs have been equipped with assumptionless bounds, many other examples still lack such guarantees (or any prediction guarantees altogether). More generally, a broadly applicable prediction theory for MRLEs is still in need.

2.1.3 Our Contribution

The contribution of this work is two-fold. First, Theorem 2.2.1 provides a general prediction guarantee for MRLEs in terms of the Kullback-Leibler loss. Besides being the first assumptionless bound in such a broad setting, the result specializes correctly to known results, such as for lasso, where the corresponding rates have been shown to be optimal up to log-factors. Second, we show that applications of the general theorem to specific examples lead to new guarantees in tensor response regression, generalized linear tensor regression, and graphical modeling. The theory thus also establishes new insights into individual cases of MRLEs.

2.2 General Theory

In this section, we present the general theory comprising the model classes, estimators, and the main result. The theory applies to an extremely wide range of data and methods; we discuss many important examples in Section 2.3. As for the models, we consider random vectors $X \in \mathcal{X}$ in a non-empty set $\mathcal{X}$ distributed according to a density $f \in \mathcal{F}$ in a general class $\mathcal{F}$. We assume that the densities in $\mathcal{F}$ can be parametrized as f_{M} with parameter $\mathrm{M} \in \mathfrak{M}$ that belongs to a convex, non-empty set $\mathfrak{M}$ in a real Hilbert space $\mathcal{H}$ and $\log f_{\mathrm{M}} - \mathbb{E}_{\mathrm{M}'} \log f_{\mathrm{M}}$ is convex in M for fixed $\mathrm{M}' \in \mathfrak{M}$. A classical example for this setup is the case of exponential families in the natural form (Berk, 1972, Lemma 2.1), see also Johansen (1979); Brown (1986). In general, however, the parametrization can be arbitrary as long as the convexity condition is fulfilled, and the parameter space can well be infinite-dimensional. In view of this very general framework, with the convexity of the parametrization being the only requirement on the models, the following theory applies to a large class of data.

The targets of our study are MRLEs in the described setup. Maximum likelihood estimation is one of the most widely accepted approaches to understand data, and regularization is a standard technique to incorporate additional structure or information. A contemporary playground for MRLEs is high-dimensional statistics, where a tremendous amount of research centers around regularization based on sparsity structures (Bühlmann and van de Geer, 2011; Giraud, 2014; Hastie et al., 2015). Given data X, we consider MRLEs of the form (assumed to exist)

$$\widehat{\mathrm{M}} \in \operatorname*{argmin}_{\mathrm{M} \in \mathfrak{M}} \big\{ -\log f_{\mathrm{M}}(X) + r u(\mathrm{M}) \big\}, \tag{2.1}$$

where $r > 0$ is a regularization parameter and $u : \mathcal{H} \mapsto [0, \infty]$ is a regularization with

properties

$$u(\mathrm{M}) = 0 \quad \Leftrightarrow \quad \mathrm{M} = 0, \tag{2.2}$$

$$u(t\mathrm{M}) = tu(\mathrm{M}) \qquad \forall \mathrm{M} \neq 0, t \geq 0. \tag{2.3}$$

These two properties allow us to formulate dual functions that generalize the classical notion of dual norms and the corresponding Hölder-like inequalities, see the definition of $\tilde{u}$ below and Lemma 2.5.1 in Section 2.5.1. Indeed, one can check readily that the properties are met by norms, including the weighted norm penalities considered in Zou (2006); van de Geer (2008); Gramfort et al. (2012); Bu and Lederer (2017) and others. However, the properties are also satisfied by the more general concept of gauges, which requires convexity in addition to (2.2) and (2.3), and which has become an increasingly popular subject of optimization theory (Friedlander and Macêdo, 2016; Aravkin et al., 2017). Furthermore, we allow for non-convex functions: for example, the category of regularizers covers ℓ_q-operators, $\ell_q(\mathrm{M}) := (\sum_{j=1}^{p} |\mathrm{M}_j|^q)^{1/q}$ for $\mathrm{M} \in \mathbb{R}^p$, even in the non-convex case $q \in (0, 1)$; we refer to Foucart and Lai (2009) for corresponding optimization techniques. More generally, it covers Minkowski functionals $u(\mathrm{M}) := \inf\{a > 0 : \mathrm{M} \in a\mathcal{K}\}$ with level set $\mathcal{K}$ that is bounded and contains an open set around the origin, but is potentially non-symmetric and non-convex. Altogether, we consider a very general class of estimators.

A standard measure to assess the accuracy of estimators is the Kullback-Leibler divergence (Huntsberger and Billingsley, 1981). This measure is particularly suited for our theory, because it can be formulated independently of the model class at hand and yet specifies to established measures in applications. For given $\mathrm{M}, \mathrm{M}' \in \mathfrak{M}$, the Kullback-Leibler divergence from f_{M} to $f_{\mathrm{M}'}$ is defined as

$$d(\mathrm{M}; \mathrm{M}') := \mathbb{E}_{\mathrm{M}'} \log\Big(\frac{f_{\mathrm{M}'}(X)}{f_{\mathrm{M}}(X)}\Big).$$

Given data X, the empirical version of $d(\mathrm{M};\mathrm{M}')$ is then

$$\widehat{d}(\mathrm{M};\mathrm{M}' \mid X) := \log\Big(\frac{f_{\mathrm{M}'}(X)}{f_{\mathrm{M}}(X)}\Big). \tag{2.4}$$

For ease of notation, we assume in the following $X \sim f_{\mathrm{M}^*}$ for the "true" parameter $\mathrm{M}^* \in \mathfrak{M}$ and set $d(\mathrm{M}) := d(\mathrm{M};\mathrm{M}^*)$ and $\widehat{d}(\mathrm{M}) := \widehat{d}(\mathrm{M};\mathrm{M}^* \mid X)$.

We can now formulate an oracle inequality for the MRLE given in (2.1). For this, the function $\tilde{u}$ at $\mathrm{M} \in \mathcal{H}$ is defined as the dual of u by

$$\tilde{u}(\mathrm{M}) := \sup\big\{\langle \mathrm{M},\, \mathrm{M}'\rangle \mid \mathrm{M}' \in \mathcal{H}, u(\mathrm{M}') \leq 1\big\},$$

where $\langle\cdot,\,\cdot\rangle$ is the inner product on $\mathcal{H}$. Moreover, $\nabla(d-\widehat{d})_{\widehat{\mathrm{M}}} \in \mathcal{H}$ denotes any subgradient of $d(\mathrm{M}) - \widehat{d}(\mathrm{M})$ at $\widehat{\mathrm{M}}$. We then find the following.

Theorem 2.2.1 (oracle inequality)**.** *For all $r \geq \tilde{u}\big(\nabla(d-\widehat{d})_{\widehat{\mathrm{M}}}\big)$, it holds that*

$$d(\widehat{\mathrm{M}}) \leq ru(\mathrm{M}^*) + ru(-\mathrm{M}^*).$$

The bound has three building blocks. First, the Kullback-Leibler loss is used as a measure of the accuracy of the MRLEs. In many examples, this loss equals a classical prediction loss. Second, the "noise term" $\tilde{u}\big(\nabla(d-\widehat{d})_{\widehat{\mathrm{M}}}\big)$ forms a lower bound on the regularization parameter. This term can typically be controlled by using bounds from empirical process theory. Finally, the (symmetrized) size of the true model $u(\mathrm{M}^*) + u(-\mathrm{M}^*)$ scales the accuracy bounds. The size is measured in terms of u, which reflects the rationale for choosing u in the first place.

Theorem 2.2.1 is in the form of an oracle inequality, which is a standard way to capture the performance of regularized estimators (Bühlmann and van de Geer, 2011). Importantly, oracle inequalities provide finite sample guarantees and are thus, as opposed to asymptotic results, of direct relevance in practice. However, the inequality also entails upper bounds on the rates of convergence. Note first that the size of the true model can be considered as a basically constant factor; indeed, in view of the motivation of regularization being that there

is a true model with reasonable size in u, largely inflating values of $u(\pm\mathrm{M}^*)$ would indicate an inappropriate choice of the regularization function. As a conclusion, one can derive bounds for the rates of convergence essentially by looking at the regularization parameter r.

An immediate question is whether the bounds in Theorem 2.2.1 are optimal. To answer this question, we first recall that for specific examples that fit our general framework, "fast rate" bounds proportional to r^2 rather than r have been derived, but despite the inaccurate nomenclature, their rates are not necessarily fast. In particular, bounds proportional to r^2 contain additional factors that can slow down the rates, and more directly for scalable estimators, the known bounds rely on strong additional assumptions. Instead, it has been shown that bounds proportional to r are optimal in lasso-type regression in the absence of further assumptions, which means that the bounds in Theorem 2.2.1 are indeed optimal in the sense that they cannot be improved in general — see Sections 2.1.2 and 2.3.1 for details.

In summary, Theorem 2.2.1 provides bounds for a wide range of models and corresponding MRLEs. Therefore, the theorem is an umbrella for bounds linear in r that have been derived for specific examples previously. The proof, however, differs from the previous ones in the way that it uses convexity arguments, Hölder-type inequalities, and connections between the log-likelihood and the Kullback-Leibler loss. Furthermore, and more importantly, Theorem 2.2.1 also entails guarantees for models and estimators that have not yet been equipped with assumptionless bounds - or any bounds at all.

2.3 Examples

We now give explicit bounds for high-dimensional tensor response regression, generalized linear tensor regression, and graphical models. The bounds are the first ones to provide assumptionless Kullback-Leibler guarantees for MRLEs in these models. An exception is linear regression with lasso-type regularization, where assumptionless guarantees have been derived before. We show that we recover the known bounds in this case.

2.3.1 Tensor Response Regression

Our first example is tensor response regression. In a standard notation (see Kolda (2006) or our Section 2.5.4 for details), tensor response regression is based on models of the form

$$Y^i = \mathrm{M}^* \times_1 \mathbf{z}^i + E^i \quad \big(i \in \{1, \ldots, n\}\big),$$

where $Y^i \in \mathbb{R}^{b_2 \times \cdots \times b_p}$ is a $(p-1)$th order tensor response, $\mathrm{M}^* \in \mathfrak{M} \subset \mathbb{R}^{b_1 \times \cdots \times b_p}$ is a pth order tensor coefficient, $\mathbf{z}^i \in \mathbb{R}^{1 \times b_1}$ is a fixed or random row-vector of covariates, and $E^i \in \mathbb{R}^{b_2 \times \cdots \times b_p}$ is random $(p-1)$th order tensor noise. The operation $\mathrm{M}^* \times_1 \mathbf{z}^i$ denotes the mode-1 product of M^* and $\mathbf{z}^i$.

Our goal is to estimate the predictive structure of the above model. Assuming that the noise tensors E^i are mutually independent, the MRLEs in (2.1) are of the form

$$\widehat{\mathrm{M}} \in \operatorname*{argmin}_{\mathrm{M} \in \mathfrak{M}} \big\{ -\sum_{i=1}^{n} \log f_{\mathrm{M}}(Y^i \mid \mathbf{z}^i) + ru(\mathrm{M}) \big\}.$$

For computational ease, $\mathfrak{M}$ is usually chosen as a set of low-rank tensors (Rabusseau and Kadri, 2016; Sun and Li, 2016). In any case, if the conditional density $f_{\mathrm{M}}(Y^i \mid \mathbf{z}^i)$ is parametrized such that $\mathrm{M} \mapsto \log f_{\mathrm{M}} - \mathbb{E}_{\mathrm{M}^*} \log f_{\mathrm{M}}$ is convex, we can derive statistical guarantees in conditional Kullback-Leibler loss from Theorem 2.2.1. Importantly, we do not impose any additional restriction on the covariates or the noise; for example, we allow the covariates to be correlated with the noise. Typical regularizers for third order tensors, for example, include the sparsity-inducing regularizer at the entry level $u(\mathrm{M}) := \sum_{i_1=1}^{b_1} \cdots \sum_{i_3=1}^{b_3} |\mathrm{M}_{i_1 i_2 i_3}|$, at the fiber level $u(\mathrm{M}) := \sum_{i_2=1}^{b_2} \sum_{i_3=1}^{b_3} \|\mathrm{M}_{\cdot i_2 i_3}\|_2$, and at the slice level $u(\mathrm{M}) := \sum_{i_3=1}^{b_3} \|\mathrm{M}_{\cdot\cdot i_3}\|_F := \sum_{i_3=1}^{b_3} \sqrt{\sum_{i_1=1}^{b_1} \sum_{i_2=1}^{b_2} |\mathrm{M}_{i_1 i_2 i_3}|^2}$ and the low-rank inducing regularizer $u(\mathrm{M}) := \|\mathrm{M}\|_*$ with $\|\cdot\|_*$ the tensor nuclear norm defined in Raskutti et al. (2015). Our framework covers all these examples.

For illustration, we consider tensor response regression with zero-mean array normal noise (Akdemir and Gupta, 2011; Hoff, 2011), the most widely-used representative of the

above model class. The conditional Lebesgue density of Y^i given by (Hoff, 2011) is

$$f_{\mathrm{M}}(Y^i \mid \mathbf{z}^i) = (2\pi)^{-b/2}\big(\prod_{k=2}^{p} |\Sigma_k|^{-b/(2b_k)}\big) \cdot \exp\big(-\frac{1}{2}\|(Y^i - \mathrm{M} \times_1 \mathbf{z}^i) \times \Sigma^{-1/2}\|^2\big),$$

where $\|\cdot\|^2$ is the array norm, $b := \prod_{j=2}^{p} b_j$, $\Sigma_k = A_k A_k^\top \in \mathbb{R}^{b_k \times b_k}$ with non-singular real matrix $A_k \in \mathbb{R}^{b_k \times b_k}$, $\Sigma^{-1/2} = \{A_2^{-1}, \ldots, A_p^{-1}\}$, and $\times$ denotes the tensor product. One can check readily that $\log f_{\mathrm{M}}(Y^i \mid \mathbf{z}^i) - \mathbb{E}_{\mathrm{M}^*} \log f_{\mathrm{M}}(Y^i \mid \mathbf{z}^i)$ is linear in M. Hence our theory applies; in particular, Theorem 2.2.1 specializes to array normal models as follows.

Lemma 2.3.1 (tensor response regression with array normal noise)**.** *For all* $r \geq \tilde{u}\big(\sum_{i=1}^{n} \big(E^i \times \Sigma^{-1} \times_1 (\mathbf{z}^i)^\top\big)\big)$*, where* $\Sigma^{-1} = \{\Sigma_2^{-1}, \ldots, \Sigma_p^{-1}\}$*, it holds that*

$$d(\widehat{\mathrm{M}}) \leq ru(\mathrm{M}^*) + ru(-\mathrm{M}^*),$$

with Kullback-Leibler loss

$$d(\widehat{\mathrm{M}}) = \frac{1}{2}\sum_{i=1}^{n} \|(\mathrm{M}^* - \widehat{\mathrm{M}}) \times_1 \mathbf{z}^i \times \Sigma^{-1/2}\|^2.$$

This bound entails that MRLEs for tensor response regression with array normal noise are consistent in average conditional Kullback-Leiber loss under minimal assumptions. Our results thus complement the known consistency guarantees, which hold for specific tensor regressions with additional constraints on the covariates (Raskutti et al., 2015; Sun and Li, 2016). In addition, Lemma 2.3.1 elucidates the interpretation of the conditional Kullback-Leibler loss as a prediction loss.

For an instantiation of the bound, assume $\sum_{i=1}^{n} (\mathbf{z}^i)_j^2/n = 1$, $j \in \{1, \ldots, b_1\}$, and $(\Sigma_k^{-1})_{i_k i_k} = h_k^2$, $h_k > 0$, $k \in \{2, \ldots, p\}$, $i_k \in \{1, \ldots, b_k\}$. Consider the sparsity-inducing regularizer at the entry level $u(\mathrm{M}) := \sum_{i_1=1}^{b_1} \cdots \sum_{i_p=1}^{b_p} |\mathrm{M}_{i_1,\ldots,i_p}|$. Then, the tuning parameter r can be

calibrated such that with probability at least $1 - 2\exp(-t^2)$, it holds that

$$d(\widehat{\mathrm{M}}) \leq 2\big(\prod_{k=2}^{p} h_k\big)\sqrt{2n\big(t^2 + \log(\prod_{j=1}^{p} b_j)\big)}\sum_{i_1=1}^{b_1}\cdots\sum_{i_p=1}^{b_p}|\mathrm{M}^*_{i_1,\ldots,i_p}|.$$

This concrete bound follows from Lemma 2.5.2. That lemma and its proof — as well as all other technical derivations — are deferred to the supplementary materials.

While the above results for tensor regression are novel, assumptionless bounds for simple regression with lasso-type estimators such as lasso (Tibshirani, 1996), group lasso (Yuan and Lin, 2006), sparse group lasso (Simon et al., 2013), and slope estimator (Bogdan et al., 2015) have been derived before. Simple linear regression is thus an ideal test case to confirm that our results specialize correctly. To this end, we first observe that for $p = 2$ and $b_2 = 1$, tensor response regression with array normal noise reduces to ordinary linear regression of the form

$$y^i = \mathbf{z}^i\beta^* + \varepsilon^i \quad \big(i \in \{1,\ldots,n\}\big),$$

where $y^i \in \mathbb{R}$ is a scalar response, $\mathbf{z}^i \in \mathbb{R}^{1\times b_1}$ is a row-vector of covariates, $\beta^* \in \mathbb{R}^{b_1}$ is the regression vector, and $\varepsilon^i \in \mathbb{R}$ is noise distributed as $\mathcal{N}(0,\sigma^2)$. For $u(\beta) := \|\beta\|_1 := \sum_{i=1}^{b_1}|\beta_i|$, the MRLE (2.1) becomes

$$\widehat{\beta} \;\in\; \operatorname*{argmin}_{\beta\in\mathbb{R}^{b_1}}\big\{\frac{1}{2\sigma^2}\sum_{i=1}^{n}\big(y^i - \mathbf{z}^i\beta\big)^2 + r\|\beta\|_1\big\},$$

which, setting $r = r'/(2\sigma^2)$, can be written in the standard lasso-form

$$\widehat{\beta} \;\in\; \operatorname*{argmin}_{\beta\in\mathbb{R}^{b_1}}\big\{\sum_{i=1}^{n}\big(y^i - \mathbf{z}^i\beta\big)^2 + r'\|\beta\|_1\big\}.$$

If $r' \geq 2\|\sum_{i=1}^{n}(\mathbf{z}^i)^\top\varepsilon^i\|_\infty$, Lemma 2.3.1 now implies the bound

$$\sum_{i=1}^{n}\big(\mathbf{z}^i\beta^* - \mathbf{z}^i\widehat{\beta}\big)^2 \leq 2r'\|\beta^*\|_1.$$

This result equals the classical penalty bound for lasso prediction, see Hebiri and Lederer (2013, Equation 3) for example. It has been shown that these bounds are essentially optimal in the absence of other assumptions (Foygel and Srebro, 2011; Zhang et al., 2017; Dalalyan et al., 2017). Along the same lines, one can also show that our bounds specify correctly for the other lasso-type estimators mentioned above (Lederer et al., 2016, Section 3), and similarly, for trace regression (Koltchinskii et al., 2011, Theorem 1).

2.3.2 Generalized Linear Tensor Regression

Our second example is generalized linear tensor regression. The corresponding models consist of two components (Zhou et al., 2013; Li et al., 2013): an exponential family distribution and a link function. The exponential family distribution reads

$$f(y^i \mid \theta^i) = \exp\big(\frac{y^i\theta^i - b(\theta^i)}{\alpha} + c(y^i, \alpha)\big) \qquad \big(i \in \{1, \dots, n\}\big),$$

where $y^i \in \mathbb{R}$ is a scalar response, $\theta^i \in \mathbb{R}$ is the natural parameter, $\alpha > 0$ is the overdispersion factor, and b, c are known real-valued functions. The link function $g : \mathbb{R} \mapsto \mathbb{R}$, assumed strictly increasing, provides a linear connection between the mean functions $\mathbb{E}(y^i \mid \theta^i)$ and tensor predictors $\mathbf{z}^i \in \mathbb{R}^{b_1 \times \cdots \times b_p}$ according to

$$g\big(\mathbb{E}(y^i \mid \theta^i)\big) = \langle \mathrm{M}^*, \mathbf{z}^i \rangle,$$

where $\mathrm{M}^* \in \mathfrak{M} \subset \mathbb{R}^{b_1 \times \cdots \times b_p}$ and $\langle \cdot, \cdot \rangle$ is the tensor inner product. One can check that $b'(\theta^i) = \mathbb{E}(y^i \mid \theta^i)$. With canonical link $g := (b')^{-1}$, it holds that $\theta^i = \langle \mathrm{M}^*, \mathbf{z}^i \rangle$ and the distribution of the response y^i conditioned on $\mathbf{z}^i$ has density

$$f_{\mathrm{M}^*}(y^i \mid \mathbf{z}^i) = \exp\big(\frac{y^i\langle \mathrm{M}^*, \mathbf{z}^i \rangle - b(\langle \mathrm{M}^*, \mathbf{z}^i \rangle)}{\alpha} + c(y^i, \alpha)\big).$$

Further, by introducing basis functions in the mean models, it is straightforward to extend this parametric setting to non-parametric frameworks. In sum, generalized linear tensor

regression provides very flexible model classes for scalar responses.

We can now turn to the corresponding MRLEs. Given n independent observations $(y^i, \mathbf{z}^i)$ and considering the canonical link, the MRLEs in (2.1) become

$$\widehat{\mathrm{M}} \in \operatorname*{argmin}_{\mathrm{M}\in\mathfrak{M}} \big\{ \frac{1}{\alpha} \sum_{i=1}^{n} \big(- y^i \langle \mathrm{M}, \mathbf{z}^i \rangle + b(\langle \mathrm{M}, \mathbf{z}^i \rangle) \big) + ru(\mathrm{M}) \big\}.$$

Similarly to tensor response regression, optimization over the full set $\mathbb{R}^{b_1 \times \cdots \times b_p}$ is computationally challenging due to high-dimensionality of the problem. Thus, $\mathfrak{M}$ is typically chosen considerably smaller, with the belief that the true parameter has some additional structure. For example, a choice proposed in Zhou et al. (2013) is $\mathfrak{M} := \{\mathrm{M} \in \mathbb{R}^{b_1 \times \cdots \times b_p} \mid \mathrm{M} = \sum_{i=1}^{m} \beta_1^{(i)} \circ \cdots \circ \beta_p^{(i)}\}$, where m is a fixed integer, $\beta_j^{(i)} \in \mathbb{R}^{b_j}$, and $\circ$ denotes the outer product.

Let us now apply Theorem 2.2.1 to equip MRLEs in generalized linear tensor regression with theoretical guarantees. For this, note that the log-parametrization here is again linear, so that the main theorem indeed applies and yields the following results.

Lemma 2.3.2 (generalized linear tensor regression with canonical link)**.** *For all* $r \geq \tilde{u}\big(\frac{1}{\alpha} \sum_{i=1}^{n} \big(y^i - \mathbb{E}_{\mathrm{M}^*}(y^i)\big)\mathbf{z}^i\big)$*, it holds that*

$$d(\widehat{\mathrm{M}}) \leq ru(\mathrm{M}^*) + ru(-\mathrm{M}^*),$$

with Kullback-Leibler loss

$$d(\widehat{\mathrm{M}}) = \frac{1}{\alpha} \sum_{i=1}^{n} \big(g^{-1}(\langle \mathrm{M}^*, \mathbf{z}^i \rangle) \cdot \langle \mathrm{M}^* - \widehat{\mathrm{M}}, \mathbf{z}^i \rangle - b(\langle \mathrm{M}^*, \mathbf{z}^i \rangle) + b(\langle \widehat{\mathrm{M}}, \mathbf{z}^i \rangle) \big),$$

and $\mathbb{E}_{\mathrm{M}^*}(y^i)$ *denotes the conditional expectation of* y^i *on* $\mathbf{z}^i$ *here.*

To the best of our knowledge, this is the first oracle inequality for regularized generalized linear tensor regression.

As a special case, Lemma 2.3.2 applies to ordinary logistic regression, where $p = 1$, the canonical link is $g(x) = \log\big(x/(1-x)\big)$, and the MRLEs with a general regularizer are in

the form of

$$\widehat{\mathrm{M}} \in \operatorname*{argmin}_{\mathrm{M}\in\mathbb{R}^{b_1}} \Big\{ \sum_{i=1}^{n} \big(- y^i \langle \mathrm{M},\, \mathbf{z}^i \rangle + \log\big(1 + e^{\langle \mathrm{M},\, \mathbf{z}^i \rangle}\big)\big) + r u(\mathrm{M}) \Big\}.$$

Lemma 2.3.2 then implies the bound

$$d(\widehat{\mathrm{M}}) \leq r u(\mathrm{M}^*) + r u(-\mathrm{M}^*)$$

for $r \geq \tilde{u}\big(\sum_{i=1}^{n} \big(y^i - e^{\langle \mathrm{M}^*,\, \mathbf{z}^i \rangle} / \big(1 + e^{\langle \mathrm{M}^*,\, \mathbf{z}^i \rangle}\big)\big)\mathbf{z}^i\big)$. For ℓ_1-regularization, the tuning parameter r can be calibrated such that with probability at least $1 - 2\exp(-t^2)$, it holds that

$$d(\widehat{\mathrm{M}}) \leq 2\sqrt{\frac{1 + 2\max_i\{p^i(1-p^i)\}}{3}\, n(t^2 + \log b_1)} \sum_{j=1}^{b_1} |\mathrm{M}_j^*|,$$

where $p^i := \mathbb{E}_{\mathrm{M}^*}(y^i)$. We refer to Lemma 2.5.3 in Section 2.5.3 for details. The bound complements results for (weighted) ℓ_1-regularized logistic regression that have been derived under additional assumptions, see van de Geer (2008) and van de Geer (2016, Chapter 12.4).

2.3.3 Graphical Models

Standard types of graphical models, such as Gaussian graphical models (Yuan and Lin, 2007; Friedman et al., 2008), non-paranormal graphical models (Gu et al., 2015), and Ising models (Lenz, 1920; Brush, 1967)

Exponential trace models are based on densities of the form

$$f_{\mathrm{M}}(\mathbf{x}) = \exp\big(- \langle \mathrm{M},\, T(\mathbf{x}) \rangle - a(\mathrm{M})\big) \qquad \big(\mathrm{M} \in \mathfrak{M}\big)$$

with respect to some σ-finite measure ν on $\mathbb{R}^p$. Here, the matrix-valued parameter M encodes the dependence structure of the random vector $X \in \mathcal{X} \subset \mathbb{R}^p$, the matrix-valued function $T(\cdot)$ on $\mathcal{X}$ determines how the data enters the model, $a(\mathrm{M})$ is the normalization, and $\mathfrak{M}$

is a convex set of matrices M with finite normalization. Given independent observations $X^1, \ldots, X^n$ of X, the MRLEs in (2.1) are of the form

$$\widehat{\mathrm{M}} \in \operatorname*{argmin}_{\mathrm{M} \in \mathfrak{M}} \big\{ \sum_{i=1}^{n} \langle \mathrm{M},\, T(X^i) \rangle + n a(\mathrm{M}) + r u(\mathrm{M}) \big\}.$$

Since the function $\log f_{\mathrm{M}} - \mathbb{E}_{\mathrm{M}^*} \log f_{\mathrm{M}}$ is linear in M, we can then apply Theorem 2.2.1 to derive the following bound.

Lemma 2.3.3 (graphical models)**.** *For all* $r \geq \tilde{u}\big(\sum_{i=1}^{n} \big(\mathbb{E}_{\mathrm{M}^*} T(X^i) - T(X^i) \big) \big)$*, it holds that*

$$d(\widehat{\mathrm{M}}) \leq r u(\mathrm{M}^*) + r u(-\mathrm{M}^*),$$

with Kullback-Leibler loss

$$d(\widehat{\mathrm{M}}) = \langle \sum_{i=1}^{n} \mathbb{E}_{\mathrm{M}^*} T(X^i),\, \widehat{\mathrm{M}} - \mathrm{M}^* \rangle - n a(\mathrm{M}^*) + n a(\widehat{\mathrm{M}}).$$

The Kullback-Leibler loss is a standard predictive risk for graphical models (Yuan and Lin, 2007; Shevlyakova and Morgenthaler, 2013) and has a geometric interpretation as the difference between $a(\widehat{\mathrm{M}})$ and the tangent approximation of $a(\widehat{\mathrm{M}})$ at M^* (Wainwright and Jordan, 2008, Chapter 5.2.2).

As a special case, Lemma 2.3.3 applies to the graphical lasso for multivariate Gaussian data. Recall that the graphical lasso (Yuan and Lin, 2007; Friedman et al., 2008) is formulated as

$$\widehat{\mathrm{M}} \in \operatorname*{argmin}_{\mathrm{M} \in S^p_{++}} \big\{ \frac{1}{n} \sum_{i=1}^{n} \operatorname{tr}\big(X^i (X^i)^\top \mathrm{M} \big) - \log\det \mathrm{M} + r' \|\mathrm{vec}(\mathrm{M})\|_1 \big\},$$

where S^p_{++} is the set of positive definite $p \times p$ matrices and $X^1, \ldots, X^n$ are i.i.d. samples from a centered Gaussian distribution with unknown covariance matrix $(\mathrm{M}^*)^{-1}$. The Kullback-Leibler loss of $\widehat{\mathrm{M}}$ reads

$$\frac{1}{n} d(\widehat{\mathrm{M}}) = \frac{1}{2} \big(\langle \widehat{\mathrm{M}},\, (\mathrm{M}^*)^{-1} \rangle - \log\det \widehat{\mathrm{M}} + \log\det \mathrm{M}^* - p \big),$$

which is equivalent (up to the factor 1/2) to Stein's loss of the centered multivariate Gaussian distribution with covariance matrix $(\widehat{\mathrm{M}})^{-1}$ (James and Stein, 1961). Thus, if $r' \geq \|\mathrm{vec}\big((\mathrm{M}^*)^{-1} - \frac{1}{n}\sum_{i=1}^n X^i(X^i)^\top\big)\|_\infty$, Lemma 2.3.3 yields

$$\langle \widehat{\mathrm{M}},\, (\mathrm{M}^*)^{-1} \rangle - \log\det \widehat{\mathrm{M}} + \log\det \mathrm{M}^* - p \leq 2r' \|\mathrm{vec}(\mathrm{M}^*)\|_1.$$

Following Lemma 2.5.4 in Section 2.5.3, the tuning parameter r' can be calibrated such that with probability at least $1 - 4\exp(-t^2)$, it holds that

$$\frac{1}{n} d(\widehat{\mathrm{M}}) \leq 160 \max_{k\in\{1,\ldots,p\}} \big((\mathrm{M}^*)^{-1}\big)_{kk} \sqrt{\frac{1}{n}\big(t^2 + \log\big(p(p-1)\big)\big)} \sum_{i_1=1}^p \sum_{i_2=1}^p |\mathrm{M}^*_{i_1 i_2}|$$

for all t such that $0 < t < \sqrt{n/4 - \log\big(p(p-1)\big)}$. Our theory thus establishes the rate $\sqrt{\log p/n}$ for the graphical lasso in Kullback-Leibler loss. This prediction rate complements the known estimation rates $\sqrt{\log p/n}$ in vectorized-matrix ℓ_∞-norm and $\sqrt{\min\{s+p, d^2\}\log p/n}$ in spectral norm (Ravikumar et al., 2011), where s is the number of non-zero elements in M^* and d is the maximum node degree. However, those results additionally require the mutual incoherence condition. Our prediction rate also complements the known estimation rate $\sqrt{(s+p)\log p/n}$ in Frobenius norm and spectral norm (Rothman et al., 2008), which requires only mild assumptions on the population covariance matrix.

2.4 Summary

We have established assumptionless oracle inequalities for a general class of maximum regularized likelihood estimators. For regression, the inequalities match known lower bounds up to log-factors. We conjecture that the same is true more generally; in particular, we believe that general counter-examples to "fast rates" can be generated similarly as in the regression case.

2.5 Supplementary Materials

2.5.1 Proof of Theorem 2.2.1

Before proving Theorem 2.2.1, we first introduce a lemma about the regularizer. Throughout, we use the convention $0 \cdot \infty := \infty$.

Lemma 2.5.1 (inner product inequality)**.** *Let* $\mathrm{M}, \mathrm{M}' \in \mathcal{H}$*. It holds that*

$$\langle \mathrm{M},\, \mathrm{M}' \rangle \leq \tilde{u}(\mathrm{M}) u(\mathrm{M}'). \tag{2.5}$$

Proof of Lemma 2.5.1. The proof consists of two steps. First, we show that Inequality (2.5) holds in the case $u(\mathrm{M}') = 0$. Second, we show that Inequality (2.5) holds in the case $u(\mathrm{M}') \neq 0$. In view of the mentioned convention, we can assume that $u(\mathrm{M}') < \infty$.

Case 1. If $u(\mathrm{M}') = 0$, we have $\mathrm{M}' = 0$ since $u(\mathrm{M}') = 0$ if and only if $\mathrm{M}' = 0$ by condition (2). Then,

$$\langle \mathrm{M},\, \mathrm{M}' \rangle = \tilde{u}(\mathrm{M}) u(\mathrm{M}') = 0.$$

In particular, $\langle \mathrm{M},\, \mathrm{M}' \rangle \leq \tilde{u}(\mathrm{M}) u(\mathrm{M}')$, as desired.

Case 2. If $u(\mathrm{M}') \neq 0$, we rewrite

$$\langle \mathrm{M},\, \mathrm{M}' \rangle = \langle \mathrm{M},\, \mathrm{M}' \rangle \cdot \frac{u(\mathrm{M}')}{u(\mathrm{M}')} = \langle \mathrm{M},\, \frac{\mathrm{M}'}{u(\mathrm{M}')} \rangle \cdot u(\mathrm{M}').$$

Next we show that $\langle \mathrm{M},\, \frac{\mathrm{M}'}{u(\mathrm{M}')} \rangle \leq \tilde{u}(\mathrm{M})$ by two observations. The first observation is that since $\mathcal{H}$ is a real vector space,

$$\frac{\mathrm{M}'}{u(\mathrm{M}')} \in \mathcal{H}.$$

The second observation is that $u(\mathrm{M}') \in (0, \infty)$ in Case 2, and therefore for regularizers that

are positive homogeneous of degree one as specified in Condition (2.3),

$$u\Big(\frac{\mathrm{M}'}{u(\mathrm{M}')}\Big) = \frac{u(\mathrm{M}')}{u(\mathrm{M}')} = 1.$$

Combining the two observations, we have

$$\langle \mathrm{M},\, \frac{\mathrm{M}'}{u(\mathrm{M}')}\rangle \leq \sup\{\langle \mathrm{M},\, \Lambda_1\rangle \mid \Lambda_1 \in \mathcal{H}, u(\Lambda_1) \leq 1\} = \tilde{u}(\mathrm{M}).$$

Since $u(\mathrm{M}') \in (0, \infty)$, it follows that $\langle \mathrm{M},\, \mathrm{M}'\rangle \leq \tilde{u}(\mathrm{M})u(\mathrm{M}')$. □

We now proceed to the proof of Theorem 2.2.1.

Proof of Theorem 2.2.1. The proof consists of three steps. First, we link the objective function of MRLEs with the regularized Kullback-Leibler loss. Second, we use the convexity of $\mathrm{M} \mapsto \log f_{\mathrm{M}} - \mathbb{E}_{\mathrm{M}^*} \log f_{\mathrm{M}}$ to obtain an enhanced basic inequality. Third, we use the properties of u and $\tilde{u}$ shown in Lemma 2.5.1 to bound the empirical process and conclude the proof.

Step 1: (Regularized Kullback-Leibler Loss) We first show that the MRLE $\widehat{\mathrm{M}}$ defined in (2.1) satisfies

$$\widehat{\mathrm{M}} \in \operatorname*{argmin}_{\mathrm{M}\in\mathfrak{M}} \big\{\widehat{d}(\mathrm{M}) + ru(\mathrm{M})\big\}.$$

Recall that MRLEs are defined as

$$\widehat{\mathrm{M}} \in \operatorname*{argmin}_{\mathrm{M}\in\mathfrak{M}} \big\{ -\log f_{\mathrm{M}}(X) + ru(\mathrm{M})\big\}.$$

Since adding constant terms does not alter the estimator, we rewrite the definition as

$$\widehat{\mathrm{M}} \in \operatorname*{argmin}_{\mathrm{M}\in\mathfrak{M}} \big\{ \log f_{\mathrm{M}^*}(X) - \log f_{\mathrm{M}}(X) + ru(\mathrm{M})\big\}.$$

The term $\log f_{\mathrm{M}^*}(X) - \log f_{\mathrm{M}}(X)$ is the empirical version of the Kullback-Leibler loss defined

in Equation (2.4). Hence we obtain an equivalent definition of MRLE in the form of

$$\widehat{\mathrm{M}} \in \operatorname*{argmin}_{\mathrm{M}\in\mathfrak{M}} \big\{\widehat{d}(\mathrm{M}) + ru(\mathrm{M})\big\}.$$

This concludes Step 1.

Step 2: (Enhanced Basic Inequality) We use Step 1 and the convexity of $\mathrm{M} \mapsto \log f_{\mathrm{M}} - \mathbb{E}_{\mathrm{M}^*} \log f_{\mathrm{M}}$ to derive the enhanced basic inequality

$$d(\widehat{\mathrm{M}}) \leq ru(\mathrm{M}^*) - ru(\widehat{\mathrm{M}}) + \langle \nabla(d - \widehat{d})_{\widehat{\mathrm{M}}},\, \widehat{\mathrm{M}}\rangle + \langle \nabla(d - \widehat{d})_{\widehat{\mathrm{M}}},\, -\mathrm{M}^*\rangle.$$

The proof of this inequality has two ingredients. The first ingredient is that $\widehat{\mathrm{M}}$ minimizes $\widehat{d}(\mathrm{M}) + ru(\mathrm{M})$, as derived in Step 1. Hence, in particular,

$$\widehat{d}(\widehat{\mathrm{M}}) + ru(\widehat{\mathrm{M}}) \leq \widehat{d}(\mathrm{M}^*) + ru(\mathrm{M}^*).$$

Rearranging the inequality yields

$$\widehat{d}(\widehat{\mathrm{M}}) \leq \widehat{d}(\mathrm{M}^*) + ru(\mathrm{M}^*) - ru(\widehat{\mathrm{M}}).$$

The second ingredient is the convexity of $\log f_{\mathrm{M}} - \mathbb{E}_{\mathrm{M}^*} \log f_{\mathrm{M}}$ in M. Since $d(\mathrm{M}) - \widehat{d}(\mathrm{M}) = (\mathbb{E}_{\mathrm{M}^*} \log f_{\mathrm{M}^*} - \log f_{\mathrm{M}^*}) + (\log f_{\mathrm{M}} - \mathbb{E}_{\mathrm{M}^*} \log f_{\mathrm{M}})$, the convexity implies that the function $d(\mathrm{M}) - \widehat{d}(\mathrm{M})$ is also convex in M. Hence, it holds that

$$d(\mathrm{M}^*) - \widehat{d}(\mathrm{M}^*) \geq d(\widehat{\mathrm{M}}) - \widehat{d}(\widehat{\mathrm{M}}) + \langle \nabla(d - \widehat{d})_{\widehat{\mathrm{M}}},\, \mathrm{M}^* - \widehat{\mathrm{M}}\rangle,$$

where $\nabla(d - \widehat{d})_{\widehat{\mathrm{M}}}$ is any subgradient of $d(\mathrm{M}) - \widehat{d}(\mathrm{M})$ at $\widehat{\mathrm{M}}$. Rearranging the equality leads to

$$d(\widehat{\mathrm{M}}) - \widehat{d}(\widehat{\mathrm{M}}) \leq d(\mathrm{M}^*) - \widehat{d}(\mathrm{M}^*) + \langle \nabla(d - \widehat{d})_{\widehat{\mathrm{M}}},\, \widehat{\mathrm{M}} - \mathrm{M}^*\rangle.$$

Combining the two ingredients and doing some algebra yield

$$d(\widehat{\mathrm{M}}) \leq d(\mathrm{M}^*) + ru(\mathrm{M}^*) - ru(\widehat{\mathrm{M}}) + \langle \nabla(d - \widehat{d})_{\widehat{\mathrm{M}}}, \widehat{\mathrm{M}} - \mathrm{M}^* \rangle.$$

By the definition of $d(\mathrm{M}^*)$, we also find

$$d(\mathrm{M}^*) = \mathbb{E}_{\mathrm{M}^*} \log\Big(\frac{f_{\mathrm{M}^*}(x)}{f_{\mathrm{M}^*}(x)}\Big) = 0.$$

We can thus remove $d(\mathrm{M}^*)$ from the inequality above and find

$$d(\widehat{\mathrm{M}}) \leq ru(\mathrm{M}^*) - ru(\widehat{\mathrm{M}}) + \langle \nabla(d - \widehat{d})_{\widehat{\mathrm{M}}}, \widehat{\mathrm{M}} - \mathrm{M}^* \rangle = ru(\mathrm{M}^*) - ru(\widehat{\mathrm{M}}) + \langle \nabla(d - \widehat{d})_{\widehat{\mathrm{M}}}, \widehat{\mathrm{M}} \rangle + \langle \nabla(d - \widehat{d})_{\widehat{\mathrm{M}}}, -\mathrm{M}$$

This concludes Step 2.

