Likelihood Analysis of Causal Models

F. Richard Guo

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

Doctor of Philosophy

University of Washington

2021

Reading Committee:

Thomas S. Richardson, Chair

Emilija Perković

Mathias Drton

Program Authorized to Offer Degree:
Statistics

University of Washington

Abstract

Likelihood Analysis of Causal Models

F. Richard Guo

Chair of the Supervisory Committee:
Professor Thomas S. Richardson
Department of Statistics

We analyze several problems in causal inference from the perspective of maximum likelihood. Two archetypal likelihoods are primarily concerned: Gaussian likelihood for continuous data and multinomial likelihood for discrete data.

In the first half of this dissertation, Gaussian likelihood is considered for testing and estimation. Motivated by the selection of causal graphs, in Chapter 2, we study testing between marginal and conditional independence in a Gaussian setting with the likelihood ratio test (LRT). We introduce a class of "envelope" distributions by taking pointwise suprema over asymptotic distribution functions of LRT. We show that these envelope distributions are well-behaved and lead to uniformly consistent model selection procedures. In Chapter 3, we consider the estimation of total causal effects under causal sufficiency and linearity. We derive a simple recursive least squares estimator as the MLE under Gaussian errors, which can consistently estimate any identified total effect, under either point or joint intervention. Further, this estimator is shown to be asymptotically efficient even beyond the Gaussian assumption, when compared to a reasonably large class of estimators.

In the latter half, we study the inference of instrumental variable (IV) models with discrete data. In Chapter 4, we develop non-asymptotic tail bounds for the likelihood ratio statistic under multinomial sampling. Such bounds are established by bounding the moment generating function of the statistic uniformly over all multinomial parameters, which can be

viewed as a finite-sample version of Wilks' theorem. Then, in Chapter 5, such bounds are combined with a convex parametrization of the IV model to streamline statistical inference as convex programming. This approach delivers strong guarantees and circumvents the difficulty in identification and post-selection inference. The approach is illustrated with a case study on the distributional effect of military service on annual earnings, using the Vietnam draft lottery as a monotone instrument. Finally, we study partial identification of the average treatment effect in a latent variable formulation and make connections to the Bell-CHSH inequalities in quantum mechanics.

Contents

List of Figures

List of Tables

GLOSSARY

ATE: Average treatment effect.

CHSH: Inequality named after Clauser, Horne, Shimony and Holt (1969).

CPDAG: Completed partially directed acyclic graph; essential graph.

DAG: Directed acyclic graph.

IV: Instrumental variable.

LATE: Local average treatment effect; the average treatment effect among "compliers".

LRT: Likelihood ratio test.

MLE: Maximum likelihood estimate.

MPDAG: Maximally oriented partially directed acyclic graph.

SEM: Structural equation model.

SWIG: Single world intervention graph.

ACKNOWLEDGMENTS

I owe this dissertation to many people. First, none of this work would have been possible without my advisor, Thomas S. Richardson. I am extremely grateful for his patience, generosity, insightfulness, passion and superb memory, as well as financial support through ONR Grant N000141912446.

I would like to thank Emilija Perković and Mathias Drton for reading and examining my thesis, as well as Jon Wellner, Michael Perlman, Robin Evans and Dmitriy Drusvyatskiy for serving on my supervisory committee.

I am grateful to the staff, faculty, and fellow students of the Department of Statistics for making me feel at home. In particular, I want to thank Jon Wellner for his dedication to teaching, Yen-Chi Chen for his encouragement and candid advice, Emilija Perković for teaching me many things on graphical models. Thanks also go to Ellen Reynolds and Marina Meilă for keeping the Ph.D. program running smoothly, and to Asa Sourdiffe for IT support.

I would like to thank my friends and family for their support.

I want to thank BBC Radio 6 Music (shout out to Gideon Coe!) for cheering me up during the lockdown.

Lastly, I thank Haohao for her patience and affections.

DEDICATION

To my grandpa,

who sat me down and taught me math while I found going to kindergarten all too dreadful...

Chapter 1

INTRODUCTION

1.1 Motivation

In this dissertation, we present analyses of several problems from causal inference, including model selection (Chapter 2), estimation (Chapter 3), partial identification and post-selection inference (Chapter 5). The solutions to these problems are based on, inspired by or related to the principle of maximum likelihood. We will use two archetypal likelihood functions: the Gaussian likelihood for continuous data (Chapters 2 and 3) and the multinomial likelihood for discrete data (Chapters 4 and 5). Although many causal models can be posited non-parametrically and there is often a concern on the misspecification of a parametric model (especially for continuous data), we argue that the analysis through Gaussian or multinomial likelihood serves as an essential first step, which can provide valuable insight from the following aspects.

First, for a new class of graphical models, it is customary to understand its parametrization in terms of Gaussian or discrete data; see, e.g., Lauritzen (1996, Chap. 4–5), Richardson and Spirtes (2002, §8), Evans and Richardson (2014, §5). Such parametrization sheds light on the smoothness of the model. If a model can be parametrized as a curved exponential family, it is considered smooth. On the other hand, if singularities are shown to be present (e.g., Drton, 2009a), the model is non-smooth and its practicality may be limited.

Second, the analysis of the multinomial likelihood is a useful guide for extending to the nonparametric case; see Chamberlain (1987).

Third, estimators that naturally arise as maximum likelihood estimates (MLE) can have optimality properties beyond the specific parametric setting. We show such a result for least squares in Chapter 3.

Finally, for certain problems, restricting the set of distributions is even necessary for making progress, such as testing conditional independence when the conditioned variable is continuous (Shah and Peters, 2020) or testing the instrumental variable assumption when the treatment is continuous (Gunsilius, 2020).

1.2 Notation

The following notations are used throughout. $\mathbb{R}_{\mathrm{PD}}^{n\times n}$ denotes $n \times n$ positive definite matrices. Typically, Θ denotes the parameter space. Symbol $\mathcal{M}$ denotes a model, which is a subset of the parameter space or a subset of laws. We use P and Q for probability measures, P_n and Q_n for sequences of measures, as well as $\mathbb{P}_n$ and $\mathbb{Q}_n$ for empirical measures. μ is reserved for Lebesgue measure. Lower-case letters p, q denote the densities of P, Q with respect to μ. P_n^n denotes the n-sample product (tensorized) measure of P_n, namely the law of $X_1, \ldots, X_n \overset{\text{iid}}{\sim} P_n$. We write $P_n \to_d P$ if P_n converges (weakly) to P in law. For $X_n(t)$ a stochastic process indexed by $t \in T$, we write $X_n \rightsquigarrow X$ if $X_n(t)$ converges weakly to $X(t)$.

For two sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n = O(b_n)$ if there exists a constant $c < \infty$ such that $a_n \leq cb_n$ for large enough n; $a_n = o(b_n)$ and $b_n = \omega(a_n)$ if $\lim_{n\to\infty} a_n/b_n = 0$; $a_n \asymp b_n$ if $a_n = O(b_n)$ and $b_n = O(a_n)$. Also, we write $x \lesssim y$ if $x \leq cy$ for some constant $c > 0$. We write $x \vee y = \max(x, y)$ and $x \wedge y = \min(x, y)$.

1.3 Organization

The rest of this dissertation is organized as follows. In Chapter 2, which is based on Guo and Richardson (2020), we consider selecting between two Gaussian graphical models that are non-nested. "Envelope" distributions are introduced to define correct critical values for non-standard likelihood ratio tests. In Chapter 3, we consider recursive least squares for estimating total effects under linearity and causal sufficiency. New efficiency bounds are derived for a restricted class of estimators. In Chapter 4, which largely follows Guo and Richardson (2021), we develop state-of-the-art tail bounds for the multinomial likelihood ratio. Finally, Chapter 5 is devoted to automating the inference of discrete instrumental variable models,

which are prescribed as convex polytopes. New results on partial identification are presented.

Chapter 2

TESTING MARGINAL VERSUS CONDITIONAL INDEPENDENCE

In this Chapter, we are concerned with selecting between two Gaussian graphical models that are non-nested, a problem that is motivated by model selection of causal DAGs. We study selection rules based on the likelihood ratio test (LRT), but in a setting where standard χ^2 asymptotics do not hold. New techniques are developed to control the probability of selecting the wrong model in a uniform sense.

More concretely, we focus on the problem of selecting between marginal independence and conditional independence in a trivariate setting. The two models are non-nested and their intersection is a union of two marginal independences. We consider two sequences of such models, one from each type of independence, that are closest to each other in the Kullback-Leibler sense as they approach the intersection. They become indistinguishable if the signal strength, as measured by the product of two correlation parameters, decreases faster than the standard parametric rate. Under local alternatives at such rate, we show that the asymptotic distribution of the likelihood ratio depends on where and how the local alternatives approach the intersection. To deal with this non-uniformity, we introduce a class of "envelope" distributions by taking pointwise suprema over asymptotic cumulative distribution functions. We show that these envelope distributions are well-behaved and lead to model selection procedures with rate-free uniform error guarantees and near-optimal power. To control the error even when the two models are indistinguishable, rather than insist on a dichotomous choice, the proposed procedure will choose either or both models.

The rest of this Chapter is organized as follows. Section 2.1 describes the setup. In Section 2.2, we derive the maximum likelihood estimates under the two types of indepen-

dence models, and obtain the loglikelihood ratio statistic in a closed form. In Section 2.3, we characterize the information-theoretic limit to distinguishing the two models, and outline two regimes on the boundary of distinguishability. Then in Section 2.4, we consider local alternative sequences in the two aforementioned regimes and establish the asymptotic distribution of the loglikelihood ratio. Section 2.4.3 provides a geometric perspective in terms of limit experiments. We then deal with non-uniformity issue of asymptotic distributions in Section 2.5 by introducing a family of envelope distributions. Next in Section 2.6 we propose model selection procedures with a uniform error guarantee. In Section 2.7, we compare the performance of several methods through simulation studies. We present a realistic example in Section 2.8 on inferring the American occupational structure. Finally some discussions are given in Section 2.9. Additional proofs are delegated to Appendix A.

2.1 Introduction

It is often of interest to test marginal or conditional independence for a set of random variables. For example, in the context of graphical modeling, the PC algorithm (Spirtes et al., 2000) for directed acyclic graph model selection determines the orientation of an unshielded triple $X - Z - Y$ based on whether the separating set of X and Y contains Z: if so, $X \perp\!\!\!\perp Y \mid Z$ and Z is not a collider; if not, $X \not\perp\!\!\!\perp Y$ and the triple is oriented as $X \to Z \leftarrow Y$. The reader is referred to Dawid (1979); Lauritzen (1996); Koller et al. (2009) and Reichenbach (1956) for more discussion.

Here we consider the simplest case, namely testing $X_1 \perp\!\!\!\perp X_2$ versus $X_1 \perp\!\!\!\perp X_2 \mid X_3$ in a trivariate Gaussian setting. For testing whether *a specific* marginal or conditional independence holds, it is common to use the correlation coefficient or partial correlation coefficient under Fisher's z-transformation (Fisher, 1924) as the test statistic. Under independence, the transformed correlation coefficient is approximately distributed as a normal distribution with zero mean and variance determined by the sample size and the number of variables being conditioned on (Hotelling, 1953; Anderson, 1984). In this Chapter, however, we assume *at least one* type of independence holds (from prior knowledge or precursory inference) and

we want to contrast the two types. To this end, we will use the likelihood ratio statistic, which often provides intuitively reasonable tests for composite hypotheses (Perlman and Wu, 1999), especially in terms of model selection.

Setup For $(X_1, X_2, X_3) \sim \mathcal{N}\{0, \Sigma = (\sigma_{ij})\}$ with parameter space Θ being the set of 3×3 real positive definite matrices $\mathbb{R}^{3\times3}_{\text{PD}}$, we consider testing

$$\mathcal{M}_0 : X_1 \perp\!\!\!\perp X_2 \quad \text{versus} \quad \mathcal{M}_1 : X_1 \perp\!\!\!\perp X_2 \mid X_3. \tag{2.1}$$

$\mathcal{M}_0$ and $\mathcal{M}_1$ are algebraic models (Drton and Sullivant, 2007) as represented by equality constraints

$$\mathcal{M}_0 : \{\sigma_{12} = 0\}, \quad \mathcal{M}_1 : \{\sigma_{12}\sigma_{33} = \sigma_{13}\sigma_{23}\} \tag{2.2}$$

imposed on Θ. They are visualized in the correlation space (ignoring the variances) in Figure 2.1.