Step 3: (Bound for the Empirical Process Term) We show that on the event where $r \geq \tilde{u}(\nabla(d - \widehat{d})_{\widehat{\mathrm{M}}})$, it holds that

$$d(\widehat{\mathrm{M}}) \leq ru(\mathrm{M}^*) + ru(-\mathrm{M}^*).$$

To this end, we first apply Lemma 2.5.1 to the last two terms on right-hand side of the result in Step 2 and find

$$d(\widehat{\mathrm{M}}) \leq ru(\mathrm{M}^*) - ru(\widehat{\mathrm{M}}) + \tilde{u}(\nabla(d - \widehat{d})_{\widehat{\mathrm{M}}}) u(\widehat{\mathrm{M}}) + \tilde{u}(\nabla(d - \widehat{d})_{\widehat{\mathrm{M}}}) u(-\mathrm{M}^*).$$

Rearranging the terms of the right-hand side yields

$$d(\widehat{\mathrm{M}}) \leq ru(\mathrm{M}^*) + \tilde{u}(\nabla(d - \widehat{d})_{\widehat{\mathrm{M}}}) u(-\mathrm{M}^*) + \big(- r + \tilde{u}(\nabla(d - \widehat{d})_{\widehat{\mathrm{M}}}) \big) u(\widehat{\mathrm{M}}).$$

Because $u(\cdot)$ is non-negative by definition, the inequality $r \geq \tilde{u}(\nabla(d - \widehat{d})_{\widehat{\mathrm{M}}})$ implies

$$\big(- r + \tilde{u}(\nabla(d - \widehat{d})_{\widehat{\mathrm{M}}})\big) u(\widehat{\mathrm{M}}) \leq 0,$$

and

$$\tilde{u}(\nabla(d - \widehat{d})_{\widehat{\mathrm{M}}}) u(-\mathrm{M}^*) \leq r u(-\mathrm{M}^*).$$

Combining the last three inequalities, we thus find on the event where $r \geq \tilde{u}(\nabla(d - \widehat{d})_{\widehat{\mathrm{M}}})$ the inequality

$$d(\widehat{\mathrm{M}}) \leq r u(\mathrm{M}^*) + r u(-\mathrm{M}^*).$$

This concludes the Step 3 and thus completes the proof of Theorem 2.2.1. □

2.5.2 Proof of Example Results

Proof of Lemma 3.1. This lemma is a specification of Theorem 2.2.1 to tensor response regression with array normal noise. The proof consists of three steps. First, we obtain the explicit form of the empirical and population version of Kullback-Leibler loss, $\widehat{d}(\mathrm{M})$ and $d(\mathrm{M})$. Second, we derive the gradient of $d(\mathrm{M}) - \widehat{d}(\mathrm{M})$ at $\widehat{\mathrm{M}}$, denoted as $\nabla(d - \widehat{d})_{\widehat{\mathrm{M}}}$. At last, we apply Theorem 2.2.1 with the derived explicit forms of $d(\widehat{\mathrm{M}})$ and $\nabla(d - \widehat{d})_{\widehat{\mathrm{M}}}$ to conclude the proof.

Step 1. We first derive the explicit form of the empirical and population version of Kullback-Leibler loss. Plugging the array normal density into the empirical Kullback-Leibler

divergence between conditional densities f_{M} and f_{M^*} yields

$$\begin{aligned}\widehat{d}(\mathrm{M}) &= \frac{1}{2}\sum_{i=1}^{n}\big(\|(Y^i - \mathrm{M}\times_1 \mathbf{z}^i)\times\Sigma^{-1/2}\|^2 - \|(Y^i - \mathrm{M}^*\times_1 \mathbf{z}^i)\times\Sigma^{-1/2}\|^2\big)\\ &= \frac{1}{2}\sum_{i=1}^{n}\big(\|(Y^i - \mathrm{M}^*\times_1 \mathbf{z}^i + \mathrm{M}^*\times_1 \mathbf{z}^i - \mathrm{M}\times_1 \mathbf{z}^i)\times\Sigma^{-1/2}\|^2 -\\ &\quad \|(Y^i - \mathrm{M}^*\times_1 \mathbf{z}^i)\times\Sigma^{-1/2}\|^2\big).\end{aligned}$$

Lemma 2.5.5 shows that $(Y^i - \mathrm{M}^*\times_1 \mathbf{z}^i + \mathrm{M}^*\times_1 \mathbf{z}^i - \mathrm{M}\times_1 \mathbf{z}^i)\times\Sigma^{-1/2} = (Y^i - \mathrm{M}^*\times_1 \mathbf{z}^i)\times\Sigma^{-1/2} + (\mathrm{M}^*\times_1 \mathbf{z}^i - \mathrm{M}\times_1 \mathbf{z}^i)\times\Sigma^{-1/2}$. We can thus reorganize the above equation as

$$\begin{aligned}\widehat{d}(\mathrm{M}) = \frac{1}{2}\sum_{i=1}^{n}\big(&\|(Y^i - \mathrm{M}^*\times_1 \mathbf{z}^i)\times\Sigma^{-1/2} + (\mathrm{M}^*\times_1 \mathbf{z}^i - \mathrm{M}\times_1 \mathbf{z}^i)\times\Sigma^{-1/2}\|^2 -\\ &\|(Y^i - \mathrm{M}^*\times_1 \mathbf{z}^i)\times\Sigma^{-1/2}\|^2\big).\end{aligned}$$

We expand the first array norm as shown in Lemma 2.5.6 and get

$$\begin{aligned}\widehat{d}(\mathrm{M}) = \frac{1}{2}\sum_{i=1}^{n}\big(&\|(Y^i - \mathrm{M}^*\times_1 \mathbf{z}^i)\times\Sigma^{-1/2}\|^2 + \|(\mathrm{M}^*\times_1 \mathbf{z}^i - \mathrm{M}\times_1 \mathbf{z}^i)\times\Sigma^{-1/2}\|^2\\ &+ 2\langle(Y^i - \mathrm{M}^*\times_1 \mathbf{z}^i)\times\Sigma^{-1/2}, (\mathrm{M}^*\times_1 \mathbf{z}^i - \mathrm{M}\times_1 \mathbf{z}^i)\times\Sigma^{-1/2}\rangle -\\ &\|(Y^i - \mathrm{M}^*\times_1 \mathbf{z}^i)\times\Sigma^{-1/2}\|^2\big).\end{aligned}$$

Canceling the first and last term of $\widehat{d}(\mathrm{M})$ yields

$$\begin{aligned}\widehat{d}(\mathrm{M}) = \frac{1}{2}\sum_{i=1}^{n}\big(&\|(\mathrm{M}^*\times_1 \mathbf{z}^i - \mathrm{M}\times_1 \mathbf{z}^i)\times\Sigma^{-1/2}\|^2\\ &+ 2\langle(Y^i - \mathrm{M}^*\times_1 \mathbf{z}^i)\times\Sigma^{-1/2}, (\mathrm{M}^*\times_1 \mathbf{z}^i - \mathrm{M}\times_1 \mathbf{z}^i)\times\Sigma^{-1/2}\rangle\big).\end{aligned}$$

Taking the expectation of $\widehat{d}(\mathrm{M})$ with respect to Y^i conditioning on $\mathbf{z}^i$ gives the conditional Kullback-Leibler divergence

$$\begin{aligned} d(\mathrm{M}) =& \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}_{\mathrm{M}^*}\big(\|(\mathrm{M}^*\times_1\mathbf{z}^i-\mathrm{M}\times_1\mathbf{z}^i)\times\Sigma^{-1/2}\|^2 \\ &+2\langle(Y^i-\mathrm{M}^*\times_1\mathbf{z}^i)\times\Sigma^{-1/2},(\mathrm{M}^*\times_1\mathbf{z}^i-\mathrm{M}\times_1\mathbf{z}^i)\times\Sigma^{-1/2}\rangle\big) \\ =& \frac{1}{2}\sum_{i=1}^{n}\|(\mathrm{M}^*\times_1\mathbf{z}^i-\mathrm{M}\times_1\mathbf{z}^i)\times\Sigma^{-1/2}\|^2 \\ &+\sum_{i=1}^{n}\mathbb{E}_{\mathrm{M}^*}\langle(Y^i-\mathrm{M}^*\times_1\mathbf{z}^i)\times\Sigma^{-1/2},(\mathrm{M}^*\times_1\mathbf{z}^i-\mathrm{M}\times_1\mathbf{z}^i)\times\Sigma^{-1/2}\rangle, \end{aligned}$$

where $\mathbb{E}_{\mathrm{M}^*}$ denotes the conditional expectation conditioning on $\mathbf{z}^i$. The expectation of the inner product term becomes

$$\begin{aligned} &\mathbb{E}_{\mathrm{M}^*}\langle(Y^i-\mathrm{M}^*\times_1\mathbf{z}^i)\times\Sigma^{-1/2},(\mathrm{M}^*\times_1\mathbf{z}^i-\mathrm{M}\times_1\mathbf{z}^i)\times\Sigma^{-1/2}\rangle \\ =&\sum_{i_2=1}^{b_2}\cdots\sum_{i_p=1}^{b_p}\mathbb{E}_{\mathrm{M}^*}\big((Y^i-\mathrm{M}^*\times_1\mathbf{z}^i)\times\Sigma^{-1/2}\big)_{i_2,\ldots,i_p}\cdot\big((\mathrm{M}^*\times_1\mathbf{z}^i-\mathrm{M}\times_1\mathbf{z}^i)\times\Sigma^{-1/2}\big)_{i_2,\ldots,i_p}. \end{aligned}$$

Further, since $\mathbb{E}_{\mathrm{M}^*}\big(Y^i\big)=\mathrm{M}^*\times_1\mathbf{z}^i$,

$$\begin{aligned} &\mathbb{E}_{\mathrm{M}^*}\big((Y^i-\mathrm{M}^*\times_1\mathbf{z}^i)\times\Sigma^{-1/2}\big)_{i_2,\ldots,i_p} \\ =&\mathbb{E}_{\mathrm{M}^*}\big(\sum_{j_2=1}^{b_2}\cdots\sum_{j_p=1}^{b_p}(Y^i-\mathrm{M}^*\times_1\mathbf{z}^i)_{j_2,\ldots,j_p}\cdot\sigma^{(2)}_{i_2j_2}\ldots\sigma^{(p)}_{i_pj_p}\big) \\ =&\sum_{j_2=1}^{b_2}\cdots\sum_{j_p=1}^{b_p}\mathbb{E}_{\mathrm{M}^*}(Y^i-\mathrm{M}^*\times_1\mathbf{z}^i)_{j_2,\ldots,j_p}\cdot\sigma^{(2)}_{i_2j_2}\ldots\sigma^{(p)}_{i_pj_p} \\ =&0, \end{aligned}$$

where $\sigma_{ij}^{(k)}$ denotes the (i,j)th element of A_k^{-1}. Thus, $\mathbb{E}_{\mathrm{M}^*}\langle(Y^i - \mathrm{M}^* \times_1 \mathbf{z}^i) \times \Sigma^{-1/2}, (\mathrm{M}^* \times_1 \mathbf{z}^i - \mathrm{M} \times_1 \mathbf{z}^i) \times \Sigma^{-1/2}\rangle = 0$. We then find

$$d(\mathrm{M}) = \frac{1}{2}\sum_{i=1}^{n} \|(\mathrm{M}^* \times_1 \mathbf{z}^i - \mathrm{M} \times_1 \mathbf{z}^i) \times \Sigma^{-1/2}\|^2.$$

By the definition of the mode-1 product, we write $\mathrm{M}^* \times_1 \mathbf{z}^i - \mathrm{M} \times_1 \mathbf{z}^i = (\mathrm{M}^* - \mathrm{M}) \times_1 \mathbf{z}^i$ and conclude that the conditional Kullback-Leibler divergence of tensor regression with array normal noise has the explicit form of

$$d(\mathrm{M}) = \frac{1}{2}\sum_{i=1}^{n} \|(\mathrm{M}^* - \mathrm{M}) \times_1 \mathbf{z}^i \times \Sigma^{-1/2}\|^2,$$

which is the prediction error.

Step 2. We derive the explicit form of $\nabla(d - \widehat{d})_{\widehat{\mathrm{M}}}$. With $d(\mathrm{M})$ and $\widehat{d}(\mathrm{M})$ derived in Step 1, we obtain

$$\begin{aligned} d(\widehat{\mathrm{M}}) - \widehat{d}(\widehat{\mathrm{M}}) = &\frac{1}{2}\sum_{i=1}^{n} \|(\mathrm{M}^* - \widehat{\mathrm{M}}) \times_1 \mathbf{z}^i \times \Sigma^{-1/2}\|^2 - \frac{1}{2}\sum_{i=1}^{n} \big(\|(\mathrm{M}^* \times_1 \mathbf{z}^i - \widehat{\mathrm{M}} \times_1 \mathbf{z}^i) \times \Sigma^{-1/2}\|^2 \\ &+ 2\langle(Y^i - \mathrm{M}^* \times_1 \mathbf{z}^i) \times \Sigma^{-1/2}, (\mathrm{M}^* \times_1 \mathbf{z}^i - \widehat{\mathrm{M}} \times_1 \mathbf{z}^i) \times \Sigma^{-1/2}\rangle\big). \end{aligned}$$

Canceling the first and second term in the above equality yields

$$d(\widehat{\mathrm{M}}) - \widehat{d}(\widehat{\mathrm{M}}) = -\sum_{i=1}^{n} \langle(Y^i - \mathrm{M}^* \times_1 \mathbf{z}^i) \times \Sigma^{-1/2}, (\mathrm{M}^* \times_1 \mathbf{z}^i - \widehat{\mathrm{M}} \times_1 \mathbf{z}^i) \times \Sigma^{-1/2}\rangle.$$

Let $\mathcal{W}^i := \langle(Y^i - \mathrm{M}^* \times_1 \mathbf{z}^i) \times \Sigma^{-1/2}, (\mathrm{M}^* \times_1 \mathbf{z}^i - \widehat{\mathrm{M}} \times_1 \mathbf{z}^i) \times \Sigma^{-1/2}\rangle$. We write $d(\widehat{\mathrm{M}}) - \widehat{d}(\widehat{\mathrm{M}}) = -\sum_{i=1}^{n} \mathcal{W}^i$ and find

$$\nabla(d - \widehat{d})_{\widehat{\mathrm{M}}} = -\sum_{i=1}^{n} \nabla\mathcal{W}^i(\widehat{\mathrm{M}}),$$

where $\nabla\mathcal{W}^i(\widehat{\mathrm{M}})$ denotes the gradient of $\mathcal{W}^i$ at $\widehat{\mathrm{M}}$.

Next we derive $\nabla \mathcal{W}^i(\widehat{\mathrm{M}})$. Expanding $\mathcal{W}^i$ by the definition of tensor inner product, we get

$$\mathcal{W}^i = \sum_{i_2=1}^{b_2} \cdots \sum_{i_p=1}^{b_p} \big((Y^i - \mathrm{M}^* \times_1 \mathbf{z}^i) \times \Sigma^{-1/2}\big)_{i_2,\ldots,i_p} \big((\mathrm{M}^* - \widehat{\mathrm{M}}) \times_1 \mathbf{z}^i \times \Sigma^{-1/2}\big)_{i_2,\ldots,i_p}.$$

Expanding the tensor operations in the second term yields

$$\mathcal{W}^i = \sum_{i_2=1}^{b_2} \cdots \sum_{i_p=1}^{b_p} \big((Y^i - \mathrm{M}^* \times_1 \mathbf{z}^i) \times \Sigma^{-1/2}\big)_{i_2,\ldots,i_p} \cdot \Big(\sum_{j_1=1}^{b_1} \cdots \sum_{j_p=1}^{b_p} (\mathrm{M}^* - \widehat{\mathrm{M}})_{j_1,\ldots,j_p} z^i_{1j_1} \sigma^{(2)}_{i_2 j_2} \ldots \sigma^{(p)}_{i_p j_p} \Big).$$

The partial derivative of $\mathcal{W}^i$ with regarding to $\widehat{\mathrm{M}}_{m_1,\ldots,m_p}$ is

$$\frac{\partial \mathcal{W}^i}{\partial \widehat{\mathrm{M}}_{m_1,\ldots,m_p}} = \sum_{i_2=1}^{b_2} \cdots \sum_{i_p=1}^{b_p} \big((Y^i - \mathrm{M}^* \times_1 \mathbf{z}^i) \times \Sigma^{-1/2}\big)_{i_2,\ldots,i_p} (-z^i_{1m_1} \sigma^{(2)}_{i_2 m_2} \ldots \sigma^{(p)}_{i_p m_p}).$$

Here $z^i_{1m_1}$ is the $(m_1, 1)$th entry of $(\mathbf{z}^i)^\top$, $\sigma^{(k)}_{i_k m_k}$ is the (m_k, i_k)th entry of $(A_k^{-1})^\top$, $k \in \{2, \ldots, p\}$. Let $(\Sigma^{-1/2})^\top := \{(A_2^{-1})^\top, \ldots, (A_p^{-1})^\top\}$, we obtain

$$\frac{\partial \mathcal{W}^i}{\partial \widehat{\mathrm{M}}_{m_1,\ldots,m_p}} = -\big((Y^i - \mathrm{M}^* \times_1 \mathbf{z}^i) \times \Sigma^{-1/2} \times (\Sigma^{-1/2})^\top \times_1 (\mathbf{z}^i)^\top\big)_{m_1,\ldots,m_p}.$$

Hence, $\nabla \mathcal{W}^i(\widehat{\mathrm{M}})$ has the form

$$\nabla \mathcal{W}^i(\widehat{\mathrm{M}}) = -(Y^i - \mathrm{M}^* \times_1 \mathbf{z}^i) \times \Sigma^{-1/2} \times (\Sigma^{-1/2})^\top \times_1 (\mathbf{z}^i)^\top.$$

Plugging the explicit form of $\nabla \mathcal{W}^i(\widehat{\mathrm{M}})$ into the equation of $\nabla(d - \widehat{d})_{\widehat{\mathrm{M}}}$ yields

$$\nabla(d - \widehat{d})_{\widehat{\mathrm{M}}} = \sum_{i=1}^{n} (Y^i - \mathrm{M}^* \times_1 \mathbf{z}^i) \times \Sigma^{-1/2} \times (\Sigma^{-1/2})^\top \times_1 (\mathbf{z}^i)^\top.$$

Under the assumed tensor regression model, $Y^i - \mathrm{M}^* \times_1 \mathbf{z}^i = E^i$, so that

$$\nabla(d - \widehat{d})_{\widehat{\mathrm{M}}} = \sum_{i=1}^{n} \big(E^i \times \Sigma^{-1/2} \times (\Sigma^{-1/2})^\top \times_1 (\mathbf{z}^i)^\top\big).$$

Expanding the Tucker product as a sequence of mode products gives

$$\nabla(d - \widehat{d})_{\widehat{\mathrm{M}}} = \sum_{i=1}^{n} \big(E^i \times_1 A_2^{-1} \times_2 \ldots \times_{p-1} A_p^{-1} \times_1 (A_2^{-1})^\top \times_2 \ldots \times_{p-1} (A_p^{-1})^\top \times_1 (\mathbf{z}^i)^\top\big).$$

The properties of the tensor mode product include $(\mathcal{T} \times_i W) \times_j V = (\mathcal{T} \times_j V) \times_i W = \mathcal{T} \times_i W \times_j V$ for $i \neq j$ and $(\mathcal{T} \times_i W) \times_i V = \mathcal{T} \times_i (VW)$ (De Lathauwer et al., 2000). Reorganizing the above equation yields

$$\nabla(d - \widehat{d})_{\widehat{\mathrm{M}}} = \sum_{i=1}^{n} \big(E^i \times_1 \big((A_2^{-1})^\top A_2^{-1}\big) \times_2 \ldots \times_{p-1} \big((A_p^{-1})^\top A_p^{-1}\big) \times_1 (\mathbf{z}^i)^\top\big).$$

Since $(A_k^{-1})^T A_k^{-1} = (A_k A_k^T)^{-1} = \Sigma_k^{-1}$,

$$\nabla(d - \widehat{d})_{\widehat{\mathrm{M}}} = \sum_{i=1}^{n} \big(E^i \times \Sigma^{-1} \times_1 (\mathbf{z}^i)^\top\big).$$

Step 3. We apply Theorem 2.2.1 with the explicit form of $d(\widehat{\mathrm{M}})$ and $\nabla(d - \widehat{d})_{\widehat{\mathrm{M}}}$, and conclude that for all $r \geq \tilde{u}\big(\sum_{i=1}^{n} \big(E^i \times \Sigma^{-1} \times_1 (\mathbf{z}^i)^\top\big)\big)$, it holds that

$$\frac{1}{2} \sum_{i=1}^{n} \|(\mathrm{M}^* - \widehat{\mathrm{M}}) \times_1 \mathbf{z}^i \times \Sigma^{-1/2}\|^2 \leq ru(\mathrm{M}^*) + ru(-\mathrm{M}^*).$$

This completes the proof. □

Proof of Lemma 3.2. This lemma is a specification of Theorem 2.2.1 to generalized linear tensor regression. The proof consists of three steps. First, we obtain the explicit form of the empirical and population version of the Kullback-Leibler loss, $\widehat{d}(\mathrm{M})$ and $d(\mathrm{M})$. Second, we derive the gradient of $d(\mathrm{M}) - \widehat{d}(\mathrm{M})$ at $\widehat{\mathrm{M}}$, denoted as $\nabla(d - \widehat{d})_{\widehat{\mathrm{M}}}$. At last, we apply

Theorem 2.2.1 with the derived explicit forms of $d(\widehat{\mathrm{M}})$ and $\nabla(d-\widehat{d})_{\widehat{\mathrm{M}}}$ to conclude the proof.

Step 1. We first derive the explicit form of the empirical and population version of the Kullback-Leibler loss. Plugging the conditional density of $y^i \mid \mathbf{z}^i$ into the empirical Kullback-Leibler divergence yields

$$\widehat{d}(\mathrm{M}) = \sum_{i=1}^{n} \big(\frac{y^i \langle \mathrm{M}^*, \mathbf{z}^i \rangle - b(\langle \mathrm{M}^*, \mathbf{z}^i \rangle)}{\alpha} + c(y^i, \alpha) - \frac{y^i \langle \mathrm{M}, \mathbf{z}^i \rangle - b(\langle \mathrm{M}, \mathbf{z}^i \rangle)}{\alpha} - c(y^i, \alpha) \big).$$

Canceling the term $c(y^i, \alpha) - c(y^i, \alpha)$ and reorganizing the terms on the right-hand side yields

$$\widehat{d}(\mathrm{M}) = \frac{1}{\alpha} \sum_{i=1}^{n} \big(y^i \langle \mathrm{M}^* - \mathrm{M}, \mathbf{z}^i \rangle - b(\langle \mathrm{M}^*, \mathbf{z}^i \rangle) + b(\langle \mathrm{M}, \mathbf{z}^i \rangle) \big).$$

Taking the expectation of $\widehat{d}(\mathrm{M})$ with respect to y^i conditioning on $\mathbf{z}^i$, we obtain the conditional Kullback-Leibler divergence in the form of

$$d(\mathrm{M}) = \frac{1}{\alpha} \sum_{i=1}^{n} \mathbb{E}_{\mathrm{M}^*} \big(y^i \langle \mathrm{M}^* - \mathrm{M}, \mathbf{z}^i \rangle - b(\langle \mathrm{M}^*, \mathbf{z}^i \rangle) + b(\langle \mathrm{M}, \mathbf{z}^i \rangle) \big),$$

where $\mathbb{E}_{\mathrm{M}^*}$ denotes the conditional expectation conditioning on $\mathbf{z}^i$. As $\mathbb{E}_{\mathrm{M}^*}(y^i) = g^{-1}(\langle \mathrm{M}^*, \mathbf{z}^i \rangle)$, we find that the conditional Kullback-Leibler divergence of the generalized linear tensor regression has the explicit form of

$$d(\mathrm{M}) = \frac{1}{\alpha} \sum_{i=1}^{n} \big(g^{-1}(\langle \mathrm{M}^*, \mathbf{z}^i \rangle) \cdot \langle \mathrm{M}^* - \mathrm{M}, \mathbf{z}^i \rangle - b(\langle \mathrm{M}^*, \mathbf{z}^i \rangle) + b(\langle \mathrm{M}, \mathbf{z}^i \rangle) \big).$$

Step 2. We derive the explicit form of $\nabla(d-\widehat{d})_{\widehat{\mathrm{M}}}$. With $d(\mathrm{M})$ and $\widehat{d}(\mathrm{M})$ derived in Step 1,

we obtain

$$d(\widehat{\mathrm{M}}) - \widehat{d}(\widehat{\mathrm{M}}) = \frac{1}{\alpha}\sum_{i=1}^{n}\big(g^{-1}(\langle \mathrm{M}^*,\, \mathbf{z}^i\rangle)\cdot\langle \mathrm{M}^* - \widehat{\mathrm{M}},\, \mathbf{z}^i\rangle - b(\langle \mathrm{M}^*,\, \mathbf{z}^i\rangle) + b(\langle \widehat{\mathrm{M}},\, \mathbf{z}^i\rangle) - y^i\langle \mathrm{M}^* - \widehat{\mathrm{M}},\, \mathbf{z}^i\rangle + b(\langle \mathrm{M}^*,\, \mathbf{z}^i\rangle) - b(\langle \widehat{\mathrm{M}},\, \mathbf{z}^i\rangle)\big).$$

Canceling the terms involving $b(\cdot)$ in the above equality yields

$$\begin{aligned} d(\widehat{\mathrm{M}}) - \widehat{d}(\widehat{\mathrm{M}}) &= \frac{1}{\alpha}\sum_{i=1}^{n}\big(g^{-1}(\langle \mathrm{M}^*,\, \mathbf{z}^i\rangle)\cdot\langle \mathrm{M}^* - \widehat{\mathrm{M}},\, \mathbf{z}^i\rangle - y^i\langle \mathrm{M}^* - \widehat{\mathrm{M}},\, \mathbf{z}^i\rangle\big) \\ &= \frac{1}{\alpha}\sum_{i=1}^{n}\big(g^{-1}(\langle \mathrm{M}^*,\, \mathbf{z}^i\rangle) - y^i\big)\langle \mathrm{M}^* - \widehat{\mathrm{M}},\, \mathbf{z}^i\rangle. \end{aligned}$$

Expanding the tensor inner product by definition, we find that the gradient $\nabla(d - \widehat{d})_{\widehat{\mathrm{M}}}$ has the explicit form

$$\begin{aligned} \nabla(d - \widehat{d})_{\widehat{\mathrm{M}}} &= \frac{1}{\alpha}\sum_{i=1}^{n}\big(g^{-1}(\langle \mathrm{M}^*,\, \mathbf{z}^i\rangle) - y^i\big)(-\mathbf{z}^i) \\ &= \frac{1}{\alpha}\sum_{i=1}^{n}\big(y^i - g^{-1}(\langle \mathrm{M}^*,\, \mathbf{z}^i\rangle)\big)\mathbf{z}^i. \end{aligned}$$

Step 3. We apply Theorem 2.2.1 with the explicit form of $d(\widehat{\mathrm{M}})$ and $\nabla(d - \widehat{d})_{\widehat{\mathrm{M}}}$ and conclude that for all $r \geq \tilde{u}\big(\frac{1}{\alpha}\sum_{i=1}^{n}\big(y^i - g^{-1}(\langle \mathrm{M}^*,\, \mathbf{z}^i\rangle)\big)\mathbf{z}^i\big)$, it holds that

$$d(\widehat{\mathrm{M}}) \leq ru(\mathrm{M}^*) + ru(-\mathrm{M}^*)$$

with Kullback-Leibler loss $d(\widehat{\mathrm{M}}) = \frac{1}{\alpha}\sum_{i=1}^{n}\big(g^{-1}(\langle \mathrm{M}^*,\, \mathbf{z}^i\rangle)\cdot\langle \mathrm{M}^* - \widehat{\mathrm{M}},\, \mathbf{z}^i\rangle - b(\langle \mathrm{M}^*,\, \mathbf{z}^i\rangle) + b(\langle \widehat{\mathrm{M}},\, \mathbf{z}^i\rangle)\big)$. To convey the idea that the choice of the regularization parameter is in the form of noise, we write the lower bound of the regularization parameter as $\tilde{u}\big(\frac{1}{\alpha}\sum_{i=1}^{n}\big(y^i - \mathbb{E}_{\mathrm{M}^*}(y^i)\big)\mathbf{z}^i\big)$. □

Proof of Lemma 3.3. The proof consists of three steps. First, we obtain the explicit form of

$\widehat{d}(\mathrm{M})$ and $d(\mathrm{M})$ in exponential trace models. Second, we derive the gradient of $d(\mathrm{M}) - \widehat{d}(\mathrm{M})$ at $\widehat{\mathrm{M}}$. At last, we apply Theorem 2.2.1 with the derived explicit forms.

Step 1. In the exponential trace model, it holds that

$$\sum_{i=1}^{n} \log\big(f_{\mathrm{M}}(X^i)\big) = -\sum_{i=1}^{n} \langle \mathrm{M},\, T(X^i) \rangle - na(\mathrm{M}).$$

The definition of inner product implies that $\langle \mathrm{M},\, T(X^i) \rangle = \langle T(X^i),\, \mathrm{M} \rangle$. Therefore, we write

$$\sum_{i=1}^{n} \log\big(f_{\mathrm{M}}(X^i)\big) = -\sum_{i=1}^{n} \langle T(X^i),\, \mathrm{M} \rangle - na(\mathrm{M}).$$

Plugging the log-likelihood into the empirical Kullback-Leibler loss yields

$$\widehat{d}(\mathrm{M}) = -\sum_{i=1}^{n} \langle T(X^i),\, \mathrm{M}^* \rangle - na(\mathrm{M}^*) + \sum_{i=1}^{n} \langle T(X^i),\, \mathrm{M} \rangle + na(\mathrm{M}).$$

Since inner product $\langle \cdot,\, \cdot \rangle$ is linear, we write

$$\widehat{d}(\mathrm{M}) = \langle \sum_{i=1}^{n} T(X^i),\, \mathrm{M} - \mathrm{M}^* \rangle - na(\mathrm{M}^*) + na(\mathrm{M}).$$

The population loss $d(\mathrm{M})$ is the expectation of the empirical loss $\widehat{d}(\mathrm{M})$. We thus obtain

$$d(\mathrm{M}) = \mathbb{E}_{\mathrm{M}^*}\big(\langle \sum_{i=1}^{n} T(X^i),\, \mathrm{M} - \mathrm{M}^* \rangle - na(\mathrm{M}^*) + na(\mathrm{M}) \big).$$

Again, $\langle \cdot,\, \cdot \rangle$ is linear, and we find

$$d(\mathrm{M}) = \langle \sum_{i=1}^{n} \mathbb{E}_{\mathrm{M}^*} T(X^i),\, \mathrm{M} - \mathrm{M}^* \rangle - na(\mathrm{M}^*) + na(\mathrm{M}).$$

Step 2. We derive the explicit form of $\nabla(d - \widehat{d})_{\widehat{\mathrm{M}}}$ in the exponential trace model. With

the forms of $d(\mathrm{M})$ and $\widehat{d}(\mathrm{M})$ derived in Step 1, we obtain

$$d(\widehat{\mathrm{M}})-\widehat{d}(\widehat{\mathrm{M}}) = \big(\langle\sum_{i=1}^{n}\mathbb{E}_{\mathrm{M}^*}T(X^i),\, \widehat{\mathrm{M}}-\mathrm{M}^*\rangle-na(\mathrm{M}^*)+na(\widehat{\mathrm{M}})\big)-\big(\langle\sum_{i=1}^{n}T(X^i),\, \widehat{\mathrm{M}}-\mathrm{M}^*\rangle-na(\mathrm{M}^*)+na(\widehat{\mathrm{M}})\big).$$

Applying the linearity of inner product and canceling $-na(\mathrm{M}^*)+na(\widehat{\mathrm{M}})+na(\mathrm{M}^*)-na(\widehat{\mathrm{M}})$, we obtain

$$d(\mathrm{M})-\widehat{d}(\mathrm{M}) = \langle\sum_{i=1}^{n}\big(\mathbb{E}_{\mathrm{M}^*}T(X^i)-T(X^i)\big),\, \widehat{\mathrm{M}}-\mathrm{M}^*\rangle.$$

The definition of inner product implies that

$$\nabla(d-\widehat{d})_{\widehat{\mathrm{M}}} = \sum_{i=1}^{n}\big(\mathbb{E}_{\mathrm{M}^*}T(X^i)-T(X^i)\big).$$

Step 3. Plugging the explicit form of $\nabla(d-\widehat{d})_{\widehat{\mathrm{M}}}$ into the lower bound for the regularization parameter in Theorem 2.2.1, we conclude for all $r \geq \tilde{u}\big(\sum_{i=1}^{n}\big(\mathbb{E}_{\mathrm{M}^*}T(X^i)-T(X^i)\big)\big)$ the inequality

$$d(\widehat{\mathrm{M}}) \leq ru(\mathrm{M}^*)+ru(-\mathrm{M}^*)$$

with Kullback-Leibler loss $d(\widehat{\mathrm{M}}) = \langle\sum_{i=1}^{n}\mathbb{E}_{\mathrm{M}^*}T(X^i),\, \widehat{\mathrm{M}}-\mathrm{M}^*\rangle - na(\mathrm{M}^*)+na(\widehat{\mathrm{M}})$. □

2.5.3 Empirical Process Terms

Lemma 2.5.2 (control of the empirical term for tensor response regression)**.** *Let z_j^i be the jth entry of row vector $\mathbf{z}^i$. Suppose $\sum_{i=1}^{n}(z_j^i)^2/n = 1$, and $(\Sigma_k^{-1})_{i_k i_k} = h_k^2$, $h_k > 0$, for $k \in \{2,\ldots,p\}$, $i_k \in \{1,\ldots,b_k\}$. Consider the case with zero-mean array normal noise in Section 2.3.1 and $u(\mathrm{M}) := \sum_{i_1=1}^{b_1}\cdots\sum_{i_p=1}^{b_p}|\mathrm{M}_{i_1,\ldots,i_p}|$. For all $t > 0$ and $r_0 = \big(\prod_{k=2}^{p}h_k\big)\sqrt{2n\big(t^2+\log(\prod_{j=1}^{p}b_j)\big)}$, it holds that*

$$P\big(\tilde{u}\big(\sum_{i=1}^{n}\big(E^i\times\Sigma^{-1}\times_1(\mathbf{z}^i)^\top\big)\big)\leq r_0\big) \geq 1-2\exp(-t^2).$$

Proof of Lemma 2.5.2. The proof consists of two steps. First, we plug in $\tilde{u}(\cdot)$ and re-express the target probability in terms of random variables $\big(\sum_{i=1}^n \big(E^i \times \Sigma^{-1} \times_1 (\mathbf{z}^i)^\top\big)_{i_1,\ldots,i_p}$. Second, we apply the Chernoff inequality to bound the tail probabilities.

Step 1. For $\mathrm{M} \in \mathbb{R}^{b_1 \times \cdots \times b_p}$, the dual of the regularizer is of the form

$$\tilde{u}(\mathrm{M}) := \sup\big\{\langle \mathrm{M},\, \mathrm{M}'\rangle \mid \mathrm{M}' \in \mathbb{R}^{b_1 \times \cdots \times b_p}, u(\mathrm{M}') \le 1\big\} = \max_{i1,\ldots,i_p} |\mathrm{M}_{i_1,\ldots,i_p}|.$$

Therefore,

$$\begin{aligned}
&P\big(\tilde{u}\big(\sum_{i=1}^n \big(E^i \times \Sigma^{-1} \times_1 (\mathbf{z}^i)^\top\big)\big) \le r_0\big) \\
&= P\big(\max_{i_1,\ldots,i_p} |\big(\sum_{i=1}^n \big(E^i \times \Sigma^{-1} \times_1 (\mathbf{z}^i)^\top\big)_{i_1,\ldots,i_p}| \le r_0\big) \\
&= 1 - P\big(\max_{i_1,\ldots,i_p} |\big(\sum_{i=1}^n \big(E^i \times \Sigma^{-1} \times_1 (\mathbf{z}^i)^\top\big)_{i_1,\ldots,i_p}| > r_0\big) \\
&\ge 1 - \sum_{i_1=1}^{b_1} \cdots \sum_{i_p=1}^{b_p} P\big(|\big(\sum_{i=1}^n \big(E^i \times \Sigma^{-1} \times_1 (\mathbf{z}^i)^\top\big)_{i_1,\ldots,i_p}| > r_0\big).
\end{aligned}$$

Step 2. We now work on the distribution of $\big(\sum_{i=1}^n \big(E^i \times \Sigma^{-1} \times_1 (\mathbf{z}^i)^\top\big)_{i_1,\ldots,i_p}$ to bound its tail probability. We first observe that

$$\big(\sum_{i=1}^n \big(E^i \times \Sigma^{-1} \times_1 (\mathbf{z}^i)^\top\big)_{i_1,\ldots,i_p} = \big(\sum_{i=1}^n \big(E^i \times \Sigma^{-1/2} \times (\Sigma^{-1/2})^\top \times_1 (\mathbf{z}^i)^\top\big)_{i_1,\ldots,i_p}.$$

The definition of the array normal distribution in (Hoff, 2011) shows that

$$E^i = N^i \times A,$$

where N^i is an array of independent standard normal entries in $\mathbb{R}^{b_2 \times \cdots \times b_p}$. Then,

$$E^i \times \Sigma^{-1/2} = (N^i \times A) \times \Sigma^{-1/2}.$$

Expanding the Tucker product in the above equation yields

$$E^i \times \Sigma^{-1/2} = (N^i \times_1 A_2 \times_2 \ldots \times_{p-1} A_p) \times_1 A_2^{-1} \times_2 \cdots \times_{p-1} A_p^{-1}.$$

Further, we apply the properties of mode products on the right-hand side and find

$$E^i \times \Sigma^{-1/2} = N^i \times_1 (A_2^{-1} A_2) \times_2 \ldots \times_{p-1} (A_p^{-1} A_p) = N^i \times_1 I^{(2)} \times_2 \ldots \times_{p-1} I^{(p)},$$

where $I^{(k)}$ is the identity matrix of dimension $b_k \times b_k$ for $k \in \{2, \ldots, p\}$. Hence,

$$E^i \times \Sigma^{-1/2} = N^i.$$

The above equality can be shown by writing out the element-wise form of $E^i \times \Sigma^{-1/2}$ as

$$\begin{aligned}
(E^i \times \Sigma^{-1/2})_{i_2,\ldots,i_p} &= \left(N^i \times_1 I^{(2)} \times_2 \ldots \times_{p-1} I^{(p)}\right)_{i_2,\ldots,i_p} \\
&= \sum_{j_2}^{b_2} \cdots \sum_{j_p}^{b_p} N^i_{j_2,\ldots,j_p} I^{(2)}_{i_2 j_2} \ldots I^{(p)}_{i_p j_p} \\
&= N^i_{i_2,\ldots,i_p}.
\end{aligned}$$

Plugging the equality between $E^i \times \Sigma^{-1/2}$ and N^i into $\left(\sum_{i=1}^n \left(E^i \times \Sigma^{-1/2} \times (\Sigma^{-1/2})^\top \times_1\right.\right.$

$(\mathbf{z}^i)^\top\big)_{i_1,\ldots,i_p}$ shows

$$\begin{aligned}\big(\sum_{i=1}^n \big(E^i \times \Sigma^{-1} \times_1 (\mathbf{z}^i)^\top\big)_{i_1,\ldots,i_p} &= \big(\sum_{i=1}^n \big(N^i \times_1 (A_2^{-1})^\top \times_2 \ldots \times_{p-1} (A_p^{-1})^\top \times_1 (\mathbf{z}^i)^\top\big)_{i_1,\ldots,i_p} \\ &= \sum_{i=1}^n \sum_{j_2=1}^{b_2} \cdots \sum_{j_p=1}^{b_p} N^i_{j_2,\ldots,j_p} (A_2^{-1})^\top_{i_2 j_2} \ldots (A_p^{-1})^\top_{i_p j_p} z^i_{i_1}.\end{aligned}$$

Since N^i is an array of independent standard normal entries, we find

$$\big(\sum_{i=1}^n \big(E^i \times \Sigma^{-1} \times_1 (\mathbf{z}^i)^\top\big)_{i_1,\ldots,i_p} \sim \mathcal{N}(0, \sum_{i=1}^n \sum_{j_2=1}^{b_2} \cdots \sum_{j_p=1}^{b_p} (z^i_{i_1} (A_2^{-1})^\top_{i_2 j_2} \ldots (A_p^{-1})^\top_{i_p j_p})^2),$$

where the variance can be re-written as

$$\big(\sum_{i=1}^n (z^i_{i_1})^2\big)\big(\sum_{j_2=1}^{b_2} (A_2^{-1})^2_{j_2 i_2}\big) \ldots \big(\sum_{j_p=1}^{b_p} (A_p^{-1})^2_{j_p i_p}\big).$$

Observe that $\sum_{j_k=1}^{b_k} (A_k^{-1})^2_{j_k i_k} = \big((A_k^\top)^{-1}(A_k)^{-1}\big)_{i_k i_k} = (\Sigma_k^{-1})_{i_k i_k} = h_k^2$, and $\sum_{i=1}^n (z^i_j)^2/n = 1$, we show

$$\big(\sum_{i=1}^n \big(E^i \times \Sigma^{-1} \times_1 (\mathbf{z}^i)^\top\big)_{i_1,\ldots,i_p} \sim \mathcal{N}(0, n \prod_{k=2}^p h_k^2).$$

We now apply the Chernoff bound for the normal distribution and control the stochastic term to find

$$\begin{aligned}&\sum_{i_1=1}^{b_1} \cdots \sum_{i_p=1}^{b_p} P\big(|\big(\sum_{i=1}^n \big(E^i \times \Sigma^{-1} \times_1 (\mathbf{z}^i)^\top\big)_{i_1,\ldots,i_p}| > r_0\big) \\ &= 2 \prod_{m=1}^p b_m \cdot P\big(\big(\sum_{i=1}^n \big(E^i \times \Sigma^{-1} \times_1 (\mathbf{z}^i)^\top\big)_{i_1,\ldots,i_p} > r_0\big) \\ &\leq 2 \prod_{m=1}^p b_m \cdot \exp(-\frac{r_0^2}{2n \prod_{k=2}^p h_k^2})).\end{aligned}$$

Plugging in r_0 yields

$$\begin{aligned}
&\sum_{i_1=1}^{b_1}\cdots\sum_{i_p=1}^{b_p} P\big(|\big(\sum_{i=1}^{n}\big(E^i\times\Sigma^{-1}\times_1(\mathbf{z}^i)^\top\big)_{i_1,\ldots,i_p}|>r_0\big)\\
&\le 2\prod_{m=1}^{p} b_m\cdot\exp(-\frac{\prod_{k=2}^{p}h_k^2\cdot 2n(t^2+\log(\prod_{j=1}^{p}b_j)}{2n\prod_{k=2}^{p}h_k^2})\\
&= 2\prod_{m=1}^{p} b_m\cdot\exp(-t^2-\log(\prod_{j=1}^{p}b_j))\\
&= 2\exp(-t^2).
\end{aligned}$$

We conclude that $P(\tilde{u}\big(\sum_{i=1}^{n}\big(E^i\times\Sigma^{-1}\times_1(\mathbf{z}^i)^\top\big)\big)\le r_0)\ge 1-2\exp(-t^2)$ as desired. □

Lemma 2.5.3 (control of the empirical term for logistic regression)**.** *Let z_j^i be the jth entry of vector $\mathbf{z}^i$. Suppose $\sum_{i=1}^{n}(z_j^i)^2/n=1$ and $u(\mathrm{M}):=\sum_{j=1}^{b_1}|\mathrm{M}_j|$. For all $t>0$ and $r_0=\sqrt{(1+2\max_i\{p^i(1-p^i)\})/3\cdot n(t^2+\log b_1)}$, where $p^i:=\mathbb{E}_{\mathrm{M}^*}y^i=e^{\langle\mathrm{M}^*,\mathbf{z}^i\rangle}/\big(1+e^{\langle\mathrm{M}^*,\mathbf{z}^i\rangle}\big)$, it holds that*

$$P\big(\tilde{u}\big(\sum_{i=1}^{n}\big(y^i-p^i\big)\mathbf{z}^i\big)\le r_0\big)\ge 1-2\exp(-t^2).$$

Proof of Lemma 2.5.3. The proof consists of two steps. First, we plug in $\tilde{u}(\cdot)$ and re-express the target probability in terms of random variables $\sum_{i=1}^{n}\big(y^i-p^i\big)z_j^i$. Second, we apply the improved Hoeffding's inequality (Bercu et al., 2015, Theorem 2.47) to bound the tail probabilities.

Step 1. For $\mathrm{M}\in\mathbb{R}^{b_1}$, the dual of the regularizer is of the form

$$\tilde{u}(\mathrm{M}):=\sup\big\{\langle\mathrm{M},\,\mathrm{M}'\rangle\mid\mathrm{M}'\in\mathbb{R}^{b_1},u(\mathrm{M}')\le 1\big\}=\max_{j\in\{1,\ldots,b_1\}}|\mathrm{M}_j|.$$

Therefore,

$$P\big(\tilde{u}\big(\sum_{i=1}^{n}\big(y^i-p^i\big)\mathbf{z}^i\big)\le r_0\big) = P\big(\max_{j\in\{1,\ldots,b_1\}}\big|\sum_{i=1}^{n}\big(y^i-p^i\big)z_j^i\big|\le r_0\big)$$
$$= 1-P\big(\max_{j\in\{1,\ldots,b_1\}}\big|\sum_{i=1}^{n}\big(y^i-p^i\big)z_j^i\big|> r_0\big)$$
$$\ge 1-\sum_{j=1}^{b_1}P\big(\big|\sum_{i=1}^{n}\big(y^i-p^i\big)z_j^i\big|> r_0\big).$$

Step 2. We now bound the tail probability of $\sum_{i=1}^{n}\big(y^i-p^i\big)z_j^i$. Let $U_i := y^i z_j^i$ and $S_n := \sum_{i=1}^{n} U_i$, we observe that

$$\sum_{i=1}^{n}\big(y^i-p^i\big)z_j^i = S_n - \mathbb{E}_{\mathrm{M}^*}S_n,$$

where $\mathbb{E}_{\mathrm{M}^*}S_n$ is the conditional expectation given z_j^i. The random variable U_i is bounded, specifically, $\min\{0,z_j^i\}\le U_i\le\max\{0,z_j^i\}$. Applying the improved Hoeffding's inequality (Bercu et al., 2015, Theorem 2.47) yields

$$P\big(\sum_{i=1}^{n}\big(y^i-p^i\big)z_j^i> r_0\big) = P\big(S_n-\mathbb{E}_{\mathrm{M}^*}S_n> r_0\big)\le\exp\big(-\frac{3r_0^2}{D_n+2V_n}\big),$$

where $D_n=\sum_{i=1}^{n}\big(\max\{0,z_j^i\}-\min\{0,z_j^i\}\big)^2=\sum_{i=1}^{n}(z_j^i)^2=n$, and $V_n=\mathbb{E}_{\mathrm{M}^*}(S_n-\mathbb{E}_{\mathrm{M}^*}S_n)^2=\sum_{i=1}^{n}(z_j^i)^2p^i(1-p^i)$. Similarly, we find

$$P\big(\sum_{i=1}^{n}\big(y^i-p^i\big)z_j^i<-r_0\big) = P\big(-S_n-\mathbb{E}_{\mathrm{M}^*}(-S_n)> r_0\big)\le\exp\big(-\frac{3r_0^2}{D_n+2V_n}\big).$$

We are left with bounding the tail probabilities. To this end, we observe that

$$\begin{aligned}\sum_{j=1}^{b_1} P(|\sum_{i=1}^{n}\big(y^i-p^i\big)z_j^i|>r_0) &\le 2b_1\exp\big(-\frac{3r_0^2}{D_n+2V_n}\big)\\ &= 2b_1\exp\big(-\frac{3r_0^2}{n+2\sum_{i=1}^n(z_j^i)^2p^i(1-p^i)}\big)\\ &\le 2b_1\exp\big(-\frac{3r_0^2}{n+2n\max_i\{p^i(1-p^i)\}}\big).\end{aligned}$$

Plugging in r_0 yields

$$\begin{aligned}&\sum_{j=1}^{b_1} P(|\sum_{i=1}^{n}\big(y^i-p^i\big)z_j^i|>r_0)\\ &\le 2b_1\exp\big(-\frac{(1+2\max_i\{p^i(1-p^i)\})\cdot n(t^2+\log b_1)}{n+2n\max_i\{p^i(1-p^i)\}}\big)\\ &= 2b_1\exp\big(-t^2-\log b_1\big)\\ &= 2\exp\big(-t^2\big).\end{aligned}$$

We conclude that $P\big(\tilde{u}\big(\sum_{i=1}^n\big(y^i-p^i\big)\mathbf{z}^i\big)\le r_0\big)\ge 1-2\exp(-t^2)$ as desired. □

Lemma 2.5.4 (control of the empirical term for gaussian graphical model)**.** *Consider* $u(\mathrm{M}):=\sum_{i_1=1}^p\sum_{i_2=1}^p|\mathrm{M}_{i_1i_2}|$ *and* $X^i\sim\mathcal{N}(0,(\mathrm{M}^*)^{-1})$. *For* $0<t<\sqrt{n/4-\log\big(p(p-1)\big)}$, *and* $r_0=80\max_k\big((\mathrm{M}^*)^{-1}\big)_{kk}\sqrt{n\big(t^2+\log\big(p(p-1)\big)\big)}$, *it holds that*

$$P\big(\tilde{u}\big(\sum_{i=1}^n\big((\mathrm{M}^*)^{-1}-X^i(X^i)^\top\big)\big)\le r_0\big)\ge 1-4\exp(-t^2).$$

Proof of Lemma 2.5.4. The proof consists of two steps. First, we plug in $\tilde{u}(\cdot)$ and re-express the target probability in terms of random variables $\big(\widehat{\mathrm{M}}^{-1}-(\mathrm{M}^*)^{-1}\big)_{i_1i_2}$, where $\widehat{\mathrm{M}}^{-1}:=\sum_{i=1}^n X^i(X^i)^\top/n$ is the sample covariance matrix. Second, we apply the concentration result for sample covariance matrix (Ravikumar et al., 2011, Lemma 1) to bound the tail probabilities.

Step 1. For $\mathrm{M} \in \mathbb{R}^{p\times p}$, the dual of the regularizer is of the form

$$\tilde{u}(\mathrm{M}) := \sup\big\{\langle \mathrm{M},\, \mathrm{M}'\rangle \mid \mathrm{M}' \in \mathbb{R}^{p\times p}, u(\mathrm{M}') \le 1\big\} = \max_{i_1,i_2} |\mathrm{M}_{i_1 i_2}|.$$

Therefore,

$$\begin{aligned}
&P\big(\tilde{u}\big(\sum_{i=1}^{n}\big((\mathrm{M}^*)^{-1} - X^i(X^i)^\top\big)\big) \le r_0\big)\\
&= P\big(\max_{i_1,i_2} |n\big((\mathrm{M}^*)^{-1} - \frac{1}{n}\sum_{i=1}^{n} X^i(X^i)^\top\big)_{i_1 i_2}| \le r_0\big)\\
&= P\big(\max_{i_1,i_2} |\big((\mathrm{M}^*)^{-1} - \frac{1}{n}\sum_{i=1}^{n} X^i(X^i)^\top\big)_{i_1 i_2}| \le \frac{r_0}{n}\big)\\
&= 1 - P\big(\max_{i_1,i_2} |\big((\mathrm{M}^*)^{-1} - \frac{1}{n}\sum_{i=1}^{n} X^i(X^i)^\top\big)_{i_1 i_2}| > \frac{r_0}{n}\big)\\
&\ge 1 - \sum_{i_1=1}^{p}\sum_{i_2=1, i_2\neq i_1}^{p} P(|(\mathrm{M}^*)^{-1} - \frac{1}{n}\sum_{i=1}^{n} X^i(X^i)^\top\big)_{i_1 i_2}| > \frac{r_0}{n}).
\end{aligned}$$

Step 2. (Ravikumar et al., 2011, Lemma 1) gives the tail bound of sample covariance matrix as

$$P(|\big(\widehat{\mathrm{M}}^{-1} - (\mathrm{M}^*)^{-1}\big)_{i_1 i_2}| > \delta) \le 4\exp\big(-\frac{n\delta^2}{80^2 \max_k \big((\mathrm{M}^*)^{-1}\big)^2_{kk}}\big),$$

for all $\delta \in (0, 40\max_k \big((\mathrm{M}^*)^{-1}\big)_{kk})$. We are now ready to work on the desired tail bound.

$$\sum_{i_1=1}^{p}\sum_{i_2=1, i_2\neq i_1}^{p} P(|(\mathrm{M}^*)^{-1} - \frac{1}{n}\sum_{i=1}^{n} X^i(X^i)^\top\big)_{i_1 i_2}| > \frac{r_0}{n}) \le 4p(p-1)\exp(-\frac{n(r_0/n)^2}{80^2 \max_k \big((\mathrm{M}^*)^{-1}\big)^2_{kk}}).$$

Plugging in r_0 yields

$$\begin{aligned}
&\sum_{i_1=1}^{p}\sum_{i_2=1,i_2\neq i_1}^{p} P(|(\mathrm{M}^*)^{-1} - \frac{1}{n}\sum_{i=1}^{n} X^i(X^i)^\top\big)_{i_1,i_2}| > \frac{r_0}{n}) \\
&\leq 4p(p-1)\exp\big(-\frac{n\big(80\max_k\big((\mathrm{M}^*)^{-1}\big)_{kk}\sqrt{n(t^2+\log\big(p(p-1)\big))}/n\big)^2}{80^2\max_k\big((\mathrm{M}^*)^{-1}\big)_{kk}^2}\big) \\
&= 1 - p(p-1)\cdot 4\exp\big(-t^2-\log\big(p(p-1)\big)\big) \\
&= 1 - 4\exp\big(-t^2\big)
\end{aligned}$$

for all $0 < r_0/n < 40\max_k\big((\mathrm{M}^*)^{-1}\big)_{kk}$. It can now be readily shown that the desired bound holds for all $0 < t < \sqrt{n/4 - \log\big(p(p-1)\big)}$. □

2.5.4 Notation and Properties of Tensor Operations

We follow the notation for tensor and tensor operations in (Kolda, 2006; Kolda and Bader, 2009). A tensor $\mathcal{T} \in \mathbb{R}^{b_1\times\cdots\times b_p}$ can simply be seen as a multi-dimensional array $(\mathcal{T}_{i_1,\ldots,i_p} : i_k \in \{1,\ldots b_k\}; k \in \{1,\ldots,p\})$. The mode-$k$ fibers of $\mathcal{T}$ are vectors obtained by fixing all indices except the kth one; for example, $\mathcal{T}_{i_1,\ldots,i_{k-1},:,i_{k+1},\ldots,i_p} \in \mathbb{R}^{b_k}$. The kth mode matricization of $\mathcal{T}$ is the matrix having the mode-k fibers of $\mathcal{T}$ as columns and is represented by $T_{(k)} \in \mathbb{R}^{b_k\times\left(b_1\ldots b_{k-1}b_{k+1}\ldots b_p\right)}$. The mode-$k$ product of tensor $\mathcal{T}$ and a matrix $C \in \mathbb{R}^{m\times b_k}$ is a tensor defined by $\mathcal{Y} = \mathcal{T}\times_k C \in \mathbb{R}^{b_1\times\cdots\times b_{k-1}\times m\times b_{k+1}\times\cdots\times b_p}$. The resulting array $\mathcal{Y}$ is from the inversion of the kth mode matricization operation on the matrix $CT_{(k)}$, that is, $Y_{(k)} = CT_{(k)}$. The entries of $\mathcal{Y}$ are given by

$$\big(\mathcal{T}\times_k C\big)_{i_1,\ldots,i_{k-1},j,i_{k+1},\ldots,i_p} = \sum_{i_k=1}^{b_k} T_{i_1,\ldots,i_{k-1},i_k,i_{k+1},\ldots,i_p} c_{ji_k}$$

for $j \in \{1,\ldots,m\}, k \in \{2,\ldots,p-1\}, i_l \in \{1,\ldots,b_l\}, l \in \{1,\ldots,k-1,k+1,\ldots,p\}$ (the case of $k \in \{1,p\}$ is similar). The Tucker product is an extension of the mode-k product, which is the product of a tensor $\mathcal{T}$ and a list of matrices $E = \{E_1,\ldots,E_p\}$ in which $E_k \in \mathbb{R}^{m_k\times b_k}$.

The (i,j)th element of E_k is denoted by $e_{ij}^{(k)}$. The Tucker product is given by

$$\mathcal{T} \times E = \mathcal{T} \times_1 E_1 \times_2 \ldots \times_p E_p,$$

or, elementwise,

$$\big(\mathcal{T} \times E\big)_{j_1,\ldots,j_p} = \sum_{i_1=1}^{b_1} \cdots \sum_{i_p=1}^{b_p} T_{i_1,\ldots,i_p} e_{j_1 i_1}^{(1)} \ldots e_{j_p i_p}^{(p)}$$

for $j_k \in \{1,\ldots,m_k\}, k \in \{1,\ldots,p\}$. The inner product of two same-sized tensors $\mathcal{W}, \mathcal{V} \in \mathbb{R}^{b_1 \times \cdots \times b_p}$ is the sum of the products of same-index elements, that is,

$$\langle \mathcal{W}, \mathcal{V} \rangle = \sum_{i_1=1}^{b_1} \cdots \sum_{i_p=1}^{b_p} w_{i_1,\ldots,i_p} v_{i_1,\ldots,i_p}.$$

The array norm of tensor $\mathcal{T}$ is the inner product of itself and given by $\|\mathcal{T}\|^2 = \langle \mathcal{T}, \mathcal{T} \rangle = \sum_{i_1} \cdots \sum_{i_p} t_{i_1,\ldots,i_p}^2$.

Lemma 2.5.5. *Let tensors $\mathcal{W}, \mathcal{V} \in \mathbb{R}^{b_1 \times \cdots \times b_p}$, list of matrices $E = \{E_1, \ldots, E_p\}$ in which $E_k \in \mathbb{R}^{m_k \times b_k}, k \in \{1,\ldots,p\}$, it holds that*

$$\big(\mathcal{W} + \mathcal{V}\big) \times E = \mathcal{W} \times E + \mathcal{V} \times E.$$

Proof of Lemma 2.5.5. This lemma can be proved by writing out the Tucker product on the

left-hand side element-wise. For $k \in \{1, \dots p\}$, and $j_k \in \{1, \dots, m_k\}$,

$$\begin{aligned}
\Big(\big(\mathcal{W} + \mathcal{V}\big) \times E \Big)_{j_1,\dots,j_p} =& \sum_{i_1=1}^{b_1} \cdots \sum_{i_p=1}^{b_p} \big(w + v\big)_{i_1,\dots,i_p} e^{(1)}_{j_1 i_1} \dots e^{(p)}_{j_p i_p} \\
=& \big(\sum_{i_1=1}^{b_1} \cdots \sum_{i_p=1}^{b_p} w_{i_1,\dots,i_p} e^{(1)}_{j_1 i_1} \dots e^{(p)}_{j_p i_p} \big) + \\
& \big(\sum_{i_1=1}^{b_1} \cdots \sum_{i_p=1}^{b_p} v_{i_1,\dots,i_p} e^{(1)}_{j_1 i_1} \dots e^{(p)}_{j_p i_p} \big) \\
=& \big(\mathcal{W} \times E\big)_{j_1,\dots,j_p} + \big(\mathcal{V} \times E\big)_{j_1,\dots,j_p}.
\end{aligned}$$

This is equivalent to the relationship $\big(\mathcal{W} + \mathcal{V}\big) \times E = \mathcal{W} \times E + \mathcal{V} \times E$. □

Lemma 2.5.6. *For two same-sized tensors $\mathcal{W}, \mathcal{V} \in \mathbb{R}^{b_1 \times \cdots \times b_p}$, it holds that*

$$\|\mathcal{W} + \mathcal{V}\|^2 = \|\mathcal{W}\|^2 + \|\mathcal{V}\|^2 + 2\langle \mathcal{W}, \mathcal{V} \rangle.$$

Proof of Lemma 2.5.6. By the definition of the array norm,

$$\begin{aligned}
\|\mathcal{W} + \mathcal{V}\|^2 =& \sum_{i_1=1}^{b_1} \cdots \sum_{i_p=1}^{b_p} \big(\mathcal{W} + \mathcal{V}\big)^2_{i_1,\dots,i_p} \\
=& \sum_{i_1=1}^{b_1} \cdots \sum_{i_p=1}^{b_p} \big(w_{i_1,\dots,i_p} + v_{i_1,\dots,i_p}\big)^2 \\
=& \sum_{i_1=1}^{b_1} \cdots \sum_{i_p=1}^{b_p} w^2_{i_1,\dots,i_p} + \sum_{i_1=1}^{b_1} \cdots \sum_{i_p=1}^{b_p} v^2_{i_1,\dots,i_p} + 2 \sum_{i_1=1}^{b_1} \cdots \sum_{i_p=1}^{b_p} w_{i_1,\dots,i_p} v_{i_1,\dots,i_p},
\end{aligned}$$

that is, $\|\mathcal{W} + \mathcal{V}\|^2 = \|\mathcal{W}\|^2 + \|\mathcal{V}\|^2 + 2\langle \mathcal{W}, \mathcal{V} \rangle$. □

Chapter 3

GRAPHICAL MODELS FOR DISCRETE AND CONTINUOUS DATA

This chapter is a slightly revised version of Zhuang et al. (2016). Dr. Johannes Lederer initiated and constructed most of the framework set-up while I primarily focus on the implementation and application.

3.1 Introduction

Gaussian graphical models (Drton and Maathuis, 2017; Lauritzen, 1996b; Wainwright and Jordan, 2008) have been increasingly popular to describe the dependence structure of multivariate normal distributions. Suppose a p-dimensional random vector $X := (X_1, ..., X_p)^T \in \mathbb{R}^p$ follows a centered normal distribution $\mathcal{N}_p(0, \Sigma)$, where $\Sigma \in \mathbb{R}^{p \times p}$ is the symmetric and positive definite population covariance matrix. The density of X has a form of

$$f_\Sigma(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}\sqrt{|\Sigma|}}\, e^{-\mathbf{x}^\top \Sigma^{-1}\mathbf{x}/2} \tag{3.1}$$

with respect to the Lebesgue measure. The dependence among vector coordinates can be represented by an undirected graph $G(V, E)$. In the undirected graph, $V := \{1, ..., p\}$ is a set of vertices that contains p nodes corresponding to $X_1, ..., X_p$ and $E := \{(i, j) : i, j \in \{1, \ldots, p\}, i \neq j, \Sigma^{-1}_{ij} \neq 0\}$ is the set of edges that meet at these vertices. Two entries X_i, X_j, $i \neq j$, are conditionally independent given all other entries if and only if $(i, j) \notin E$. Thus, precision matrix $\mathbf{\Sigma}^{-1}$ encodes the dependence structure of X and is sufficient for recovering the Gaussian graphical model. A natural and straightforward estimator of Σ^{-1} is the maximum likelihood estimator. Given an independent and identically distributed sample of n ($n > p$) observations, the maximum likelihood estimator would be the inverse of the

sample covariance matrix.

However, the assumption of Gaussian distribution is restrictive. The normality does not always hold in practice, in which case estimation and inference based on Gaussian methods can be misleading (Xue and Zou, 2012). For example, data of interest may be highly skewed, restricted to specific ranges or count-valued. To relax the Gaussian assumption, researchers have explored in various directions. For example, recent developed methods include copula-based models (Gu et al., 2015; Liu et al., 2012, 2009; Xue and Zou, 2012), score matching approach (Lin et al., 2016; Yu et al., 2019), Ising models (Brush, 1967; Lenz, 1920), and multinomial extensions of the Ising models (Loh and Wainwright, 2013). These methods are typically derived on a case-by-case basis. Instead, we are interested in a broadly applicable framework that comprises different data types and ensures a rigid theoretical structure.

In this chapter, we review the exponential trace model framework (Zhuang et al., 2016) in Section 3.2 with a variety of examples. Further, we establish a sampling-based approximation algorithm for computing the maximum likelihood estimator in Section 3.3. We show simulation results for different types of data in Section 3.4 and apply to neural spike data in Section 3.5. At last, we conclude the chapter with a summary in Section 3.6. We refer readers to the archived paper (Zhuang et al., 2016) for detailed properties of exponential trace model framework.

Notation

For matrices $A, B \in \mathbb{R}^{s\times t}$, $s, t \in \{1, 2, \dots\}$, we represent the trace inner product by

$$\langle A,\, B\rangle_{\mathrm{tr}} := \mathrm{tr}\left(A^\top B\right) = \sum_{i=1}^{s}\sum_{j=1}^{t} A_{ij}B_{ij}.$$

We denote random vectors and their realizations by upper case letters such as X and arguments of functions by lower case, boldface letters such as $\mathbf{x}$. Given n i.i.d. data samples $X^1, \dots, X^n$, we represent the data by $\underline{X} := (X^1, \dots, X^n)$ and the corresponding function argument by $\underline{\mathbf{x}} := (\mathbf{x}^1, \dots, \mathbf{x}^n)$ for $\mathbf{x}^1, \dots, \mathbf{x}^n \in \mathcal{D}$. Given a set $S \subset \{1, \dots, p\}$, we denote

by $X_S \in \mathbb{R}^{|S|}$ the vector that consists of the coordinates of X with indices in S, and we set $X_{-S} := X_{S^c} \in \mathbb{R}^{p-|S|}$. Independence of two elements X_i and X_j, $i \neq j$, is denoted by $X_i \perp X_j$; conditional independence of X_i and X_j given all other elements is denoted by $X_i \perp X_j | X_{-\{i,j\}}$.

3.2 Framework

We first review the framework of exponential trace models and study graphical model examples.

3.2.1 Exponential Trace Models

Exponential trace models are probabilistic models for vector-valued observations when the dependence structure among vector coordinates is of interest. For non-empty finite or continuous domains $\mathcal{D} \subset \mathbb{R}^p$ and random vectors $X \in \mathcal{D}$, the model has densities in the form of

$$f_{\mathrm{M}}(\mathbf{x}) := e^{-\langle \mathrm{M}, T(\mathbf{x}) \rangle_{\mathrm{tr}} + \xi(\mathbf{x}) - a(\mathrm{M})} \tag{3.2}$$

with respect to some σ-finite measure ν on $\mathcal{D}$.

The model has four key components. First, the model parameter M encodes the conditional dependence structure of X. We define $\mathrm{M} \in \mathfrak{M}$, where

$$\mathfrak{M} \subset \mathfrak{M}^* := \mathrm{interior}\{\mathrm{M} \in \mathbb{R}^{p \times p} : a(\mathrm{M}) < \infty\}.$$

Second, a matrix-valued function T mapping from $\mathcal{D}$ to $\mathbb{R}^{p \times p}$ determines the sufficient statistics for the model. The introduction of function T yields the flexibility of the model and allows the integrability condition $a(\mathrm{M}) < \infty$ to be feasible. The setup avoids stringent constraints on the parameter space. At last, ξ is a read-valued function from $\mathcal{D}$ to $\mathbb{R}$ and $a(\mathrm{M})$

is the normalization defined as

$$a(\mathrm{M}) := \log \int_{\mathcal{D}} e^{-\langle \mathrm{M}, T(\mathbf{x})\rangle_{\mathrm{tr}} + \xi(\mathbf{x})}\, d\nu\,.$$

To ensure the parameter M is identifiable and has a compact and "full-dimensional" neighborhood in an affine subspace of $\mathbb{R}^{p\times p}$, we consider two mild technique assumptions on the parameter space: (1) the function $\mathrm{M} \mapsto f_{\mathrm{M}}$ of $\mathfrak{M}$ to the densities with respect to the measure ν is bijective; (2) $\mathfrak{M}$ is convex and that $\mathfrak{M}$ is open with respect to an affine subspace of $\mathbb{R}^{p\times p}$.

Given the above setup, it is shown that the set $\mathfrak{M}^*$ is convex and the coordinates of $T(X)$ have moments of all orders with respect to f_{M} for any $\mathrm{M} \in \mathfrak{M}^*$ (Zhuang et al., 2016, Lemma 2.1)

3.2.2 Undirected Graphical Model Examples

The goal of this section is to discuss graphical model examples under the exponential trace framework. By convention of undirected graphical models, we consider $\xi(\mathbf{x}) = \sum_{j=1}^{p} \xi_j(x_j)$, and symmetric matrices M. We then define the associated undirected graph $G(V, E)$, where node set $V := \{1, \ldots, p\}$ and edge set $E := \{(i, j) : i, j \in V, i \neq j, \mathrm{M}_{ij} \neq 0\}$. The graph is undirected in the sense that $(i, j) \in E$ if and only if $(j, i) \in E$. It can be shown that the graph G encodes the conditional dependence structure by utilizing the Hammersley-Clifford theorem (Besag, 1974; Grimmett, 1973). The exponential trace model density $f_{\mathrm{M}}(\mathbf{x}) = f^1_{\mathrm{M}}(\mathbf{x}_{-i}) f^2_{\mathrm{M}}(\mathbf{x}_{-j})$ with positive functions $f^1_{\mathrm{M}}, f^2_{\mathrm{M}}$ if and only if $\mathrm{M}_{ij} = 0$. Therefore, we have

$$X_i \perp X_j | X_{-\{i,j\}} \qquad \text{if and only if} \qquad (i, j) \notin E\,.$$

Figure 3.1 gives an illustration of the corresponding relationship between the parameter M and the graph G. Next, we discuss example graphical models encompassed by the exponential trace model for specific data types.

$$M = \begin{bmatrix} 1.2 & -0.2 & 0 & 0 \\ -0.2 & 1.5 & 0 & 0.1 \\ 0 & 0 & 1 & 0 \\ 0 & 0.1 & 0 & 0.5 \end{bmatrix}$$

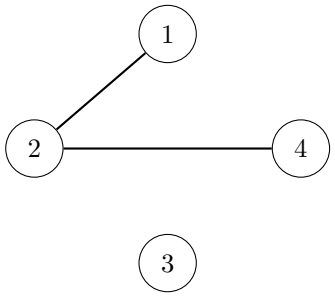

Figure 3.1: Example of a matrix M (left) and the corresponding graph G (right). The node set of the graph is $V = \{1, 2, 3, 4\}$, the edge set of the graph is $E = \{(1, 2), (2, 1), (2, 4), (4, 2)\}$. For example, X_1 and X_4 are conditionally independent given the other elements of X (indicated by $(1, 4) \notin E$).

Gaussian The well-studied Gaussian graphical models (Lauritzen, 1996b) serves as an important example. For centered multivariate normal random vectors $X \sim \mathcal{N}_p(0, \Sigma)$, where Σ is a positive definite matrix. The Gaussian graphical models can be recovered by setting $\mathcal{D} = \mathbb{R}^p$, $\mathrm{M} = \Sigma^{-1}$, $T_{ij}(\mathbf{x}) = x_i x_j/2$, and $\xi \equiv 0$ in the exponential framework (3.2). It holds that

$$a(\mathrm{M}) < \infty \quad \text{for all positive definite } \mathrm{M} \in \mathbb{R}^{p \times p}\,, \tag{3.3}$$

that is, the integrability condition is satisfied for all positive definite matrices. Since the set of all positive definite matrices in $\mathbb{R}^{p \times p}$ is open, we have

$$\mathfrak{M}^* \supset \{\mathrm{M} \in \mathbb{R}^{p \times p} : \mathrm{M} \text{ positive definite}\}\,.$$

Poisson Multivariate count data with infinite range is of high relevance in practice. For example, this type of data includes the abundance of microbe in system biology, radioactive decay of particles in particle physics, number of crimes and arrests in criminalistics. However, existing graphical models for this type of data are still imperfect so far. For example, standard approach (extend independent case by adding additional pairwise interaction terms) requires all the direct interactions to be negative to ensure a valid joint density. We can model count-valued data using framework (3.2) by specifying $\mathcal{D} = \{0, 1, \dots\}^p$, $\xi(\mathbf{x}) = -\sum_{j=1}^p \log(x_j!)$,

and functions T that satisfy $T_{ii}(\mathbf{x}) = T_{ii}(x_i) = x_i$ and $T_{ij}(\mathbf{x}) = T_{ij}(x_i, x_j) \leq c(x_i + x_j)$ for some $c \in (0, \infty)$. The choice $T_{ij} = \sqrt{x_i x_j}$ considered in Inouye et al. (2016) is a special case of such transformations. Without imposing unreasonable assumptions on the parameter space, it holds that

$$a(\mathrm{M}) < \infty \text{ for all matrices } \mathrm{M} \in \mathbb{R}^{p \times p}$$

(We defer the technique details to Section 3.7.1.) Meanwhile, the node conditional approximately follows a Poisson distribution when the interaction indexed by M is small. It can be revealed by the node conditional in the form of

$$f_{\mathrm{M}}(x_j | \mathbf{x}_{-j}) \sim e^{-\mathrm{M}_{jj} x_j - \log(x_j!)} L_{\mathrm{Int}} \,,$$

where $L_{\mathrm{Int}} = e^{-\sum_{k \in \mathcal{N}(j)} (\mathrm{M}_{jk} + \mathrm{M}_{kj}) T_{jk}(x_j, x_k)}$.

Exponential Similarly as in Poisson case, the exponential trace model framework can model exponential data while avoiding the integrability issues of standard extensions of independent case. We consider $\mathcal{D} = [0, \infty)^p$, $\xi \equiv 0$, and the square-root transformation $T_{ij}(\mathbf{x}) = \sqrt{x_i x_j}$. Applying the fact that eigenvalues of positive definite matrix are positive, one can show that the integrability condition holds for all positive definite M (See Section 3.7.1 for details.) We note that additional transformations T would satisfy the integrability property. In addition, the framework preserves the exponential flavor for node conditionals. The node conditional has a format of

$$f_{\mathrm{M}}(x_j | \mathbf{x}_{-j}) \sim e^{-\mathrm{M}_{jj} x_j} L_{\mathrm{Int}} \,,$$

where $L_{\mathrm{Int}} = e^{-\sqrt{x_j} \sum_{k \in \mathcal{N}(j)} (\mathrm{M}_{jk} + \mathrm{M}_{kj}) \sqrt{x_k}}$. When the interactions of node j with others are small in scale, $L_{\mathrm{Int}} \approx 1$ and the node conditional approximates an exponential random variable.

Composite Models Composite models refer to the model specialized for data involving more than one type. Since count-valued data is one of the most challenging and exciting data types, we consider the model for a composite of Poisson and Bernoulli as well as the model for a composite of Poisson and Gaussian. These two examples represent the composite of different discrete data and that of discrete and continuous data, respectively. We consider $p_1, p_2 \in \{1, 2, \dots\}$, $p_1 + p_2 = p$, the first p_1 elements of the random vector X are Poisson, while the other p_2 elements are Bernoulli (or Gaussian). In addition, we consider the choice $T_{ij} = t_i(x_i)t_j(x_j)$, where $t_i(x_i) = \sqrt{x_i}$ for non-Gaussian coordinates and $t_i(x_i) = x_i$ for Gaussian coordinates. For the composite of discrete data, $a(\mathrm{M}) < 0$ for all matrices $\mathrm{M} \in \mathbb{R}^{p\times p}$. For the composite of discrete and Gaussian data, the integrability condition is satisfied for all matrices $\mathrm{M} \in \mathbb{R}^{p\times p}$ such that the "Gaussian" block is positive definite. One can verify that the node conditional approximates the corresponding desirable form (Poisson or Bernoulli or Gaussian as pertained) following the same line as in the Poisson case.

3.3 Estimation

We consider estimation based on maximum likelihood and propose a sampling-based approximation algorithm for computing the maximum likelihood estimator.

Given the model class and n i.i.d. data samples $X^1, \dots, X^n$ from a distribution of the form (3.2), the negative joint log-likelihood function $-\ell_{\mathrm{M}}$ can be expressed by

$$-\ell_{\mathrm{M}}(\underline{\mathbf{x}}) = n\langle \mathrm{M}, \overline{T}(\underline{\mathbf{x}})\rangle_{\mathrm{tr}} + na(\mathrm{M}) + c\,,$$

where $\overline{T}(\underline{\mathbf{x}}) := \frac{1}{n}\sum_{i=1}^{n} T(\mathbf{x}^i)$, and $c \in \mathbb{R}$ does not depend on M. The maximum likelihood estimator of M is thus defined as

$$\widehat{\mathrm{M}} := \underset{\widetilde{\mathrm{M}}\in\mathfrak{M}}{\operatorname{argmin}} \left\{ -\ell_{\widetilde{\mathrm{M}}}(\underline{X}) \right\} = \underset{\widetilde{\mathrm{M}}\in\mathfrak{M}}{\operatorname{argmin}} \left\{ \langle \widetilde{\mathrm{M}}, \overline{T}(\underline{X})\rangle_{\mathrm{tr}} + a(\widetilde{\mathrm{M}}) \right\}. \tag{3.4}$$

The objective function is convex with explicit derivatives (Zhuang et al., 2016, Lemma 3.2,

3.3). Under our assumption that $\mathrm{M} \in \mathfrak{M} \subset \mathfrak{M}^*$ and $\mathfrak{M}$ is open and convex, and the minimizer exists for n sufficiently large, cf. (Berk, 1972).