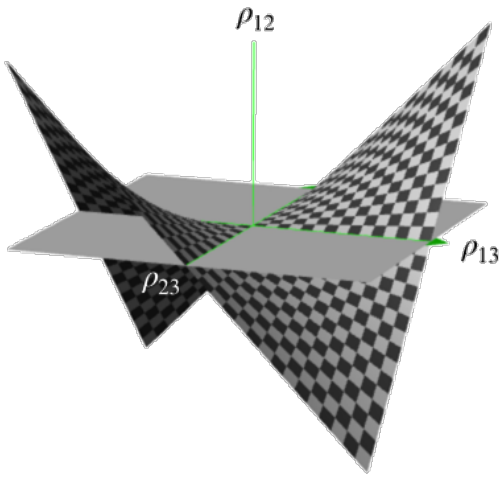


Figure 2.1: The two models visualized in the correlation space: $\mathcal{M}_0 : \rho_{12} = 0$ (grey plane) and $\mathcal{M}_1 : \rho_{12} = \rho_{13}\rho_{23}$ (checkerboard). $\mathcal{M}_0 \cap \mathcal{M}_1$ consists of the ρ_{13} and ρ_{23} axes; they intersect at the origin $\mathcal{M}_{\text{sing}}$. See also Evans (2020, Figure 3).

$\mathcal{M}_0$ and $\mathcal{M}_1$ are non-nested and they further intersect at the origin

$$\mathcal{M}_{\text{sing}} : \{\sigma_{12} = \sigma_{13} = \sigma_{23} = 0\}, \tag{2.3}$$

which is a *singularity* within $\mathcal{M}_0 \cap \mathcal{M}_1$ that corresponds to diagonal covariances. At $\mathcal{M}_{\text{sing}}$ the likelihood ratio statistic is not regular in the sense that the tangent cones (linear approximations to the parameter space; see Bertsekas et al. (2003, Chap. 4.6)) of the two models coincide. As pointed out by Evans (2020), we will see that the equivalence of local geometry between the two models presents a challenge for model selection. It is also worth mentioning that, in the setting of nested model selection, the behavior of the likelihood ratio of testing $\mathcal{M}_0 \cap \mathcal{M}_1$ against a saturated model, especially at the singularity, has been studied by Drton (2006a); Drton and Sullivant (2007); Drton (2009b).

2.2 Maximum likelihood

The log-likelihood of a Gaussian graphical model under sample size n (Lauritzen, 1996, Chap. 5) is

$$\ell_n(\Sigma) = \frac{n}{2}(-\log|\Sigma| - \text{Tr}(S_n\Sigma^{-1})), \tag{2.4}$$

where S_n is the sample covariance computed with respect to mean zero (i.e., the scatter matrix divided by n). A model can be scored by its log-likelihood maximized within the model contrasted against the *saturated* model

$$\lambda_n^{(i)} := 2\left(\sup_{\Sigma\in\Theta} \ell_n(\Sigma) - \sup_{\Sigma\in\Theta_i} \ell_n(\Sigma)\right) \geq 0 \tag{2.5}$$

for $i = 0, 1$, which is the quantity considered in nested model selection. The saturated model attains maximal likelihood when $\Sigma = S_n$, yielding

$$\begin{aligned}\ell_n^{\text{sat}} &:= \sup_{\Sigma\in\Theta} \ell_n(\Sigma) \\ &= -\frac{n}{2}\left(\log\left(s_{11}s_{22}s_{33} + 2s_{12}s_{23}s_{13} - s_{11}s_{23}^2 - s_{13}^2 s_{22} - s_{12}^2 s_{33}\right) + 3\right).\end{aligned}$$

To contrast $\mathcal{M}_0$ and $\mathcal{M}_1$, we instead consider

$$\begin{aligned}\lambda_n^{(0:1)} &:= \lambda_n^{(1)} - \lambda_n^{(0)} \\ &= 2\left(\sup_{\Sigma\in\Theta_0} \ell_n(\Sigma) - \sup_{\Sigma\in\Theta_1} \ell_n(\Sigma)\right) = 2\left(\ell_n(\hat{\Sigma}_n^{(0)}) - \ell_n(\hat{\Sigma}_n^{(1)})\right),\end{aligned} \tag{2.6}$$

where $\hat{\Sigma}_n^{(0)}$ and $\hat{\Sigma}_n^{(1)}$ are MLEs within the two models. Intuitively, a positive value of $\lambda_n^{(0:1)}$ prefers $\mathcal{M}_0$, and a negative value prefers $\mathcal{M}_1$.

2.2.1 MLE within $\mathcal{M}_0$

By $X_1 \perp\!\!\!\perp X_2$, we can factorize the likelihood of $\mathcal{M}_0$ as

$$\begin{aligned}p(X_1, X_2, X_3) &= p(X_1)p(X_2)p(X_3 \mid X_1, X_2) \\ &= \mathcal{N}(X_1; 0, \sigma_{11})\mathcal{N}(X_2; 0, \sigma_{22})\mathcal{N}(X_3; \beta_{32\cdot1}X_1 + \beta_{31\cdot2}X_2, \sigma_{33\cdot12}).\end{aligned}$$

where the parameters $\sigma_{11}, \sigma_{22}, \beta_{32\cdot1}, \beta_{31\cdot2}, \sigma_{33\cdot12}$ are variation independent (Barndorff-Nielsen, 2014, Chap. 10.2). The MLEs for them are given by

$$\hat{\sigma}_{11}^{(0)} = s_{11}, \quad \hat{\sigma}_{22}^{(0)} = s_{22}, \tag{2.7}$$

$$\hat{\beta}_{32\cdot1}^{(0)} = \frac{s_{22}s_{13} - s_{12}s_{23}}{s_{11}s_{22} - s_{12}^2}, \quad \hat{\beta}_{31\cdot2}^{(0)} = \frac{s_{11}s_{23} - s_{12}s_{13}}{s_{11}s_{22} - s_{12}^2},$$

and

$$\hat{\sigma}_{33\cdot12}^{(0)} = s_{33} - \frac{s_{22}s_{13}^2 - 2s_{12}s_{23}s_{13} + s_{11}s_{23}^2}{s_{11}s_{22} - s_{12}^2}.$$

Mapping them back to the original parameters via relations $\beta_{32\cdot1} = \sigma_{13}/\sigma_{11}$, $\beta_{31\cdot2} = \sigma_{23}/\sigma_{22}$ and $\sigma_{33\cdot12} = \sigma_{33} - \sigma_{13}^2/\sigma_{11} - \sigma_{23}^2/\sigma_{22}$, in addition to Eq. (2.7) we have the MLEs as

$$\hat{\sigma}_{13}^{(0)} = \frac{s_{11}\left(s_{22}s_{13} - s_{12}s_{23}\right)}{s_{11}s_{22} - s_{12}^2}, \quad \hat{\sigma}_{23}^{(0)} = \frac{s_{22}\left(s_{11}s_{23} - s_{12}s_{13}\right)}{s_{11}s_{22} - s_{12}^2} \tag{2.8}$$

and

$$\hat{\sigma}_{33}^{(0)} = s_{33} - \frac{2s_{12}\left(s_{12}s_{13} - s_{11}s_{23}\right)\left(s_{12}s_{23} - s_{13}s_{22}\right)}{\left(s_{11}s_{22} - s_{12}^2\right)^2}. \tag{2.9}$$

This derivation is essentially the same as executing the iterative conditional fitting algorithm of Chaudhuri et al. (2007) in the order of X_1, X_2, X_3. Plugging Eqs. (2.7) to (2.9) into Eq. (2.4), we have the following *closed-form* expression of maximized log-likelihood of $\mathcal{M}_0$

$$\ell_n^{(0)} = -\frac{n}{2}\left[\log\left(s_{11}s_{22}\left(\frac{s_{22}\hat{s}_{13}^2 - 2s_{12}s_{23}s_{13} + s_{11}s_{23}^2}{s_{12}^2 - s_{11}s_{22}} + s_{33}\right)\right) + 3\right]. \tag{2.10}$$

2.2.2 MLE within $\mathcal{M}_1$

The MLE within $\mathcal{M}_1$ in the covariance parametrization is simpler. By writing $\sigma_{12} = \sigma_{13}\sigma_{23}/\sigma_{33}$ and simplifying the score condition, we obtain

$$\hat{\sigma}_{11}^{(1)} = s_{11}, \quad \hat{\sigma}_{22}^{(1)} = s_{22}, \quad \hat{\sigma}_{33}^{(1)} = s_{33}, \quad \hat{\sigma}_{13}^{(1)} = s_{13}, \quad \hat{\sigma}_{23}^{(1)} = s_{23}, \tag{2.11}$$

all of which are their sample counterparts. Plugging into Eq. (2.4), we have

$$\ell_n^{(1)} = -\frac{n}{2}\left[\log\left(\frac{(s_{13}^2 - s_{11}s_{33})\,(s_{23}^2 - s_{22}s_{33})}{s_{33}}\right) + 3\right]. \tag{2.12}$$

2.2.3 Likelihood ratio

Finally, $\mathcal{M}_0$ and $\mathcal{M}_1$ are contrasted with

$$\lambda_n^{(0:1)} = 2(\ell_n^{(0)} - \ell_n^{(1)}) = n\log\left(\frac{(s_{13}^2 - s_{11}s_{33})\,(s_{23}^2 - s_{22}s_{33})}{s_{33}}\right) - n\log\left(s_{11}s_{22}\left(\frac{s_{22}s_{13}^2 - 2s_{12}s_{23}s_{13} + s_{11}s_{23}^2}{s_{12}^2 - s_{11}s_{22}} + s_{33}\right)\right). \tag{2.13}$$

2.3 Optimal error

We study the information-theoretic limit to distinguishing the two models. Specifically, consider two sequences of sampling distributions — one within $\mathcal{M}_0$ and the other within $\mathcal{M}_1$, as they approach the same limit in $\mathcal{M}_0 \cap \mathcal{M}_1$. Let P_n be the sequence in $\mathcal{M}_0$ under covariance $\Sigma_n^{(0)} \in \mathcal{M}_0 \setminus \mathcal{M}_1$, and let Q_n be the sequence in $\mathcal{M}_1$ under covariance $\Sigma_n^{(1)} \in \mathcal{M}_1 \setminus \mathcal{M}_0$. Further, let P_n^n and Q_n^n be the product measures of n independent copies of P_n and Q_n respectively.

The fundamental limit to distinguishing two distributions P and Q is characterized by their total variation distance $\mathrm{d_{TV}}(P, Q) := \sup_A\{P(A) - Q(A)\}$. We have the following classical result on testing two simple hypotheses, where the minimum total error is achieved by the likelihood ratio test.

Lemma 2.1 (Theorem 13.1.1 of Lehmann and Romano (2006))**.** *For testing $H_0 : X \sim P$ versus $H_1 : X \sim Q$, the minimum sum of type-I and type-II errors is $1 - \mathrm{d}_{TV}(P, Q)$.*

The optimal error above does not permit a tractable formula. The analysis for a product measure is more tractable in terms of the Hellinger squared distance $H^2(P,Q) := (1/2)\int (p^{1/2}-q^{1/2})^2 \,\mathrm{d}\mu$, for which it holds that

$$H^2(P_n^n, Q_n^n) = 1 - \left\{1 - H^2(P_n, Q_n)\right\}^n. \tag{2.14}$$

The total variation is related to Hellinger by Le Cam's inequality (Tsybakov, 2009, Lemma 2.3)

$$H^2(P_n^n, Q_n^n) \le \mathrm{d}_{\mathrm{TV}}(P_n^n, Q_n^n) \le H(P_n^n, Q_n^n)\left\{2 - H^2(P_n^n, Q_n^n)\right\}^{1/2}. \tag{2.15}$$

Lemma 2.2 (see also Theorem 13.1.3 of Lehmann and Romano (2006))**.** *It holds that*

$$1 - \mathrm{d}_{TV}(P_n^n, Q_n^n) \to \begin{cases} 0, & H^2(P_n, Q_n) = \omega(n^{-1}) \\ 1, & H^2(P_n, Q_n) = o(n^{-1}) \end{cases}.$$

And when $nH^2(P_n, Q_n) \to h > 0$, it holds that

$$\begin{aligned} 0 < 1 - \{1 - \exp(-2h)\}^{1/2} &\le \liminf_{n\to\infty} \{1 - \mathrm{d}_{TV}(P_n^n, Q_n^n)\} \\ &\le \limsup_{n\to\infty} \{1 - \mathrm{d}_{TV}(P_n^n, Q_n^n)\} \le \exp(-h) < 1. \end{aligned}$$

Proof. Using Eq. (2.14) and Eq. (2.15), we have

$$\begin{aligned} 1 - \mathrm{d}_{\mathrm{TV}}(P_n^n, Q_n^n) \le 1 - H^2(P_n^n, Q_n^n) &= \left\{1 - H^2(P_n, Q_n)\right\}^n \\ &= \exp\left[n \log\left\{1 - H^2(P_n, Q_n)\right\}\right]. \end{aligned}$$