The main challenge in computing the maximum likelihood estimator is the unconventional normalization term. We address this challenge by approximating the objective function using a sampling-based technique. In this section, we describe the corresponding algorithm.

We denote the objective function of the maximum likelihood estimator in Equation (3.4) as

$$g(\widetilde{\mathrm{M}}) := \langle \widetilde{\mathrm{M}},\, \overline{T}(\underline{X}) \rangle_{\mathrm{tr}} + a(\widetilde{\mathrm{M}}).$$

The normalization term $a(\widetilde{\mathrm{M}})$, in general, does not have a closed-form formula and, therefore, makes the objective function hard to compute exactly. However, we show in the following that it can be feasibly approximated. Adding the constant term $a(\mathrm{M}_0)$, where M_0 is a pre-specified constant parameter matrix, to the objective function of (3.4) yields an equivalent definition of the maximum likelihood estimator as

$$\widehat{\mathrm{M}} = \underset{\widetilde{\mathrm{M}} \in \mathfrak{M}}{\operatorname{argmin}} \left\{ \langle \widetilde{\mathrm{M}},\, \overline{T}(\underline{X}) \rangle_{\mathrm{tr}} + a(\widetilde{\mathrm{M}}) - a(\mathrm{M}_0) \right\}. \tag{3.5}$$

Algebraic transformation reveals

$$a(\widetilde{\mathrm{M}}) - a(\mathrm{M}_0) = \log \mathbb{E}_{\mathrm{M}_0} \Big(e^{-\langle \widetilde{\mathrm{M}} - \mathrm{M}_0,\, T(\mathbf{x}) \rangle_{\mathrm{tr}}} \Big).$$

The finite expectation can be approximated by an empirical mean based on some sample set Y from the distribution $f_{\mathrm{M}_0}(\mathbf{x}) := \exp(-\langle \mathrm{M}_0,\, T(\mathbf{x}) \rangle_{\mathrm{tr}} + \xi(\mathbf{x}) - a(\mathrm{M}_0))$. By the strong law of large numbers, when the cardinality of the set Y (denoted as $|Y|$) goes to infinity, the empirical mean approximates the expectation well:

$$\log \frac{1}{|Y|} \sum_{Z \in Y} e^{-\langle \mathrm{M} - \mathrm{M}_0,\, T(Z) \rangle_{\mathrm{tr}}} \xrightarrow{a.s} \log \mathbb{E}_{\mathrm{M}_0} \Big(e^{-\langle \mathrm{M} - \mathrm{M}_0,\, T(\mathbf{x}) \rangle_{\mathrm{tr}}} \Big).$$

Hence, in practice, we solve the approximate problem

$$\widehat{\mathrm{M}} \approx \operatorname*{argmin}_{\widetilde{\mathrm{M}} \in \mathfrak{M}} \left\{ \widetilde{g}(\widetilde{\mathrm{M}}) \right\}, \tag{3.6}$$

where

$$\widetilde{g}(\widetilde{\mathrm{M}}) := \langle \widetilde{\mathrm{M}},\, \overline{T}(\underline{X}) \rangle_{\mathrm{tr}} + \log \frac{1}{|Y|} \sum_{Z \in Y} e^{-\langle \widetilde{\mathrm{M}} - \mathrm{M}_0,\, T(Z) \rangle_{\mathrm{tr}}}.$$

We apply gradient descent to solve the problem of (3.6). The gradient with respect to $\widetilde{\mathrm{M}}$ is

$$\nabla \widetilde{g}(\widetilde{\mathrm{M}}) = \overline{T}(\underline{X}) - \frac{\sum_{Z \in Y} T(Z) e^{-\langle \widetilde{\mathrm{M}} - \mathrm{M}_0,\, T(Z) \rangle_{\mathrm{tr}}}}{\sum_{Z \in Y} e^{-\langle \widetilde{\mathrm{M}} - \mathrm{M}_0,\, T(Z) \rangle_{\mathrm{tr}}}}. \tag{3.7}$$

Remark 3.3.1. *The gradient in* (3.7) *can be considered as an instantiation of self-normalized importance sampling. To see the link, we first introduce importance sampling and a particular extension — self-normalized importance sampling.*

Suppose our problem is to estimate $\mu := \mathbb{E}(T(X); p)$ *for* X *distributed with the density function of* p*. When it is convenient to sample from* p*, a natural empirical estimator of* μ *is* $\widehat{\mu} = \frac{1}{n} \sum_{i=1}^{n} T(X^i)$ *where* $(X^1, \ldots, X^n)$ *are identically and independent samples from* p*. When it is difficult to sample from* p*, importance sampling draws samples from a biased distribution* q *and obtained the desired expectation by adjusting the weights for the drawn samples. In particular, the idea of importance sampling is based on*

$$\mu = \int T(X) p(X) dX = \int \frac{T(X) p(X)}{q(X)} q(X) dX.$$

And the importance sampling estimator is in the form of

$$\widehat{\mu}_q := \frac{1}{n} \sum_{i=1}^{n} \frac{p(X^i) T(X^i)}{q(X^i)},$$

where X^i *are sampled from* q*. There are two merits of the approach: first,* $\widehat{\mu}_q$ *is an unbiased estimator of* μ*; second, (even if* p *is easy to sample from), a good sampling distribution* q

would lead to a smaller variance $Var(\widehat{\mu}_q)$ *than* $Var(\widehat{\mu})$ *(Owen, 2013, Theorem 9.1)*

Sometimes the density p *is only known up to a normalization term. Suppose* $p(x) = c_p p_0(x), q(x) = c_q q_0(x)$,*where* c_p *is an unknown constant. To tackle this challenge, a further derivation along the line of importance sampling gives*

$$\int T(X)p(X)dX = \frac{\int \frac{T(X)p_0(X)}{q_0(X)}q(X)dX}{\int \frac{p_0(X)}{q_0(X)}q(X)dX} = \frac{\int T(X)w(X)q(X)dX}{\int w(X)q(X)dX},$$

where $w(X) = p(X)/q(X)$*. An estimator with weights normalized by their sum is developed as*

$$\widetilde{\mu}_q := \sum_{i=1}^{n} \frac{w(X^i)T(X^i)}{\sum_{i=1}^{n} w(X^i)},$$

and called self-normalized self-importance sampling.

Lemma 3.3 of Zhuang et al. (2016) proves that $g(\widetilde{\mathrm{M}})$ *is differentiable with a gradient in the form of*

$$\nabla g(\widetilde{\mathrm{M}}) = \overline{T}(\underline{X}) - \mathbb{E}_{\widetilde{\mathrm{M}}}\overline{T}(\underline{X}) = \overline{T}(\underline{X}) - \mathbb{E}_{\widetilde{\mathrm{M}}}T(X)\,.$$

In the general setting of exponential trace models, we observe: $\mathbb{E}_{\widetilde{\mathrm{M}}}T(X)$ *lacks an algebraic expression; the density function* $f_{\widetilde{\mathrm{M}}}$ *is inconvenient to sample from; and the density function is only known up to a constant factor* $a(\mathrm{M})$*. The observations indicate that self-importance sampling would be useful to approximate the expectation term* $\mathbb{E}_{\widetilde{\mathrm{M}}}T(X)$*.*

Actually, Equation (3.7) *reflects the same idea of self-normalized importance sampling and uses the pre-specified* f_{M_0} *as the biased sampling distribution. The corresponding self-normalized weights are*

$$\frac{e^{-\langle \widetilde{\mathrm{M}}-\mathrm{M}_0, T(Z)\rangle_{\mathrm{tr}}}}{\sum_{Z\in Y} e^{-\langle \widetilde{\mathrm{M}}-\mathrm{M}_0, T(Z)\rangle_{\mathrm{tr}}}}, \quad Z \sim f_{\mathrm{M}_0}.$$

Although sharing similar ideas and concluding the same formula for the gradient, the two derivation processes yet have significant differences and would lead to attempts in different

directions. The derivation in the main text approximates the objective function using Monte Carlo samples from a biased sampling distribution and then optimize the approximate objective function as a routine. This approach only requires one set of Monte Carlo samples, which reduces the difficulty and computational expense of sampling. In contrast, applying self-normalized importance sampling on the gradient in Lemma 3.2 sticks to the original objective function but approximates the optimization direction at each iteration. This approach falls to the same line of stochastic gradient descent. However, the stochasticity at each iteration brings in additional challenges. For example, how to quantify the distance between the approximate solution and the actual minimizer theoretically, how to find the appropriate step size in a computationally effective way. Importantly, as we will see in the later sections, the choice of the sampling distribution is critical for our implementation. Repeating the sampling for each iteration will not be fun. Based on the above considerations, we choose the derivation and implementation idea in the main text.

However, the approaches of importance sampling are well-developed and equipped with extensive theoretical and practical results regarding the choice of biased sampling distributions. The link to self-normalized importance sampling allows us using the results of importance sampling to guide our implementation.

The choice of M_0 is essential for the finite-sample performance of the approximation. Our two main considerations are: First, the sampling distribution f_{M_0} should be straightforward for generating samples. Second, it should lead to balanced weights and a small variance of $\widehat{\mathbb{E}}_{\widetilde{\mathrm{M}}} T(X)$; if weights concentrate in just a few samples, we have effectively only got these observations, resulting in large variability of the approximation (Owen, 2013). To incorporate the two considerations, we propose

$$\mathrm{M}_0 := \underset{\widetilde{\mathrm{M}} \in \mathfrak{M}}{\operatorname{argmin}} \left\{ \langle \widetilde{\mathrm{M}}, \overline{T}(\underline{X}) \rangle_{\mathrm{tr}} + a(\widetilde{\mathrm{M}}) \right\} \text{ subject to } \widetilde{\mathrm{M}}_{kl} = 0, \quad \forall k \neq l. \tag{3.8}$$

We restrict the parameter to a diagonal matrix, by which we presume mutual independence among all coordinates and disentangle the objective function. Hence, both the optimizing

problem (3.8) and the task of sampling from f_{M_0} can be handled for each coordinate separately, and each coordinate reduces to a standard univariate exponential family distribution. Besides, M_0 is the diagonal matrix closest to the actual parameter. When the off-diagonal entries of the actual parameter matrix are small, weights of the drawn samples are expected to be reasonable.

Algorithm 1 summarizes our computational pipeline. In the full version, which is stated in Section 3.7.2, a backtracking line search selects the step size adaptively and incorporates the domain constraint as needed (for example, the positive definite requirement for the case of continuous measures).

Algorithm 1: Solving for the maximum likelihood estimator

`// ` η`: step size`
Input : $\overline{T}(\underline{X}), \eta > 0$
Output: $\widehat{\mathrm{M}}$
`// Solve for` M_0
$\mathrm{M}_0 \leftarrow \mathbf{0}_{p\times p}$;
for $i = 1, \ldots, p$ **do**
  $(\mathrm{M}_0)_{ii} \leftarrow \mathrm{argmin}_{m\in\mathbb{R}} \Big\{ m\overline{T}_{ii}(\underline{X}) + \log \int \exp\big(- mT_{ii}(\mathbf{x}) + \xi(x_i)\big) dx_i \Big\}$;
`// Generate sample set` Y `from` $f_{\mathrm{M}_0} = \prod_{i=1}^{p} f_{(\mathrm{M}_0)_{ii}}(x_i)$
for $i = 1, \ldots, p$ **do**
  Generate 10,000 random samples from $f_{(\mathrm{M}_0)_{ii}}(x_i)$ for the i-th coordinate;
`// Apply gradient descent with constant step size to` (3.6)
$k \leftarrow 0$;
$\widetilde{\mathrm{M}}_k \leftarrow \mathrm{M}_0$;
repeat
  $k \leftarrow k + 1$;
  $\nabla\widetilde{g}(\widetilde{\mathrm{M}}_{k-1}) \leftarrow \overline{T}(\underline{X}) - \frac{\sum_{Z\in Y} T(Z) e^{-\langle \widetilde{\mathrm{M}}_{k-1} - \mathrm{M}_0,\, T(Z)\rangle_{\mathrm{tr}}}}{\sum_{Z\in Y} e^{-\langle \widetilde{\mathrm{M}}_{k-1} - \mathrm{M}_0,\, T(Z)\rangle_{\mathrm{tr}}}}$;
  $\widetilde{\mathrm{M}}_k \leftarrow \widetilde{\mathrm{M}}_{k-1} - \eta\nabla\widetilde{g}(\widetilde{\mathrm{M}}_{k-1})$;
until $\big|\widetilde{g}(\widetilde{\mathrm{M}}_k) - \widetilde{g}(\widetilde{\mathrm{M}}_{k-1})\big| < 10^{-4}$;
$\widehat{\mathrm{M}} \leftarrow \widetilde{\mathrm{M}}_k$;

Remark 3.3.2. *The maximum likelihood estimator of* M *is asymptotically normal with co-*

variance equal to the inverse Fisher information. In particular, consistency and asymptotic normality of the maximum likelihood estimator can be proved following the arguments in the classical paper (Berk, 1972), see especially (Berk, 1972, Theorems 4.1 and 6.1). We refer to that paper for details.

3.4 Simulation Studies

Exponential trace models apply to a large variety of multivariate data that have correlated coordinates. The framework is especially useful for data that are discrete, heavy-tailed, or composed of different data types. We consider the following four model-types in simulations: Poisson, Exponential, Poisson-Bernoulli (a composite of Poisson and Bernoulli), and Poisson-Gaussian (a composite of Poisson and Gaussian). Such types of models are useful in practice but, to date, have proven challenging to characterize and estimate.

3.4.1 Settings

We follow the discussion of Section 3.2.2 and consider the square-root transformation for non-Gaussian data as one example, that is, we set $T_{ij}(\mathbf{x}) = t_i(x_i)t_j(x_j)$, where $t_i(x_i) = \sqrt{x_i}$ for non-Gaussian coordinates and $t_i(x_i) = x_i$ for Gaussian coordinates.

In the following, we describe the specific graph structures for discrete data, continuous data, and composite data of both types. The discrete data category also covers composite data of different discrete types (for example, Poisson-Bernoulli).

Discrete Data

We consider Erdős-Rényi random graphs and generate the corresponding $p \times p$ parameter matrix M in the following manner. Let c_0 and c_1 be two constants ($c_1 \neq 0$). We set the diagonal entries to $\mathrm{M}_{ii} = c_0$. For $i \neq j$, the off-diagonal entries are independent and

identically distributed as

$$\mathrm{M}_{ij} = \mathrm{M}_{ji} = \begin{cases} c_1 & \text{with probability } 1/p, \\ -c_1 & \text{with probability } 1/p, \\ 0 & \text{with probability } 1 - 2/p. \end{cases} \tag{3.9}$$

The corresponding Erdős-Rényi random graph has $p - 1$ edges in expectation, among which half represent positive interactions and the other half negative ones.

Continuous Data

We consider again Erdős-Rényi random graphs. In addition, we generate strictly diagonally dominant matrices with positive diagonal entries for the parameter matrix M. It can be shown that the M's are positive definite and satisfy the integrability condition. More specifically, we first generate a $p \times p$ adjacency matrix with i.i.d. off-diagonal entries

$$A_{ij} = A_{ji} = \begin{cases} 1 & \text{with probability } 1/p, \\ -1 & \text{with probability } 1/p, \\ 0 & \text{with probability } 1 - 2/p, \end{cases}$$

We denote the maximum node degree by $s := \max_i \sum_{j \neq i} |A_{ij}|$. Then, a positive definite M can be generated by setting the diagonal entries to 1 and the off-diagonal entries to

$$\mathrm{M}_{ij} = \mathrm{M}_{ji} = 1/(s + 0.1)A_{ij}.$$

The corresponding Erdős-Rényi random graph has $p - 1$ edges in expectation, among which half represent positive interactions and the other half negative ones.

A Composite of Discrete and Continuous Data

We consider even values of p, with the first $p_1 = p/2$ coordinates discrete, and the remaining $p_2 = p/2$ coordinates continuous. The parameter matrix in blockwise format is

$$\mathrm{M} = \begin{bmatrix} \mathrm{M}_{11} & \mathrm{M}_{12} \\ \mathrm{M}_{12}^{\top} & \mathrm{M}_{22} \end{bmatrix},$$

where $\mathrm{M}_{11}, \mathrm{M}_{22}$ represent the conditional dependences among p_1 discrete and p_2 continuous coordinates, respectively. The remaining block M_{12} describes the conditional dependences between discrete and continuous coordinates. When the continuous node conditionals follow a Gaussian distribution, the integrability condition is satisfied for all $\mathrm{M} \in \mathbb{R}^{p \times p}$ such that M_{22} is positive definite. Therefore, we generate M_{22} in the above described continuous case to guarantee its positive definiteness while generating M_{11} and M_{12} in the above described discrete case.

3.4.2 Data Generation

Here we describe how to generate samples from an exponential trace model in the form of

$$f_{\mathrm{M}}(\mathbf{x}) = e^{-\langle \mathrm{M}, T(\mathbf{x}) \rangle_{\mathrm{tr}} + \xi(\mathbf{x}) - a(\mathrm{M})}.$$

Generating data from a multivariate distribution directly is difficult for moderate node numbers. Instead, we consider a Gibbs sampler to sample from the conditional distribution at each iteration. Conditioning on all the other variables $\mathbf{x}_{-j}$, the density of x_j is

$$f_{\mathrm{M}}(x_j \mid \mathbf{x}_{-j}) \sim e^{-\mathrm{M}_{jj} T_{jj} - 2\sum_{k \neq j} \mathrm{M}_{jk} T_{jk} - \xi(x_j)}.$$

We generate the conditional distribution using a slice sampler Neal (2003) with R package `MfUSampler` Mahani and Sharabiani (2014).

Note that the slice sampler was designed for continuous variables. For discrete coor-

dinates, we uniformly spread the probability in the spike at an integer c into the interval between c and $c+1$. This defines a continuous density

$$f_{\mathrm{M}}(y_j \mid \mathbf{x}_{-j}) \sim e^{-\mathrm{M}_{jj}\lfloor y_j \rfloor - 2\sum_{k\neq j}\mathrm{M}_{jk}T_{jk}(\lfloor y_j \rfloor, x_k) - \xi(\lfloor y_j \rfloor)},$$

where $\lfloor y_j \rfloor$ represents the largest integer less than y_j. We sample from the above continuous density and take $\lfloor y_j \rfloor$ as the realization of a discrete x_j.

3.4.3 Results

We evaluate the maximum likelihood estimators in terms of edge recovery by studying average ROC (receiver operating characteristic) curves based on thresholding maximum likelihood estimates. Each average ROC curve is taken over 50 individual ROC curves that correspond to 50 different Erdős-Rényi random graphs. ROC curves are combined using horizontal-averaging via the `R` package `ROCR` (Sing et al., 2005). Since Gaussian graphical models are currently widely used, even in cases with obvious misspecification (eg. count data) (Zhao and Duan, 2019), we compare the maximum likelihood estimators of the exponential trace model to that of the (misspecified) Gaussian graphical model.

The graph structures are described in Section 3.4.1. Details regarding the data generation approaches are deferred to Appendix 3.4.2. We use $n = 250$ independent observations to recover the conditional dependence of $p = 20$ variables. The diagonal entry is $c_0 = -1$ and the off-diagonal entry is $c_1 = 0.3$. We show the average ROC curves of Poisson, Exponential, Poisson-Bernoulli, and Poisson-Gaussian in Figures 3.2 and 3.4. In addition, we explore scenarios with diagonal entry $c_0 \in \{0, -0.5, -1, -1.5, -2\}$ and off-diagonal entry $c_1 \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$: we show the differences in AUC of average ROC curves between the exponential trace model and the Gaussian graphical model in Figures 3.3 and 3.5.

In the pure data type scenarios: The exponential trace model outperforms the Gaussian graphical model substantially—the expoential trace model's ROC curves lie entirely above the Gaussian graphical model's ROC curves—in the scenario of small interaction and small

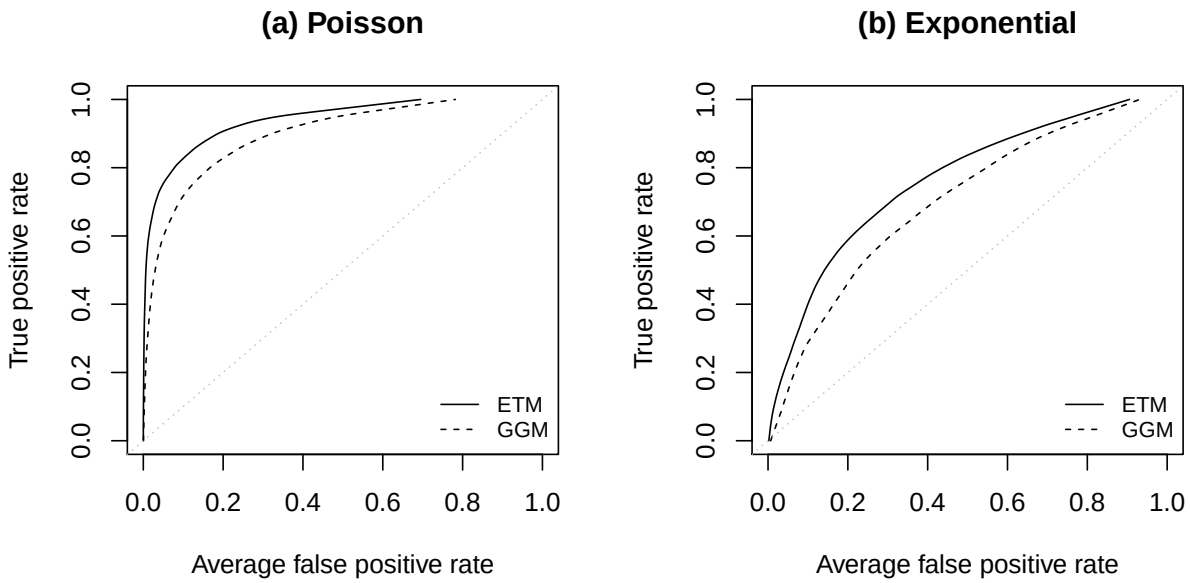

Figure 3.2: Average ROC curves for pure data. ETM stands for the Exponential trace mode and GGM stands for Gaussian graphical model.

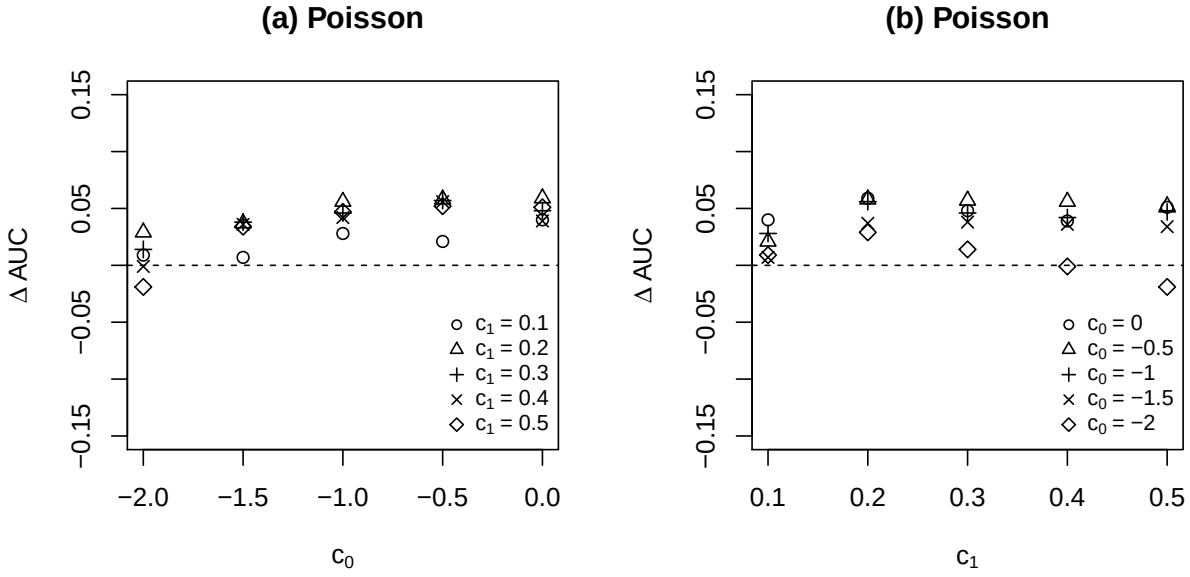

Figure 3.3: Differences in AUC of average ROC curve between exponential trace model and Gaussian graphical model for Poisson data.

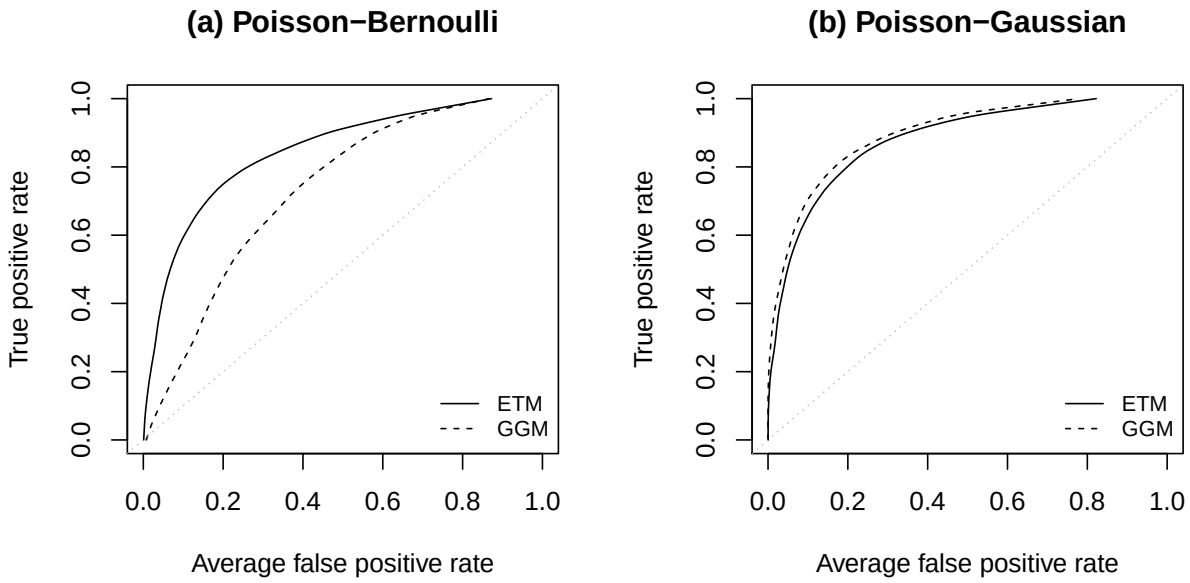

Figure 3.4: Average ROC curves for composite data. ETM stands for the Exponential trace mode and GGM stands for Gaussian graphical model.

sufficient statistics $T(\mathbf{x})$ (see Figure 3.2). In addition, we look into more configurations of Poisson type scenarios by varying c_0 and c_1: the performance of the sampling-based approximation determines that of the exponential trace model (see Figure 3.3). When the interaction term and sufficient statistics are small to moderate, the exponential trace model has a larger AUC than the Gaussian graphical model. But when the interaction term and sufficient statistics are large, for example, $c_0 = -2$ and $c_1 = 0.5$, the improvement is not guaranteed. For the composite data type scenarios: The exponential trace model shows improved performance for Poisson-Bernoulli data but not for Poisson-Gaussian data (see Figures 3.4 and 3.5). Bernoulli data have very small sufficient statistics so that the improvement in Poisson-Bernoulli data is substantial for a large range of interaction terms, but Poisson-Gaussian data involves Gaussian coordinates, so that the performance of the exponential trace model is mixed. The comparative performance is not only determined by the performance of the approximation algorithm but also the degree to which the conditional

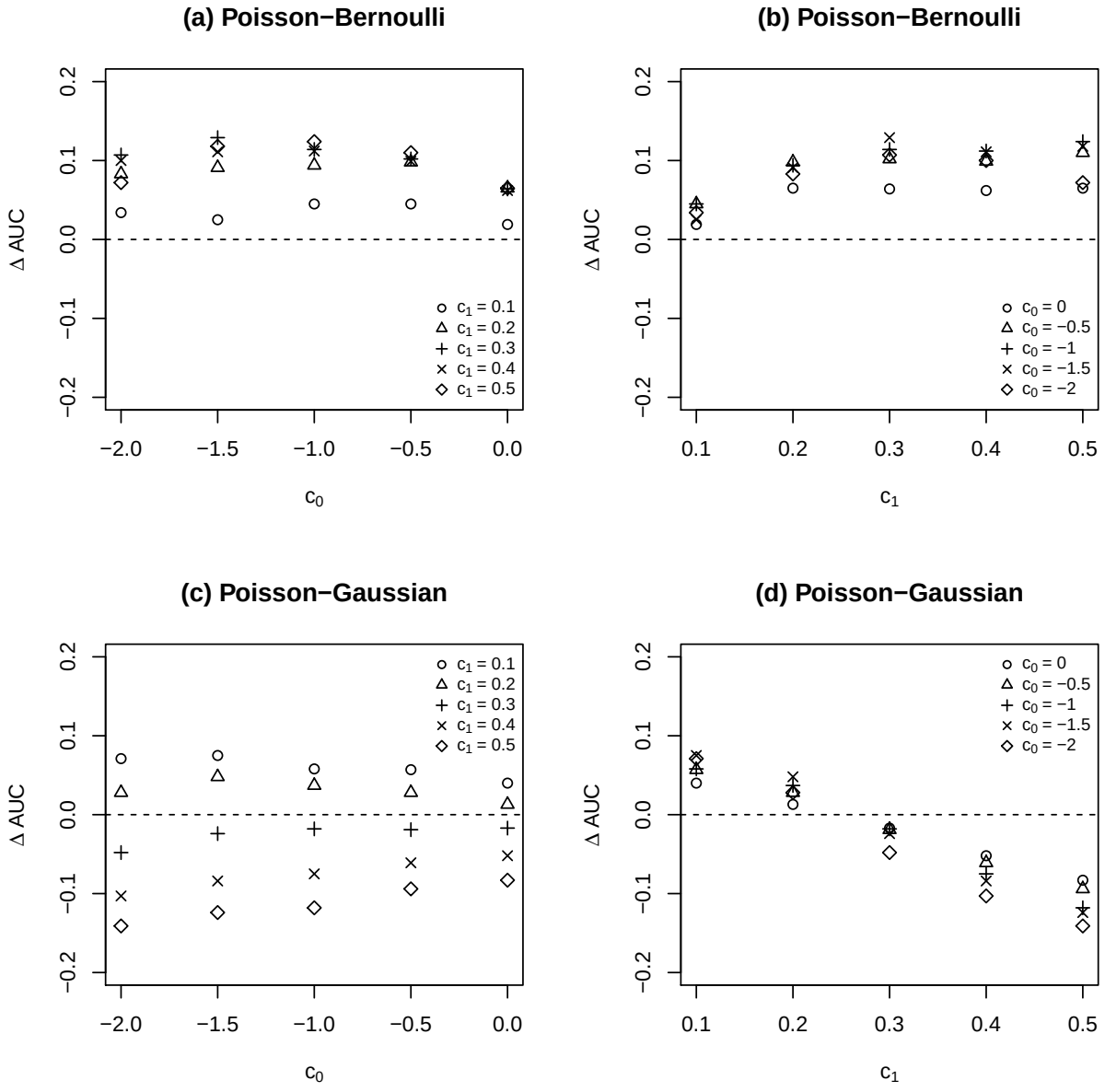

Figure 3.5: Differences in AUC of average ROC curves between exponential trace model and Gaussian graphical model for composite data.

dependence structure resembles the zero pattern of the precision matrix.

In conclusion, the simulation study shows that: (1) the exponential trace model can improve on the Gaussian graphical model for non-Gaussian data, especially when the sufficient statistics and small interactions are small; (2) the approximation approach can struggle when sufficient statistics and interactions are large. Based on this, we believe that our modeling approach has substantial potential: Some of this potential is realized by our current implementation, however some of the potential might require additional computational insights.

3.5 Application to Neural Spike Data

In this section, we apply the proposed exponential trace model to neural spike data. The temporal and spatial patterns of neural spikes capture the concurrent activity of neurons. Understand this is essential for learning neural circuits. Neural spike data is usually formulated as spike counts in a short time bin and modeled by a Poisson distribution (Theis et al., 2016). We consider a data set of multi-electrode array recordings of spike trains in mouse retina (Demas et al., 2003) obtained from the Retinal Wave Repository (Eglen et al., 2014). We transform the spike time data into spike counts in time bins of $40\,ms$ following conventions in neural science (Theis et al., 2016). The short time interval captures the instantaneous characteristics of neuron firing. The recording covers an $800 \times 800\,\mu m$ surface area and provides locations of each recorded unit in the form of (x, y)-coordinates. The number of recorded units ranges from 12 to 22 in different mice.

The spike counts of each recorded unit range from 0 to 13 and roughly follow the mean-variance relationship of a Poisson distribution. The small counts imply: (1) the exponential trace model with square-root transformation is close to the Poisson one; (2) the properties of the data fit the scenario for which the proposed algorithm is appropriate. These two observations render the exponential trace model with a square-root transformation appropriate.

In Figures 3.6, we present the recovered connections of recorded units for a 6-week-old wild-type mouse. To obtain sparse graphs, the maximum likelihood estimator is thresholded to 30 edges. In the plotted graphs, the positions of the nodes correspond to to [slightly

jittered] (x, y)-coordinates of the neurons in the recording. This allows us to consider spatial summaries, eg. the average physical length of edges in our graph.

The exponential trace model finds a more centralized graph than the Gaussian graphical model, that is, neurons located together tend to connect in the exponential trace model approach while the estimated connections in the graph are longer in the Gaussian approach. To evaluate this difference quantitatively, we compute the mean Euclidean distance between all pairs of directly connected neurons. In the 6-week-old wild-type mouse, the mean distance is $169\,\mu m$ (SE: $24\,\mu m$) under the exponential trace model and $252\,\mu m$ (SE: $36\,\mu m$) under the Gaussian graphical model. In addition, the exponential trace model recovers a main connection component and an isolated neuron unit (see Unit 7 in Figure 3.6). The isolated neuron unit is physically far away from the primary component and may belong to another functional group. In contrast, the Gaussian graphical model does not distinguish the location separation clearly but instead connects almost every units.

These characteristics found in the exponential trace model but not the Gaussian graphical model align with our biological understanding. In particular, neurons transmit signals to others through synapses, which are physical connections between two neurons (Lodish et al., 2008). This biological mechanism favors direct coordination of closely located neurons. Specifically, previous studies find that the degree two units spike together within some small time window decays with the distance separating the neurons in retina (Cutts and Eglen, 2014; Wong et al., 1993; Xu et al., 2011).

3.6 Summary

In this chapter, we propose the exponential family-based framework for characterizing undirected graphical models. This framework allows for a wide range of data types, as highlighted in Section 3.2.2. The models are amenable to estimation based on maximum likelihood via a sampling-based approximation technique.

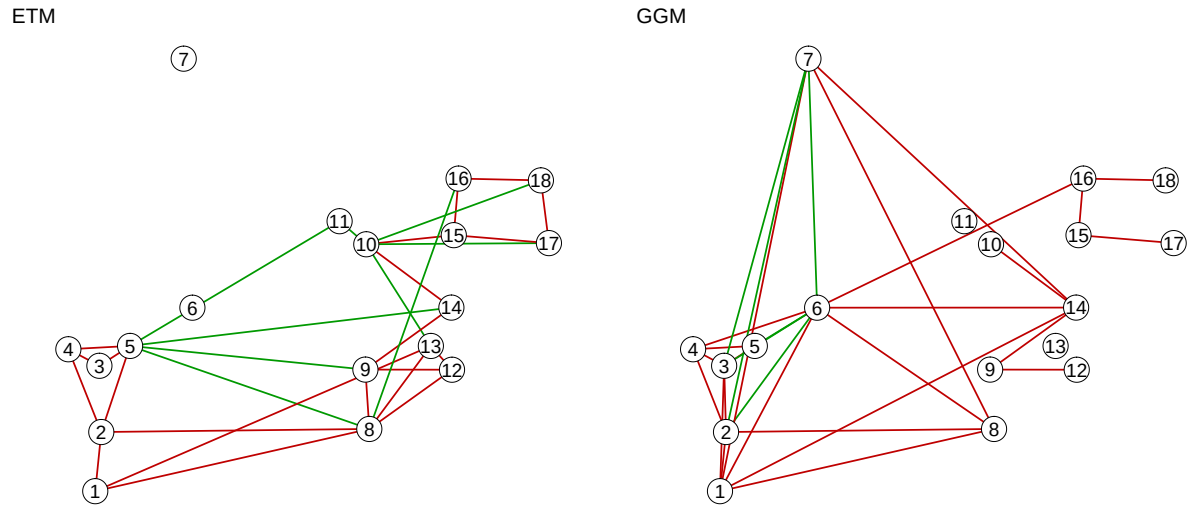

Figure 3.6: Retinal connections recovered for a 6-week-old wild-type mouse of the Demas et al. 2003 data. The positive interactions are drawn in red,and the negative ones in green. The left panel displays the connection recovered by the exponential model with square-root transformation; the right panel displays the connection recovered by the Gaussian graphical model.

3.7 Supplementary Materials

3.7.1 Integrability Property for Poisson and Exponential Models

For Poisson models described in 3.2.2, the integrability property holds in the sense of

$$a(\mathrm{M}) = \sum_{x_1,\ldots,x_p=0}^{\infty} \frac{e^{-\sum_j \mathrm{M}_{jj}x_j - \sum_{i,j:i\neq j} \mathrm{M}_{ij}T_{ij}(x_i,x_j)}}{\prod_j x_j!} < \infty\,,$$

for all matrices $\mathrm{M} \in \mathbb{R}^{p\times p}$. For $T_{ij}(x_i, x_j) \leq c(x_i + x_j)$, we have

$$-\mathrm{M}_{ij}T_{ij}(x_i, x_j) \leq \tilde{c}x_i + \tilde{c}x_j,$$

where $\tilde{c} := c \max_{i,j} |\mathrm{M}_{ij}|$. That implies

$$a(\mathrm{M}) \leq \sum_{x_1,\ldots,x_p=0}^{\infty} \frac{e^{-\sum_j (\mathrm{M}_{jj}-2(p-1)\tilde{c})x_j}}{\prod_j x_j!}.$$

We define $C(\mathrm{M}, j) := e^{-(\mathrm{M}_{jj}-2(p-1)\tilde{c})}$, $j \in \{1, \ldots, p\}$. By the fact that $C(\mathrm{M}, j) \in (0, \infty)$, we have

$$\prod_{j=1}^{p} e^{C(\mathrm{M},j)} < \infty\,.$$

Combining the above steps together, we show

$$\begin{aligned} a(\mathrm{M}) &= \sum_{x_1,\ldots,x_p=0}^{\infty} \frac{e^{-\sum_j \mathrm{M}_{jj}x_j - \sum_{i,j:i\neq j} \mathrm{M}_{ij}T_{ij}(x_i,x_j)}}{\prod_j x_j!} < \infty \\ &\leq \sum_{x_1,\ldots,x_p=0}^{\infty} \prod_{j=1}^{p} \frac{C(\mathrm{M}, j)^{x_j}}{x_j!} \\ &= \prod_{j=1}^{p} e^{C(\mathrm{M},j)} \\ &< \infty\,. \end{aligned}$$

Then the integrability condition for any M.

In the following, we prove that $a(\mathrm{M}) < \infty$ for all matrices $\mathrm{M} \in \{\mathrm{M} \in \mathbb{R}^{p\times p} : \mathrm{M} \text{ positive definite}\}$ for the exponential model with square-root transformation described in 3.2.2. We expand the exponentiation of the normalization term as

$$e^{a(\mathrm{M})} = \int_0^{\infty} \cdots \int_0^{\infty} e^{-\sum_{i,j=1}^{p} \mathrm{M}_{ij}\sqrt{x_i x_j}}\, dx_1 \ldots dx_p$$

Note an important property of positive definite matrices is that the smallest ℓ_2-eigenvalue

of M (denoted as $\kappa(\mathrm{M})$) is positive, that is

$$\frac{\mathrm{M}_{ij}\sqrt{x_i x_j}}{\|\mathbf{x}\|_1} \geq \kappa(\mathrm{M}) > 0.$$

We find

$$\begin{aligned} e^{a(\mathrm{M})} &\leq \int_0^\infty \cdots \int_0^\infty e^{-\kappa(\mathrm{M})\,\|\mathbf{x}\|_1}\; dx_1 \ldots dx_p \\ &= \Big(\int_0^\infty e^{-\kappa(\mathrm{M})x} dx \Big)^p \\ &= \kappa(\mathrm{M})^{-p} \;< \infty\,. \end{aligned}$$

Hence, $a(\mathrm{M}) < \infty$ for all positive definite matrices M.

3.7.2 Full Algorithm for the Maximum Likelihood Estimator

In this section, we present the full algorithm, which utilizes a backtracking line search to adaptively select step sizes and incorporate the applicable domain constraint.

Backtracking line search is an inexact line search approach for preventing the step size from being too large. When step size η is too large in the sense that $\widetilde{g}$ is not decreasing "enough", it is reduced by a factor of β $(0 < \beta < 1)$ until $\widetilde{g}(\mathrm{M} - \eta\nabla\widetilde{g}(\mathrm{M})) \leq \widetilde{g}(\mathrm{M}) - \alpha\eta\|\nabla\widetilde{g}(\mathrm{M})\|^2$ is satisfied. Our choice of the backtracking parameters is $\alpha = 0.3$ and $\beta = 0.5$. When additional domain constraint is applicable, we set $g(\widetilde{\mathrm{M}})$ to be infinite for $\widetilde{\mathrm{M}} \notin \mathfrak{M}$ by convention. The inequality of backtracking line search requires $\widetilde{\mathrm{M}}_{k-1} - \eta\nabla\widetilde{g}(\widetilde{\mathrm{M}}_{k-1}) \in \mathfrak{M}$. In a practical implementation, we multiple η by β until $\widetilde{\mathrm{M}}_{k-1} - \eta\nabla\widetilde{g}(\widetilde{\mathrm{M}}_{k-1}) \in \mathfrak{M}$ and then check the inequality.

Algorithm 2: Solving for the maximum likelihood estimator with a backtracking line search

`// ` η `: initial step size`
`// ` α, β`: backtracking parameters`
Input : $\overline{T}(\underline{X}), \eta > 0, \alpha \in (0, 0.5), \beta \in (0, 1)$
Output: $\widehat{\mathrm{M}}$
`// Solve for` M_0
$\mathrm{M}_0 \leftarrow \mathbf{0}_{p\times p}$;
for $i = 1, \ldots, p$ **do**
  $(\mathrm{M}_0)_{ii} \leftarrow \mathrm{argmin}_{m\in\mathbb{R}} \Big\{ m\overline{T}_{ii}(\underline{X}) + \log \int \exp\big(- mT_{ii}(\mathbf{x}) + \xi(x_i)\big) dx_i \Big\}$;
`// Generate sample set` Y `from` $f_{\mathrm{M}_0} = \prod_{i=1}^{p} f_{(\mathrm{M}_0)_{ii}}(x_i)$
for $i = 1, \ldots, p$ **do**
  Generate 10,000 random samples from $f_{(\mathrm{M}_0)_{ii}}(x_i)$ for the i-th coordinate;
`// Apply gradient descent with a backtracking line search to` (3.6)
$k \leftarrow 0$;
$\widetilde{\mathrm{M}}_k \leftarrow \mathrm{M}_0$;
repeat
  $k \leftarrow k + 1$;
  $\nabla\widetilde{g}(\widetilde{\mathrm{M}}_{k-1}) \leftarrow \overline{T}(\underline{X}) - \frac{\sum_{Z\in Y} T(Z) e^{-\langle \widetilde{\mathrm{M}}_{k-1} - \mathrm{M}_0,\, T(Z)\rangle_{\mathrm{tr}}}}{\sum_{Z\in Y} e^{-\langle \widetilde{\mathrm{M}}_{k-1} - \mathrm{M}_0,\, T(Z)\rangle_{\mathrm{tr}}}}$;
  `// Select the stepsize adaptively using a backtracking line search`
  repeat
    $\eta \leftarrow \beta\eta$;
  until $\widetilde{g}(\widetilde{\mathrm{M}}_{k-1} - \eta\nabla\widetilde{g}(\widetilde{\mathrm{M}}_{k-1})) \le \widetilde{g}(\widetilde{\mathrm{M}}_{k-1}) - \alpha\eta \left\|\nabla\widetilde{g}(\widetilde{\mathrm{M}}_{k-1})\right\|^2$;
  $\widetilde{\mathrm{M}}_k \leftarrow \widetilde{\mathrm{M}}_{k-1} - \eta\nabla\widetilde{g}(\widetilde{\mathrm{M}}_{k-1})$;
until $\left|\widetilde{g}(\widetilde{\mathrm{M}}_k) - \widetilde{g}(\widetilde{\mathrm{M}}_{k-1})\right| < 10^{-4}$;
$\widehat{\mathrm{M}} \leftarrow \widetilde{\mathrm{M}}_k$;

Chapter 4

MEASURING SURROGACY IN CLINICAL RESEARCH: WITH AN APPLICATION TO STUDYING SURROGATE MARKERS FOR HIV TREATMENT-AS-PREVENTION

This chapter is a revised version of my work published in Statistics in Biosciences (Zhuang and Chen, 2019) combined with unpublished work.

4.1 Introduction

A surrogate marker in clinical trials is considered to be "a laboratory measurement or physical sign used as a substitute for a clinically meaningful endpoint that measures directly how a patient feels, functions, or survives and that is expected to predict the effect of the therapy" (FDA, 1992). Applications of surrogate markers are motivated by replacing a rare or distant outcome by an earlier or easier-accessed surrogate. In this context, how to validate surrogate markers for a clinically meaningful endpoint are especially important.

In a seminal paper, Prentice (Prentice, 1989) provided a qualitative criterion which defines a surrogate marker to be "a response variable for which a test of the null hypothesis of no relationship to the treatment groups under comparison is also a valid test of the corresponding null hypothesis based on the true endpoint." This criterion was further operationalized. To be specific, let Z, S, and T represent the intervention, the marker, and the clinical endpoint, respectively. It has been shown that as long as

$$P(T \mid S, Z) = P(T \mid S) \tag{4.1}$$

and

$$P(T \mid S) \neq P(T), \tag{4.2}$$

departure from the null hypothesis $P(T \mid Z) = P(T)$ is equivalent to departure from the null hypothesis $P(S \mid Z) = P(S)$ (Prentice, 1989; Freedman et al., 1992). Thus equations (4.1) and (4.2) can be used as the operational criterion for surrogate markers, which is known as the Prentice's criterion. Conceptually, this requires a surrogate marker to fully capture the treatment effect on the clinical endpoint. However, in randomized clinical trials, even when they are well-designed and well-conducted, different surrogacy levels occur (Wang and Taylor, 2002; Kobayashi and Kuroki, 2015):

1. The marker S is a perfect surrogate marker if it captures all the dependence of T on Z in the sense of $P(T \mid S, Z) = P(T \mid S)$, while the distribution of T given Z still depends on the marker S, i.e., $P(T \mid Z, S) \neq P(T \mid Z)$ and S depends on Z, i.e., $P(S \mid Z) \neq P(S)$. This condition of perfect surrogacy is equivalent to Prentice's criterion (Kobayashi and Kuroki, 2015).

2. The marker S is a useless surrogate marker if it does not account for any dependence of T on Z in the sense of $P(T \mid Z, S) = P(T \mid Z)$ or the marker S is independent of the treatment Z, i.e., $P(S \mid Z) = P(S)$.

3. The marker S is a partial surrogate marker if it captures some of the treatment effect on the clinical endpoint but not all of that.

A large percentage of candidate markers may capture only part of the treatment effect in practice. Some of the most common causes include that there are multiple pathways of the disease process, or unanticipated mechanisms of the intervention (Fleming and DeMets, 1996). Thus, from the practical point of view, the degree to which a marker captures the treatment effect on the clinical endpoint provides essential information for evaluating the surrogacy level of a candidate marker. In addition to the qualitative Prentice's criterion, it is important to have quantitative statistical measures to identify and rank the valid markers. In this context, a variety of measures were proposed to quantify the degree of candidate markers' surrogacy level. Contrary to a common misconception, high correlation with the clinical

endpoint does not guarantee a good surrogate marker (Baker and Kramer, 2003; Fleming and DeMets, 1996). A proper quantitative surrogate measure should take into account the multi-aspect dependence among intervention, markers, and the clinical endpoint (Fleming and Powers, 2012).

In this chapter, we first review the main quantitative surrogate measures in Section 4.2. Then we suggest a new surrogacy measure, termed "population surrogacy fraction of treatment effect" (PSF), or simply the ρ-measure in Section 4.3. In essence, the ρ-measure is the proportion of risks associated with the marker-mediate treatment effect. It is closely related with prominent surrogate measures, for example, the proportion of treatment effect explained (PTE) (Freedman et al., 1992; Wang and Taylor, 2002), and the public health concept of population attributable fraction (PAF) (Levin, 1953). Importantly, the ρ-measure enjoys desirable numerical properties and carries an appealing public health interpretation. In addition, we consider a survival setting and define a particular model-free PTE (i.e., F-measure) in Section 4.4. At last, we apply the new measures to HPTN 052 study in Section 4.5, and then conclude this chapter with discussion remarks in Section 4.6. Proofs and additional results are deferred to Section 4.7.

4.2 Review of Quantitative Surrogate Measures

In this section, we review surrogacy quantification methods in the literature. Most of the publications point out that statistical evidence alone is not sufficient to validate a surrogate endpoint; biological evidence is indispensable in the evaluation process. Beyond the commonality, these quantitative methods focus on different aspects of the multi-facet surrogate endpoints. By their rationale, we group the main methods into three frameworks:

- **Adjusted approaches** compare the unadjusted and adjusted model (by surrogate marker distribution) in the context of a single randomized clinical trial. These approaches combine the information of biological mechanism in all subgroups, which measures the population impact and be appealing to public health scientists (Alonso

et al., 2006).

- **Meta-analytic approaches** evaluate how precise the treatment effect on a surrogate predicts that on the clinical endpoints. They focus on the perspective of prediction accuracy and precision.

- **Causal-effect approaches** start from the perspective of causality and infer the causal effect of the marker-mediated pathway. They reflect how a surrogate marker is involved in the intervention's functioning process.

We review the three frameworks in the following subsections.

4.2.1 Adjusted Approaches

PTE (Freedman et al., 1992) is an influential quantitative measure evaluating the extent to which a marker explains the treatment benefit in a controlled trial. In the case with a binary clinical endpoint, it compares the treatment effects estimated from unadjusted and adjusted logistic models in the form of

$$\log \frac{P(T = 1 \mid Z)}{1 - P(T = 1 \mid Z)} = \mu_1 + \beta Z,$$

and

$$\log \frac{P(T = 1 \mid Z, S)}{1 - P(T = 1 \mid Z, S)} = \mu_2 + \beta_s Z + \phi_z S,$$

respectively. Here the notation is the same as that in the Prentice's criterion. The adjusted model assumes no interaction between the surrogate marker and the intervention. Then the PTE, denoted by π, is defined as

$$\pi = \frac{\beta - \beta_s}{\beta}.$$

Conceptually, PTE measures the proportion of treatment effect explained by the surrogate marker. It is readily to show that $\pi = 1$ for a perfect surrogate marker such that $f(T \mid$

$Z, S) = f(T \mid S)$, and $\pi = 0$ for a useless surrogate marker such that $f(T \mid Z, S) = f(T \mid Z)$. However, the value of PTE alone can not determine the surrogacy level of potential markers. A value of PTE near 1 implies that the marker is a good surrogate only if it is known that the intervention effect is primarily mediated by the marker. Values of PTE close to 1 can also happen when the intervention has harmful or toxic effects that are not mediated via the marker (Mildvan et al., 1997). Moreover, the value of PTE is not necessarily restricted to the interval of $[0, 1]$ (Buyse and Molenberghs, 1998; Bycott and Taylor, 1998). For example, if adjusting for the surrogate marker changes the direction of treatment effect on the clinical endpoint (i.e., $\beta_s\beta < 0$), π is then greater than 1. Without the understanding of biological mechanisms, PTE lacks an appropriate interpretation. Furthermore, PTE's denominator is the marginal treatment effect. This format largely increases the estimation variability (Lin et al., 1997; De Gruttola et al., 1997; Bycott and Taylor, 1998; Wang and Taylor, 2002). A highly significant treatment effect, which may not be very likely in practice, is necessary to get a precise estimate of PTE (Freedman et al., 1992; Freedman, 2001).

Beyond the original definition, PTE has been extended in various aspects, for example, generalizing to other models, extending to a model-free version and considering the setting of multiple markers. Lin et al. (Lin et al., 1997) extended PTE to the setting of Cox regression models with time-to-event outcomes; Li et al. (Li et al., 2001) extended PTE to the generalized linear model framework. Interestingly, Li et al. considered the model

$$g(\text{risk}) = \beta_{20} + \beta_{21}Z + \beta_{22}S,$$

and defined PTE on the scale of risk reduction. For example, with a link function $g(x) = \log x$, PTE is defined as the proportion of risk reduction explained by the surrogate marker in the form of

$$\pi = \frac{1 - \exp\left(\beta_{22}E(\Delta_s)\right)}{1 - \exp\left(\beta_{21} + \beta_{22}E(\Delta_s)\right)},$$

where $E(\Delta_s)$ is the expected difference in surrogate values between the treatment and control arms.

The model dependence of PTE is disputable. The adjusted and unadjusted models do not hold simultaneously in general model classes (Lin et al., 1997; Bycott and Taylor, 1998; Wang and Taylor, 2002). For example, Lin, Fleming and De Gruttola showed that the assumption of adjusted Cox model conflicts with the proportional hazard assumption of the unadjusted one (Lin et al., 1997). In fact, neither of the models can be true but serves only as a working model. Furthermore, the adjusted model assumes no interaction between the surrogate marker and the intervention. Although Freedman and Graubard argued that violation of the assumption itself was evidence against the validity of the surrogate marker (Freedman et al., 1992), the assumption weakens the application of PTE as a quantitative measure of surrogacy.

As a route to avoid the model-dependence, Wang and Taylor (Wang and Taylor, 2002) proposed a model-free version for PTE, which is the so-called F-measure. They considered treatment groups A and B for $Z = 1$ and $Z = 0$, respectively. Let $P_A(s)$, $P_B(s)$ be the distributions of surrogate value $S = s$, $g_A(s)$, $g_B(s)$ be the conditional distributions of T given $S = s$ in group A and B, respectively. With a monotonic function $h(\cdot)$, the F-measure is defined as

$$F = \frac{AA - AB}{AA - BB},$$

where

$$\begin{aligned} AA &= h\Big(\textstyle\int_{\Omega_S} g_A(s) dP_A(s) \Big), \\ BB &= h\Big(\textstyle\int_{\Omega_S} g_B(s) dP_B(s) \Big), \\ AB &= h\Big(\textstyle\int_{\Omega_S} g_A(s) dP_B(s) \Big), \end{aligned}$$

with respect to the integration domain of S, Ω_S. Here, $h(\cdot)$, $g_A(\cdot)$ and $g_B(\cdot)$ are user-defined such that $AA - BB$ measures the treatment effect on the clinically meaningful endpoint T. For illustration, example choices were given for specific clinically meaningful endpoints. When T is binary, for example, it was chosen that $h(u) = u$, $g_A(s) = P(T = 1 \mid Z = 1, S =$

s), and $g_B(s) = P(T = 1 \mid Z = 0, S = s)$. In this case, we have

$$\begin{aligned} AA &= \int_{\Omega_S} P(T = 1 \mid Z = 1, S = s) dP(s \mid Z = 1) = P(T = 1 \mid Z = 1), \\ BB &= \int_{\Omega_S} P(T = 1 \mid Z = 0, S = s) dP(s \mid Z = 0) = P(T = 1 \mid Z = 0), \\ AB &= \int_{\Omega_S} P(T = 1 \mid Z = 1, S = s) dP(s \mid Z = 0) := P^*(T = 1 \mid Z = 1), \end{aligned}$$

thus

$$F = \frac{P(T = 1 \mid Z = 1) - P^*(T = 1 \mid Z = 1)}{P(T = 1 \mid Z = 1) - P(T = 1 \mid Z = 0)}.$$

The adjusted probability $P^*(T = 1 \mid Z = 1)$ represents the expected risk in group A if the distribution of S is the same as that in group B. For a perfect surrogate marker, $P(T = 1 \mid Z = 0, S = s) = P(T = 1 \mid Z = 1, S = s)$, then $AB = BB$, and thus $F = 1$. For a useless surrogate marker, $P(T = 1 \mid Z = z, S = s_1) = P(T = 1 \mid Z = z, S = s_2)$ or $P(S = s \mid Z = z) = P(S = s)$, then $AA = AB$, and thus $F = 0$.

An intervention usually acts through multiple pathways so that a single surrogate marker may not be enough to capture all the treatment effect. Instead, multiple markers may play essential roles as a network herein. In the context of multiple markers, overall PTE stays the same except for that the full model adjusts for all the markers now. An interesting question is how we decompose and compare the PTE for each marker among the multiple ones. Using an idea of information separation (Zografos, 1998), Chen et al. (Chen et al., 2003) writes the full model adjusting for multiple markers as

$$g(\text{risk}) = \gamma_0 + (\gamma_Z + \sum_{j=1}^{m} c_j \gamma_j) Z + \sum_{j=1}^{m} \gamma_j (S_j - c_j Z),$$

where c_j is the tuning parameter such that the coefficient of Z and those of S_j are uncorrelated. Then the information of overall treatment effect would be largely separated from that by surrogate markers. The quantity $\gamma_Z + \sum_{j=1}^{m} c_j \gamma_j$ can be considered as the overall treatment effect, and $c_j \gamma_j$ be the treatment effect explained by the j-th marker. The PTE

for the j-th marker is decomposed as

$$\pi = \frac{c_j \gamma_j}{\gamma_Z + \sum_{j=1}^{m} c_j \gamma_j}.$$

Following the information separation idea, Qu and Case (Qu and Case, 2006) further took into account the cause-effect relationship among the potential surrogate markers. For example, if mutually unrelated markers $\{S_2, \ldots, S_m\}$ are believed to have a causal effect on marker S_1, the full model can be decomposed as

$$\begin{aligned} g(\text{risk}) = \gamma_0 &+ \Big(\gamma_Z + a_1\gamma_1 + \sum_{j=2}^{m} b_j(\gamma_j + a_j\gamma_1)\Big) Z + \gamma_1 \Big(S_1 - a_1 Z - \sum_{j=2}^{m} a_j S_j\Big) \\ &+ \sum_{j=2}^{m} (\gamma_j + a_j\gamma_1)(S_j - b_j Z). \end{aligned}$$

Two set of parameters $\{a_j, b_j\}$ are used to separate the information matrix. Then the overall treatment effect $\gamma_Z + a_1\gamma_1 + \sum_{j=2}^{m} b_j(\gamma_j + a_j\gamma_1)$ can be decomposed into three parts: γ_Z, the treatment effect not explained by the markers; $a_1\gamma_1$, the treatment effect explained by S_1; $b_j(\gamma_j + a_j\gamma_1)$, the treatment effect explained by S_j $(j = 2, \ldots, m)$. Furthermore, within the treatment effect by S_j $(j = 2, \ldots, m)$, $b_j\gamma_j$ is the treatment effect independent of S_1 and $b_j a_j \gamma_1$ is that dependent of S_1. Therefore, the PTE for each pathway can be obtained by comparing the corresponding treatment effect with the overall one. It is also a generalization of path analysis (Retherford and Choe, 2011) under generalized linear models and Cox regression models.

4.2.2 Meta-analytic Approaches

Meta-analytic approaches are another category of methods for evaluating surrogate endpoints. A large number of observations is usually necessary for adjusted approaches to provide a definite conclusion of markers' surrogacy (Freedman et al., 1992; Freedman, 2001; Hughes et al., 1995). Assuming the relationship among the triplet (Z, S, T) stays the same,

data from multiple trials are desirable. Beyond that, meta-analytic approaches emphasize the predictive aspect of surrogate endpoints; that is, the treatment effect on clinical endpoints can be predicted by the observed effect on valid surrogate markers. In this perspective, quantification of surrogacy can be achieved by measuring how precise the treatment effect on a surrogate endpoint predicts that on the clinical endpoint.

A straightforward meta-analytic approach is to assess the relationship between the treatment effect on potential markers and that on the clinical endpoint among a set of randomized controlled trials. The idea was first pursued via meta-regression (Boissel et al., 1992; Fleming, 1994; Hughes et al., 1995). Regressing each trial's treatment effect on the clinical outcome versus that on the potential marker portrays the trial-level relationship between the two treatment effects (Daniels and Hughes, 1997). Specifically, let β_i and α_i be the treatment effect on the clinical outcome and that on the potential marker in trial i respectively, it can be assumed that

$$\beta_i \mid \alpha_i \sim \mathcal{N}(\gamma_1 + \gamma_2 \alpha_i, \tau^2).$$

A Bayesian approach was proposed to derive the prediction interval of the treatment effect on the clinical endpoint for a given treatment effect on the surrogate marker (Daniels and Hughes, 1997). The precision of the prediction indicates the surrogate's validity on the trial level. If $\tau^2 = 0$ and $\gamma_2 \neq 0$, the treatment effect on the clinical outcome can be perfectly predicted by that on the surrogate marker. Baker (Baker, 2005) proposed an alternative approach to the linear model. When both the clinical endpoint and the surrogate are binary, the observed treatment effect on the clinical endpoint in trial j can be written as

$$\Delta_{(\text{obs})_j} = \Big(\theta_{j11}\pi_{j1} + \theta_{j10}(1-\pi_{j1})\Big) - \Big(\theta_{j01}\pi_{j0} + \theta_{j00}(1-\pi_{j0})\Big),$$

where $\theta_{jzs} := P(T = 1 \mid S = s, Z = z, \text{trial } j)$ and $\pi_{jz} := P(S = 1 \mid Z = z, \text{trial } j)$. For a trial j that only observes the treatment effect on the surrogate, (i.e., π_{j0} and π_{j1}), the treatment effect on the clinical endpoint can be naturally predicted with information from

another trial i. We write the prediction for j based on i as

$$\widehat{\Delta}_{ij} = \Big(\theta_{i11}\pi_{j1} + \theta_{i10}(1-\pi_{j1})\Big) - \Big(\theta_{i01}\pi_{j0} + \theta_{i00}(1-\pi_{j0})\Big).$$

Overall, the treatment effect of trial j can be predicted by a weighted average of $\widehat{\Delta}_{ij}$ for i ranging over all the previous trials of pertinent. To validate surrogate markers, the average prediction error for the predicted effect of intervention (APEP) was computed. A marker with a small APEP compared to the clinically meaningful treatment difference is considered as a valid surrogate marker.

After that, researchers proposed another idea for meta-analytic method, the concept of trial-level and individual-level surrogacy (Buyse and Molenberghs, 1998; Buyse et al., 2000), to fully describe the potential surrogate marker. In the context of a single study: the ratio of the two treatment effects β/α is defined as the relative effect (RE), which measures the trial-level surrogacy; the coefficient of the marker in the adjusted model is defined as the adjusted association (AA), which measures the individual-level surrogacy (Buyse and Molenberghs, 1998). With normally distributed endpoints, it can be shown that AA/RE = PTE. In general, RE and AA correspond to the trial-level and individual-level component of the composite measure PTE (Buyse and Molenberghs, 1998; Molenberghs et al., 2002). In the context of multiple studies with normal endpoints (Buyse et al., 2000), random effect models are assumed as

$$\begin{aligned} S_{ij} \mid Z_{ij} &= (\mu_s + m_{si}) + (\alpha + a_i)Z_{ij} + \epsilon_{sij}, \\ T_{ij} \mid Z_{ij} &= (\mu_t + m_{ti}) + (\beta + b_i)Z_{ij} + \epsilon_{tij}, \end{aligned}$$

where S_{ij}, T_{ij}, and Z_{ij} are the marker, the clinical endpoint and the intervention assignment, respectively, for the j-th subject in the i-th trial; μ_s and μ_t are fixed intercepts, α and β are fixed effects, m_{si}, m_{ti} and a_i, b_i are random intercepts and effects, respectively. The error

terms $(\epsilon_{sij}, \epsilon_{tij})$ and the random components $(m_{si}, m_{ti}, a_i, b_i)$ are normally distributed as

$$\begin{pmatrix} \epsilon_{sij} \\ \epsilon_{tij} \end{pmatrix} \sim \mathcal{N}\left[\mathbf{0}, \quad \begin{pmatrix} \sigma_{ss} & \sigma_{st} \\ & \sigma_{tt} \end{pmatrix}\right], \quad \begin{pmatrix} m_{si} \\ m_{ti} \\ a_i \\ b_i \end{pmatrix} \sim \mathcal{N}\left[\mathbf{0}, \quad \begin{pmatrix} d_{ss} & d_{st} & d_{sa} & d_{sb} \\ & d_{tt} & d_{ta} & d_{tb} \\ & & d_{aa} & d_{ab} \\ & & & d_{bb} \end{pmatrix}\right].$$

For a new trial k, the treatment effect on the clinical endpoint, i.e., $\beta + b_k$, has

$$\text{var}(\beta + b_k \mid m_{sk}, a_k) = d_{bb} - \begin{pmatrix} d_{sb} \\ d_{ab} \end{pmatrix}^\top \begin{pmatrix} d_{ss} & d_{sa} \\ d_{sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_{sb} \\ d_{ab} \end{pmatrix}.$$

A measure capturing the trial-level surrogacy in the perspective of prediction precision is defined as

$$R^2_{\text{trial}} := \frac{\begin{pmatrix} d_{sb} \\ d_{ab} \end{pmatrix}^\top \begin{pmatrix} d_{ss} & d_{sa} \\ d_{sa} & d_{aa} \end{pmatrix}^{-1} \begin{pmatrix} d_{sb} \\ d_{ab} \end{pmatrix}}{d_{bb}}.$$

A marker with $R^2_{\text{trial}} = 1$ perfectly predicts the treatment effect on the clinical endpoint and is valid on the trial level. The measure is also the coefficient of determinant $R^2_{b_i|m_{si},a_i}$ for the linear model

$$b_i = \lambda_0 + \lambda_1 m_{si} + \lambda_2 a_i + \epsilon_i,$$

where ϵ_i is a normal error. At the individual level, the squared correlation between S and T after adjusting for the treatment is defined as the measure of individual-level surrogacy and denoted as

$$R^2_{\text{indiv}} := \frac{\sigma^2_{st}}{\sigma_{ss}\sigma_{tt}}.$$

The framework of trial-level and individual-level surrogacy has been extended in various directions, including employing alternative modeling approaches, generalizing to other endpoints types, and considering more complex settings with repeated measurements (Burzykowski

et al., 2001; Molenberghs et al., 2001; Renard et al., 2002; Alonso et al., 2003; Burzykowski et al., 2004; Alonso et al., 2006; Renard et al., 2010). The central idea is to assess the trial-level surrogacy by a coefficient of determinant and assess the individual-level surrogacy by an adjusted correlation.

The above meta-analytical methods employ the variability of prediction directly to capture the trial-level surrogacy. Besides, the concept of the surrogate threshold effect (STE) (Burzykowski and Buyse, 2006) provides another pragmatic and interpretable angle to this problem. STE is defined as the treatment effect on S such that the lower bound of the prediction interval attains zero. Naturally, STE can be interpreted as the minimal treatment effect on the surrogate to predict a statistically significant treatment effect on the clinical endpoint. The size of STE depends on the prediction variance and can be viewed as another realization of the trial-level surrogacy.

At the end of this subsection, we discuss an approach bridging the meta-analytical paradigms of assessing individual-level surrogacy and the adjusted approaches originated from the Prentice's criterion. The likelihood reduction factor (LRF) (Alonso et al., 2004) extends the spirit of adjusted approaches to the meta-analytical context. As shown in the previous section, the key idea of adjusted approaches is to capture the surrogacy by comparing the unadjusted model and that adjusting for the surrogate. In the context of multiple trials, it is straightforward to apply the comparison within every single trial. Whereas, how to combine the information is a challenge. Alonso et al. considered two generalized linear models for the trial i and the subject j,

$$g\big(E(T_{ij} \mid Z_{ij})\big) = \mu_{ti} + \beta_i Z_{ij},$$
$$g\big(E(T_{ij} \mid Z_{ij}, S_{ij})\big) = \theta_{0i} + \theta_{1i} Z_{ij} + \theta_{2i} S_{ij}.$$

LRF combines the information from N trials as defined in the form of

$$\text{LRF} = 1 - \frac{1}{N} \sum_{i=1}^{N} \exp\Big(- \frac{G_i^2}{n_i} \Big),$$

where n_i is the number of subjects, and G_i is the log-likelihood ratio statistic comparing the unadjusted model to the adjusted one in trial i. For every single trial, the corresponding G_i summarizes the information gain about the clinical endpoint via using the surrogate. Moreover, the term $1 - \exp(-G_i^2/n_i)$ is a generalized correlation between T and S after adjusting for Z (Kent, 1983). When the endpoints are normal, LRF reduces to R^2_{indiv}, which is the adjusted correlation for assessing individual-level surrogacy in the meta-analytic framework. A follow-up paper proposed a similar measure called the proportion of the information gain (PIG), which compares the full model to the reduced model that only includes the surrogate marker (Qu and Case, 2007). PIG is more relevant to the concept of PTE and sharing its characteristics as a composite measure that combines both the trial-level and individual-level surrogacy.

In summary, the meta-analytic approaches provide a way to combine information from multiple studies, but the application still has some challenges. The analysis requires the trials be of similar interventions, the same marker measurements, and the same clinical outcome. However, the number of trials meeting a strict requirement may be limited. The lack of access to a large number of high-quality studies may lead to substantial estimation errors in the meta-analysis (Burzykowski and Buyse, 2006; Hughes, 2008; Ciani et al., 2017). Even if many previous studies are included in the meta-analysis, the between-study variation may still lead to severe precision loss of the estimated treatment effect on the clinical endpoint (Gail et al., 2000).

4.2.3 Causal-effect Approaches

Causal-effect approaches identify the indirect causal effect mediated by the marker and assess the marker's surrogacy from the perspective of causality. In this section, we review two main approaches under this framework: counterfactual approach (Robins and Greenland, 1992) and principal stratification approach (Frangakis and Rubin, 2002).

The method by Robin and Greenland (Robins and Greenland, 1992) classifies the subjects into different types, based on their potential outcomes under each combination of the

intervention assignment and the response of the intermediate marker. For simplicity, they considered the case when both the intervention and the marker were beneficial to the endpoint. With an additional assumption of no interaction between the intervention and the marker, the direct and indirect effects separate by presenting in different types of subjects. In this way, the classification enables the isolation of the marker-mediated indirect causal effect. Since there is no way to observe all the potential outcomes, the type of each subject is non-identifiable in general. Robin and Greenland showed that the following "exchangeability" conditions are required to identify the indirect effect,

E1 The expected endpoint incidence among the treated subjects without marker signal equals the incidence among the treated subjects with marker signal would have had if the marker had been prevented.

E2 The expected endpoint incidence among the control subjects without marker signal equals the incidence among the control subjects with marker signal would have had if the marker had been prevented.

The exchangeability conditions imply that the subjects of the same intervention group are comparable regardless of their marker level. The conditions rule out the existence of uncontrolled confounding factors between the marker and the endpoint, so that validate the approach of estimating potential incidences by the observed ones. For example, in the context of the women's health initiative study (Writing Group for the Women's Health Initiative Investigators, 2002), the estrogen level mediates the effect of postmenopausal hormone use on coronary heart disease. With the confounding of years after menopause, the group with high or low estrogen level are not expected to be comparable. In the presence of uncontrolled confounding factors, it would be biased to estimate the potential proportions by the observed ones. The authors further generalized the direct effect computation to the case with measured confounding variables. If the assumption E1 and E2 may hold within the stratum by the confounding factors, the fraction of the treatment-induced endpoint that

could be prevented by controlling the marker can be estimated by taking into account both the stratum adjustment and the proportion of each stratum. Note that the standard adjustment approaches are also valid in such an ideal situation. The key of the above model is the stratification based on the potential outcome under each configuration of the intervention assignment and the marker response. It involves nonexistent outcome values and implicitly treats the marker as manipulable. For example, in the above framework, one type is defined as subjects satisfying $S(Z=0)=1, S(Z=1)=1, Y(Z=1,S=1)=1, Y(Z=1,S=0)=1, Y(Z=0,S=1)=1, Y(Z=0,S=0)=1$. But the scenario of $(Z=z,S=0)$ is actually nonexistent if S is not manipulable.

Alternatively, the principal stratification approach (Frangakis and Rubin, 2002) classifies subjects based on the potential marker-responses $(S(Z=0), S(Z=1))$. Specifically, there are three principal strata for a binary marker S, $(S(Z=0), S(Z=1)) = (0,0), (S(Z=0), S(Z=1)) = (0,1), (S(Z=0), S(Z=1)) = (1,1)$. These strata are not affected by the intervention so that the stratified effects are equipped with causal interpretations. However, the price of the "rough" classification is that the principal stratification approach cannot distinguish direct and indirect effects. Instead, the authors introduced the associative and dissociative effects which are more relaxing concepts. The comparison between the ordered set $\{Y_i(Z=0) : S_i(Z=0) \neq S_i(Z=1)\}$ and $\{Y_i(Z=1) : S_i(Z=0) \neq S_i(Z=1)\}$ is called the associative effect, while the comparison between the ordered set $\{Y_i(Z=0) : S_i(Z=0) = S_i(Z=1)\}$ and $\{Y_i(Z=1) : S_i(Z=0) = S_i(Z=1)\}$ is called the dissociative effect. If the dissociative effect is zero, the marker is defined as a principal surrogate. Heuristically, the relative magnitude of the associative and dissociative effects indicate the level of principal surrogacy. Stemmed from this conceptual framework, many works have been inspired and developed for vaccine research and surrogacy quantification (Follmann, 2006; Li et al., 2010; Huang et al., 2013).

In summary, causal-effect approaches provide a perspective on surrogate validation from a causal relationship. A core challenge is the non-identifiability of causal parameters involving missing potential outcomes. Estimation of these parameters via data-completing techniques

requires additional assumptions, which are usually unverifiable. Besides, there is a trade-off between the precision of estimation and the plausibility of assumptions. Strong and less plausible assumptions are typically required for definite conclusions while weak and plausible ones may only lead to wide bounds of the causal effect (Rubin, 2004).

4.3 Population Surrogacy Fraction of Treatment Effect

We suggest a surrogacy measure called "population surrogacy fraction of treatment effect" (PSF). The new measure describes the proportion of risks reduced by eliminating the treatment effect on the surrogate marker. It gauges the population impact and is clinically meaningful in prevention trials. Besides, it is a model-free measure with desirable numerical properties. And interestingly, it reveals a close relationship between the surrogate quantification in clinical research and the mediation analysis in public health.

4.3.1 Motivation and Definition of PSF

Population attributable fraction (PAF) is an epidemiological concept and defined as

$$\varphi = \frac{P(D=1) - P(D=1 \mid E=0)}{P(D=1)},$$

where D and E are the indicator of disease and risk factor exposure, respectively. It measures the potential impact of removing a risk factor on the entire population and is essential to guide policy-making decisions about population-based prevention. PAF is intrinsically close to the approach for validating intermediate endpoints. Schatzkin illustrated the relationship $r_D = \varphi_0 \cdot r_E$ for perfect surrogate endpoints in clinical trials (Schatzkin et al., 1990), where r_D is the percentage reduction of cancer incidence by treatment, φ_0 is the PAF of the control arm, and r_E is the percent reduction of intermediate endpoint exposure by treatment. The equation implies that the intervention works entirely through the intermediate endpoint. However, the connection has not been pursued further. We reveal that F-measure is equivalent to $r_E/r_D \cdot \varphi_0$ and serves as a straightforward realization of using PAF to validate intermediate

endpoints in clinical trials. As a continuation of the idea, we formulate a measure PSF and denote it as the ρ-measure. It captures the magnitude of the treatment effect mediated by the surrogate marker with meaningful signs. Specifically, the ρ-measure is defined as

$$\rho = \frac{P(T = 1 \mid Z = 1) - P^*(T = 1 \mid Z = 1)}{P(T = 1 \mid Z = 1)}, \tag{4.3}$$

where $P^*(T = 1 \mid Z = 1) = \int_{\Omega_S} P(T = 1 \mid Z = 1, s) dP(s \mid Z = 0)$. It is the risk reduction in the treatment group associated with the marker distribution, as $P^*(T = 1 \mid Z = 1)$ is considered to be the direct adjustment of risk according to the marker distribution in the control group. The ρ-measure provides absolute information of the mediated outcome risk. It complements the relative measures, such as the proportion of treatment effect explained that represents the mediated fraction of treatment effect.