It follows that

$$\begin{aligned} \limsup_{n\to\infty} \{1 - \mathrm{d}_{\mathrm{TV}}(P_n^n, Q_n^n)\} &\le \limsup_{n\to\infty} \exp\left[n \log\left\{1 - H^2(P_n, Q_n)\right\}\right] \\ &= \begin{cases} 0, & H^2(P_n, Q_n) = \omega(n^{-1}) \\ \exp(-h), & nH^2(P_n, Q_n) \to h > 0 \end{cases}. \end{aligned}$$

Similarly, we also have

$$\begin{aligned} \liminf_{n\to\infty} \{1 - \mathrm{d}_{\mathrm{TV}}(P_n^n, Q_n^n)\} &\ge \liminf_{n\to\infty} 1 - H(P_n^n, Q_n^n)\left\{2 - H^2(P_n^n, Q_n^n)\right\}^{1/2} \\ &= \begin{cases} 1, & H^2(P_n, Q_n) = o(n^{-1}) \\ 1 - \{1 - \exp(-2h)\}^{1/2}, & nH^2(P_n, Q_n) \to h > 0 \end{cases}. \end{aligned}$$

The proof is finished by combining the previous two displays with the fact that

$$\liminf_n \{1 - \mathrm{d_{TV}}(P_n^n, Q_n^n)\} \leq \limsup_n \{1 - \mathrm{d_{TV}}(P_n^n, Q_n^n)\}$$

and noting $\mathrm{d_{TV}} \in [0, 1]$. □

Corollary 2.1. *Under $nH^2(P_n, Q_n) \to h > 0$, the optimal power of an asymptotic α-level procedure satisfies*

$$1 - \exp(-h) \leq \textit{optimal asymptotic power} \leq \alpha + \{1 - \exp(-2h)\}^{1/2}. \tag{2.16}$$

Proof. This directly follows from Lemma 2.2 since (1 − optimal power) + type-I error = $1 - \mathrm{d_{TV}}(P_n^n, Q_n^n)$ for type-I error asymptotically between 0 and α, and then passing to the limit. □

By Lemma 2.2, the asymptotic error converges to zero (exponentially fast) if P_n and Q_n are separated by a distance that is decreasing more slowly than rate $n^{-1/2}$. For example, when $P_n = P$, $Q_n = Q$ are *fixed* distributions from which we observe n independent samples, that is, when $P_n^n = P^n$ and $Q_n^n = Q^n$. The analysis above shows that the ability to differentiate P_n and Q_n based on n samples depends on the distance between P_n and Q_n. The consideration of P_n and Q_n as $n \to \infty$ is necessitated by the development of asymptotic results that are applicable in a specific analysis with a fixed n. In particular, here we want to investigate what happens when the sample size is small compared to the signal strength, or equivalently, when signal strength is weak under a given sample size. This is modeled by the regime that yields a non-trivial optimal error strictly between 0 and 1. By Lemma 2.2, we need to choose sequences $\Sigma_n^{(0)}$ and $\Sigma_n^{(1)}$ such that $H^2\left(P_{\Sigma_n^{(0)}}, P_{\Sigma_n^{(1)}}\right) \asymp n^{-1}$. More specifically, we choose $\Sigma_n^{(1)}$ that is the most difficult to distinguish from $\Sigma_n^{(0)}$. That is, we choose $\Sigma_n^{(1)}$ to minimize $\mathcal{D}_{\mathrm{KL}}\left(P_{\Sigma_n^{(0)}} \| P_{\Sigma_n^{(1)}}\right)$, i.e., $\Sigma_n^{(1)}$ is the MLE projection of $\Sigma_n^{(0)}$ in $\mathcal{M}_1$ by Eq. (2.11). The two

sequences take the form of

$$\Sigma_n^{(0)} = \begin{pmatrix} \sigma_{11,n} & 0 & \rho_{13,n}\sqrt{\sigma_{11,n}\sigma_{33,n}} \\ 0 & \sigma_{22,n} & \rho_{23,n}\sqrt{\sigma_{11,n}\sigma_{33,n}} \\ \rho_{13,n}\sqrt{\sigma_{11,n}\sigma_{33,n}} & \rho_{23,n}\sqrt{\sigma_{11,n}\sigma_{33,n}} & \sigma_{33,n} \end{pmatrix},$$
$$\Sigma_n^{(1)} = \begin{pmatrix} \sigma_{11,n} & \rho_{13,n}\rho_{23,n}\sqrt{\sigma_{11,n}\sigma_{22,n}} & \rho_{13,n}\sqrt{\sigma_{11,n}\sigma_{33,n}} \\ \rho_{13,n}\rho_{23,n}\sqrt{\sigma_{11,n}\sigma_{22,n}} & \sigma_{22,n} & \rho_{23,n}\sqrt{\sigma_{22,n}\sigma_{33,n}} \\ \rho_{13,n}\sqrt{\sigma_{11,n}\sigma_{33,n}} & \rho_{23,n}\sqrt{\sigma_{22,n}\sigma_{33,n}} & \sigma_{33,n} \end{pmatrix}. \tag{2.17}$$

Both of them converge to $\Sigma^* \in \mathcal{M}_0 \cap \mathcal{M}_1$ as $n \to \infty$. We assume the variances $\sigma_{ii,n} \to \sigma_{ii} > 0$ for $i = 1, 2, 3$. For $H^2\left(P_{\Sigma_n^{(0)}}, P_{\Sigma_n^{(1)}}\right) \to 0$, it is necessary that either (or both) $\rho_{13,n}$ and $\rho_{23,n}$ converges to zero. The squared Hellinger distance is calculated as

$$\begin{aligned} H^2\left(P_{\Sigma_n^{(0)}}, P_{\Sigma_n^{(1)}}\right) &= 1 - \frac{|\Sigma_n^{(0)}|^{1/4}|\Sigma_n^{(1)}|^{1/4}}{|(\Sigma_n^{(0)} + \Sigma_n^{(1)})/2|^{1/2}} \\ &= \begin{cases} \rho_{13,n}^2\rho_{23,n}^2/8 + O(\rho_{13,n}^4 + \rho_{23,n}^4), & \rho_{13,n}, \rho_{23,n} \to 0 \\ \rho_{23}^2(1-\rho_{23}^2)^{-1}\rho_{13,n}^2/8 + O(\rho_{13,n}^4), & \rho_{13,n} \to 0, \rho_{23,n} \to \rho_{23} \neq 0 \\ \rho_{13}^2(1-\rho_{13}^2)^{-1}\rho_{23,n}^2/8 + O(\rho_{23,n}^4), & \rho_{23,n} \to 0, \rho_{13,n} \to \rho_{13} \neq 0 \end{cases}. \end{aligned} \tag{2.18}$$

The calculation reveals that $H^2(P_{\Sigma_n^{(0)}}, Q_{\Sigma_n^{(1)}}) \asymp 1/n$ if and only if $\rho_{13,n}\rho_{23,n} \asymp n^{-1/2}$. This entails two distinct regimes.

The weak-strong regime Between $\rho_{13,n}$ and $\rho_{23,n}$, one (the weak edge) converges to zero at $n^{-1/2}$ rate, and the other (the strong edge) converges to a non-zero limit $\rho \in (-1, 1)$. The limiting model is on $\mathcal{M}_0 \cap \mathcal{M}_1 \setminus \mathcal{M}_{\text{sing}}$, namely one of the axes excluding the origin in Fig. 2.1.

The weak-weak regime $\rho_{13,n}, \rho_{23,n} \to 0$ and $\sqrt{n}\rho_{13,n}\rho_{23,n} \to \delta \neq 0$. The limiting model is on $\mathcal{M}_{\text{sing}}$, namely the origin in Fig. 2.1.

Remark 2.1. The result can be rephrased as the sample size required to distinguish $\mathcal{M}_0$ and $\mathcal{M}_1$. Consider distinguishing $\mathcal{M}_0$ and $\mathcal{M}_1$ in a Euclidean $m^{-1/2}$-neighborhood of $\Sigma^* \in$

$\mathcal{M}_0 \cap \mathcal{M}_1$ as $m \to \infty$. The sample size required is m^2 if $\Sigma^* \in \mathcal{M}_{\text{sing}}$, and m if $\Sigma^* \notin \mathcal{M}_{\text{sing}}$. This phenomenon is described by Evans (2020) in terms of equivalence of local geometry. $\mathcal{M}_0$ and $\mathcal{M}_1$ are 1-equivalent at $\Sigma^* \in \mathcal{M}_{\text{sing}}$ in the sense that their tangent cones coincide; and they are 1-near-equivalent at $\Sigma^* \notin \mathcal{M}_{\text{sing}}$ in the sense that they have distinct tangent cones. See Evans (2020, Theorem 2.8).

Proposition 2.1. *In testing $\mathcal{M}_0$ versus $\mathcal{M}_1$, the sample complexity required is*

$$\begin{cases} n = \omega\left(\frac{1}{\rho_{13}^2\rho_{23}^2}\right), & \text{for consistent model selection} \\ n \asymp \left(\frac{1}{\rho_{13}^2\rho_{23}^2}\right), & \text{for asymptotic total error} \in (0,1) \end{cases}. \tag{2.19}$$

2.4 Local asymptotics

In this section, we analyze the asymptotic distribution of the log-likelihood ratio statistic $\lambda_n^{(0:1)} = 2(\ell_n^{(0)} - \ell_n^{(1)})$ under the two regimes outlined earlier.

2.4.1 Weak-strong regime

Without loss of generality, we choose $\rho_{13,n} = \gamma/\sqrt{n}$ as the weak edge and $\rho_{23,n} \to \rho \neq 0$ as the strong edge. $\gamma \in \mathbb{R}$ characterizes the size of the local asymptotic, and is also referred to as a *local parameter* (van der Vaart, 2000, Chapter 9). We consider asymptotics under local sequences $\Sigma_n^{(0)}$ and $\Sigma_n^{(1)}$ approaching the limiting covariance

$$\Sigma^* = \begin{pmatrix} \sigma_{11} & 0 & 0 \\ 0 & \sigma_{22} & \rho\sqrt{\sigma_{22}\sigma_{33}} \\ 0 & \rho\sqrt{\sigma_{22}\sigma_{33}} & \sigma_{33} \end{pmatrix} \in \mathcal{M}_0 \cap \mathcal{M}_1 \setminus \mathcal{M}_{\text{sing}}. \tag{2.20}$$

We consider the following local alternatives of size γ on the correlation scale. Again, $\Sigma_n^{(1)}$ is the KL-projection (i.e., MLE-projection) of $\Sigma_n^{(0)}$.

$$\Sigma_n^{(0)} = \begin{pmatrix} \sigma_{11} & 0 & \gamma\sqrt{\sigma_{11}\sigma_{33}}/\sqrt{n} \\ 0 & \sigma_{22} & \rho\sqrt{\sigma_{22}\sigma_{33}} \\ \gamma\sqrt{\sigma_{11}\sigma_{33}}/\sqrt{n} & \rho\sqrt{\sigma_{22}\sigma_{33}} & \sigma_{33} \end{pmatrix} \in \mathcal{M}_0 \setminus \mathcal{M}_1, \tag{2.21}$$

$$\Sigma_n^{(1)} = \begin{pmatrix} \sigma_{11} & \gamma\rho\sqrt{\sigma_{11}\sigma_{22}}/\sqrt{n} & \gamma\sqrt{\sigma_{11}\sigma_{33}}/\sqrt{n} \\ \gamma\rho\sqrt{\sigma_{11}\sigma_{22}}/\sqrt{n} & \sigma_{22} & \rho\sqrt{\sigma_{22}\sigma_{33}} \\ \gamma\sqrt{\sigma_{11}\sigma_{33}}/\sqrt{n} & \rho\sqrt{\sigma_{22}\sigma_{33}} & \sigma_{33} \end{pmatrix} \in \mathcal{M}_1 \setminus \mathcal{M}_0. \tag{2.22}$$

At the limit, both models are correct (intersection); see Figure 2.3. However, the sequence approaches the limit *only* on one of the models, and the size of violation of the other model is $|\gamma| n^{-1/2}$. To ensure positive definiteness, we require $|\rho| < 1$.

Proposition 2.2. *Under local alternative* $\Sigma_n^{(0)}$,

$$\lambda_n^{(0:1)} \to_d \rho \left[\left(Z_1 + \frac{\gamma}{\sqrt{2(1-\rho)}} \right)^2 - \left(Z_2 + \frac{\gamma}{\sqrt{2(1+\rho)}} \right)^2 \right]; \tag{2.23}$$

and under local alternative $\Sigma_n^{(1)}$,

$$\lambda_n^{(0:1)} \to_d \rho \left[\left(Z_1 + \gamma\sqrt{\frac{1-\rho}{2}} \right)^2 - \left(Z_2 + \gamma\sqrt{\frac{1+\rho}{2}} \right)^2 \right], \tag{2.24}$$

where Z_1, Z_2 *are two independent standard normal variables.*

We leave the proof to Appendix A, which is done geometrically. Alternatively, the distribution can be derived by a change of measure with Le Cam's third lemma; see van der Vaart (2000, Example 6.7).