Table 4.1: A contingency table showing the notation of subjects in each category

	Marker S	$T = 0$	$T = 1$	Subtotal
Control group	0	n_{000}	n_{001}	n_0
$Z = 0$	1	n_{010}	n_{011}	
Treatment group	0	n_{100}	n_{101}	n_1
$Z = 1$	1	n_{110}	n_{111}	

Table 4.1 summarizes the observations in the two independent intervention groups. The numbers of subjects in the group of $z = 0, 1$ can be assumed to follow a multinomial distribution in the form of

$$(n_{z00}, n_{z01}, n_{z10}, n_{z11}) \sim M_4\Big(n_z, \boldsymbol{p}_z^\top = (p_{z00}, p_{z01}, p_{z10}, p_{z11})\Big).$$

Thus, the empirical estimator of the joint probabilities $\boldsymbol{p} := (\boldsymbol{p}_0^\top, \boldsymbol{p}_1^\top)^\top$ is asymptotically

normally distributed. As n_0, n_1 go to infinity,

$$\widehat{\boldsymbol{p}} \xrightarrow{d} \mathcal{N}\left[\boldsymbol{p}, \begin{pmatrix} \left(\mathrm{diag}(\boldsymbol{p_0}) - \boldsymbol{p_0}\boldsymbol{p_0}^T\right)/n_0 & \mathbf{0} \\ \mathbf{0} & \left(\mathrm{diag}(\boldsymbol{p_1}) - \boldsymbol{p_1}\boldsymbol{p_1}^T\right)/n_1 \end{pmatrix}\right].$$

Let $\widehat{\theta}_1$ and $\widehat{\theta}_2$ be an empirical estimator of $P(T = 1 \mid Z = 1) - P^*(T = 1 \mid Z = 1)$ and $P(T = 1 \mid Z = 1)$, respectively, a natural estimator of the ρ-measure is $\widehat{\rho} := \widehat{\theta}_1/\widehat{\theta}_2$. Specifically,

$$\widehat{\theta}_1 = \widehat{p}_{101} + \widehat{p}_{111} - \frac{\widehat{p}_{000} + \widehat{p}_{001}}{\widehat{p}_{100} + \widehat{p}_{101}}\widehat{p}_{101} - \frac{\widehat{p}_{010} + \widehat{p}_{011}}{\widehat{p}_{111} + \widehat{p}_{110}}\widehat{p}_{111}, \quad \widehat{\theta}_2 = \widehat{p}_{101} + \widehat{p}_{111}.$$

Since the ρ-measure is essentially a ratio and $(\widehat{\theta}_1, \widehat{\theta}_2)$ has an asymptotically bivariate normal distribution, the confidence interval for the ρ-measure can be computed by applying the Fieller's theorem (Fieller, 1940). The covariance matrix of $(\widehat{\theta}_1, \widehat{\theta}_2)$ is consistently estimated by the delta method as

$$\begin{pmatrix} \partial\theta_1(\widehat{\boldsymbol{p}}) \\ \partial\theta_2(\widehat{\boldsymbol{p}}) \end{pmatrix} \begin{pmatrix} \left(\mathrm{diag}(\boldsymbol{p_0}) - \boldsymbol{p_0}\boldsymbol{p_0}^T\right)/n_0 & \mathbf{0} \\ \mathbf{0} & \left(\mathrm{diag}(\boldsymbol{p_1}) - \boldsymbol{p_1}\boldsymbol{p_1}^T\right)/n_1 \end{pmatrix} \begin{pmatrix} \partial\theta_1(\widehat{\boldsymbol{p}}) \\ \partial\theta_2(\widehat{\boldsymbol{p}}) \end{pmatrix}^\top.$$

Let σ_{11}^2 and σ_{22}^2 denote the variances of $\widehat{\theta}_1$ and $\widehat{\theta}_2$, and σ_{12}^2 denote the covariance, a $(1-\alpha)\%$ Fieller confidence interval of ρ satisfies

$$\frac{\left(\widehat{\theta}_1 - \rho\widehat{\theta}_2\right)^2}{\widehat{\sigma}_{11}^2 - 2\rho\widehat{\sigma}_{12}^2 + \rho^2\widehat{\sigma}_{22}^2} \leq Z_{1-\alpha/2}^2$$

with the resulting confidence bounds as (Beyene and Moineddin, 2005)

$$\widehat{\rho} + \frac{k}{1-k}\left(\widehat{\rho} - \frac{\widehat{\sigma}_{12}^2}{\widehat{\sigma}_{22}^2}\right) \pm \frac{Z_{1-\alpha/2}}{\widehat{\theta}_2(1-k)}\sqrt{\widehat{\sigma}_{11}^2 - 2\widehat{\rho}\widehat{\sigma}_{12}^2 + \widehat{\rho}^2\widehat{\sigma}_{22}^2 - k\left(\widehat{\sigma}_{11}^2 - \frac{\widehat{\sigma}_{12}^4}{\widehat{\sigma}_{22}^2}\right)},$$

where $k = Z_{1-\alpha/2}^2\widehat{\sigma}_{22}^2/\widehat{\theta}_2^2$.

4.3.2 The Connection With the Proportion of Treatment Effect Explained

The ρ-measure captures equivalent information with two metrics of the proportion of treatment effect explained, namely, the PTE π-measure and the F-measure. First, the ρ-measure is mathematically equivalent to a counterpart definition of π-measure in terms of relative risks. Consider the log linear models

$$\log P(T = 1 \mid Z) = \mu_1 + \beta Z, \tag{4.4}$$

and

$$\log P(T = 1 \mid Z, S) = \mu_2 + \beta_s Z + \phi_z S. \tag{4.5}$$

As in the original definition of PTE, we assume there is no interaction between Z and S. The π-measure in terms of relative risks can be defined as

$$\pi = \frac{RR - RR_s}{RR}, \tag{4.6}$$

where the relative risk $RR = P(T = 1 \mid Z = 1)/P(T = 1 \mid Z = 0) = e^{\beta}$ and the marker adjusted relative risk

$$RR_s = \frac{P(T = 1 \mid Z = 1, S = 1)}{P(T = 1 \mid Z = 0, S = 1)} = \frac{P(T = 1 \mid Z = 1, S = 0)}{P(T = 1 \mid Z = 0, S = 0)} = e^{\beta_s}.$$

It can be shown that

$$\pi = \frac{P(T = 1 \mid Z = 1) - \int_{\Omega_S} P(T = 1 \mid Z = 1, S = s) dP(s \mid Z = 0)}{P(T = 1 \mid Z = 1)} = \rho.$$

Similarly, the ρ-measure is equivalent to the F-measure up to a factor of relative risk, as shown by

$$\rho = \frac{P(T = 1 \mid Z = 1) - \int_{\Omega_S} P(T = 1 \mid Z = 1, S = s) dP(s \mid Z = 0)}{P(T = 1 \mid Z = 1) - P(T = 1 \mid Z = 0)} \times \frac{RR - 1}{RR}.$$

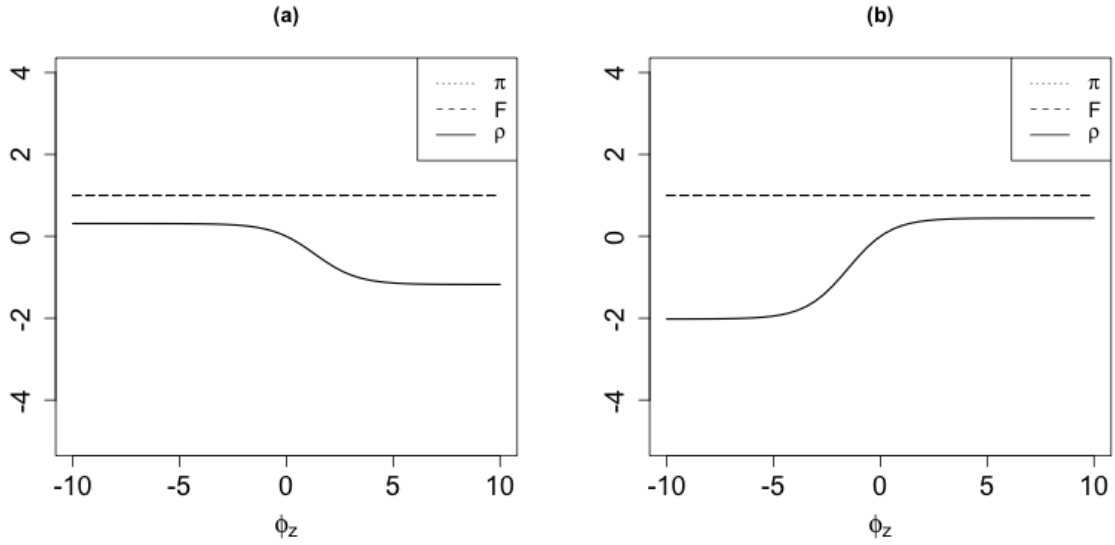

Figure 4.1: The numerical behaviors of the π-measure, F-measure and ρ-measure for perfect surrogate markers. (a) $\alpha_1 < 0$, $\beta_s = 0$; (b) $\alpha_1 > 0$, $\beta_s = 0$.

The detailed derivation is deferred to Section 4.7.1.

In the following, we compare these metrics' numerical characteristics in representative examples. We assume the distributions of treatment Z, marker S and primary endpoint T follow the models

$$\log \frac{P(T = 1 \mid Z, S)}{1 - P(T = 1 \mid Z, S)} = \beta_0 + \beta_s Z + \phi_z S,$$

and

$$\log \frac{P(S = 1 \mid Z)}{1 - P(S = 1 \mid Z)} = \alpha_0 + \alpha_1 Z.$$

The ρ-measure, the PTE π-measure and the F-measure are herein equipped with closed-form formulas. The values for perfect, useless and partial markers are shown in Figure 4.1, 4.2, and 4.3, respectively.

A perfect surrogate marker has a zero-valued adjusted treatment effect (i.e., $\beta_s = 0$). In Figure 4.1, we show that both the π-measure and the F-measure have the value of 1. The ρ-measure equals $(RR - 1)/RR$. It has an upper bound of 1 but is not necessarily positive.

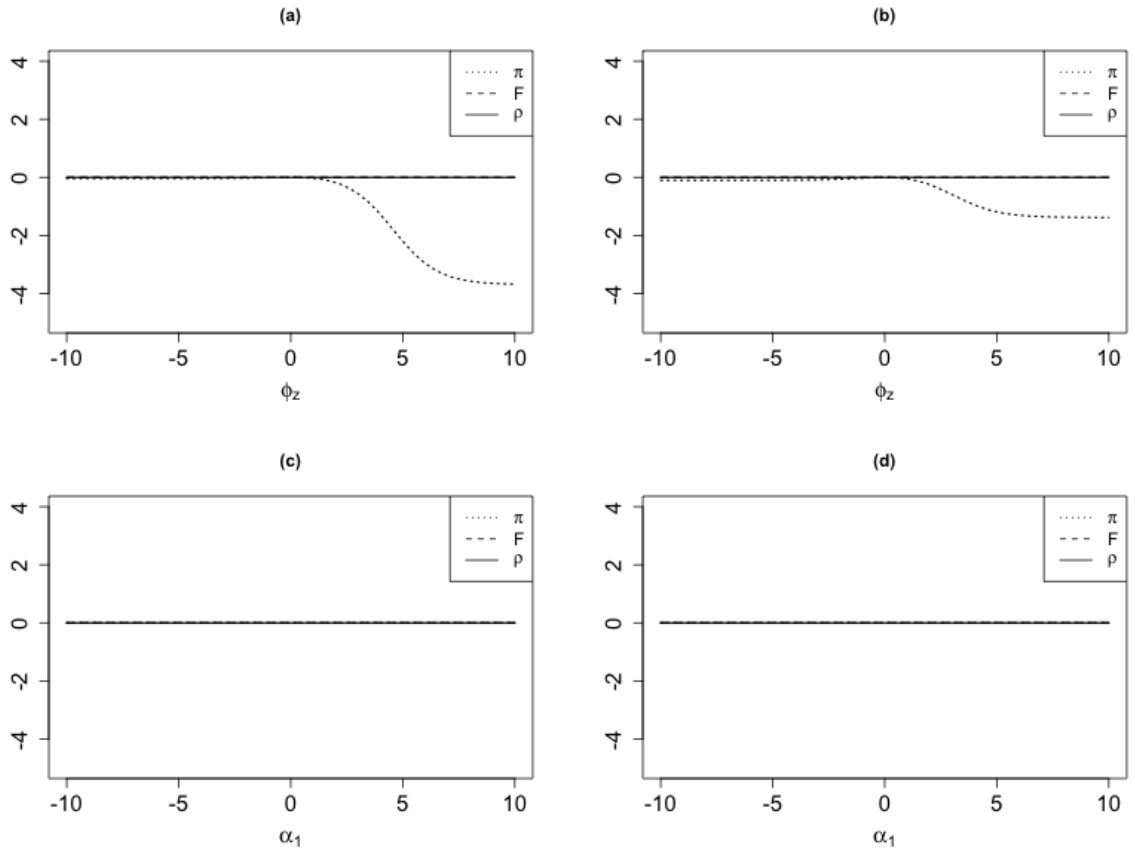

Figure 4.2: The numerical behaviors of the π-measure, F-measure and ρ-measure for useless surrogate markers. (a) $\alpha_1 = 0$, $\beta_s < 0$; (b) $\alpha_1 = 0$, $\beta_s > 0$; (c) $\phi_z = 0$, $\beta_s < 0$; (d) $\phi_z = 0$, $\beta_s > 0$.

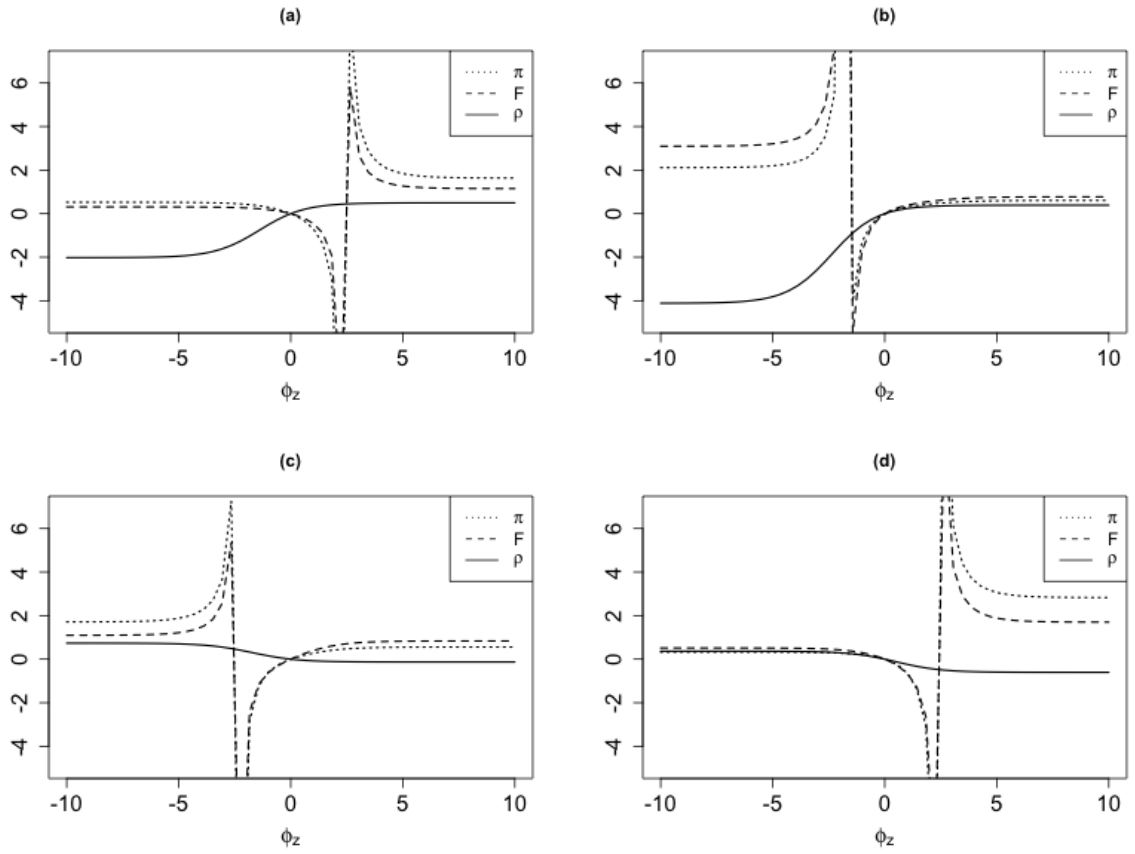

Figure 4.3: The numerical behaviors of the π-measure, F-measure and ρ-measure for partial surrogate markers with typical configurations. (a) $\alpha_1 > 0$, $\beta_s < 0$; (b) $\alpha_1 > 0$, $\beta_s > 0$; (c) $\alpha_1 < 0$, $\beta_s < 0$; (d) $\alpha_1 < 0$, $\beta_s > 0$.

Figure 4.2 illustrates the numerical behavior of a useless marker. The useless marker S is either conditionally independent of the endpoint T given Z (i.e., $\phi_z = 0$), or independent of the treatment Z (i.e., $\alpha_1 = 0$). For the case $\phi_z = 0$, we show $\pi = F = \rho = 0$. For the case $\alpha_1 = 0$, $F = \rho = 0$ still holds; but the PTE π-measure is not necessarily 0. Beyond the above categories, a partial marker has a variety of possible settings. Figure 4.3 covers the typical configurations and curve shapes for partial surrogate markers. Specifically, the marker and the treatment is positively associated (i.e., $\alpha_1 > 0$) in Figure 4.3 (a) and (b), while negatively associated (i.e., $\alpha_1 < 0$) in Figure 4.3 (c) and (d). The adjusted treatment effect is negative (i.e., $\beta_s < 0$) in Figure 4.3 (a) and (c), while positive (i.e., $\beta_s > 0$) in Figure 4.3 (b) and (d). In Figure 4.3, the PTE π-measure and the F-measure have similar patterns. Both of them have the unadjusted treatment effect as the denominator so that they are extremely volatile when the unadjusted treatment effect is close to zero. Besides, when the direct treatment effect and the marker-mediated indirect effect in the same direction, both the π-measure and the F-measure lie in the interval of [0, 1]. Otherwise, they are intractable and not appropriate to rank the surrogacy of markers. On the contrary, the ρ-measure is a smooth function of the adjusted association of the marker and the endpoint. It can be shown that $P(T = 1 \mid Z = 1) - P^*(T = 1 \mid Z = 1) = \big(P(S = 1 \mid Z = 1) - P(S = 1 \mid Z = 0)\big) \times \big(P(T = 1 \mid Z = 1, S = 1) - P(T = 1 \mid Z = 1, S = 0)\big)$. Thus, the ρ-measure's sign indicates the direction of the indirect treatment effect mediated by the surrogate marker. Meanwhile, its absolute value represents the magnitude of the indirect treatment effect. Hence, we interpret the ρ-measure as the proportion of risk reduced if the treatment effect on the surrogate marker is eliminated, which is appropriate to rank candidate markers for public health prevention.

4.3.3 Causal Interpretation

Mediation analysis (VanderWeele, 2016) studying the causal effects among an exposure, mediator, and outcome is a resemblance to surrogate validation. Among the methods for mediation analysis, the counterfactual-based ones receive much recent attention. Under the

counterfactual framework, the total effect (TE) of the exposure can be decomposed into the natural direct effect (NDE) and natural indirect effect (NIE) (Pearl, 2001; Robins and Greenland, 1992). The natural direct effect is defined as the expected change of the outcome when holding the distribution of the mediator but increasing the exposure by one unit. In contrast, the natural indirect effect is defined as the expected change of the outcome when holding the exposure level but setting the mediator's distribution as it would have obtained with the exposure level increased by one unit. Mathematically, the effects of changing exposure from z to z' are written in the form of

$$\begin{aligned} TE_{zz'} &= E\big(T(z', S(z'))\big) - E\big(T(z, S(z))\big), \\ NDE_{zz'} &= E\big(T(z', S(z))\big) - E\big(T(z, S(z))\big), \\ NIE_{zz'} &= E\big(T(z, S(z'))\big) - E\big(T(z, S(z))\big). \end{aligned}$$

In general, $TE_{zz'} = NDE_{zz'} - NIE_{z'z}$ (Pearl, 2012). Assuming a confounding-free scenario, the effects can be estimated from the population data by formulas

$$\begin{aligned} \widetilde{TE}_{zz'} &= E(T \mid z') - E(T \mid z), \\ \widetilde{NDE}_{zz'} &= \int_{\Omega_S} \Big(E(T \mid z', s) - E(T \mid z, s)\Big) dP(s \mid z), \\ \widetilde{NIE}_{zz'} &= \int_{\Omega_S} E(T \mid z, s) d\Big(P(s \mid z') - P(s \mid z)\Big). \end{aligned}$$

The formulas are generally applicable for any non-linear system (Judea, 2010). Recalling the definition of the ρ-measure in (4.3), the numerator can be written as

$$\begin{aligned} \widetilde{TE}_{01} - \widetilde{NDE}_{01} \;=\; & \big(P(T = 1 \mid Z = 1) - P(T = 1 \mid Z = 0)\big) - \\ & \int_{\Omega_S} \Big(P(T = 1 \mid Z = 1, s) - P(T = 1 \mid Z = 0, s)\Big) dP(s \mid Z = 0). \end{aligned}$$

It captures the indirect effect attributable to the marker-mediated pathway. Furthermore, instead of comparing to $P(T = 1 \mid Z = 0)$ to draw conclusions regarding treatment effect,

we can compare $P(T = 1 \mid Z = 1)$ to 0, the situation with no event, i.e., $P(T = 1) = 0$. We observe

$$\rho = \frac{\big(P(T = 1 \mid Z = 1) - 0\big) - \big(\sum_s P(T = 1 \mid Z = 1, s)P(s \mid Z = 0) - 0\big)}{P(T = 1 \mid Z = 1) - 0}.$$

The ρ-measure describes the proportion of risks that are attributable to the marker-mediated treatment effect. It entails the natural indirect effect interpretation.

4.4 The time-varying F-measure

In this section, we bring in the time dimension and define an F-measure for time-to-event outcomes and time-varying internal surrogate markers explicitly. In Sections 4.4.1 and 4.4.2, we introduce the time-varying F-measure and show that it can be estimated using a consistent and asymptotically normal estimator in a non-parametric manner. In Section 4.4.3, we give examples to visualize the change of F-measure with time and conduct Monte Carlo simulation studies to evaluate the proposed non-parametric estimation and inference.

4.4.1 Definition

We consider intervention groups $Z = 1$ and $Z = 0$. Let T represent the time-to-event outcome, X_t represents the value of a candidate marker measured at time t (after randomization). The time-varying F-measure is formulated to evaluate the marker when survival status at time c ($c > t$) is of primary interest. We choose $h(u) = u,\ g_z(x) = P(T \geq c \mid X_t = x, T \geq t, Z = z),\ P_z(s) = P(T \geq c \mid T \geq t, Z = z)$. Then, the F-measure for a time-to-event outcome T writes

$$F(c, t) = \frac{AA - AB}{AA - BB},$$

where

$$AA = P(T \geq c \mid T \geq t, Z = 1),$$
$$BB = P(T \geq c \mid T \geq t, Z = 0),$$
$$AB = \sum_x P(T \geq c \mid X_t = x, T \geq t, Z = 1) P(X_t = x \mid T \geq t, Z = 0).$$

It is a function of time c when survival status is of primary interest, and time t when the surrogate marker is measured. The definition and estimation do not necessarily rely on any model assumption and are exempt from model misspecification.

The time-varying F-measure reflects Prentice's criterion. Namely, the scenarios of perfect markers, in which a marker mediates all the treatment effect, lead to $F(c,t) = 1$; the scenarios of useless markers, in which a marker does not mediate any treatment effect or is independent of intervention in the group of interest, leads to $F(c,t) = 0$. In addition, when the treatment effect mediated by the marker is consistent with the direct treatment effect, the F-measure for a partial marker is guaranteed to be bounded within (0,1). A value outside the ideal bound indicates treatment effects via different pathways are not in the same direction so that the marker is not an appropriate surrogate. (Theoretic results are deferred to Section 4.7.2.)

In summary, the time-varying F-measure evaluates the relative position of the survival probability adjusted by eliminating the treatment effect on a biomarker. It serves as a model-free metric for assessing the proportion of treatment effect explained by the marker.

4.4.2 Estimation and Inference

In the time-varying F-measure, survival probabilities can be estimated by the non-parametric Kaplan-Meier estimator (Kaplan and Meier, 1958). Under the assumption of random censoring, the conditional probability $p_{x|0} := P(X_t = x \mid T \geq t, Z = 0)$ can be estimated by the empirical distribution. Naturally, we propose a plug-in estimator for the defined time-varying

F-measure

$$\widehat{F} = \frac{\widehat{s_1} - \sum_x \widehat{s_{1x}} \cdot \widehat{p_{x|0}}}{\widehat{s_1} - \widehat{s_0}}, \tag{4.7}$$

where $\widehat{s_z}(z = 0, 1)$ and $\widehat{s_{1x}}$ are the Kaplan-Meier estimator for $P(T \geq c \mid T \geq t, Z = z)$ and $P(T \geq c \mid X_t = x, T \geq t, Z = 1)$, respectively. Let $u_{z1} < u_{z2} < \ldots$ be the ordered, distinct times observed on arm z; $n_z(\tau)$ be the number of subjects at risk set at time τ on arm z; $d_z(\tau)$ be the number of events at time τ on arm z. The Kaplan-Meier estimator of survival probabilities reads

$$\widehat{s_z} = \prod_{t<u_{zk}\leq c} \Big(1 - \frac{d_z(u_{zk})}{n_z(u_{zk})}\Big).$$

Similarly, $s_{1x} := Pr(T \geq c \mid X_t = x, T \geq t, Z = 1)$ can be estimated by the Kaplan-Meier estimator in the strata by $Z = 1$ and $X_t = x$ as

$$\widehat{s_{1x}} = \prod_{t<u_{1xk}\leq c} \Big(1 - \frac{d_{1x}(u_{1xk})}{n_{1x}(u_{1xk})}\Big).$$

Under the assumption of random censoring, $p_{x|0} := P(X_t = x \mid T \geq t, Z = 0)$ can be estimated by the empirical distribution as

$$\widehat{p_{x|0}} = \frac{n_{0x}(t)}{n_0(t)},$$

where $n_{0x}(t) = \sum_{i=1}^{n} I(X_t = x, T \geq t, Z = 0)$.

We show the proposed estimator converges weakly to a Gaussian process under the following regularity conditions.

A1. The time c is in a range of (t, τ) for some constant $t > 0$, $0 < \tau < \infty$ such that $s_1(\tau)s_0(\tau) > 0$ and $1 - (1 - H(\tau_-))(1 - G(\tau_-)) < 1$, where H is the distribution function of time-to-event T and G is the distribution function of censoring time U.

A2. Survival probabilities $s_1(\cdot) \neq s_0(\cdot)$ on $(0, \tau)$.

A3. Random censoring: the censoring time U is independent of both the failure time T and

time-varying covariates X_t on $(0, \tau)$.

Theorem 4.4.1. *Under regularity conditions A1-A3, given a time t, $\sqrt{n_t}(\widehat{F}(c) - F(c))$ converges weakly to a zero-mean Gaussian process with covariance function $E\big(\zeta(c)\zeta(c')\big)$ between times c and c', where*

$$\zeta(c) = \frac{s_1 - \sum_x p_{x|0} s_{1x}}{(s_1 - s_0)^2} \cdot \eta_0 + \frac{\sum_x p_{x|0} s_{1x} - s_0}{(s_1 - s_0)^2} \cdot \eta_1 + \frac{1}{s_0 - s_1} \sum_x p_{x|0} \cdot \eta_{1x} + \frac{1}{s_0 - s_1} \sum_x s_{1x} \eta^p_{0x},$$

and

$$\begin{aligned}
\eta_0 &= -s_0(c) I(Z = 0 \mid T \geq t) \int_t^c \frac{dN(u) - Y(u) d\Lambda_0(u)}{E\big(I(Z = 0 \mid T \geq t) Y(u)\big)}, \\
\eta_1 &= -s_1(c) I(Z = 1 \mid T \geq t) \int_t^c \frac{dN(u) - Y(u) d\Lambda_1(u)}{E\big(I(Z = 1 \mid T \geq t) Y(u)\big)}, \\
\eta_{1x} &= -s_1(c | X_t = x, T \geq t) I(Z = 1, X_t = x \mid T \geq t) \int_t^c \frac{dN(u) - Y(u) d\Lambda_{1x}(u)}{E\big(I(Z = 1, X_t = x \mid T \geq t) Y(u)\big)}, \\
\eta^p_{0x} &= \frac{1}{p_0}\big(I(X_t = x, Z = 0 \mid T \geq t) - p_{0x}\big) - \frac{p_{0x}}{p_0^2}\big(I(Z = 0 \mid T \geq t) - p_0\big).
\end{aligned}$$

In the above equations, $N(u) := I(T \leq U, T \leq u)$ denote the observed counting process and $Y(u) := I(T \geq u, U \geq u)$ the at-risk process. The covariance function $E\big(\zeta(c)\zeta(c')\big)$ can be consistently estimated by $1/n_t \sum_{i=1}^{n_t} \hat{\zeta}_i(c)\hat{\zeta}_i(c')$, where $\hat{\zeta}_i(\cdot)$ is the sample versions of $\zeta(\cdot)$.

4.4.3 Numerical Studies

To assess the proposed surrogate measure, we conduct numerical studies motivated by the HIV Prevention Trial Network. The plasma HIV-1 viral load represents the degree of viral burden and is believed to play a crucial role in mediating the benefit of antiretroviral therapy (ART) on HIV-related disease progression and transmission. We consider a viral load measurement dichotomized by a threshold of 1,000 copies per cubic millimeter as the biomarker of interest. In a hypothetical scenario, participants have some HIV-1 exposure at enrollment. The viral load level may increase fast in the follow-up while an effective inter-

vention could delay virus proliferation and further suspend the failure time. We express the above scenario in the following mathematical models. The dichotomous viral load level at time t is modeled as $X_t = I(t \geq t_s)$, where t_s denotes the time when one's viral load shifts from level 0 to 1 after enrollment. We assume t_s follows an exponential distribution with mean μ_z in intervention group z, and a time-varying Cox-Weibull model

$$h(t \mid X_t, Z) = h_0(t) \exp(b_1 Z + b_2 X_t), \tag{4.8}$$

$$h_0(t) = \lambda v t^{v-1}, \tag{4.9}$$

where Z is Bernoulli with success probability of 0.5.

4.4.4 Numerical Examples

In Figure 4.4, we explore the numerical behavior of the time-varying F-measure under the motivation scenario described above. In particular, we assume Model (4.8) with a constant baseline hazard where $h_0(t) = 0.2$. Without loss of generality, we assume $b_1 \leq 0,\ b_2 \geq 0$, and $c = 5$. With the model assumption, the F-measure has a closed-form formula with details in Section 4.7.4. Figure 4.4 (a) has $b_1 = -1$, $t_0 = 0.5$, $t_1 = 2$, varying b_2 visualized F-measures curves from a useless marker with $b_2 = 0$ to partial markers with $b_2 > 0$. Figures 4.4 (b) has $b_2 = 1$, $t_0 = 0.5$, $t_1 = 2$, varying b_1 gives the F-measure curves from a perfect marker with $b_1 = 1$ to partial markers with $b_1 < 0$.

4.4.5 Monte-Carlo Simulation

In this section, we describe our Monte-Carlo simulation to evaluate the proposed non-parametric estimator. We generate failure times for each z-group based on a closed-form approach described in Austin (2012): first, a random value u is generated from the Uniform(0, 1) distribution and the subject-specific shift time t_s is generated from an exponential distribution with mean μ_z; second, if $-\log(u) < \lambda \exp(b_1 z) t_s^v$, we let failure time $T = \Big(-\log(u) / \big(\lambda \exp(b_1 z)\big) \Big)^{1/v}$; otherwise, $T = \Big(\big(-\log(u) - \lambda \exp(b_1 z) t_s^v +$

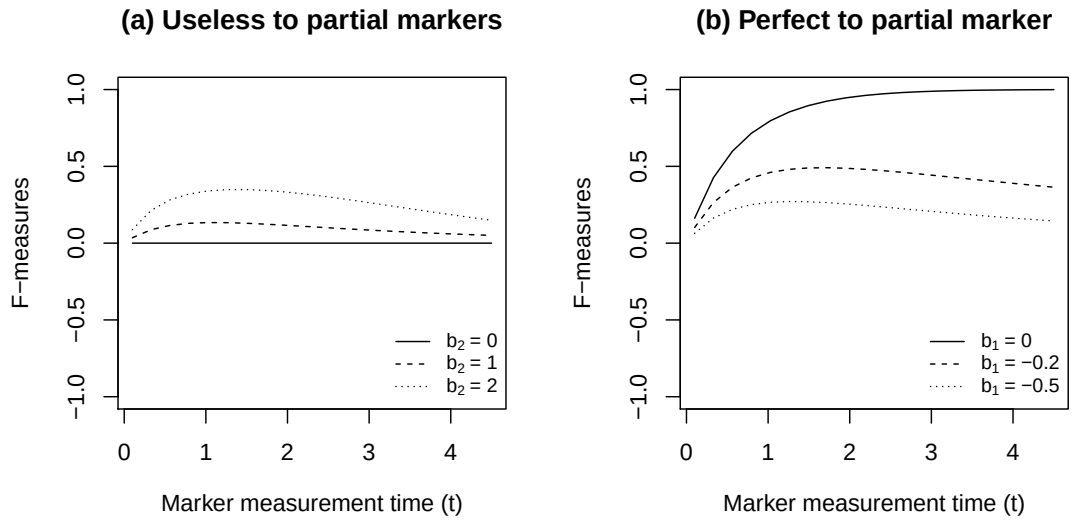

Figure 4.4: F-measure curves describing the surrogacy level for survival status at Year 5.

$\lambda \exp(b_2) \exp(b_1 z) t_s^v \big) / \big(\lambda \exp(b_2) \exp(b_1 z) \big) \Big)^{1/v}$. In addition, we generate the censoring times from Uniform$(0, \tau)$, in which τ is chosen to give a censoring rate of 20%. The censoring is independent of the failure time T, the covariates Z and X_t.

Table 4.2 summarizes the simulation results. We consider the scale parameter v to be 0.8, 1, and 1.2, representing when the hazard is decreasing, constant and increasing with time, respectively. We are interested in the surrogacy level of the t-th year marker measurement for the treatment effect on the c-th year survival probability. For each setting of v, we choose $c = 5$ (years) and $t = 0.25$, 0.5, 1, 2 (years). We show typical scenarios when a surrogate marker is perfect, useless, or partial. A perfect surrogate marker explains all the treatment effect on the clinical endpoint, i.e., $b_1 = 0$; a useless maker is conditionally independent of the failure time given Z, i.e., $b_2 = 0$; a partial marker, the most common scenario in practice, is beyond the above extreme situations. Without loss of generality, we consider a treatment delaying the failure time by both directly affecting the clinical endpoint and suppressing a harmful marker. That is, $b_1 \leq 0$, $b_2 \geq 0$, and $\mu_0 \leq \mu_1$. Specifically, here are

the configurations for the three scenarios in Table 4.2: (1) a perfect marker: $\lambda = 0.02$, $b_1 = 0$, $b_2 = 3$, $t_0 = 3$ months, $t_1 = 30$ months; (2) a useless marker: $\lambda = 0.3$, $b_1 = -1$, $b_2 = 0$, $t_0 = 3$ months, $t_1 = 30$ months; (3) a partial marker: $\lambda = 0.2$, $b_1 = -0.5$, $b_2 = 0.5$, $t_0 = 3$ months, $t_1 = 30$ months. We replicate 1,000 times with 20,000 subjects. Under a large sample size, the estimator is unbiased; its variance accurately reflects the sampling variation; the coverage of 95% Wald-type confidence intervals is close to the nominal probability. One limitation for the non-parametric estimator is lack of efficiency, which is the price for avoiding model misspecification.

Table 4.2: Simulation results under Cox-Weibull distribution. The sample size of the study is 20,000 subjects and the coverage probability is obtained by 1,000 replicates.

c=5, v=0.8						
Scenario	Marker time[a]	True value	Bias	Sampling SE	Mean of SE	Coverage
Perfect	0.25	0.747	0.003	0.052	0.051	0.941
	0.5	0.932	0.002	0.054	0.055	0.956
	1	0.995	0.003	0.059	0.059	0.953
	2	1.000	0.007	0.081	0.080	0.948
Useless	0.25	0.000	0.001	0.034	0.034	0.948
	0.5	0.000	0.001	0.031	0.030	0.951
	1	0.000	0.001	0.023	0.023	0.949
	2	0.000	0.001	0.017	0.016	0.944
Partial	0.25	0.197	0.003	0.051	0.051	0.955
	0.5	0.229	0.001	0.047	0.047	0.953
	1	0.213	0.002	0.038	0.037	0.952
	2	0.167	0.003	0.030	0.030	0.957
c=5, v=1						
Scenario	Marker time[a]	True value	Bias	Sampling SE	Mean of SE	Coverage

Continued on next page

Table 4.2 – continued from previous page

Scenario	Marker time[a]	True value	Bias	Sampling SE	Mean of SE	Coverage
Perfect	0.25	0.743	0.002	0.043	0.044	0.951
	0.5	0.931	0.002	0.044	0.046	0.954
	1	0.995	0.002	0.047	0.048	0.949
	2	1.000	0.004	0.062	0.063	0.945
Useless	0.25	0.000	0.001	0.033	0.034	0.964
	0.5	0.000	0.000	0.030	0.030	0.958
	1	0.000	0.000	0.022	0.022	0.954
	2	0.000	0.001	0.015	0.016	0.954
Partial	0.25	0.204	0.002	0.051	0.051	0.951
	0.5	0.241	0.000	0.046	0.046	0.953
	1	0.228	0.000	0.036	0.036	0.960
	2	0.181	0.001	0.028	0.029	0.951
c=5, v=1.2						
Scenario	Marker time[a]	True value	Bias	Sampling SE	Mean of SE	Coverage
Perfect	0.25	0.742	0.002	0.036	0.036	0.949
	0.5	0.930	0.002	0.036	0.037	0.956
	1	0.995	0.001	0.037	0.038	0.952
	2	1.000	0.003	0.049	0.048	0.940
Useless	0.25	0.000	0.001	0.037	0.037	0.953
	0.5	0.000	0.000	0.032	0.032	0.955
	1	0.000	0.000	0.023	0.023	0.943
	2	0.000	0.000	0.016	0.016	0.953
Partial	0.25	0.219	0.003	0.055	0.054	0.948
	0.5	0.262	0.000	0.049	0.049	0.956
	1	0.252	0.000	0.037	0.038	0.947

Continued on next page

Table 4.2 – continued from previous page

Scenario	Marker time[a]	True value	Bias	Sampling SE	Mean of SE	Coverage
	2	0.203	0.001	0.030	0.030	0.952

[a] the time (year) measuring the marker.

4.5 Surrogate Markers for HIV Prevention Trials

We apply the introduced surrogate measures to the HIV Prevention Trial Network (HPTN) 052 Study. The study enrolled a total of 1763 serodiscordant couples in which one partner was HIV-positive (index participants), and the other was HIV-negative. The index participants were randomized to initiate ART either immediately at enrollment or delayed. For those randomized to the delayed arm, the median delay time is about 2.3 years. In the study, a genetically-linked HIV transmission event in the HIV-negative partners was the primary prevention endpoint. The occurrence of a severe AIDS event, e.g., death, in the index participants was the primary clinical endpoint. The earlier occurrence of either critical clinical outcomes in the index participants or the HIV transmission to the uninfected partners was the key monitoring endpoint (Chen et al., 2012). Here, the composite monitoring endpoint in the five-year follow-up was used as the clinically meaningful endpoint to instantiate the surrogate measures. We perform the statistical analysis on an intention-to-treat basis of the randomization assignment.

The CD4+ count and the viral load in the index participants are generally considered strongly related with the HIV-associated disease progression and transmission. In particular, a high CD4+ count or low viral load indicate of low risk for the monitoring endpoint. The CD4+ count is the first and most widely used biomarker for assessing the degree of immunodeficiency and risk of disease progression in HIV-positive individuals (Mildvan et al., 1997). However, despite being able to predict the disease progression, treatment-induced increasing in CD4+ counts does not capture much of the clinical benefit of antiretroviral therapies (De Gruttola et al., 1993; Lin et al., 1993; Tsiatis et al., 1995; Fleming and DeMets,

1996). Hence, the CD4+ count is not a reliable surrogate endpoint. In contrast, the plasma HIV-1 viral load represents the degree of viral burden and is believed to play a crucial role in mediating the intervention benefit on disease progression and transmission (Katzenstein et al., 1996; Murray et al., 1999).

4.5.1 Application of ρ-measure

In this section, we evaluate the surrogacy of high CD4+ count (> 250 per cubic millimeter) and low viral load (≤ 400 or ≤ 1000 copies per cubic millimeter) in the index participants. We choose the thresholds based on clinical consensus. Besides, considering the complexity of the disease process, candidate surrogate markers are not necessarily a single natural biomarker. It can be composed of multiple ones. As for HIV-associated diseases, the CD4+ count and viral load interact closely with each other and may mediate the treatment effect together (Chen et al., 2010). Here, we also evaluate three composite markers: composite marker I, a health status indicator for a CD4+ count greater than 350 and a viral load less than 1,000; composite marker II, a concerned health status indicator for a CD4+ count less than 250 and a viral load greater than 10,000; composite marker III, a three-level health category. For the composite marker III, an adverse health condition with a CD4+ count less than 250 and a viral load greater than 10,000 is the reference level, a good health condition with a CD4+ count greater than 350 and a viral load less than 1,000 is level 2, and the remaining intermediate condition is level 1. The surrogate measures are naturally extendable for categorical markers.

Table 4.3 shows the markers' prevalence and effect on the clinical endpoint. During the follow-up period, the CD4+ count is slightly increased, and the viral load is effectively suppressed. The immediate ART therapy is beneficial in controlling the viral burden, especially at the beginning of the follow-up. The delayed arm caught up when more participants initiated ART later. Also, the results show that low viral load in the index participants significantly decreased the risk of the composite outcome on the immediate arm. In this study, the composite marker I is dominated by the criterion of the CD4+ count so that it behaves like the marker of the CD4+ count. Similarly, the composite marker II act like the

marker of the viral load due to the viral load's predominating role.

We compute the ρ-measure along with the proportion of treatment effect explained (i.e., the PTE π-measure and the F-measure) for the candidate surrogate markers measured at Year 1 to 4. In Table 4.4, the π-measure for all the candidate markers are estimated to be close to zero with a non-significant confidence interval. So, if judged by the PTE π-measure, the surrogacy of the viral load and the CD4+ count for the monitoring endpoint is almost non-existent. However, the conclusion is contradictory to common knowledge. The contradiction suggests that the model assumptions posited by the π-measure might be of concern here. Likelihood ratio test of the interaction between the candidate marker and the intervention shows significance for all the cases when the π-measure differs much from the model-free surrogate measures. The model-free measures, F and ρ, perform more reasonably and show that the low viral load level at Year 2, 3 and 4 captured some treatment effect on the composite outcome. The biological mechanism of ART and the role of the potential markers in HIV-associated diseases are relatively clear. There is no much concern about the unexpected adverse effect. We believe that ART suppresses the viral load, in turn, the viral suppression prevents the CD4+ count decreasing, the disease progression, and HIV transmission. In this case, the F-measure could be used as a cross-reference supporting the findings of the ρ-measure. Comparing the model-free measures over the four years in Table 3, the ρ-measure for the composite marker I and III have the most substantial value of -1.26 at Year 2. That means the risk of the composite outcome on the immediate ART arm would substantially increase by 126% if the distribution of the composite marker stays the same as on the delayed arm. The ρ-measure of the single marker of a low viral load is very close to that of the composite marker I and III. The finding accords with the biological understanding that the surrogacy mainly owes to the viral load suppression. As a cross-reference, the F-measure for a low viral load level and the composite marker I and III at Year 2 is estimated to be close to 1, which indicates a nearly perfect level of surrogacy. The results that markers at Year 2 have the most substantial value correctly reflect a clinical reality that the median delay time for ART-initiation on the delayed arm is about 2.3 years in the HPTN 052 Study.

Before the median ART-initiation time, the change of viral loads due to ART-initiation on the delayed arm is relatively less impactful; but after the second year of the follow-up, more and more index participants would initiate ART, and the arm would then become similar to the immediate one. The level of surrogacy combines the treatment effect on the marker and the marker's impact on the clinical endpoint. It is expected that the viral load conceptually reaches its largest surrogacy at Year 2.

Table 4.3: The prevalence of HIV-1 RNA level suppression and CD4+ lymphocyte count elevation of the index participants and their effect on the composite outcome during the first five-year follow-up. These markers are measured at each year from Year 1 to 4.

Year	Marker	Delayed ART		Immediate ART	
		Marker prevalence	Log odds ratio (95% CI)	Marker prevalence	Log odds ratio (95% CI)
1	CD4+ > 250	0.95	-0.20 (-0.99, 0.59)	1.00	-
	VL [a] ≤ 400	0.12	0.36 (-0.16, 0.89)	0.89	-0.65 (-1.34, 0.04)
	VL [a] ≤ 1000	0.15	0.35 (-0.15, 0.84)	0.90	-0.74 (-1.44, -0.05)
	Composite I [b]	0.11	0.26 (-0.32, 0.83)	0.87	-0.65 (-1.30, 0.00)
	Composite II [c]	0.04	0.55 (-0.26, 1.36)	0.00	-
	Composite III [d]				
	Level 1	0.85	-0.58 (-1.40, 0.23)	0.13	-
	Level 2	0.11	-0.29 (-1.25, 0.66)	0.87	-
2	CD4+ > 250	0.95	-0.63 (-1.37, 0.10)	0.99	-1.30 (-3.51, 0.91)
	VL [a] ≤ 400	0.34	0.20 (-0.21, 0.61)	0.91	-1.46 (-2.11, -0.82)
	VL [a] ≤ 1000	0.36	0.13 (-0.28, 0.53)	0.92	-1.47 (-2.16, -0.79)
	Composite I [b]	0.28	0.14 (-0.29, 0.56)	0.90	-1.41 (-2.06, -0.77)
	Composite II [c]	0.03	0.72 (-0.16, 1.60)	0.01	1.30 (-0.91, 3.51)

Continued on next page

Table 4.3 – continued from previous page

Year	Marker	Delayed ART		Immediate ART	
		Marker prevalence	Log odds ratio (95% CI)	Marker prevalence	Log odds ratio (95% CI)
	Composite III [d]				
	Level 1	0.68	-0.78 (-1.67, 0.11)	0.09	-0.12 (-2.38, 2.15)
	Level 2	0.28	-0.59 (-1.52, 0.34)	0.90	-1.52 (-3.74, 0.69)
3	CD4+ > 250	0.96	-0.69 (-1.50, 0.13)	1.00	-1.62 (-3.90, 0.66)
	VL [a] ≤ 400	0.62	0.10 (-0.32, 0.53)	0.92	-1.20 (-1.95, -0.46)
	VL [a] ≤ 1000	0.66	0.12 (-0.31, 0.56)	0.94	-1.14 (-1.95, -0.32)
	Composite I [b]	0.56	0.13 (-0.28, 0.55)	0.92	-1.11 (-1.85, -0.36)
	Composite II [c]	0.03	0.27 (-0.83, 1.36)	0.00	2.03 (-0.39, 4.45)
	Composite III [d]				
	Level 1	0.41	-0.36 (-1.48, 0.76)	0.08	-1.12 (-3.62, 1.38)
	Level 2	0.56	-0.20 (-1.31, 0.91)	0.92	-2.15 (-4.57, 0.27)
4	CD4+ > 250	0.97	-0.58 (-1.59, 0.43)	0.99	-2.11 (-3.53, -0.69)
	VL [a] ≤ 400	0.81	0.02 (-0.52, 0.55)	0.93	-1.54 (-2.28, -0.8)
	VL [a] ≤ 1000	0.82	0.02 (-0.52, 0.57)	0.94	-1.72 (-2.47, -0.97)
	Composite I [b]	0.75	-0.01 (-0.49, 0.46)	0.92	-1.57 (-2.26, -0.87)
	Composite II [c]	0.02	-0.68 (-2.73, 1.36)	0.00	3.48 (1.06, 5.90)
	Composite III [d]				
	Level 1	0.23	0.72 (-1.36, 2.80)	0.08	-2.27 (-4.75, 0.22)
	Level 2	0.75	0.67 (-1.38, 2.72)	0.92	-3.68 (-6.11, -1.26)

Continued on next page

Table 4.3 – continued from previous page

Year	Marker	Delayed ART		Immediate ART	
		Marker prevalence	Log odds ratio (95% CI)	Marker prevalence	Log odds ratio (95% CI)

[a] Plasma HIV-1 viral load.

[b] CD4+ count > 350 and viral load < 1000.

[c] CD4+ count < 250 and viral load > 10000. The marker's prevalence and odds ratio are not available at Year 1 since there is no participant on the immediate arm satisfying the bad health condition.

[d] Composite III is a categorical marker with three levels: reference level, CD4+ count < 250 and viral load > 10000; level 2, CD4+ count > 350 and viral load < 1000; level1, the remaining. The Year 1 information is not available for the same reason of composite marker II.

Table 4.4: The proportion of treatment effect explained and the population surrogacy fraction of treatment effect relating the composite outcome during the first five-year follow-up to HIV-1 RNA level suppression and CD4+ lymphocyte count elevation in the index participants at Year 1 to 4.

Year	Marker	π		F		ρ	
		Estimate	95% CI [e]	Estimate	95% CI	Estimate	95% CI
1	CD4+ > 250	0.01	-0.05, 0.07	-0.05	-0.09, -0.02	0.05	0.04, 0.07
	VL [a] ≤ 400	-0.01	-0.53, 0.57	0.51	-0.15, 1.41	-0.57	-1.32, 0.17
	VL [a] ≤ 1000	-0.01	-0.49, 0.53	0.60	-0.09, 1.55	-0.66	-1.43, 0.11
	Composite I [b]	0.13	-0.38, 0.75	0.51	-0.10, 1.34	-0.56	-1.24, 0.11
	Composite II [c]	0.03	-0.03, 0.11	-	-	-	-
	Composite III [d]	0.14	-0.37, 0.78	-	-	-	-

Continued on next page

Table 4.4 – continued from previous page

Year	Marker	π		F		ρ	
		Estimate	95% CI [e]	Estimate	95% CI	Estimate	95% CI
2	CD4+ $>$ 250	0.04	-0.01, 0.12	0.08	-0.13, 0.33	-0.10	-0.37, 0.16
	VL [a] $\leq$ 400	0.13	-0.15, 0.52	1.00	0.33, 2.09	-1.22	-2.01, -0.47
	VL [a] $\leq$ 1000	0.14	-0.14, 0.52	1.02	0.30, 2.18	-1.25	-2.12, -0.43
	Composite I [b]	0.20	-0.13, 0.66	1.03	0.33, 2.17	-1.26	-2.09, -0.47
	Composite II [c]	0.03	-0.01, 0.09	0.05	-0.08, 0.20	-0.06	-0.22, 0.10
	Composite III [d]	0.19	-0.13, 0.65	1.03	0.32, 2.16	-1.26	-2.08, -0.47
3	CD4+ $>$ 250	0.05	-0.01, 0.13	0.10	-0.14, 0.39	-0.12	-0.40, 0.16
	VL [a] $\leq$ 400	0.06	-0.1, 0.26	0.45	0.05, 1.08	-0.51	-0.97, -0.07
	VL [a] $\leq$ 1000	0.03	-0.12, 0.21	0.39	-0.01, 1.00	-0.45	-0.91, 0.01
	Composite I [b]	0.05	-0.13, 0.27	0.47	0.03, 1.16	-0.54	-1.05, -0.04
	Composite II [c]	0.02	-0.03, 0.07	0.10	-0.11, 0.35	-0.11	-0.36, 0.12
	Composite III [d]	0.05	-0.13, 0.27	0.50	0.04, 1.21	-0.57	-1.10, -0.06
4	CD4+ $>$ 250	0.03	-0.01, 0.08	0.08	-0.03, 0.21	-0.09	-0.21, 0.03
	VL [a] $\leq$ 400	0.06	-0.01, 0.18	0.24	0.05, 0.55	-0.28	-0.51, -0.07
	VL [a] $\leq$ 1000	0.07	-0.01, 0.19	0.29	0.07, 0.65	-0.34	-0.60, -0.10
	Composite I [b]	0.09	-0.01, 0.25	0.34	0.10, 0.77	-0.40	-0.69, -0.13
	Composite II [c]	0.01	-0.02, 0.04	0.12	-0.02, 0.31	-0.14	-0.32, 0.02
	Composite III [d]	0.09	-0.01, 0.25	0.39	0.13, 0.84	-0.46	-0.78, -0.17

[a] Plasma HIV-1 viral load.

[b] CD4+ count $>$ 350 and viral load $<$ 1000.

[c] CD4+ count $<$ 250 and viral load $>$ 10000. The model-free measures at Year 1 at not estimable since there is no participant on the immediate arm satisfying the bad health condition.

[d] Composite III is a categorical marker with three levels: reference level, CD4+ count $<$ 250 and viral load $>$ 10000; level 2, CD4+ count $>$ 350 and viral load $<$ 1000; level1, the remaining. The model-free measures at Year 1 are not estimable for the same reason of composite marker II.

[e] The confidence interval of π is obtained by non-parametric bootstrap of 5000 times.

4.5.2 Application of the Time-varying F-measure

We also apply the proposed time-varying F-measure to the HIV Prevention Trial Network (HPTN) 052 study (Cohen et al., 2016). In this application, we consider the survival setting and use plasma viral load (specifically, indicator of a viral load greater than 1,000 copies per cubic millimeter) as a candidate marker and evaluate its surrogacy level on the composite monitoring endpoint in a 3-year follow-up. We estimate the time-varying F-measure for the viral load measured at each of the 2nd to 7th quarter after randomization. Table 4.5 shows the results of the application. Comparing the prevalence of a high viral load between the two arms reveals that ART was very effective in suppressing viral proliferation. In addition, a low viral load significantly decreases the hazard of the composite endpoint on the immediate arm before the treatment effect kicks in on the delayed arm. The time-varying F-measure gradually increases until reaching its maximum at the 6th quarter. This temporal pattern reflects the fact that surrogacy level is a combination of the treatment effect on the marker and the marker effect on the clinical endpoint. On the one hand, it takes time to realize the effect of viral load suppression. On the other hand, as more and more patients on the delayed arm started ART, the difference in marker distribution between two arms are getting smaller. The time-varying F-measure correctly reflects the temporal pattern and the biological mechanism of ART.

Table 4.5: Application to an HIV prevention trial HPTN 052. The proposed time-varying F-measure captures the proportion of treatment effect explained by the plasma HIV-1 viral load.

Marker time[a]	Delayed ART arm		Immediate ART arm		F-measure[d]	
	Prevalence of high VL[b]	Hazard ratio[c]	Prevalence of high VL[b]	Hazard ratio[c]	Estimator	95% CI
2	0.88	1.39	0.08	2.20	0.18	-0.03, 0.39
3	0.88	0.94	0.08	3.21*	0.41	0.13, 0.70

Continued on next page

Table 4.5 – continued from previous page

Marker time[a]	Delayed ART arm		Immediate ART arm		F-measure[d]	
	Prevalence of high viral load	Hazard ratio[c]	Prevalence of high viral load	Hazard ratio[c]	Estimator	95% CI
4	0.87	1.00	0.09	4.49*	0.52	0.09, 0.95
5	0.85	1.59	0.08	5.59*	0.72	0.10, 1.34
6	0.81	2.51*	0.07	4.49*	1.12	-0.42, 2.67
7	0.75	2.11	0.08	6.55*	0.81	-0.92, 2.54

[a] the time point (quarter) when plasma HIV-1 viral load was measured.

[b] plasma viral load $\geq 1,000$ copies per cubic millimeter.

[c] hazard ratio between groups with a viral load higher and lower than 1,000 copies per cubic millimeter. Significant results at the level of 0.05 are marked with $\star$.

[d] the proposed time-varying F-measure.

4.6 Discussion

A surrogate endpoint is a biomarker that can be used in place of a clinically meaningful endpoint for evaluating an intervention. Valid surrogate endpoints that can be measured earlier or easier are desirable for clinical trials with a rare or distal endpoint. However, challenges exist to identify a reliable marker. One particular statistical difficulty arises on how to measure and rank the surrogacy of potential markers quantitatively. In this chapter, we review the main types of statistical surrogate measures, propose a new one rooted in the population attributable fraction (PAF), and extend model-free F-measure to the survival setting.

In general, the challenges of evaluating surrogate markers arise from the complexity of the pathophysiologic process and the intervention mechanism. An intervention may affect the clinical outcome via multiple pathways, either known or unknown. The unknown adverse effect not mediated by the marker is especially problematic, since it may cause the

inconsistency of the treatment effect on the marker and that on the clinical endpoint. The inconsistency has been summarized as "surrogate paradox," in which situation the treatment effect on the surrogate is positive, the surrogate and clinical endpoint is positively correlated, but the treatment effect on the clinical endpoint is negative (Chen et al., 2007). In the paradoxical situation, the results of using surrogate endpoints would be misleading. As discussed in (VanderWeele, 2013), the approaches of surrogacy quantification are potentially vulnerable to the surrogate paradox. In-depth clinical understanding and empirical evidence are needed to avoid the paradox.

4.7 Supplementary Materials

4.7.1 Connection of the ρ-measure with the proportion of treatment effect explained

In the following, we show that both the PTE π-measure and the F-measure can be expressed by the ρ-measure. The π-measure in terms of relative risk in (4.6) can be written as

$$\pi = \frac{P(T=1 \mid Z=1)/P(T=1 \mid Z=0) - RR_s}{P(T=1 \mid Z=1)/P(T=1 \mid Z=0)}.$$

Multiplying $P(T=1 \mid Z=0)$ in both the numerator and denominator gives

$$\pi = \frac{P(T=1 \mid Z=1) - P(T=1 \mid Z=0)RR_s}{P(T=1 \mid Z=1)}. \tag{4.10}$$

Under model (4.4) and (4.5), $RR_s = P(T=1 \mid Z=1, S=s)/P(T=1 \mid Z=0, S=s)$ and is independent of s. We can write

$$\begin{aligned} P(T=1 \mid Z=0)RR_s &= \int_{\Omega_S} P(T=1 \mid Z=0, S=s)dP(s \mid Z=0) \cdot RR_s \\ &= \int_{\Omega_S} P(T=1 \mid Z=0, S=s) \cdot \frac{P(T=1 \mid Z=1, S=s)}{P(T=1 \mid Z=0, S=s)} dP(s \mid Z=0) \\ &= \int_{\Omega_S} P(T=1 \mid Z=1, S=s)dP(s \mid Z=0). \end{aligned} \tag{4.11}$$

Plugging (4.11) into (4.10) reveals

$$\pi = \frac{P(T=1 \mid Z=1) - \int_{\Omega_S} P(T=1 \mid Z=1, S=s) dP(s \mid Z=0)}{P(T=1 \mid Z=1)} = \rho.$$

To show ρ-measure in terms of the F-measure,

$$\rho = \frac{P(T=1 \mid Z=1) - \int_{\Omega_S} P(T=1 \mid Z=1, S=s) dP(s \mid Z=0)}{P(T=1 \mid Z=1) - P(T=1 \mid Z=0)} \times \frac{P(T=1 \mid Z=1) - P(T=1 \mid Z=0)}{P(T=1 \mid Z=1)} = F \times \frac{RR-1}{RR}.$$

Thus $\rho = F \times (RR-1)/RR$.

4.7.2 Range of F-measure

Perfect marker When the marker mediates all the treatment effect, we have $P(T \geq c \mid X_t = x, T \geq t, Z=1) = P(T \geq c \mid X_t = x, T \geq t, Z=0)$. It implies $\sum_x P(T \geq c \mid X_t = x, T \geq t, Z=1) P(X_t = x \mid T \geq t, Z=0) = \sum_x P(T \geq c \mid X_t = x, T \geq t, Z=0) P(X_t = x \mid T \geq t, Z=0)$, and furthermore, $F(c,t) = 1$.

Useless marker When the marker does not mediate any treatment effect, we have $P(T \geq c \mid X_t = x_1, T \geq t, Z=1) = P(T \geq c \mid X_t = x_2, T \geq t, Z=1)$; when the intervention is independent of X_t in the risk set at time point t, we have $P(X_t = x \mid T \geq t, Z=1) = P(X_t = x \mid T \geq t, Z=0)$. Either of the above useless marker conditions leads to $\sum_x P(T \geq c \mid X_t = x, T \geq t, Z=1) P(X_t = x \mid T \geq t, Z=1) = \sum_x P(T \geq c \mid X_t = x, T \geq t, Z=1) P(X_t = x \mid T \geq t, Z=0)$, and furthermore, $F(c,t) = 0$.

Partial marker Without loss of generality, we consider the case $AA - BB > 0$. The following theorem and proof are naturally extendable for the counterpart case $AA - BB < 0$. To give interpretability and links to common instances in clinical trials, we impose three mild assumptions:

B1. X_t in the treatment group and that in the control group are stochastically ordered, $P(X_t \leq x \mid T \geq t, Z = 1) \prec P(X_t \leq x \mid T \geq t, Z = 0)\ \forall x$, or $P(X_t \leq x \mid T \geq t, Z = 1) \succ P(X_t \leq x \mid T \geq t, Z = 0)\ \forall x$.

B2. $P(T \geq c \mid X_t = x, T \geq t, Z = z)$ is monotone with x in the same direction for any given z.

B3. $P(T \geq c \mid X_t = x, T \geq t, Z = z)$ is monotone with z in the same direction for any given x.

In addition, we formulate three conditions:

C1. $P(T \geq c \mid X_t = x, T \geq t, Z = 1) - P(T \geq c \mid X_t = x, T \geq t, Z = 0) > 0$,

C2. $P(X_t \leq x \mid T \geq t, Z = 1) \prec P(X_t \leq x \mid T \geq t, Z = 0)\ \forall x$ and $P(T \geq c \mid X_t = x, T \geq t, Z = 1)$ is increasing with x.

C3. $P(X_t \leq x \mid T \geq t, Z = 1) \succ P(X_t \leq x \mid T \geq t, Z = 0)\ \forall x$ and $P(T \geq c \mid X_t = x, T \geq t, Z = 1)$ is decreasing with x.

Theorem 4.7.1. *With Assumptions B1-B3, if Condition C1 is satisfied, then $F < 1$; if either Condition C2 or C3 is satisfied, then $F > 0$.*

Proof. The proof consists of two steps. First, we show the sufficiency of Condition C1 for $F < 1$. Second, we show the sufficiency of either Condition C2 or C3 for $F > 0$.

Step 1: We first expand $AB - BB$ as $\sum_x \Big(P(T \geq c \mid X_t = x, T \geq t, Z = 1) - P(T \geq c \mid X_t = x, T \geq t, Z = 0)\Big) P(X_t = x \mid T \geq t, Z = 0)$. If Condition C1 is satisfied, we have $AB - BB > 0$. Simple algebra reveals $(AA - AB) - (AA - BB) < 0$. Given $AA - BB > 0$, we conclude

$$F = \frac{AA - AB}{AA - BB} < 1.$$

Step 2: If X_t in the treatment group is stochastically greater than that in the control group, $E_A\big(f(X_t)\big) > E_B\big(f(X_t)\big)$ for bounded and increasing function f. When $P(T \geq c \mid$

$X_t = x, T \geq t, Z = 1)$ is increasing with x, we have $\sum_x P(T \geq c \mid X_t = x, T \geq t, Z = 1)P(X_t \leq x \mid T \geq t, Z = 1) > \sum_x P(T \geq c \mid X_t = x, T \geq t, Z = 1)P(X_t \leq x \mid T \geq t, Z = 0)$, that is, $AA > AB$. Use the same argument, if X_t in the control group is stochastically greater than that in the treatment group and $P(T \geq c \mid X_t = x, T \geq t, Z = 1)$ is decreasing with x, we have $-\sum_x P(T \geq c \mid X_t = x, T \geq t, Z = 1)P(X_t \leq x \mid T \geq t, Z = 0) > -\sum_x P(T \geq c \mid X_t = x, T \geq t, Z = 1)P(X_t \leq x \mid T \geq t, Z = 1)$, that is $-AB > -AA$. Given $AA - BB > 0$, we conclude

$$F = \frac{AA - AB}{AA - BB} > 0.$$

In summary, if Condition C1 and C2 (or C3) are satisfied, F-measure is bounded within (0,1). □

The conditions portray the scenarios that the treatment effect mediated by the marker is in the same direction as the direct treatment effect on the primary endpoint, and furthermore the marginal treatment effect $AA - BB$.

4.7.3 Proof of Theorem 4.4.1

Proof. The proof consists of three steps. First, we decompose $\sqrt{n_t}(\widehat{F} - F)$ as multiple empirical processes. Second, the convergence of each empirical process is derived. Third, we combine the asymptotic results and conclude the proof.

Step 1: We first write $\sqrt{n_t}(\widehat{F} - F)$ as

$$\sqrt{n_t}(\widehat{F} - F) = \sqrt{n_t}\big(\widehat{F}(\widehat{p_{x|0}}) - F(\widehat{p_{x|0}})\big) + \sqrt{n_t}\big(F(\widehat{p_{x|0}}) - F(p_{x|0})\big),$$

and tackle the parts one by one. For the first part, we plug in the estimator (4.7) and obtain

$$\sqrt{n_t}\big(\widehat{F}(\widehat{p_{x|0}}) - F(\widehat{p_{x|0}})\big) = \sqrt{n_t}\Big(\frac{\widehat{s_1} - \sum_x \widehat{s_{1x}} \cdot \widehat{p_{x|0}}}{\widehat{s_1} - \widehat{s_0}} - \frac{s_1 - \sum_x s_{1x}\widehat{p_{x|0}}}{s_1 - s_0}\Big).$$

Rearranging the terms yields

$$\begin{aligned}\sqrt{n_t}\big(\widehat{F}(\widehat{p_{x|0}}) - F(\widehat{p_{x|0}})\big) &= \frac{\sqrt{n_t}}{(\widehat{s_1}-\widehat{s_0})(s_1-s_0)}\Big(s_1(\widehat{s_0}-s_0) - s_0(\widehat{s_1}-s_1) - \\ &\sum_x \widehat{p_{x|0}}(s_1-s_0)(\widehat{s_{1x}}-s_{1x}) - \sum_x \widehat{p_{x|0}}s_{1x}(\widehat{s_0}-s_0) + \sum_x \widehat{p_{x|0}}s_{1x}(\widehat{s_1}-s_1)\Big).\end{aligned}$$

Then we collect the terms and write the equation in the form of $\sqrt{n_t}(\widehat{s_0}-s_0)$, $\sqrt{n_t}(\widehat{s_1}-s_1)$ and $\sqrt{n_t}(\widehat{s_{1x}}-s_{1x})$ as

$$\begin{aligned}\sqrt{n_t}\big(\widehat{F}(\widehat{p_{x|0}}) - F(\widehat{p_{x|0}})\big) &= \frac{s_1 - \sum_x \widehat{p_{x|0}}s_{1x}}{(\widehat{s_1}-\widehat{s_0})(s_1-s_0)}\cdot\sqrt{n_t}(\widehat{s_0}-s_0) + \frac{\sum_x \widehat{p_{x|0}}s_{1x} - s_0}{(\widehat{s_1}-\widehat{s_0})(s_1-s_0)}\cdot\sqrt{n_t}(\widehat{s_1}-s_1) + \\ &\sum_x \frac{\widehat{p_{x|0}}(s_0-s_1)}{(\widehat{s_1}-\widehat{s_0})(s_1-s_0)}\cdot\sqrt{n_t}(\widehat{s_{1x}}-s_{1x}).\end{aligned}$$

Similarly, we write the second part as

$$\sqrt{n_t}(F(\widehat{p_{x|0}}) - F(p_{x|0})) = \sqrt{n_t}(\frac{s_1 - \sum_x s_{1x}\widehat{p_{x|0}}}{s_1-s_0} - \frac{s_1 - \sum_x s_{1x}p_{x|0}}{s_1-s_0}) = \frac{1}{s_0-s_1}\sum_x s_{1x}\cdot\sqrt{n_t}(\widehat{p_{x|0}} - p_{x|0}).$$

Step 2: In this step, we derive the convergence of each empirical process. Under the assumption of random censoring, Kaplan-Meier estimator $\widehat{S}(c)$ satisfies (Fleming and Harrington, 2011)

$$\sqrt{n_t}\big(\widehat{S}(c) - S(c)\big) = \sqrt{n_t}(\mathcal{P}_n - \mathcal{P})\Big(-S(c)\int_0^c \frac{dM(u)}{EY(u)}\Big) + o_p(1),$$

where $M(u) := N(u) - \int_0^u Y(a)d\Lambda(a)$ represents the counting process martingale. Therefore,

the Kaplan-Meier estimator $\widehat{s_0}$, $\widehat{s_1}$ and $\widehat{s_{1x}}$ satisfy

$$\begin{aligned}
\sqrt{n_t}\big(\widehat{s_0}(c) - s_0(c)\big) &\stackrel{d}{=} \sqrt{n_t}(\mathcal{P}_n - \mathcal{P})\Big(- s_0(c) I(Z = 0 \mid T \geq t) \int_t^c \frac{dN(u) - Y(u) d\Lambda_0(u)}{E\big(I(Z = 0 \mid T \geq t) Y(u)\big)}\Big), \\
\sqrt{n_t}\big(\widehat{s_1}(c) - s_1(c)\big) &\stackrel{d}{=} \sqrt{n_t}(\mathcal{P}_n - \mathcal{P})\Big(- s_1(c) I(Z = 1 \mid T \geq t) \int_t^c \frac{dN(u) - Y(u) d\Lambda_1(u)}{E\big(I(Z = 1 \mid T \geq t) Y(u)\big)}\Big), \\
\sqrt{n_t}\big(\widehat{s_{1x}}(c) - s_{1x}(c)\big) &\stackrel{d}{=} \sqrt{n_t}(\mathcal{P}_n - \mathcal{P}) \\
&\quad \Big(- s_{1x}(c) I(Z = 1, X_t = x \mid T \geq t) \int_t^c \frac{dN(u) - Y(u) d\Lambda_{1x}(u)}{E\big(I(Z = 1, X_t = x \mid T \geq t) Y(u)\big)}\Big),
\end{aligned}$$

where $I(\cdot)$ is an indicator function for subjects' group.

Next we write $\sqrt{n_t}(\widehat{p_{x|0}} - p_{x|0})$ in the form of

$$\begin{aligned}
\sqrt{n_t}(\widehat{p_{x|0}} - p_{x|0}) &= \sqrt{n_t}\big(\widehat{P}(X_t = x \mid T \geq t, Z = 0) - P(X_t = x \mid T \geq t, Z = 0)\big) \\
&= \sqrt{n_t}\Big(\frac{\widehat{P}(X_t = x, Z = 0 \mid T \geq t)}{\widehat{P}(Z = 0 \mid T \geq t)} - \frac{P(X_t = x, Z = 0 \mid T \geq t)}{P(Z = 0 \mid T \geq t)}\Big).
\end{aligned}$$

It can be readily shown that

$$\begin{aligned}
\sqrt{n_t}(\widehat{p_{x|0}} - p_{x|0}) &= \frac{1}{\widehat{P}(Z = 0 \mid T \geq t)} \sqrt{n_t}\big(\widehat{P}(X_t = x, Z = 0 \mid T \geq t) - P(X_t = x, Z = 0 \mid T \geq t)\big) - \\
&\quad \frac{P(X_t = x, Z = 0 \mid T \geq t)}{\widehat{P}(Z = 0 \mid T \geq t) P(Z = 0 \mid T \geq t)} \sqrt{n_t}\big(\widehat{P}(Z = 0 \mid T \geq t) - P(Z = 0 \mid T \geq t)\big).
\end{aligned}$$

Let $p_0 := P(Z = 0 \mid T \geq t)$ and $p_{0x} := P(X_t = x, Z = 0 \mid T \geq t)$. Since $\widehat{p_0}$ is a consistent estimator of p_0, we obtain

$$\sqrt{n_t}(\widehat{p_{x|0}} - p_{x|0}) \stackrel{d}{=} \sqrt{n_t}(\mathcal{P}_n - \mathcal{P})\Big(\frac{1}{p_0}\big(I(X_t = x, Z = 0 \mid T \geq t) - p_{0x}\big) - \frac{p_{0x}}{p_0^2}\big(I(Z = 0 \mid T \geq t) - p_0\big)\Big)$$

as a consequence of Slutsky's theorem.

Step 3: In this step, we combine the above results and conclude the convergence of

$\sqrt{n_t}\big(\widehat{F} - F\big)$. We first introduce some notation,

$$\eta_0 = -s_0(c)I(Z=0 \mid T \geq t)\int_t^c \frac{dN(u) - Y(u)d\Lambda_0(u)}{E\big(I(Z=0 \mid T \geq t)Y(u)\big)}, \tag{4.12}$$

$$\eta_1 = -s_1(c)I(Z=1 \mid T \geq t)\int_t^c \frac{dN(u) - Y(u)d\Lambda_1(u)}{E\big(I(Z=1 \mid T \geq t)Y(u)\big)}, \tag{4.13}$$

$$\eta_{1x} = -s_{1x}(c)I(Z=1, X_t = x \mid T \geq t)\int_t^c \frac{dN(u) - Y(u)d\Lambda_{1x}(u)}{E\big(I(Z=1, X_t = x \mid T \geq t)Y(u)\big)}, \tag{4.14}$$

$$\eta_{0x}^p = \frac{1}{p_0}\big(I(X_t = x, Z=0 \mid T \geq t) - p_{0x}\big) - \frac{p_{0x}}{p_0^2}\big(I(Z=0 \mid T \geq t) - p_0\big). \tag{4.15}$$

Combining the above results and apply Slutsky's theorem, we can write

$$\begin{aligned}\sqrt{n_t}(\widehat{F}(c) - F(c)) = \sqrt{n_t}(\mathcal{P}_n - \mathcal{P})\Big(&\frac{s_1 - \sum_x p_{x|0}s_{1x}}{(s_1 - s_0)^2}\cdot\eta_0 + \frac{\sum_x p_{x|0}s_{1x} - s_0}{(s_1 - s_0)^2}\cdot\eta_1 + \\ &\frac{1}{s_0 - s_1}\sum_x p_{x|0}\cdot\eta_{1x} + \frac{1}{s_0 - s_1}\sum_x s_{1x}\eta_{0x}^p\Big) + o_p(1).\end{aligned}$$

It follows that $\sqrt{n_t}(\widehat{F}(c) - F(c))$ converges weakly to a zero-mean Gaussian process with covariance function $\mathbb{E}\{\zeta(c)\zeta(c')\}$ between time points c and c', where

$$\zeta(c) = \frac{s_1 - \sum_x p_{x|0}s_{1x}}{(s_1 - s_0)^2}\cdot\eta_0 + \frac{\sum_x p_{x|0}s_{1x} - s_0}{(s_1 - s_0)^2}\cdot\eta_1 + \frac{1}{s_0 - s_1}\sum_x p_{x|0}\cdot\eta_{1x} + \frac{1}{s_0 - s_1}\sum_x s_{1x}\eta_{0x}^p.$$

The covariance function $\mathbb{E}\{\zeta(c)\zeta(c')\}$ can be consistently estimated by $1/n_t\sum_{i=1}^{n_t}\hat{\zeta}_i(c)\hat{\zeta}_i(c')$ with

$$\hat{\zeta}_i(c) = \frac{s_1 - \sum_x p_{x|0}s_{1x}}{(s_1 - s_0)^2}\cdot\eta_{0i} + \frac{\sum_x p_{x|0}s_{1x} - s_0}{(s_1 - s_0)^2}\cdot\eta_{1i} + \frac{1}{s_0 - s_1}\sum_x p_{x|0}\cdot\eta_{1xi} + \frac{1}{s_0 - s_1}\sum_x s_{1x}\eta_{0xi}^p,$$

where $\eta_{0i}, \eta_{1i}, \eta_{1xi}$ and η_{0xi}^p is the subject i's realization of (4.12)-(4.15), respectively. The

specific forms are

$$\begin{aligned}
\eta_{0i} &= -s_0(c)I(Z_i = 0 \mid T_i \geq t) \int_t^c \frac{dN_i(u) - Y_i(u)d\Lambda_0(u)}{E\big(I(Z_i = 0 \mid T_i \geq t)Y_i(u)\big)} \\
\eta_{1i} &= -s_1(c)I(Z_i = 1 \mid T_i \geq t) \int_t^c \frac{dN_i(u) - Y_i(u)d\Lambda_1(u)}{E\big(I(Z_i = 1 \mid T_i \geq t)Y_i(u)\big)} \\
\eta_{1xi} &= -s_{1x}(c)I(Z_i = 1, X_{ti} = x \mid T_i \geq t) \int_t^c \frac{dN_i(u) - Y_i(u)d\Lambda_{1x}(u)}{E\big(I(Z_i = 1, X_{ti} = x \mid T_i \geq t)Y(u)\big)} \\
\eta^p_{0xi} &= \frac{1}{p_0}\big(I(X_{ti} = y, Z_i = 0 \mid T_i \geq t) - p_{0x}\big) - \frac{p_{0x}}{p_0^2}\big(I(Z_i = 0 \mid T_i \geq t) - p_0\big).
\end{aligned}$$

□

4.7.4 F-measure Under Time-varying Cox-Weibull Model

To facilitate the exploration and understanding of F-measure, we calculate its true value under a time-varying Cox-Weibull model as an illustrative example. We follow the notation described before: Z denotes treatment assignment (control=0; treatment=1); X_t denotes the value of the marker at time t; T denotes the failure time; c denotes the pre-specified time of interest for survival.

We consider the time-varying Cox-Weibull model

$$\begin{aligned}
h(t) &= h_0(t)\exp(b_1 Z + b_2 X_t), \\
h_0(t) &= \lambda v t^{v-1},
\end{aligned}$$

where the marker value satisfies $X_t = I(t \geq t_s)$, and t_s follows an exponential distribution with mean μ_z. The definition of F-measure reads

$$F(c,t) = \frac{P(T \geq c \mid T \geq t, Z = 1) - \sum_x Pr(T \geq c \mid X_t = x, T \geq t, Z = 1)P(X_t = x \mid T \geq t, Z = 0)}{P(T \geq c \mid T \geq t, Z = 1) - P(T \geq c \mid T \geq t, Z = 0)}.$$

With the Bayes rule, the conditional survival probability can be written in the form of

$$P(T \geq c \mid T \geq t, Z = z) = \frac{P(T \geq c \mid Z = z)}{P(T \geq t \mid Z = z)}.$$

In general, for $\tau \in (0, \infty)$,

$$P(T \geq \tau \mid Z = z) = P(T \geq \tau, X_\tau = 1 \mid Z = z) + P(T \geq \tau, X_\tau = 0 \mid Z = z). \quad (4.16)$$

The first term of (4.16)

$$\begin{aligned} & P(T \geq \tau, X_\tau = 1 \mid Z = z) = \int_0^\tau P(T \geq \tau \mid t_s, Z = z) f(t_s \mid Z = z) dt_s \\ = & \int_0^\tau \exp\Big(- \int_0^{t_s} \exp(b_1 z) h_0(u) du - \int_{t_s}^\tau \exp(b_1 z + b_2) h_0(t) du \Big) \cdot f(t_s \mid Z = z) dt_s \end{aligned}$$

Plugging in $\int_a^b h_0(u)du = \lambda(b^v - a^v)$ and the density $f(t_s \mid Z = z) = 1/\mu_z \exp(-t_s/\mu_z)$ leads to

$$P(T \geq \tau, X_\tau = 1 \mid Z = z) = \int_0^\tau \exp\big(- e^{b_1 z} \lambda t_s^v - e^{b_1 z + b_2} \lambda(\tau^v - t_s^v)\big) \frac{1}{\mu_z} \exp(-\frac{t_s}{\mu_z}) dt_s. \quad (4.17)$$

For a general Cox-Weibull model with $v > 0$, there is no closed form formula and we need to refer to numerical evaluation. Similarly, the second term of (4.16)

$$\begin{aligned} & P(T \geq \tau, X_\tau = 0 \mid Z = z) = \int_\tau^\infty P(T \geq \tau \mid t_s, Z = z) f(t_s \mid Z = z) dt_s \\ & = \int_\tau^\infty \exp\Big(- \int_0^\tau \exp(b_1 z) h_0(u) du \Big) \cdot f(t_s \mid Z = z) dt_s \\ & = \int_\tau^\infty \exp\big(- e^{b_1 z} \lambda \tau^v \big) \frac{1}{\mu_z} \exp(-\frac{t_s}{\mu_z}) dt_s \\ & = \exp\Big(- \tau^v \lambda e^{b_1 z} - \tau / t_z \Big). \end{aligned} \quad (4.18)$$

So far, with the equations (4.17) and (4.18), we can calculate $P(T \geq c \mid T \geq t, Z = 1)$ and $P(T \geq c \mid T \geq t, Z = 0)$.