Asymptotically the log-likelihood ratio statistic is distributed as a scaled difference of two independent non-central χ_1^2 variables, with non-centralities scaled by γ and weighted by ρ differently, depending on the true model. Note that the distribution only depends on the absolute values of γ and ρ. The asymptotic distributions under the two types of sequences (truths) are visualized in Fig. 2.2. We can see that the mean is positive under $\mathcal{M}_0 \setminus \mathcal{M}_1$ and negative under $\mathcal{M}_1 \setminus \mathcal{M}_0$. However, a pair of these distributions are not symmetric to each other in terms of shape. They are further separated apart (more easily distinguished) as $|\gamma|$ or $|\rho|$ becomes bigger, and only become identical (distributed as $\rho(Z_1^2 - Z_2^2)$) when $\gamma \to 0$.

Remark 2.2. The models are locally asymptotically normal at $\Sigma^* \notin \mathcal{M}_{\text{sing}}$. By regularity, replacing the constant elements in Eqs. (2.21) and (2.22) with sequences in n that converge to the corresponding limits does not alter the asymptotic distribution of $\lambda_n^{(0:1)}$.

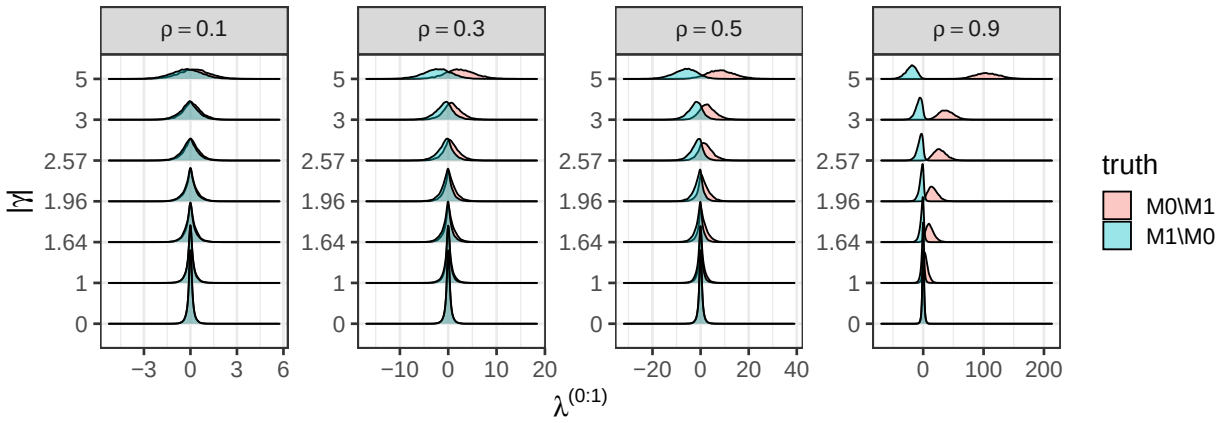


Figure 2.2: Asymptotic distributions of $\lambda_n^{(0:1)}$ under $\Sigma_n^{(0)} \in \mathcal{M}_0 \setminus \mathcal{M}_1$ and $\Sigma_n^{(1)} \in \mathcal{M}_1 \setminus \mathcal{M}_0$ in the weak-strong regime.

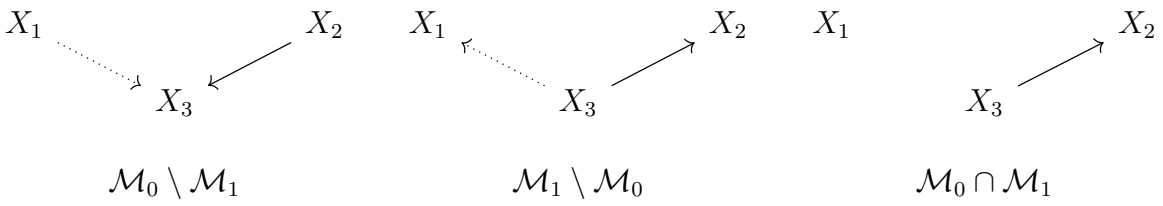


Figure 2.3: Two types of local sequences and their common limit.

2.4.2 Weak-weak regime

Now we study the asymptotic under $\rho_{13,n}, \rho_{23,n} \to 0$ and $\sqrt{n}\rho_{13,n}\rho_{23,n} \to \delta$. The limiting covariance is $\Sigma^* = \mathrm{diag}(\sigma_{11}, \sigma_{22}, \sigma_{33}) \in \mathcal{M}_{\text{sing}}$, towards which we consider two local sequences

$$\Sigma_n^{(0)} = \begin{pmatrix} \sigma_{11} & 0 & \rho_{13,n}\sqrt{\sigma_{11}\sigma_{33}} \\ 0 & \sigma_{22} & \rho_{23,n}\sqrt{\sigma_{11}\sigma_{33}} \\ \rho_{13,n}\sqrt{\sigma_{11}\sigma_{33}} & \rho_{23,n}\sqrt{\sigma_{11}\sigma_{33}} & \sigma_{33} \end{pmatrix} \in \mathcal{M}_0 \setminus \mathcal{M}_1 \tag{2.25}$$

and

$$\Sigma_n^{(1)} = \begin{pmatrix} \sigma_{11} & \rho_{13,n}\rho_{23,n}\sqrt{\sigma_{11}\sigma_{22}} & \rho_{13,n}\sqrt{\sigma_{11}\sigma_{33}} \\ \rho_{13,n}\rho_{23,n}\sqrt{\sigma_{11}\sigma_{22}} & \sigma_{22} & \rho_{23,n}\sqrt{\sigma_{22}\sigma_{33}} \\ \rho_{13,n}\sqrt{\sigma_{11,n}\sigma_{33}} & \rho_{23,n}\sqrt{\sigma_{22}\sigma_{33}} & \sigma_{33} \end{pmatrix} \in \mathcal{M}_1 \setminus \mathcal{M}_0. \tag{2.26}$$

Proposition 2.3. *Given* $\rho_{13,n}\rho_{23,n} = \delta n^{-1/2} + o(n^{-1/2})$ *for* $\delta \neq 0$ *and* $\rho_{13,n}, \rho_{23,n} \to 0$. *Under* $\Sigma_n^{(i)} \in \mathcal{M}_i \setminus \mathcal{M}_{1-i}$ *for* $i = 0, 1$, *we have*

$$\lambda_n^{(0:1)} \to_d \delta(2Z + (-1)^i\delta) =_d \mathcal{N}\left((-1)^i\delta^2, (2\delta)^2\right). \tag{2.27}$$

The limit is a centered Gaussian shifted and then scaled by δ. Plots for a few values of δ are given by Fig. 2.4.

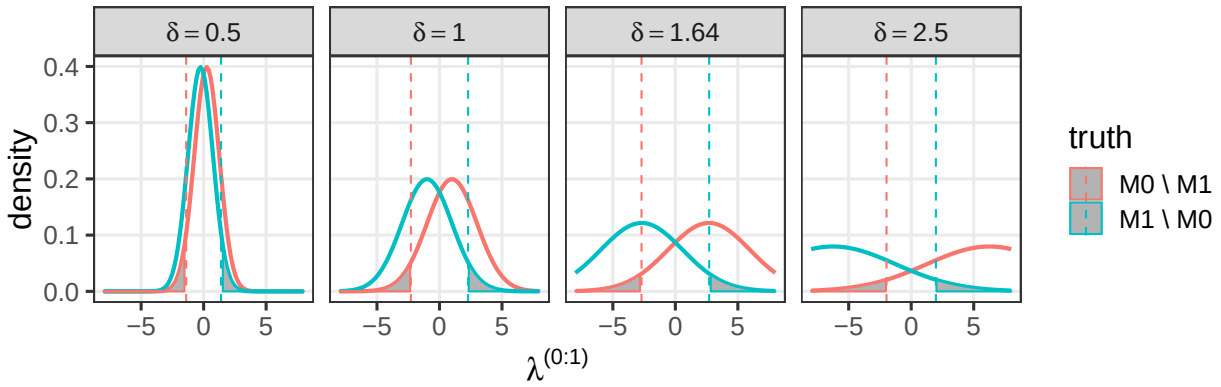


Figure 2.4: Asymptotic distribution of $\lambda_n^{(0:1)}$ in the weak-weak regime under $\mathcal{M}_0 \setminus \mathcal{M}_1$ (red) and $\mathcal{M}_1 \setminus \mathcal{M}_0$ (blue). The vertical lines and shaded areas correspond to 95% upper/lower quantiles.

Remark 2.3. The Gaussian asymptotic in Proposition 2.3 does *not* depend on how $\rho_{13,n}$ and $\rho_{23,n}$ approach zero *individually*. We verify it with simulations shown in Figure 2.5. We simulate under $n = 10,000$ for $5,000$ replicates. We set $\rho_{13,n} = rn^{-a}$ and $\rho_{23,n} = tn^{-(1/2-a)}$ such that $\rho_{13,n}\rho_{23,n} = \delta n^{-1/2}$ for $\delta = rt$ under different values of a.

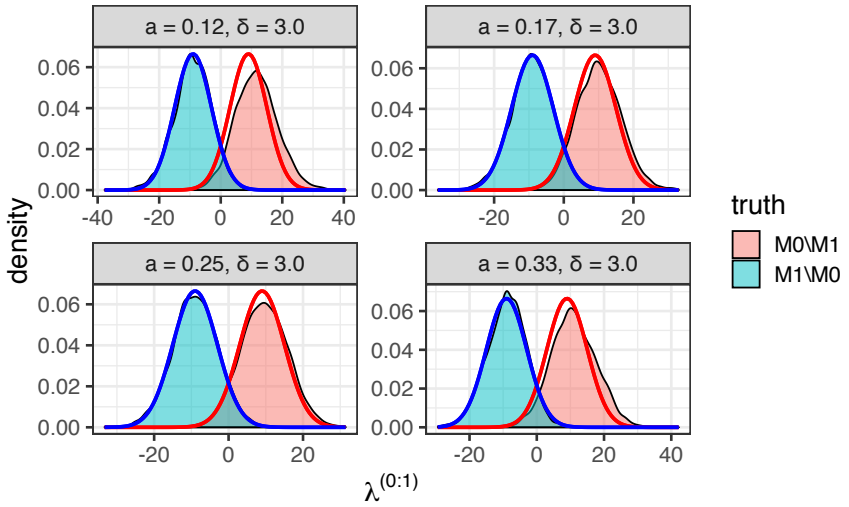


Figure 2.5: Simulated distribution of log-likelihood ratio under $\rho_{13,n} = rn^{-a}$ and $\rho_{23,n} = tn^{-(1/2-a)}$ such that $\rho_{13,n}\rho_{23,n} = \delta n^{-1/2}$ for $\delta = rt$. Red and blue solid curves are theoretical distributions.

2.4.3 Limit experiments

We establish the equivalence of testing the two models local asymptotics to that of a limit experiment, which sheds light on the form of the asymptotic distribution. As we will see, the limit experiments are Gaussian location experiments and the problem is asymptotically equivalent to testing the location between two lines from a single normal observation. Further by weak convergence, $\lambda_n^{(0:1)}$ is asymptotically distributed as the likelihood ratio statistic arising from the limit experiment. The reader is referred to van der Vaart (2000, Chapter 7, 9 and 16) for more background.

The weak-strong regime We characterize the limit experiment in the weak-strong regime.

Proposition 2.4. *The family of distributions* $\{P_{\mathbf{\Sigma}^* + \mathsf{G}\, h/\sqrt{n}} : h \in \mathbb{R}^2\}$ *is locally asymptotically*

normal, where

$$\Sigma^* = \begin{pmatrix} \sigma_{11} & 0 & 0 \\ 0 & \sigma_{22} & \sigma_{23} \\ 0 & \sigma_{23} & \sigma_{33} \end{pmatrix}, \quad h = (\mathsf{h}_1, \mathsf{h}_2)^\top,$$

and $\mathsf{G} : \mathbb{R}^2 \to \mathbb{R}^{3\times 3}$ *is a linear operator*

$$\mathsf{G}\, h := \begin{pmatrix} 0 & \mathsf{h}_1 & \mathsf{h}_2 \\ \mathsf{h}_1 & 0 & 0 \\ \mathsf{h}_2 & 0 & 0 \end{pmatrix}.$$

Proof. The Gaussian model P_Σ is differentiable in quadratic mean at Σ^*. The result follows from van der Vaart (2000, 7.14 and 7.15). □

The limit experiment of a LAN (locally asymptotically normal) family is a normal location experiment.

Proposition 2.5. *The sequence of experiments indexed by the local parameter h converges to the following normal location experiment*

$$\left(P_{\Sigma^* + \mathsf{G}\, h/\sqrt{n}}\right)_{h\in\mathbb{R}^2} \rightsquigarrow \left(\mathcal{N}(h, I_{\Sigma^*}^{-1})\right)_{h\in\mathbb{R}^2}, \tag{2.28}$$

where

$$I_{\Sigma^*}^{-1} = \sigma_{11} \begin{pmatrix} \sigma_{22} & \rho\sqrt{\sigma_{22}\sigma_{33}} \\ \rho\sqrt{\sigma_{22}\sigma_{33}} & \sigma_{33} \end{pmatrix}.$$

Proof. $\{P_{\boldsymbol{\Sigma}^* + \mathsf{G}\, h/\sqrt{n}} : h \in \mathbb{R}^2\}$ is LAN with non-singular Fisher information I_Σ^*, which is the conditional information matrix of $(\sigma_{12}, \sigma_{13})$ under P_Σ given $(\sigma_{11}, \sigma_{22}, \sigma_{33}, \sigma_{23})$, corresponding to $(\mathsf{h}_1, \mathsf{h}_2)$. The result then follows from van der Vaart (2000, Corollary 9.5). □

The local sequences Eq. (2.21) and Eq. (2.22) can be identified as $\Sigma^* + \mathsf{G}\, h/\sqrt{n}$ with h taking value of

$$h_0 = (0, \gamma\sqrt{\sigma_{11}\sigma_{33}})^\top, \quad h_1 = (\gamma\rho\sqrt{\sigma_{11}\sigma_{22}}, \gamma\sqrt{\sigma_{11}\sigma_{33}})^\top \tag{2.29}$$

respectively. Models $\mathcal{M}_0$ and $\mathcal{M}_1$ correspond to the set of h_0 and h_1 respectively as γ varies in $\mathbb{R}$. That is, $\mathcal{M}_0$ and $\mathcal{M}_1$ are represented by local parameter spaces

$$H_0 = \{0\} \times \mathbb{R}, \quad H_1 = \{(\gamma\rho\sqrt{\sigma_{11}\sigma_{22}}, \gamma\sqrt{\sigma_{11}\sigma_{33}})^\top : \gamma \in \mathbb{R}\}, \tag{2.30}$$

which consist of all limits of $\sqrt{n}\,\mathsf{G}^{-1}(\Sigma_n^{(i)} - \Sigma^*)$ for $i = 0, 1$ (see van der Vaart (2000, Chapter 7.4)). Note H_0 and H_1 are lines in $\mathbb{R}^2$ (affine) and they correspond to tangent cones from $\mathcal{M}_0$ and $\mathcal{M}_1$ at Σ^* under Chernoff regularity; see also Drton (2009b) and Geyer (1994).

Proposition 2.6. *Suppose $I_{\Sigma^*}^{-1} = LL^\top$. For $i = 0, 1$, under $P^n_{\Sigma^* + \mathsf{G}\,h/\sqrt{n}}$ for $h = h_i$, it holds that $(-1)^i \lambda_n^{(0:1)}$ is asymptotically distributed as the likelihood ratio statistic of testing*

$$\mu \in L^{-1}H_i \quad \textit{versus} \quad \mu \in L^{-1}(H_{1-i} - h_i) \tag{2.31}$$

from a single observation $Z \sim \mathcal{N}(\mu = \mathbf{0}, I_2)$.

Now we derive limit experiments based on Proposition 2.6. The Cholesky decomposition gives

$$L = \sqrt{\sigma_{11}} \begin{pmatrix} \sqrt{\sigma_{22}} & 0 \\ \rho\sqrt{\sigma_{33}} & \sqrt{(1-\rho^2)\sigma_{33}} \end{pmatrix},$$
$$L^{-1} = \frac{1}{\sqrt{\sigma_{11}}} \begin{pmatrix} 1/\sqrt{\sigma_{22}} & 0 \\ -\rho/\sqrt{(1-\rho^2)\sigma_{22}} & 1/\sqrt{(1-\rho^2)\sigma_{33}} \end{pmatrix}.$$

We have, when $h = h_0$

$$L^{-1}H_0 = \{0\} \times \mathbb{R}, \quad L^{-1}(H_1 - h) = \left\{ \begin{pmatrix} 0 \\ \frac{-\gamma}{\sqrt{1-\rho^2}} \end{pmatrix} + u \begin{pmatrix} \rho \\ \sqrt{1-\rho^2} \end{pmatrix} : u \in \mathbb{R} \right\}, \tag{2.32}$$

and when $h = h_1$

$$L^{-1}(H_0 - h) = \{-\gamma\rho\} \times \mathbb{R}, \quad L^{-1}H_1 = \left\{ u \begin{pmatrix} \rho \\ \sqrt{1-\rho^2} \end{pmatrix} : u \in \mathbb{R} \right\}. \tag{2.33}$$

They are visualized in Figure 2.6. The limit experiments Eq. (2.32) and Eq. (2.33) are of the same type as they are both characterized by an *angle* and an *intercept*. The two have the same angle $\theta = \arcsin\rho$ and their intercepts are related by a factor of $1/\sqrt{1-\rho^2}$.

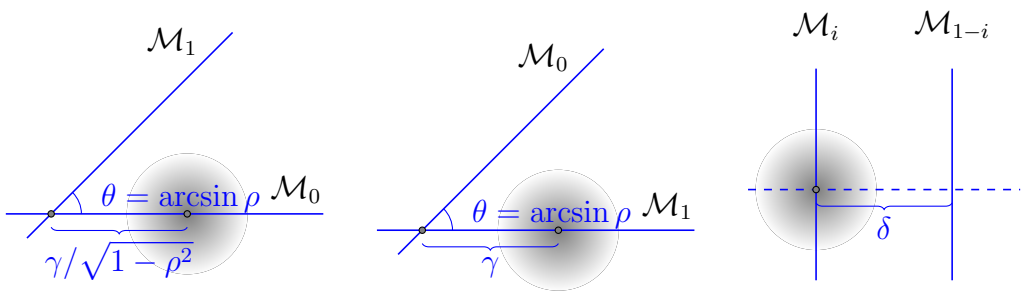


Figure 2.6: Three limit experiments: (1) $\mathcal{M}_0 \setminus \mathcal{M}_1$ in the weak-strong regime, (2) $\mathcal{M}_1 \setminus \mathcal{M}_0$ in the weak-strong regime, and (3) $\mathcal{M}_i \setminus \mathcal{M}_{1-i}$ in the weak-weak regime for $i = 0, 1$.

The weak-weak regime The Gaussian limit in Proposition 2.3 shows that the limit experiment of the weak-weak regime is testing the location of a *univariate* normal between two points; see the last panel of Fig. 2.6.

Corollary 2.2. *Testing $\mathcal{M}_0$ versus $\mathcal{M}_1$ under $\sqrt{n}\rho_{12,n}\rho_{13,n} \to \delta$ for $\delta \neq 0$ with $\rho_{12,n}, \rho_{13,n} \to 0$ is asymptotically equivalent to testing $H_0 : \mu = 0$ versus $H_1 : \mu = \delta$ from a single observation $Z \sim \mathcal{N}(\mu, 1)$.*

2.5 Envelope distributions

Though it may at first appear otherwise, the asymptotic distributions as obtained in Proposition 2.2 and Proposition 2.3 are not directly applicable to forming decision rules. This is due to the *non-uniformity* of the asymptotics.

Firstly, the asymptotic depends on the *regime*: weak-strong versus weak-weak, namely *where* the local sequence converges to. And the law is *discontinuous* between the two regimes. That is, the law in the weak-strong regime (scaled difference of noncentral chi-squares) does not converge to that of the weak-weak regime (Gaussian) as $\rho \to 0$. Furthermore, a procedure that firstly estimates the regime and then uses the corresponding distribution to form decision boundary, is susceptible to irregularity issues. Additionally, it is difficult to judge if an edge is weak based on whether its confidence interval contains zero without further assumptions,

as illustrated by the following example.

Example 2.1. Suppose $X_i \overset{\text{iid}}{\sim} \mathcal{N}(\gamma/\sqrt{n}, \sigma^2)$ for $i = 1, \cdots, n$. The usual $(1-\alpha)$-level confidence interval for the mean of X is $\bar{X}_n \pm z_{\alpha/2}\hat{\sigma}_n/\sqrt{n}$. The probability that it contains zero is

$$\begin{aligned}\Pr\left(0 \in (\bar{X}_n \pm z_{\alpha/2}\hat{\sigma}_n/\sqrt{n})\right) &= \Pr\left(\sqrt{n}\bar{X}_n/\hat{\sigma}_n \in (\pm z_{\alpha/2})\right) \\ &\to \Pr(Z + \gamma \in (\pm z_{\alpha/2})) < 1 - \alpha\end{aligned}$$

for $\gamma \neq 0$ and $Z \sim \mathcal{N}(0,1)$. A large enough γ can be chosen to make this probability arbitrarily small.

Secondly, given the regime, the distribution depends on the value of a *local parameter* (γ for strong-weak and δ for weak-weak), which determines *how* the local sequence converges. Due to the $\sqrt{n}$ factor, the standard error for its estimator does not vanish and in general the local parameter cannot be consistently estimated. The reader is referred to Berger and Boos (1994); Andrews (2001) for discussions in the literature on the treatment of asymptotic distributions involving nuisance parameters that are not point-identified. Here we take a different approach, presented as follows.

The non-uniformity of asymptotic distributions motivates us to seek a procedure that *circumvents* the inference on the regime and the local parameter. In this section, we study the "extremal" distributions arising from the asymptotic distributions as the local parameter varies in $\mathbb{R}$.

Definition 2.1. *Given a family of distribution functions $\{F_h : h \in \mathcal{H}\}$ on $\mathbb{R}$, define*

$$\bar{F}^*(x) = \sup_{h \in \mathcal{H}} F_h(x),$$

and

$$\bar{F}(x) = \begin{cases} \bar{F}^*(x), & \bar{F}^* \text{ is continuous at } x \\ \lim_{y \to x^+} \bar{F}^*(y), & \bar{F}^* \text{ is discontinuous at } x \end{cases}. \tag{2.34}$$

We call $\bar{F}$ the envelope distribution of $\{F_h : h \in \mathcal{H}\}$ if $\bar{F}$ is a valid distribution function.

Lemma 2.3. *$\bar{F}^*(x)$ is left-continuous if every $F_h(x)$ for $h \in \mathcal{H}$ is continuous.*

Proof. Fix any x and $\delta > 0$, for $\epsilon > 0$ we have $|\bar{F}^*(x) - \bar{F}^*(x-\epsilon)| = \sup_h F_h(x) - \sup_h F_h(x-\epsilon)$. By definition of supremum, there exists $h' \in \mathcal{H}$ such that $F_{h'}(x) \geq \sup_h F_h(x) - \delta/2$. Hence, $|\bar{F}^*(x) - \bar{F}^*(x-\epsilon)| \leq \delta/2 + F_{h'}(x) - F_{h'}(x-\epsilon)$. By continuity of $F_{h'}$, choosing $\epsilon > 0$ such that $F_{h'}(x) - F_{h'}(x-\epsilon) \leq \delta/2$ shows that $\bar{F}^*(x)$ is left-continuous. □

Lemma 2.4. *If $\bar{F}^*(x) \to 0$ as $x \to -\infty$, then $\bar{F}(x)$ is a valid distribution function.*

Proof. Given any $x \leq x'$, $\sup_h F_h(x) \leq \sup_h F_h(x')$ by monotonicity of every F_h. Since $\bar{F}^*$ is non-decreasing, by Folland (1999, Theorem 3.23), the set of points at which $\bar{F}^*$ is discontinuous is countable. By redefining the function value at these points to be their right limits, $\bar{F}$ is right continuous. Also, $\bar{F}(x) \geq \bar{F}^*(x) \to 1$ as $x \to +\infty$ since every $F_h(x) \to 1$. Finally, as $x \to -\infty$ if $\bar{F}^*(x) \to 0$, then $\bar{F}(x) \to 0$. $\bar{F}$ is a distribution function. □

2.5.1 Weak-weak regime

Proposition 2.7. *Let $G_\delta = \{\mathcal{N}(\delta^2, (2\delta)^2) : \delta \in \mathbb{R}\}$ be the asymptotic distributions for the weak-weak regime under $\mathcal{M}_0 \setminus \mathcal{M}_1$. The envelope of $\{G_\delta\}$ is an equal-probability mixture of $(-\chi_1^2)$ and a point mass at zero, namely*

$$\bar{G}(x) = \frac{1}{2}\left(1 - F_{\chi_1^2}(-x)\right)\mathbb{I}_{x<0} + \frac{1}{2}\mathbb{I}_{x\geq 0} \tag{2.35}$$

The corresponding envelope under $\mathcal{M}_1 \setminus \mathcal{M}_0$ is distributed as its negation.