Next, the adjusted probability in the numerator of the F-measure $\sum_x P(T \geq c \mid X_t = x, T \geq t, Z = 1)P(X_t = x \mid T \geq t, Z = 0) = \sum_x P(T \geq c \mid X_t = x, T \geq t, Z = 1)P(X_t = x \mid T \geq t, Z = 0)$. We use Bayes rule and obtain

$$P(T \geq c \mid X_t = 1, T \geq t, Z = 1) = \frac{P(T \geq c, X_t = 1 \mid Z = 1)}{P(T \geq t, X_t = 1 \mid Z = 1)}. \tag{4.19}$$

The denominator of equation (4.19) is shown in (4.17). The numerator of equation (4.19) satisfies

$$\begin{aligned} & P(T \geq c, X_t = 1 \mid Z = 1) = P(T \geq c, t_s \leq t \mid Z = 1) \\ = & \int_0^t P(T \geq c \mid t_s, Z = 1) f(t_s \mid Z = 1) dt_s \\ = & \int_0^t \exp\Big(- \int_0^{t_s} \exp(b_1) h_0(u) du - \int_{t_s}^c \exp(b_1 + b_2) h_0(u) du \Big) \cdot f(t_s \mid Z = 1) dt_s \end{aligned}$$

Plugging in $\int_a^b h_0(u)du = \lambda(b^v - a^v)$ and the density $f(t_s \mid Z = z) = 1/\mu_z \exp(-t_s/\mu_z)$ leads to

$$P(T \geq c, X_t = 1 \mid Z = 1) = \int_0^t \exp\big(- e^{b_1} \lambda t_s^v - e^{b_1 + b_2} \lambda (c^v - t_s^v) \big) \cdot \frac{1}{\mu_1} \exp(-\frac{t_s}{\mu_1}) dt_s. \tag{4.20}$$

Following the same lines, we work on

$$P(T \geq c \mid X_t = 0, T \geq t, Z = 1) = \frac{P(T \geq c, X_t = 0 \mid Z = 1)}{P(T \geq t, X_t = 0 \mid Z = 1)}. \tag{4.21}$$

The numerator of equation (4.21) reads

$$P(T \geq c, X_t = 0 \mid Z = 1) = P(T \geq c \mid Z = 1) - P(T \geq c, X_t = 1 \mid Z = 1), \tag{4.22}$$

and the denominator reads

$$P(T \geq t, X_t = 0 \mid Z = 1) = P(T \geq t \mid Z = 1) - P(T \geq t, X_t = 1 \mid Z = 1). \tag{4.23}$$

To simplify the maths, we introduce the following notation

$$\begin{aligned}
A(\tau, z) &:= P(T \geq \tau, X_\tau = 1 \mid Z = z) = \int_0^\tau \exp\big(- e^{b_1 z}\lambda t_s^v - e^{b_1 z + b_2}\lambda(\tau^v - t_s^v)\big)\frac{1}{\mu_z}\exp(-\frac{t_s}{\mu_z})dt_s, \\
B(\tau, z) &:= P(T \geq \tau, X_\tau = 0 \mid Z = z) = \exp\Big(- \tau^v \lambda e^{b_1 z} - \tau/\mu_z\Big), \\
C &:= P(T \geq c, X_t = 1 \mid Z = 1) = \int_0^t \exp\big(- e^{b_1}\lambda t_s^v - e^{b_1 + b_2}\lambda(c^v - t_s^v)\big)\frac{1}{\mu_1}\exp(-\frac{t_s}{\mu_1})dt_s.
\end{aligned}$$

With the above notation, equation (4.16) writes

$$P(T \geq \tau \mid Z = z) = A(\tau, z) + B(\tau, z),$$

equation (4.19) writes

$$P(T \geq c \mid X_t = 1, T \geq t, Z = 1) = \frac{C}{A(t, 1)},$$

and equation (4.21) writes

$$P(T \geq c \mid X_t = 0, T \geq t, Z = 1) = \frac{A(c, 1) + B(c, 1) - C}{A(t, 1) + B(t, 1) - A(t, 1)} = \frac{A(c, 1) + B(c, 1) - C}{B(t, 1)}.$$

The remaining parts in the F-measure definition are

$$\begin{aligned}
P(X_t = 1 \mid T \geq t, Z = 0) &= \frac{P(X_t = 1, T \geq t \mid Z = 0)}{P(T \geq t \mid Z = 0)} = \frac{A(t, 0)}{A(t, 0) + B(t, 0)}, \\
P(X_t = 0 \mid T \geq t, Z = 0) &= 1 - P(X_t = 1 \mid T \geq t, Z = 0) = \frac{B(t, 0)}{A(t, 0) + B(t, 0)}.
\end{aligned}$$

Gathering all the pieces together, we have

$$
\begin{aligned}
&P(T \geq c \mid T \geq t, Z = 1) = \frac{A(c,1) + B(c,1)}{A(t,1) + B(t,1)},\\
&P(T \geq c \mid T \geq t, Z = 0) = \frac{A(c,0) + B(c,0)}{A(t,0) + B(t,0)},\\
&Pr(T \geq c \mid X_t = 1, T \geq t, Z = 1)P(X_t = 1 \mid T \geq t, Z = 0) = \frac{CA(t,0)}{A(t,1)\big(A(t,0) + B(t,0)\big)},\\
&Pr(T \geq c \mid X_t = 0, T \geq t, Z = 1)P(X_t = 0 \mid T \geq t, Z = 0) = \frac{\big(A(c,1) + B(c,1) - C\big)B(t,0)}{B(t,1)\big(A(t,0) + B(t,0)\big)}.
\end{aligned}
$$

When $v = 1$ (i.e., the failure time follows an exponential distribution), terms A and C are equipped with closed-form formula in the form of

$$
\begin{aligned}
A(\tau, z) &:= \frac{\exp(-\tau\lambda e^{b_1 z + b_2}) - \exp(-\tau(\lambda e^{b_1 z} + 1/\mu_z))}{1 + \lambda\mu_z e^{b_1 z}(1 - e^{b_2})},\\
B(\tau, z) &:= \exp\Big(- \tau\big(\lambda e^{b_1 z} + 1/t_z\big)\Big),\\
C &:= \frac{\exp(-c\lambda e^{b_1 + b_2})\Big(1 - \exp\big(- \frac{t}{\mu_1} + \lambda t e^{b_1}(-1 + e^{b_2})\big)\Big)}{1 + \lambda\mu_1 e^{b_1}(1 - e^{b_2})}.
\end{aligned}
$$

BIBLIOGRAPHY

Akdemir, D. and Gupta, A. (2011). Array Variate Random Variables with Multiway Kronecker Delta Covariance Matrix Structure. *J. Algebr. Stat.*, 2(1):98–113.

Alonso, A., Geys, H., Molenberghs, G., Kenward, M. G., and Vangeneugden, T. (2003). Validation of Surrogate Markers in Multiple Randomized Clinical Trials with Repeated Measurements. *Biom. J.*, 45(8):931–945.

Alonso, A., Molenberghs, G., Burzykowski, T., Renard, D., Geys, H., Shkedy, Z., Tibaldi, F., Abrahantes, J. C., and Buyse, M. (2004). Prentice's Approach and the Meta-analytic Paradigm: a Reflection on the Role of Statistics in the Evaluation of Surrogate Endpoints. *Biometrics*, 60(3):724–728.

Alonso, A., Molenberghs, G., Geys, H., Buyse, M., and Vangeneugden, T. (2006). A Unifying Approach for Surrogate Marker Validation Based on Prentice's Criteria. *Stat. Med.*, 25(2):205–221.

Aravkin, A., Burke, J., Drusvyatskiy, D., Friedlander, M., and MacPhee, K. (2017). Foundations of Gauge and Perspective Duality. *arXiv:1702.08649*.

Austin, P. C. (2012). Generating Survival Times to Simulate Cox Proportional Hazards Models with Time-varying Covariates. *Stat. Med.*, 31(29):3946–3958.

Baker, S. G. (2005). A Simple Meta-analytic Approach for Using a Binary Surrogate Endpoint to Predict the Effect of Intervention on True Endpoint. *Biostatistics*, 7(1):58–70.

Baker, S. G. and Kramer, B. S. (2003). A Perfect Correlate Does not a Surrogate Make. *BMC Med. Res. Methodol.*, 3(1):605.

Benjamin, P., Zeestraten, E., Lambert, C., Chis Ster, I., Williams, O. A., Lawrence, A. J., Patel, B., MacKinnon, A. D., Barrick, T. R., and Markus, H. S. (2016). Progression of MRI Markers in Cerebral Small Vessel Disease: Sample Size Considerations for Clinical Trials. *J. Cereb. Blood Flow Metab.*, 36(1):228–240.

Bercu, B., Delyon, B., and Rio, E. (2015). *Concentration Inequalities for Sums and Martingales*. Springer.

Berk, R. (1972). Consistency and Asymptotic Normality of MLE's for Exponential Models. *Ann. Math. Stat.*, 43(1):193–204.

Besag, J. (1974). Spatial Interaction and the Statistical Analysis of Lattice Systems. *J. Roy. Statist. Soc. Ser. B*, 36:192–236. With discussion by D. R. Cox, A. G. Hawkes, P. Clifford, P. Whittle, K. Ord, R. Mead, J. M. Hammersley, and M. S. Bartlett and with a reply by the author.

Beyene, J. and Moineddin, R. (2005). Methods for Confidence Interval Estimation of a Ratio Parameter With Application to Location Quotients. *BMC Med. Res. Methodol.*, 5(1):32.

Bogdan, M., van den Berg, E., Sabatti, C., Su, W., and Candès, E. (2015). SLOPE-adaptive Variable Selection via Convex Optimization. *Ann. Appl. Stat.*, 9(3):1103–1140.

Boissel, J. P., Collet, J. P., Moleur, P., and Haugh, M. (1992). Surrogate Endpoints: A Basis for a Rational Approach. *Eur. J. Clin. Pharmacol.*, 43(3):235–244.

Brown, L. (1986). Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory. *Lecture Notes-Monograph Series*, 9:i–279.

Brush, S. (1967). History of the Lenz-Ising Model. *Rev. Mod. Phys.*, 39(4):883.

Bu, Y. and Lederer, J. (2017). Integrating Additional Knowledge Into Estimation of Graphical Models. *arXiv:1704.02739v2*.

Bühlmann, P. (2013). Statistical Significance in High-dimensional Linear Models. *Bernoulli*, 19(4):1212–1242.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-dimensional Data: Methods, Theory and Applications*. Springer.

Buhr, K. A. (2012). Surrogate End Points in Secondary Analyses of Cardiovascular Trials. *Prog. Cardiovasc. Dis.*, 54(4):343–350.

Bunea, F., Tsybakov, A., and Wegkamp, M. (2007a). Aggregation for Gaussian Regression. *Ann. Statist.*, 35(4):1674–1697.

Bunea, F., Tsybakov, A., and Wegkamp, M. (2007b). Sparsity Oracle Inequalities for the Lasso. *Electron. J. Stat.*, 1:169–194.

Burzykowski, T. and Buyse, M. (2006). Surrogate Threshold Effect: an Alternative Measure for Meta-analytic Surrogate Endpoint Validation. *Pharm. Stat.*, 5(3):173–186.

Burzykowski, T., Molenberghs, G., and Buyse, M. (2004). The Validation of Surrogate End Points by Using Data From Randomized Clinical Trials: a Case-study in Advanced Colorectal Cancer. *J. Roy. Statist. Soc. Ser. A*, 167(1):103–124.

Burzykowski, T., Molenberghs, G., Buyse, M., Geys, H., and Renard, D. (2001). Validation of Surrogate End Points in Multiple Randomized Clinical Trials With Failure Time End Points. *J. Roy. Statist. Soc. Ser. C*, 50(4):405–422.

Buyse, M. and Molenberghs, G. (1998). Criteria for the Validation of Surrogate Endpoints in Randomized Experiments. *Biometrics*, 54(3):1014.

Buyse, M., Molenberghs, G., Burzykowski, T., Renard, D., and Geys, H. (2000). The Validation of Surrogate Endpoints in Meta-analyses of Randomized Experiments. *Biostatistics*, 1(1):49–67.

Buyse, M., Sargent, D. J., Grothey, A., Matheson, A., and De Gramont, A. (2010). Biomarkers and Surrogate End Points — the Challenge of Statistical Validation. *Nat. Rev. Clin. Oncol.*, 7(6):309–317.

Bycott, P. W. and Taylor, J. M. G. (1998). An Evaluation of a Measure of the Proportion of the Treatment Effect Explained by a Surrogate Marker. *Control. Clin. Trials*, 19(6):555–568.

Chakravarty, A. (2005). *The Evaluation of Surrogate Endpoints*. Springer.

Chatterjee, S. (2013). Assumptionless Consistency of the Lasso. *arXiv:1303.5817*.

Chatterjee, S. (2014). A New Perspective on Least Squares Under Convex Constraint. *Ann. Statist.*, 42(6):2340–2381.

Chen, C., Wang, H., and Snapinn, S. M. (2003). Proportion of Treatment Effect (PTE) Explained by a Surrogate Marker. *Stat. Med.*, 22(22):3449–3459.

Chen, H., Geng, Z., and Jia, J. (2007). Criteria for Surrogate End Points. *J. Roy. Statist. Soc. Ser. B*, 69(5):919–932.

Chen, Y. Q., Masse, B., Wang, L., Ou, S.-S., Li, X., Donnell, D., McCauley, M., Gamble, T., Ribauldo, H. J., Cohen, M. S., and Fleming, T. R. (2012). Statistical Considerations for the HPTN 052 Study to Evaluate the Effectiveness of Early Versus Delayed Antiretroviral Strategies to Prevent the Sexual Transmission of HIV-1 in Serodiscordant Couples. *Contemp. Clin. Trials*, 33(6):1280–1286.

Chen, Y. Q., Young, A., Brown, E. R., Chasela, C. S., Fiscus, S. A., Hoffman, I. F., Valentine, M., Emel, L., Taha, T. E., Goldenberg, R. L., et al. (2010). Population Attributable Fractions for Late Postnatal Mother-to-child Transmission of HIV-1 in Sub-Saharan Africa. *J. Acquir. Immune Defic. Syndr.*, 54(3):311.

Ciani, O., Buyse, M., Drummond, M., Rasi, G., Saad, E. D., and Taylor, R. S. (2017). Time to Review the Role of Surrogate End Points in Health Policy: State of the Art and the Way Forward. *Value in Health*, 20(3):487–495.

Cohen, M. S., Chen, Y. Q., McCauley, M., Gamble, T., Hosseinipour, M. C., Kumarasamy, N., Hakim, J. G., Kumwenda, J., Grinsztejn, B., Pilotto, J. H., et al. (2016). Antiretroviral Therapy for the Prevention of HIV-1 Transmission. *N. Engl. J. Med.*, 375(9):830–839.

Colburn, W., DeGruttola, V. G., DeMets, D. L., Downing, G. J., Hoth, D. F., Oates, J. A., Peck, C. C., Schooley, R. T., Spilker, B. A., Woodcock, J., and Zeger, S. L. (2001). Biomarkers and Surrogate Endpoints: Preferred Definitions and Conceptual Framework. Biomarkers Definitions Working Group. *Clin. Pharmacol. Ther.*, 69:89–95.

Cook, T. D. and DeMets, D. L. (2007). *Introduction to Statistical Methods for Clinical Trials*. CRC Press.

Cutts, C. S. and Eglen, S. J. (2014). Detecting Pairwise Correlations in Spike Trains: An Objective Comparison of Methods and Application to the Study of Retinal Waves. *J. Neurosci.*, 34(43):14288–14303.

Dalalyan, A., Hebiri, M., and Lederer, J. (2017). On the Prediction Performance of the Lasso. *Bernoulli*, 23(1):552–581.

Dalalyan, A. and Tsybakov, A. (2007). Aggregation by Exponential Weighting and Sharp Oracle Inequalities. In *Learning theory*, volume 4539, pages 97–111.

Dalalyan, A. and Tsybakov, A. (2012a). Mirror Averaging with Sparsity Priors. *Bernoulli*, 18:914–944.

Dalalyan, A. and Tsybakov, A. (2012b). Sparse Regression Learning by Aggregation and Langevin Monte-Carlo. *J. Comput. System Sci.*, 78:1423–1443.

Daniels, M. J. and Hughes, M. D. (1997). Meta-analysis for the evaluation of potential surrogate markers. *Stat. Med.*, 16(17):1965–82.

De Gruttola, V., Fleming, T., Lin, D. Y., and Coombs, R. (1997). Perspective: Validating Surrogate Markers—Are We Being Naive? *J. Infect. Dis.*, 175(2):237–246.

De Gruttola, V., Wulfsohn, M., FischI, M. A., and Tsiatis, A. (1993). Modeling the Relationship Between Survival and CD4 Lymphocytes in Patients with AIDS and AIDS-related Complex. *J. Acquir. Immune Defic. Syndr.*, 6(4):359–365.

De Lathauwer, L., De Moor, B., and Vandewalle, J. (2000). A Multilinear Singular Value Decomposition. *SIAM J. Matrix Anal. Appl.*, 21(4):1253–1278.

Demas, J., Eglen, S. J., and Wong, R. O. (2003). Developmental Loss of Synchronous Spontaneous Activity in the Mouse Retina Is Independent of Visual Experience. *J. Neurosci.*, 23(7):2851–2860.

Downing, N. S., Aminawung, J. A., Shah, N. D., Krumholz, H. M., and Ross, J. S. (2014). Clinical Trial Evidence Supporting FDA Approval of Novel Therapeutic Agents, 2005-2012. *J. Am. Med. Assoc.*, 311(4):368–377.

Drton, M. and Maathuis, M. H. (2017). Structure Learning in Graphical Modeling. *Annu. Rev. Stat. Appl.*, 4:365–393.

Echt, D. S., Liebson, P. R., Mitchell, L. B., Peters, R. W., Obias-Manno, D., Barker, A. H., Arensberg, D., Baker, A., Friedman, L., Greene, H., and Huther, M. L. (1991). Mortality and Morbidity in Patients Receiving Encainide, Flecainide, or Placebo: the Cardiac Arrhythmia Suppression Trial. *N. Engl. J. Med.*, 324(12):781–788.

Eglen, S. J., Weeks, M., Jessop, M., Simonotto, J., Jackson, T., and Sernagor, E. (2014). Home Page for the Retinal Wave Repository. `http://www.damtp.cam.ac.uk/user/eglen/waverepo`.

Ellenberg, S. S. and Hamilton, J. M. (1989). Surrogate Endpoints in Clinical Trials: Cancer. *Stat. Med.*, 8(4):405–413.

FDA (1992). New drug, antibiotic and biological drug product regulations: accelerated approval. *Federal Register*, 57:13234–13242.

Fieller, E. C. (1940). The Biological Standardization of Insulin. *Supplement to the J. Roy. Statist. Soc.*, 7(1):1–64.

Fleming, T. R. (1994). Surrogate Markers in AIDS and Cancer Trials. *Stat. Med.*, 13(13-14):1423–1435.

Fleming, T. R. and DeMets, D. L. (1996). Surrogate End Points in Clinical Trials: Are We Being Misled? *Ann. Intern. Med.*, 125(7):605–613.

Fleming, T. R. and Harrington, D. P. (2011). *Counting processes and survival analysis*, volume 169. John Wiley & Sons.

Fleming, T. R. and Powers, J. H. (2012). Biomarkers and Surrogate Endpoints in Clinical Trials. *Stat. Med.*, 31(25):2973–2984.

Follmann, D. (2006). Augmented Designs to Assess Immune Response in Vaccine Trials. *Biometrics*, 62(4):1161–1169.

Foucart, S. and Lai, M. (2009). Sparsest Solutions of Underdetermined Linear Systems via ℓ_q-minimization for $0 < q \leq 1$. *Appl. Comput. Harmon. Anal.*, 26(3):395–407.

Foygel, R. and Srebro, N. (2011). Fast-rate and Optimistic-rate Error Bounds for ℓ_1-regularized Regression. *arXiv:1108.0373*.

Frangakis, C. E. and Rubin, D. B. (2002). Principal Stratification in Causal Inference. *Biometrics*, 58(1):21–29.

Freedman, L. S. (2001). Confidence Intervals and Statistical Power of the 'Validation' Ratio for Surrogate or Intermediate Endpoints. *J. Statist. Plann. Inference*, 96(1):143–153.

Freedman, L. S., Graubard, B. I., and Schatzkin, A. (1992). Statistical Validation of Intermediate Endpoints for Chronic Diseases. *Stat. Med.*, 11(2):167–178.

Friedlander, M. and Macêdo, I. (2016). Low-rank Spectral Optimization via Gauge Duality. *SIAM J. Sci. Comput.*, 38(3):A1616–A1638.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse Inverse Covariance Estimation with the Graphical Lasso. *Biostatistics*, 9(3):432–441.

Gail, M. H., Pfeiffer, R., Van Houwelingen, H. C., and Carroll, R. J. (2000). On Meta-analytic Assessment of Surrogate Outcomes. *Biostatistics*, 1(3):231–246.

Gallin, J. I., Malech, H. L., Weening, R. S., Curnutte, J. T., Quie, P. G., Jaffe, H. S., and Ezekowitz, R. A. B. (1991). A Controlled Trial of Interferon Gamma to Prevent Infection in Chronic Granulomatous Disease. *N. Engl. J. Med.*, 324(8):509–516.

Giraud, C. (2014). *Introduction to High-dimensional Statistics*. CRC Press.

Gramfort, A., Kowalski, M., and Hämäläinen, M. (2012). Mixed-norm Estimates for the M/EEG Inverse Problem Using Accelerated Gradient Methods. *Phys. Med. Biol.*, 57(7):1937–1961.

Greenshtein, E. and Ritov, Y. (2004). Persistence in High-dimensional Linear Predictor Selection and the Virtue of Overparametrization. *Bernoulli*, 10(6):971–988.

Grimes, D. A. and Schulz, K. F. (2005). Surrogate End Points in Clinical Research: Hazardous to Your Health. *Obstet. Gynecol.*, 105(5, Part 1):1114–1118.

Grimmett, G. R. (1973). A Theorem about Random Fields. *Bull. London Math. Soc.*, 5:81–84.

Gu, Q., Cao, Y., Ning, Y., and Liu, H. (2015). Local and Global Inference for High Dimensional Nonparanormal Graphical Models. *arXiv:1502.02347*.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC press.

Hebiri, M. and Lederer, J. (2013). How Correlations Influence Lasso Prediction. *IEEE Trans. Inf. Theory*, 59(3):1846–1854.

Hillis, A. and Seigel, D. (1989). Surrogate Endpoints in Clinical Trials: Ophthalmologic Disorders. *Stat. Med.*, 8(4):427–430.

Hoff, P. (2011). Separable Covariance Arrays via the Tucker Product, with Applications to Multivariate Relational Data. *Bayesian Anal.*, 6(2):179–196.

Huang, J. and Zhang, C. (2012). Estimation and Selection via Absolute Penalized Convex Minimization and its Multistage Adaptive Applications. *J. Mach. Learn. Res.*, 13:1839–1864.

Huang, Y., Gilbert, P. B., and Wolfson, J. (2013). Design and Estimation for Evaluating Principal Surrogate Markers in Vaccine Trials. *Biometrics*, 69(2):301–309.

Hughes, M., DeGruttola, V., and Welles, S. (1995). Evaluating Surrogate Markers. *J. Acquir. Immune Defic. Syndr.*, 10:S1–8.

Hughes, M. D. (2008). Practical Issues Arising in an Exploratory Analysis Evaluating Progression-free Survival as a Surrogate Endpoint for Overall Survival in Advanced Colorectal Cancer. *Stat. Methods Med. Res.*, 17(5):487–495.

Huntsberger, D. and Billingsley, P. (1981). *Elements of Statistical Inference (fifth ed.)*. Allyn Bacon.

Inouye, D. I., Ravikumar, P., and Dhillon, I. S. (2016). Square Root Graphical Models: Multivariate Generalizations of Univariate Exponential Families that Permit Positive Dependencies. *Proceedings of the International Conference on Machine Learning*.

James, W. and Stein, C. (1961). Estimation with Quadratic Loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 361–379.

Johansen, S. (1979). *Introduction to the Theory of Regular Exponential Families*. Lecture Notes, Institute of Mathematical Statistics, University of Copenhagen.

Jordan, M., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233.

Judea, P. (2010). An Introduction to Causal Inference. *Int. J. Biostat.*, 6(2):1–62.

Kaplan, E. L. and Meier, P. (1958). Nonparametric Estimation From Incomplete Observations. *J. Amer. Statist. Assoc.*, 53(282):457–481.

Katzenstein, D. A., Hammer, S. M., Hughes, M. D., Gundacker, H., Jackson, J. B., Fiscus, S., Rasheed, S., Elbeik, T., Reichman, R., Japour, A., et al. (1996). The Relation of Virologic and Immunologic Markers to Clinical Outcomes After Nucleoside Therapy in HIV-infected Adults with 200 to 500 CD4 Cells per Cubic Millimeter. *N. Engl. J. Med.*, 335(15):1091–1098.

Kent, J. T. (1983). Information Gain and a General Measure of Correlation. *Biometrika*, 70(1):163–173.

Kobayashi, F. and Kuroki, M. (2015). Causal Measures of the Treatment Effect Captured by Candidate Surrogate Endpoints. *J. Agric. Biol. Environ. Stat.*, 20(3):409–430.

Kolda, T. (2006). Multilinear Operators for Higher-order Decompositions. Technical Report SAND2006-2081, Sandia National Laboratories.

Kolda, T. and Bader, B. (2009). Tensor Decompositions and Applications. *SIAM rev.*, 51(3):455–500.

Koltchinskii, V., Lounici, K., and Tsybakov, A. (2011). Nuclear-norm Penalization and Optimal Rates for Noisy Low-rank Matrix Completion. *Ann. Statist.*, 39(5):2302–2329.

Landau, W. M. (1990). Clinical Neuromythology IX. Pyramid Sale in the Bucket Shop: DATATOP bottoms out. *Neurology*, 40(9):1337–1339.

Lauritzen, S. L. (1996a). *Graphical Models*. Oxford University Press.

Lauritzen, S. L. (1996b). *Graphical models*, volume 17 of *Oxford Statistical Science Series*. The Clarendon Press, Oxford University Press, New York. Oxford Science Publications.

Lederer, J., Yu, L., and Gaynanova, I. (2016). Oracle Inequalities for High-dimensional Prediction. *arXiv:1608.00624*.

Lenz, W. (1920). Beiträge zum Verständnis der magnetischen Eigenschaften in festen Körpern. *Physikalische Zeitschrift*, 21:613–615.

Levin, M. L. (1953). The Occurrence of Lung Cancer in Man. *Acta-Unio Internationalis Contra Cancrum*, 9(3):531–541.

Li, X., Zhou, H., and Li, L. (2013). Tucker Tensor Regression and Neuroimaging Analysis. *arXiv:1304.5637*.

Li, Y., Taylor, J. M., and Elliott, M. R. (2010). A Bayesian Approach to Surrogacy Assessment Using Principal Stratification in Clinical Trials. *Biometrics*, 66(2):523–531.

Li, Z., Meredith, M. P., and Hoseyni, M. S. (2001). A method to Assess the Proportion of Treatment Effect Explained by a Surrogate Endpoint. *Stat. Med.*, 20(21):3175–3188.

Lin, D. Y., Fischl, M. A., and Schoenfeld, D. A. (1993). Evaluating the Role of CD4-lymphocyte Counts as Surrogate Endpoints in Human Immunodeficiency Virus Clinical Trials. *Stat. Med.*, 12(9):835–842.

Lin, D. Y., Fleming, T. R., and DeGruttola, V. (1997). Estimating the Proportion of Treatment Effect Explained by a Surrogate Marker. *Stat. Med.*, 16(13):1515–1527.

Lin, L., Drton, M., and Shojaie, A. (2016). Estimation of high-dimensional graphical models using regularized score matching. *Electron. J. Stat.*, 10(1):806.

Liu, H., Han, F., Yuan, M., Lafferty, J., and Wasserman, L. (2012). High-dimensional semiparametric Gaussian copula graphical models. *Ann. Statist.*, 40(4):2293–2326.

Liu, H., Lafferty, J., and Wasserman, L. (2009). The Nonparanormal: Semiparametric Estimation of High Dimensional Undirected Graphs. *J. Mach. Learn. Res.*, 10(Oct):2295–2328.

Lodish, H., Darnell, J. E., Berk, A., Kaiser, C. A., Krieger, M., Scott, M. P., Bretscher, A., Ploegh, H., and Matsudaira, P. (2008). *Molecular cell biology*. Macmillan.

Loh, P. and Wainwright, M. J. (2012). Structure Estimation for Discrete Graphical Models: Generalized Covariance Matrices and Their Inverses. In *Advances in Neural Information Processing Systems*, pages 2087–2095.

Loh, P.-L. and Wainwright, M. J. (2013). Structure Estimation for Discrete Graphical Models: Generalized Covariance Matrices and Their Inverses. *Ann. Statist.*, 41(6):3022–3049.

Lonn, E. (2001). The Use of Surrogate Endpoints in Clinical Trials: Focus on Clinical Trials in Cardiovascular Diseases. *Pharmacoepidemiol. Drug Saf.*, 10(6):497–508.

Mahani, A. S. and Sharabiani, M. T. (2014). Multivariate-from-univariate MCMC sampler: R Package MfUSampler. *arXiv:1412.7784*.

Massart, P. and Meynet, C. (2011). The Lasso as an ℓ_1-ball Model Selection Procedure. *Electron. J. Stat.*, 5:669–687.

Mildvan, D., Landay, A., De Gruttola, V., Machado, S. G., and Kagan, J. (1997). An Approach to the Validation of Markers for Use in AIDS Clinical Trials. *Clin. Infect. Dis.*, 24(5):764–774.

Molenberghs, G., Buyse, M., Geys, H., Renard, D., Burzykowski, T., and Alonso, A. (2002). Statistical Challenges in the Evaluation of Surrogate Endpoints in Randomized Trials. *Control. Clin. Trials*, 23(6):607–625.

Molenberghs, G., Geys, H., and Buyse, M. (2001). Evaluation of Surrogate Endpoints in Randomized Experiments with Mixed Discrete and Continuous Outcomes. *Stat. Med.*, 20(20):3023–3038.

Murray, J., MR, E., lacono Connors, L., Cvetkovich, T., and Struble, K. (1999). The Use of Plasma HIV RNA as a Study Endpoint in Efficacy Trials of Antiretroviral Drugs. *AIDS*, 13(7):797–804.

Neal, R. M. (2003). Slice Sampling. *Ann. Statist.*, 31(3):705–767.

Owen, A. B. (2013). *Monte Carlo Theory, Methods and Examples.*

Pearl, J. (2001). Direct and Indirect Effects. In *Proceedings of the seventeenth conference on uncertainty in artificial intelligence*, pages 411–420. Morgan Kaufmann Publishers Inc.

Pearl, J. (2012). The Causal Mediation Formulaa Guide to the Assessment of Pathways and Mechanisms. *Prev. Sci.*, 13(4):426–436.

Prentice, R. L. (1989). Surrogate Endpoints in Clinical Trials: Definition and Operational Criteria. *Stat. Med.*, 8(4):431–440.

Qu, Y. and Case, M. (2006). Quantifying the Indirect Treatment Effect via Surrogate Markers. *Stat. Med.*, 25(2):223–231.

Qu, Y. and Case, M. (2007). Quantifying the Effect of the Surrogate Marker by Information Gain. *Biometrics*, 63(3):958–960.

Rabusseau, G. and Kadri, H. (2016). Low-rank Regression with Tensor Responses. In *Advances in Neural Information Processing Systems 29*, pages 1867–1875.

Raskutti, G., Yuan, M., and Chen, H. (2015). Convex Regularization for High-dimensional Multi-response Tensor Regression. *arXiv:1512.01215*.

Ravikumar, P., Wainwright, M., Raskutti, G., and Yu, B. (2011). High-dimensional Covariance Estimation by Minimizing ℓ_1-penalized Log-determinant Divergence. *Electron. J. Stat.*, 5:935–980.

Renard, D., Geys, H., Molenberghs, G., Burzykowski, T., and Buyse, M. (2002). Validation of Surrogate Endpoints in Multiple Randomized Clinical Trials with Discrete Outcomes. *Biom. J.*, 44(8):921–935.

Renard, D., Geys, H., Molenberghs, G., Burzykowski, T., Buyse, M., Vangeneugden, T., and Bijnens, L. (2010). Validation of a Longitudinally Measured Surrogate Marker for a Time-to-event Endpoint. *J. Appl. Stat.*, 30(2):235–247.

Retherford, R. D. and Choe, M. K. (2011). *Statistical Models for Causal Analysis*. John Wiley & Sons.

Rigollet, P. and Tsybakov, A. (2011). Exponential Screening and Optimal Rates of Sparse Estimation. *Ann. Statist.*, 39(2):731–771.

Robins, J. M. and Greenland, S. (1992). Identifiability and Exchangeability for Direct and Indirect Effects. *Epidemiology*, pages 143–155.

Rothman, A., Bickel, P., Levina, E., and Zhu, J. (2008). Sparse Permutation Invariant Covariance Estimation. *Electron. J. Stat.*, 2:494–515.

Rubin, D. B. (2004). Direct and Indirect Causal Effects via Potential Outcomes. *Scand. J. Stat.*, 31(2):161–170.

Schatzkin, A. (2000). Intermediate Markers as Surrogate Endpoints in Cancer Research. *Hematol. Oncol. Clin. North Am.*, 14(4):887–905.

Schatzkin, A., Freedman, L. S., Schiffman, M. H., and Dawsey, S. M. (1990). Validation of Intermediate End Points in Cancer Research. *J. Natl. Cancer Inst.*, 82(22):1746–1752.

Shevlyakova, M. and Morgenthaler, S. (2013). Identifying Graphical Models. *arXiv:1309.5740*.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A Sparse-group Lasso. *J. Comput. Graph. Statist.*, 22(2):231–245.

Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). ROCR: Visualizing Classifier Performance in R. *Bioinformatics*, 21(20):3940–3941.

Sun, W. and Li, L. (2016). Sparse Tensor Response Regression and Neuroimaging Analysis. *arXiv:1609.04523*.

Temple, R. (1995). A regulatory authority's opinion about surrogate endpoints. *Clinical measurement in drug evaluation*, pages 1–22.

Temple, R. (1999). Are Surrogate Markers Adequate to Assess Cardiovascular Disease Drugs? *J. Am. Med. Assoc.*, 282(8):790–795.

Theis, L., Berens, P., Froudarakis, E., Reimer, J., Rosón, M. R., Baden, T., Euler, T., Tolias, A. S., and Bethge, M. (2016). Benchmarking Spike Rate Inference in Population Calcium Imaging. *Neuron*, 90(3):471–482.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 58:267–288.

Tsiatis, A. A., Degruttola, V., and Wulfsohn, M. S. (1995). Modeling the Relationship of Survival to Longitudinal Data Measured with Error. Applications to Survival and CD4 Counts in Patients with AIDS. *J. Amer. Statist. Assoc.*, 90(429):27–37.

van de Geer, S. (2008). High-dimensional Generalized Linear Models and the Lasso. *Ann. Statist.*, 36:614–645.

van de Geer, S. (2016). *Estimation and Testing Under Sparsity*. Springer.

VanderWeele, T. J. (2013). Surrogate Measures and Consistent Surrogates. *Biometrics*, 69(3):561–565.

VanderWeele, T. J. (2016). Mediation Analysis: a Practitioner's Guide. *Annu. Rev. Public Health*, 37:17–32.

Verzelen, N. (2012). Minimax Risks for Sparse Regressions: Ultra-high Dimensional Phenomenons. *Electron. J. Stat.*, 6:38–90.

Wainwright, M. and Jordan, M. (2008). Graphical Models, Exponential Families, and Variational Inference. *Found. Trends. Machine Learning*, 1(1-2):1–305.

Wang, Y. and Taylor, J. M. G. (2002). A Measure of the Proportion of Treatment Effect Explained by a Surrogate Marker. *Biometrics*, 58(4):803–12.

Wittes, J., Lakatos, E., and Probstfield, J. (1989). Surrogate Endpoints in Clinical Trials: Cardiovascular Diseases. *Stat. Med.*, 8(4):415–425.

Wong, R. O., Meister, M., and Shatz, C. J. (1993). Transient Period of Correlated Bursting Activity During Development of the Mammalian Retina. *Neuron*, 11(5):923–938.

Writing Group for the Women's Health Initiative Investigators (2002). Risks and Benefits of Estrogen Plus Progestin in Healthy Postmenopausal Women: Principal Results From the Women's Health Initiative randomized controlled trial. *JAMA*, 288(3):321–333.

Xu, H.-p., Furman, M., Mineur, Y. S., Chen, H., King, S. L., Zenisek, D., Zhou, Z. J., Butts, D. A., Tian, N., Picciotto, M. R., and Crair, M. C. (2011). An Instructive Role for Patterned Spontaneous Retinal Activity in Mouse Visual Map Development. *Neuron*, 70(6):1115–1127.

Xue, L. and Zou, H. (2012). Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *Ann. Statist.*, 40(5):2541–2571.

Yu, S., Drton, M., and Shojaie, A. (2019). Generalized score matching for non-negative data. *J. Mach. Learn. Res.*, 20(76):1–70.

Yuan, M. and Lin, Y. (2006). Model Selection and Estimation in Regression with Grouped Variables. *J. Roy. Statist. Soc. Ser. B*, 68(1):49–67.

Yuan, M. and Lin, Y. (2007). Model Selection and Estimation in the Gaussian Graphical Model. *Biometrika*, 94:19–35.

Zhang, Y., Wainwright, M., and Jordan, M. (2017). Optimal Prediction for Sparse Linear Models? Lower Bounds for Coordinate-separable M-estimators. *Electron. J. Stat.*, 11(1):752–799.

Zhao, H. and Duan, Z.-H. (2019). Cancer Genetic Network Inference Using Gaussian Graphical Models. *Bioinform. Biol. Insights*, 13:1–9.

Zhou, H., Li, L., and Zhu, H. (2013). Tensor Regression with Applications in Neuroimaging Data Analysis. *J. Amer. Statist. Assoc.*, 108(502):540–552.

Zhuang, R. and Chen, Y. Q. (2019). Measuring Surrogacy in Clinical Research. *Stat. Biosci.*, pages 1–29.

Zhuang, R. and Lederer, J. (2018). Maximum Regularized Likelihood Estimators: A general Prediction Theory and Applications. *Stat*, 7(1):e186.

Zhuang, R., Simon, N., and Lederer, J. (2016). Graphical Models for Discrete and Continuous Data. *arXiv:1609.05551v3*.

Zografos, K. (1998). On a Measure of Dependence Based on Fisher's Information Matrix. *Comm. Statist. Theory Methods*, 27(7):1715–1728.

Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *J. Amer. Statist. Assoc.*, 101(476):1418–1429.