Proof. It suffices to consider $\delta \geq 0$. Given any $x < 0$,

$$\sup_\delta \Pr(\delta^2 + 2\delta Z \leq x) = \sup_{\delta > 0} \Phi\left(\frac{x - \delta^2}{2\delta}\right) = \sup_{\delta > 0} \Phi\left(-\left[\frac{-x}{2\delta} + \frac{\delta}{2}\right]\right) = \Phi(-\sqrt{-x}),$$

where $\delta^* = \sqrt{-x}$ is the maximizer; Given any $x \geq 0$, $\delta = 0$ maximizes the probability to one. Hence, the envelope CDF is

$$\bar{G}(x) = \begin{cases} \Phi(-\sqrt{-x}), & x < 0 \\ 1, & x \geq 0 \end{cases},$$

from which it follows that

$$\bar{g}(x) = \bar{G}'(x) = \frac{1}{2} f_{\chi_1^2}(-x)\mathbb{I}_{x<0} + \frac{1}{2}\delta_0(x).$$

The envelope for $\mathcal{M}_1 \setminus \mathcal{M}_0$ follows from symmetry. □

Since when $\mathcal{M}_0$ is true, the region for rejecting $\mathcal{M}_0$ should take the form $(-\infty, r)$ for some $r < 0$, only the negative part of $\bar{G}$ is relevant for decision making. It follows from Proposition 2.7 that the negative part of $\bar{G}$ is distributed as χ_1^2. The formation of the envelope is visualized in Fig. 2.7, which aligns with the behavior observed in Fig. 2.4, where as δ grows, the quantiles for $\alpha = 0.05$ first moves outward for $\delta \in (0.5, 1.64)$ and then moves inward for $\delta \in (1.64, \infty)$.

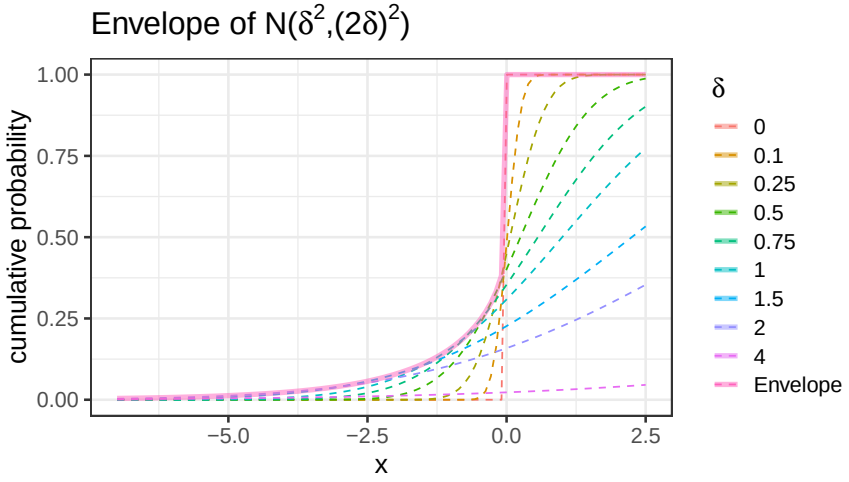


Figure 2.7: The envelope CDF $\bar{G}$ for the weak-weak regime.

2.5.2 Weak-strong regime

Now we study the envelope distributions under the weak-strong regime. We first observe that the envelope distributions, if they exist, must be symmetric for Eq. (2.23) and Eq. (2.24), in the sense that they are distributed as the negation of each other. The symmetry holds

because the two local parameters are related by a factor of $1/\sqrt{1-\rho^2}$ (see Fig. 2.6), and hence the suprema are taken over the same set of laws up to a difference in the sign. Fix ρ, let $\{F_{\rho,\gamma} : \gamma \in \mathbb{R}\}$ be the family of asymptotic distributions in the weak-strong regime under $\mathcal{M}_0 \setminus \mathcal{M}_1$ as given in Eq. (2.23). Let $\bar{F}_\rho$ be its envelope distribution function.

Proposition 2.8. *$\bar{F}_\rho$ is a valid distribution function for $|\rho| \in (0, 1]$.*

The following result shows $F_{\rho,\gamma=0}$ constitutes the envelope for the positive part of $\bar{F}_\rho$.

Proposition 2.9. *The positive part of $\bar{F}_\rho$ for $|\rho| \in (0, 1]$ is distributed as the positive part of $\rho(Z_1^2 - Z_2^2)$ for $Z_1, Z_2 \overset{iid}{\sim} \mathcal{N}(0, 1)$.*

Corollary 2.3. $\bar{F}_\rho(0) \equiv 1/2$.

Unfortunately, we do not have an analytic form of the distribution for the negative part of $\bar{F}_\rho$, which is the part relevant for decision making, except for $\rho \to 0$ and $\rho = 1$.

Proposition 2.10 (Bessel envelope)**.** *$\bar{F}_{\rho=1} =_d Z_1^2 - Z_2^2$ for $Z_1, Z_2 \overset{iid}{\sim} \mathcal{N}(0, 1)$.*

The distribution in Proposition 2.10 is a difference between two independent χ_1^2 variables. The density, as plotted in Fig. 2.8, is

$$p_B(u) = \frac{1}{2\pi} K_0(|u|/2),$$

where K_0 is a modified Bessel function of the second kind. It is referred to as a K-form Bessel distribution in the literature; see Johnson et al. (1995, Chapter 4.4), Bhattacharyya (1942) and Simon (2007, Page 25).

Proposition 2.11 (Continuity of envelope)**.** *$\bar{F}_\rho \to_d \bar{G}$ as $\rho \to 0$, where $\bar{G}$ is the envelope distribution for the weak-weak regime given in Proposition 2.7.*

Perhaps surprisingly, Proposition 2.11 shows that the asymptotic envelope is *continuous* between the two regimes, which bridges the discontinuity of the asymptotic distributions of $\lambda_n^{(0:1)}$ as presented in Propositions 2.2 and 2.3. Therefore, taking the envelope resolves the

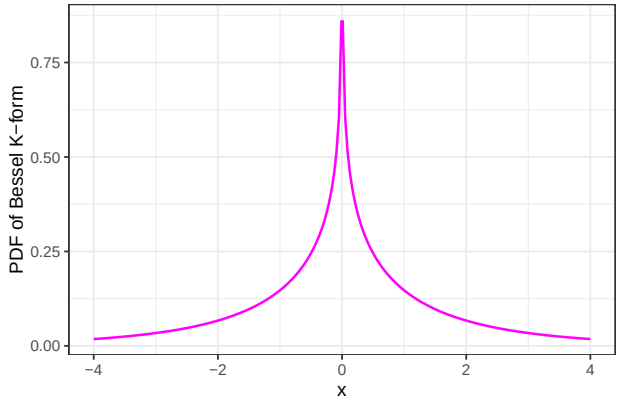


Figure 2.8: The density for $\bar{F}_{\rho=1} =_d Z_1^2 - Z_2^2$.

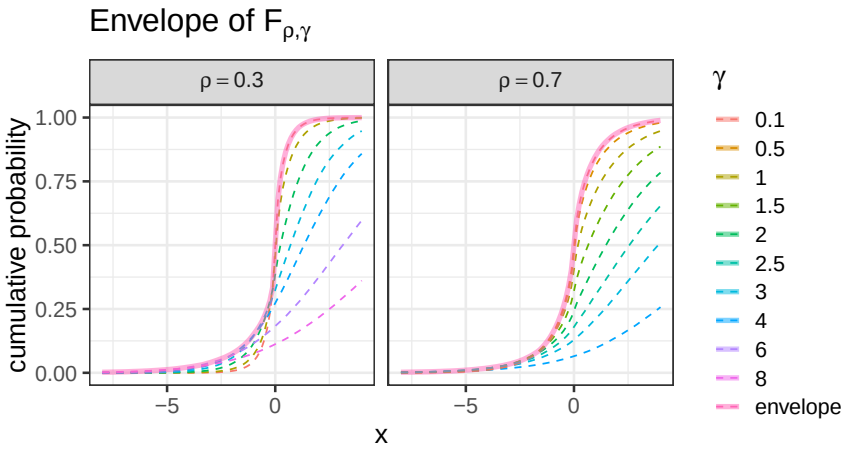


Figure 2.9: The envelope distribution $\bar{F}_\rho$ under the strong-weak regime for $\rho = 0.3, 0.7$.

non-uniformity issue in terms of *both* the regime and the local parameter. Now with this we extend the definition of the envelope $\bar{F}_\rho$ to $\rho \in [0, 1]$ by writing $\bar{F}_{\rho=0} = \bar{G}$.

Figure 2.9 showcases two envelopes. In the absence of an analytic form for $\rho \in (0, 1)$, the envelopes can be numerically simulated by taking the supremum over a grid of values for γ. We observe from simulations that there exists $\gamma^*(x, \rho) \in (0, \infty)$ such that $\pm\gamma^*$ uniquely maximizes $F_{\gamma,\rho}(x)$.

Finally, we conclude this section by noting the following *envelope of envelopes*. See Figure 2.10 for an illustration. This result will be used in the next section to form simple decision rules based on the Bessel distribution.

Proposition 2.12 (Envelope of envelopes)**.** *The negative part of the envelope of* $\{\bar{F}_\rho : \rho \in [0,1]\}$ *is distributed as the negative part of* $\bar{F}_{\rho=1}$ *(Bessel).*

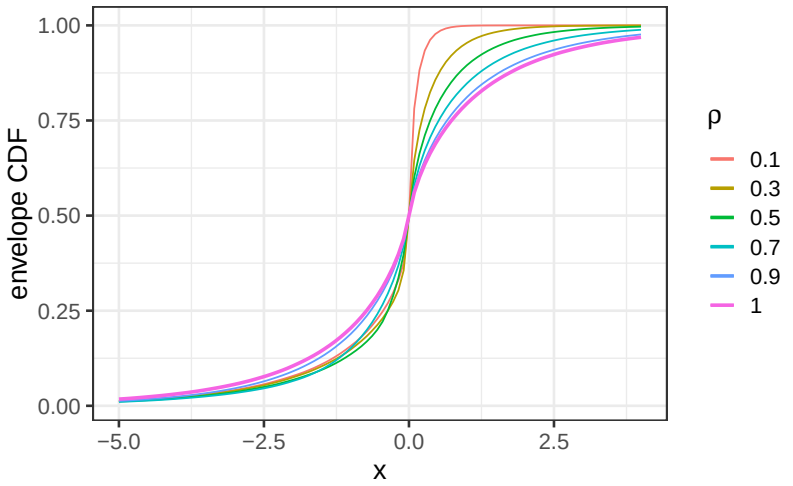


Figure 2.10: The negative part of envelope of $\{\bar{F}_\rho : \rho \in [0,1]\}$ is the negative part of $\bar{F}_{\rho=1}$.

2.6 Model selection procedures

Since we are selecting between two non-nested models, we want to refrain from choosing one of them as the default (the null hypothesis). By treating $\mathcal{M}_0$ and $\mathcal{M}_1$ symmetrically, however, a procedure that takes output value in $\{\mathcal{M}_0, \mathcal{M}_1\}$ cannot simultaneously control both types of error under a given tolerance. It can be seen from Figs. 2.2 and 2.5 that there are cases where the asymptotic distributions of $\lambda_n^{(0:1)}$ under $P_{\Sigma_n^{(0)}}$ and $P_{\Sigma_n^{(1)}}$ significantly overlap. In such cases, insisting on a dichotomous choice will inevitably result in a high probability of error under at least one model.

To deal with this possible indistinguishability, we opt for a procedure with *three options*: if two models can be sufficiently distinguished, it selects one of them; otherwise it refrains from commitment by selecting *both models*, formally denoted as the union $\mathcal{M}_0 \cup \mathcal{M}_1$. It is worth stressing that we always assume at least one of the two models is true. By such a design, when the procedure does not output the union, we are ensured that the probability of choosing the wrong model is small, being controlled below a given tolerance α. In contrast, in the usual hypothesis testing framework where supposedly $\mathcal{M}_0$ is the null and $\mathcal{M}_1$ is the alternative, one typically cannot simultaneously control both type-I and type-II error. In other words, our procedure selects model with "confidence". Recently the same notion has been investigated by Lei (2014) in a classification setting; Robins et al. (2003) also allows a test to make no decision when faced with ambiguity. We formalize the concept as follows.

Suppose $(X_1, X_2, X_3) \in \mathbb{R}^{n\times 3}$ consists of n independent samples from $\mathcal{N}(\mathbf{0}, \Sigma_n)$, where $\Sigma_n \in \mathcal{M}_0 \cup \mathcal{M}_1$ is allowed to change with n and $\Sigma_n \to \Sigma^*$. The sequence Σ_n models signal strength relative to the sample size. We consider a deterministic decision rule

$$\phi_n(S_n) : \mathbb{R}^{3\times 3}_{\text{PSD}} \to \{\mathcal{M}_0, \mathcal{M}_1, \mathcal{M}_0 \cup \mathcal{M}_1\}, \tag{2.36}$$

where sample covariance S_n is the sufficient statistic. For a given sequence of $\Sigma_n \in \mathcal{M}_i \backslash \mathcal{M}_{1-i}$ with $\Sigma_n \to \Sigma^* \in \mathcal{M}_i \cup \mathcal{M}_{1-i}$, we define the asymptotic (type-I) error of ϕ_n as the large-sample probability of rejecting the true model, i.e.,

$$p_{\text{err}}((\Sigma_n)) := \limsup_{n\to\infty} \Pr(\phi_n(S_n) = \mathcal{M}_{1-i}), \tag{2.37}$$

where the probability is taken under $P^n_{\Sigma_n}$. Similarly, the asymptotic power is defined as

$$p_{\text{pow}}((\Sigma_n)) := \liminf_{n\to\infty} \Pr(\phi_n(S_n) = \mathcal{M}_i). \tag{2.38}$$

We say that the error is *uniformly* controlled below a given size $\alpha > 0$, if

$$p^{(0)}_{\text{err}} := \sup_{(\Sigma^{(0)}_n)} p_{\text{err}}((\Sigma^{(0)}_n)) \le \alpha \quad \text{and} \quad p^{(1)}_{\text{err}} := \sup_{(\Sigma^{(1)}_n)} p_{\text{err}}((\Sigma^{(1)}_n)) \le \alpha, \tag{2.39}$$

where for $i = 0, 1$ the supremum for $(\Sigma^{(i)}_n)$ is taken over all converging sequences of $\Sigma^{(i)}_n$ within $\mathcal{M}_i \setminus \mathcal{M}_{1-i}$ (the limit can be in either $\mathcal{M}_i \setminus \mathcal{M}_{1-i}$ or $\mathcal{M}_i \cap \mathcal{M}_{1-i}$). In general, the

power $p_{\text{pow}}((\Sigma_n))$ depends on the sequence considered and we do not seek power optimality or guarantee in a uniform sense. In the next section, we will compare the power of several proposed procedures to the theoretical optimal for Σ_n considered in the two regimes of local asymptotics.

By construction, using the α-quantile of the envelope as the decision boundary achieves uniform error control. Based on the envelope of envelopes, a simple *uniform rule* is

$$\phi_n^{\text{unif}} = \begin{cases} \mathcal{M}_0, & \lambda_n^{(0:1)} > -\bar{F}_{\rho=1}^{-1}(\alpha) \\ \mathcal{M}_1, & \lambda_n^{(0:1)} < \bar{F}_{\rho=1}^{-1}(\alpha) \\ \mathcal{M}_0 \cup \mathcal{M}_1, & \text{otherwise} \end{cases} . \tag{2.40}$$

To gain more power, since $\bar{F}_\rho$ is continuous in ρ and ρ can be consistently estimated (recall that $\rho = \rho_{\text{strong}}$ in the weak-strong regime, and $\rho = 0$ in the weak-weak regime), an *adaptive rule* can be formed as

$$\phi_n^{\text{ada}} = \begin{cases} \mathcal{M}_0, & \lambda_n^{(0:1)} > -\bar{F}_{\hat{\rho}_n}^{-1}(\alpha) \\ \mathcal{M}_1, & \lambda_n^{(0:1)} < \bar{F}_{\hat{\rho}_n}^{-1}(\alpha) \\ \mathcal{M}_0 \cup \mathcal{M}_1, & \text{otherwise} \end{cases} , \tag{2.41}$$

where $\hat{\rho}_n = |\hat{\rho}_{13,n}| \vee |\hat{\rho}_{23,n}|$ is the MLE for $|\rho|$. If it is desired to report a p-value, consider a potentially conservative p-value $= \bar{F}_\rho(-|\lambda_n^{(0:1)}|)$. For $\rho = 1$ and $\rho = \hat{\rho}_n$ respectively, the uniform rule and the adaptive rule can be then restated as

$$\phi_n = \begin{cases} \mathcal{M}_0, & \lambda_n^{(0:1)} > 0 \text{ and } p\text{-value} < \alpha \\ \mathcal{M}_1, & \lambda_n^{(0:1)} < 0 \text{ and } p\text{-value} < \alpha \\ \mathcal{M}_0 \cup \mathcal{M}_1, & \text{otherwise} \end{cases} .$$

The conservative p-value can be computed numerically by Monte Carlo and then taking the maximum over a grid of values for γ.

Theorem 2.1. *The adaptive rule ϕ_n^{ada} controls asymptotic error uniformly below α for $0 < \alpha < 1/2$.*

Proof. We show error guarantee when $\mathcal{M}_0$ is true. The same argument holds when $\mathcal{M}_1$ is true. It suffices to show for any converging sequence $\Sigma_n \in \mathcal{M}_0 \setminus \mathcal{M}_1$,

$$p_{\text{err}}((\Sigma_n)) = \limsup_{n\to\infty} \Pr(\phi_n(S_n) = \mathcal{M}_1) \le \alpha,$$

where the probability is measured under $P^n_{\Sigma_n}$. Suppose $\Sigma_n \to \Sigma^* \in \mathcal{M}_0 \cup \mathcal{M}_1$. If $\Sigma^* \notin \mathcal{M}_0 \cap \mathcal{M}_1$, then $\lambda_n^{(0:1)}$ is unbounded in probability towards $+\infty$. Hence $\Pr(\phi_n(S_n) = \mathcal{M}_0) \to 1$ and $p_{\text{err}}(\Sigma_n) = 0$. In the following we prove the claim for $\Sigma^* \in \mathcal{M}_0 \cap \mathcal{M}_1$. Suppose $\hat{\rho}_{ij,n}$, $\rho_{ij,n}$ and ρ_{ij} respectively denote the corresponding correlation coefficient of S_n, Σ_n and Σ^*. We have three cases depending on the rate at which Σ_n converges.

1. When $|\rho_{13,n}\rho_{23,n}| \asymp 1/\sqrt{n}$, there are two regimes depending on Σ^*.

   (a) In the weak-strong regime, without loss of generality suppose $\sqrt{n}\rho_{13,n} \to \gamma \neq 0$ and $\rho_{23,n} \to \rho \neq 0$. By consistency $\hat{\rho}_n = |\hat{\rho}_{13,n}| \vee |\hat{\rho}_{23,n}| \to_p |\rho|$ and the definition of envelope, we have

   $$\limsup_n \Pr(\lambda_n^{(0:1)} < \bar{F}^{-1}_{\hat{\rho}_n}(\alpha)) = F_{\gamma,\rho}(\bar{F}^{-1}_{\rho}(\alpha)) \le F_{\gamma,\rho}(F^{-1}_{\gamma,\rho}(\alpha)) = \alpha.$$

   (b) In the weak-weak regime, suppose $\sqrt{n}\rho_{13,n}\rho_{23,n} \to \delta \neq 0$. We have $\hat{\rho}_n = |\hat{\rho}_{13,n}| \vee |\hat{\rho}_{23,n}| = (|\rho_{13,n}| \vee |\rho_{23,n}|) + O_p(1/\sqrt{n}) \to_p 0$ since both $\rho_{13,n}, \rho_{23,n} \to 0$. By Proposition 2.11, we have

   $$\begin{aligned}\limsup_n \Pr(\lambda_n^{(0:1)} < \bar{F}^{-1}_{\hat{\rho}_n}(\alpha)) &= G_\delta(\bar{F}^{-1}_{\rho=0}(\alpha)) \\ &= G_\delta(\bar{G}^{-1}(\alpha)) \le G_\delta(G_\delta^{-1}(\alpha)) = \alpha.\end{aligned}$$

2. When $|\rho_{13,n}\rho_{23,n}| = o(1/\sqrt{n})$, we have $\lambda_n^{(0:1)} \to_p 0$. Since $F_\rho^{-1}(\alpha) < c < 0$ for $\alpha < 1/2$, we have $\Pr(\phi_n(S_n) = \mathcal{M}_0 \cup \mathcal{M}_1) \to 1$.

3. When $|\rho_{13,n}\rho_{23,n}| = \omega(1/\sqrt{n})$, we have $\Pr(\lambda_n^{(0:1)} > c) \to 1$ for any constant c and hence $\Pr(\phi_n(S_n) = \mathcal{M}_0) \to 1$.

□

As can be seen from the proof, the consistency of model selection based on the loglikelihood (or AIC/BIC since in this case $\mathcal{M}_0$ and $\mathcal{M}_1$ have the same dimensions) is a special case when $|\rho_{13,n}\rho_{23,n}| = \omega(n^{-1/2})$, i.e., under strong signal or large enough sample size. However, under $|\rho_{13,n}\rho_{23,n}| = O(n^{-1/2})$, simply choosing the model with the highest loglikelihood (or the lowest AIC/BIC) can lead to large errors, as we will illustrate in the next section. Note that Theorem 2.1 provides a "rate-free" guarantee, in the sense that it does *not* require any *a priori* assumption on the rate of signal strength relative to the sample size. The envelope of envelopes leads to the same guarantee for the uniform rule.

Corollary 2.4. *The decision rule ϕ_n^{unif} controls asymptotic error uniformly below α for $0 < \alpha < 1/2$.*

Proof. It follows from Theorem 2.1 and Proposition 2.12. □

The uniform rule can be easily applied by comparing the difference in log-likelihoods to a single number, e.g., 3.19 for $\alpha = 0.05$ and 5.97 for $\alpha = 0.01$. The adaptive rule can be implemented by numerically evaluating $\bar{F}_\rho^{-1}(\alpha)$ via Monte Carlo on a grid of ρ and interpolating. Some values are plotted in Fig. 2.11 and tabulated in Table 2.1 based on 10^7 samples. It is interesting to notice that $\bar{F}_\rho^{-1}(\alpha)$ is not monotonic in $\rho \in [0, 1]$.

Table 2.1: Envelope quantiles $-\bar{F}_\rho^{-1}(\alpha)$ (Monte Carlo standard errors ≤ 0.01)

ρ	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
$\alpha = 0.05$	2.71	2.71	2.68	2.65	2.58	2.48	2.42	2.39	2.58	2.90	3.19
$\alpha = 0.01$	5.41	5.41	5.40	5.34	5.27	5.21	5.11	5.05	5.02	5.40	5.97

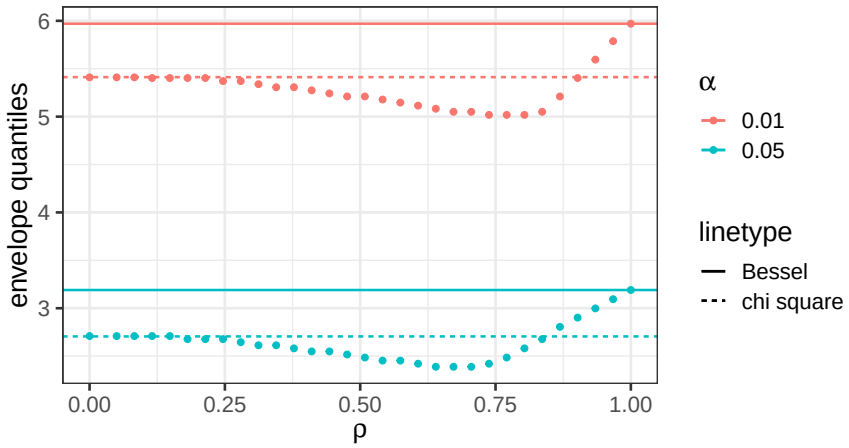


Figure 2.11: Negated α-quantiles of $\bar{F}_\rho$ evaluated on a grid.

2.7 Simulations

In this section we conduct numerical simulations to assess the performance of the adaptive and uniform decision rules proposed in the previous section. In subsequent simulations we use $\alpha = 0.05$. In addition to the two methods we propose, we also consider the following methods for comparison.

Naive The naive procedure selects the model with a higher likelihood (or equivalently, a lower AIC/BIC, since $\mathcal{M}_0$ and $\mathcal{M}_1$ have the same dimensions), namely

$$\phi_n^{\text{naive}} = \begin{cases} \mathcal{M}_0, & \lambda_n^{(0:1)} > 0 \\ \mathcal{M}_1, & \lambda_n^{(0:1)} < 0 \end{cases}.$$

This is effectively choosing a single model based on AIC/BIC since the penalty terms cancel out as $\mathcal{M}_0$ and $\mathcal{M}_1$ have the same dimension.

Interval Selection This method is adapted from Drton and Perlman (2004). We construct $(1-\alpha)$-level non-simultaneous confidence intervals on correlation coefficients ρ_{12} and $\rho_{12\cdot 3}$

with Fisher's z-transform (Fisher, 1924). The decision rule is

$$\phi_n^{\text{interval}} = \begin{cases} \mathcal{M}_0, & 0 \in \text{C.I.}(\rho_{12}) \text{ and } 0 \notin \text{C.I.}(\rho_{12\cdot 3}) \\ \mathcal{M}_1, & 0 \in \text{C.I.}(\rho_{12\cdot 3}) \text{ and } 0 \notin \text{C.I.}(\rho_{12}) \\ \mathcal{M}_0 \cup \mathcal{M}_1, & \text{otherwise} \end{cases}. \tag{2.42}$$

Note that the interval selection method controls asymptotic error below α. For example, when $\mathcal{M}_0$ is true,

$$\limsup_n \Pr(\phi_n^{\text{interval}} = \mathcal{M}_1) \leq \limsup_n \Pr(0 \notin \text{C.I.}(\rho_{12})) \leq \alpha.$$

We conduct numerical simulations in the following three settings.

2.7.1 Local hypotheses

We simulate under $\Sigma_n^{(0)}$ and $\Sigma_n^{(1)}$ (variances are set to unity) for the two regimes considered in Section 2.4. The power is compared to the theoretically optimal. Since exact values of the total variation distance are intractable, we plot bounds given by Eq. (2.16) in grey curves. We perform 4,000 replications for each point on the graphs.

See Figures 2.12 and 2.13 for the size and power in the weak-strong regime (Eqs. (2.21) and (2.22)) under $n = 1,000$. Smaller sample sizes $n = 100, 200, \ldots$ generate very similar results. See Figure 2.14 for the size and power in the weak-weak regime (Eqs. (2.25) and (2.26)), where we set $\rho_{13,n} = n^{-a/4}$, $\rho_{23,n} = n^{-1/2+a/4}$ and let a vary. We observe that (i) the naive method does not control error at all; (ii) the other three methods control error uniformly even under relatively small n. We also observe that the relation "adaptive" $>$ "uniform" $>$ "interval" holds in general in terms of both size and power. By comparing to the grey curves, we regard the adaptive rule as achieving near-optimal power in these settings.

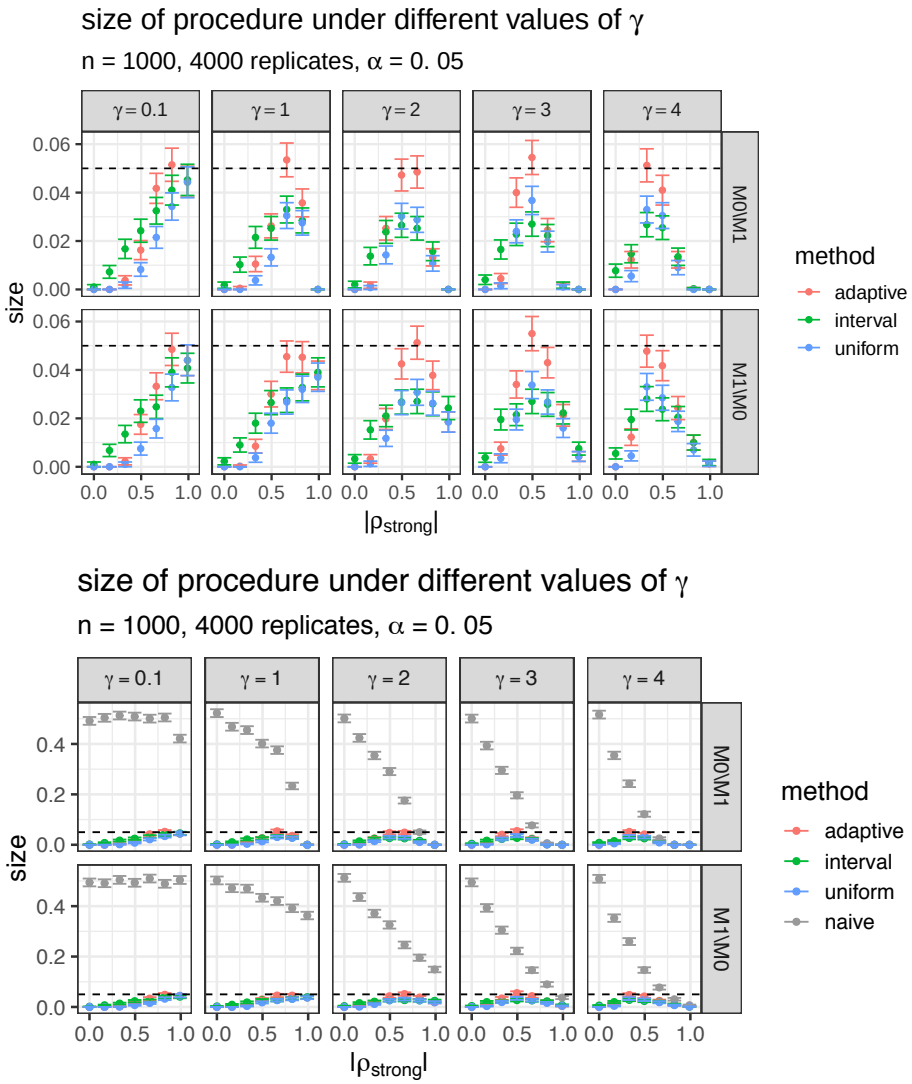


Figure 2.12: Size $\Pr(\phi_n = \mathcal{M}_{1-i}|\mathcal{M}_i)$ of the procedures (with 95% confidence intervals) under the weak-strong regime of local hypotheses. $\alpha = 0.05$ is marked as dashed. The naive method is only included in the second plot for better visualization.

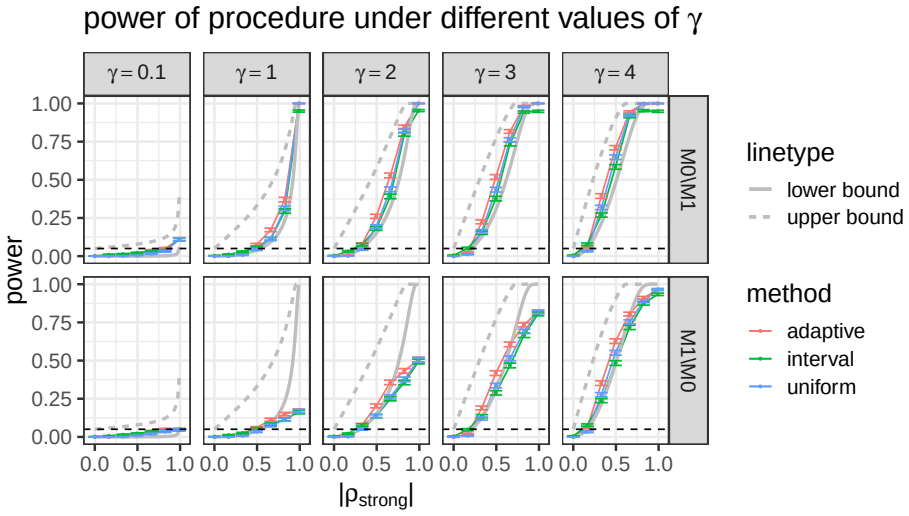


Figure 2.13: Power $\Pr(\phi_n = \mathcal{M}_i | \mathcal{M}_i)$ of the procedures (with 95% confidence intervals) under the weak-strong regime of local hypotheses. $\alpha = 0.05$ is marked as dashed. Grey curves are bounds on the theoretically optimal power.

2.7.2 Projected Wishart

We generate a covariance matrix by firstly drawing $\tilde{\Sigma}$ from the Wishart distribution (with the scale matrix chosen as $\sigma_{ij} = (-1/2)^{|i-j|}$) and then projecting $\tilde{\Sigma}$ into $\mathcal{M}_0$ or $\mathcal{M}_1$ respectively by finding the MLE under each model. Then we perform model selection based on two sets of zero-mean Gaussian samples generated with the two projected covariances respectively. We vary the degrees of freedom for the Wishart distribution. See Figure 2.15 for the results.

2.7.3 Conditional on covariates

We consider the common regression setting where two types of independences are contrasted conditional on a set of covariates $X \in \mathbb{R}^p$. In other words, we want to select between $\mathcal{M}_0 : Y_1 \perp\!\!\!\perp Y_2 \mid X$ and $\mathcal{M}_1 : Y_1 \perp\!\!\!\perp Y_2 \mid Y_3, X$. We generate instances by

$$(Y_1, Y_2, Y_3) = X^\top (\beta_1, \beta_2, \beta_3) + E, \quad E \sim \mathcal{N}(\mathbf{0}, \Sigma), \tag{2.43}$$

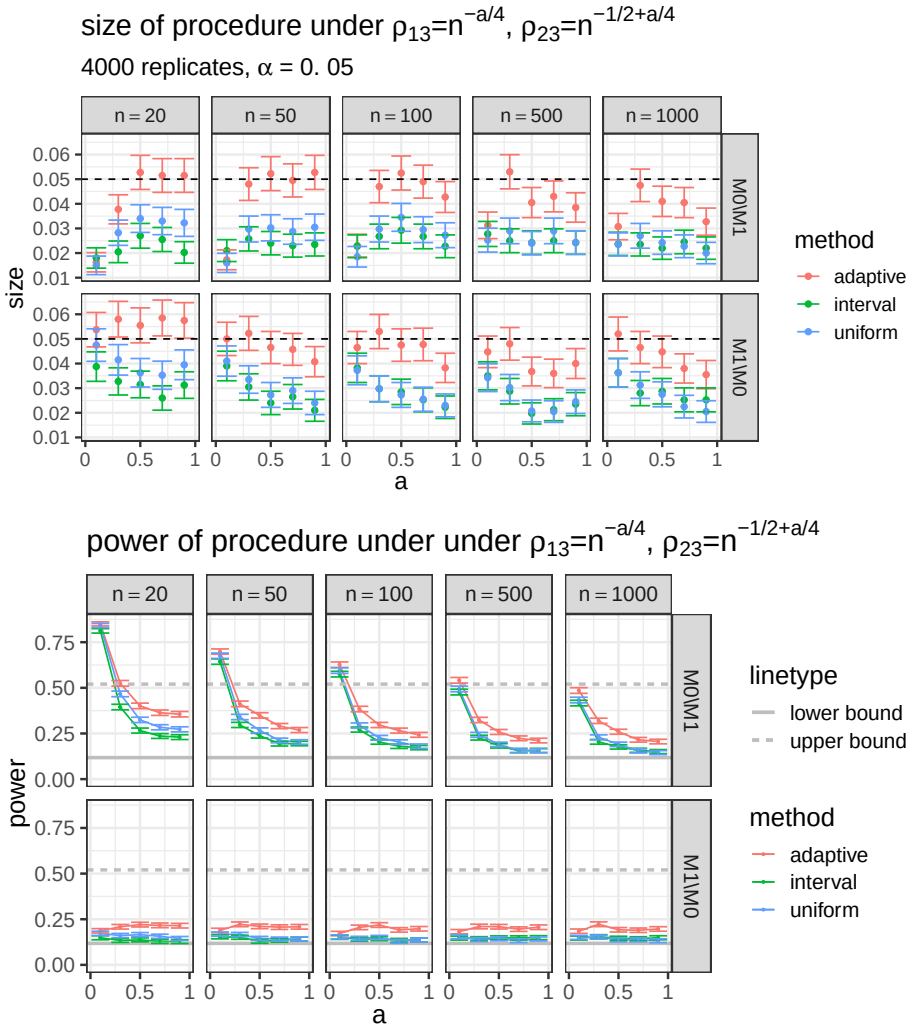


Figure 2.14: Size $\Pr(\phi_n = \mathcal{M}_{1-i}|\mathcal{M}_i)$ and power $\Pr(\phi_n = \mathcal{M}_i|\mathcal{M}_i)$ of the procedures (with 95% confidence intervals) under the weak-weak regime of local hypotheses ($\rho_{13,n}\rho_{23,n} = n^{-1/2}$). $\alpha = 0.05$ is marked as dashed. Grey lines are bounds on the theoretically optimal power in the second plot. The naive method is excluded due to its large type-I error.

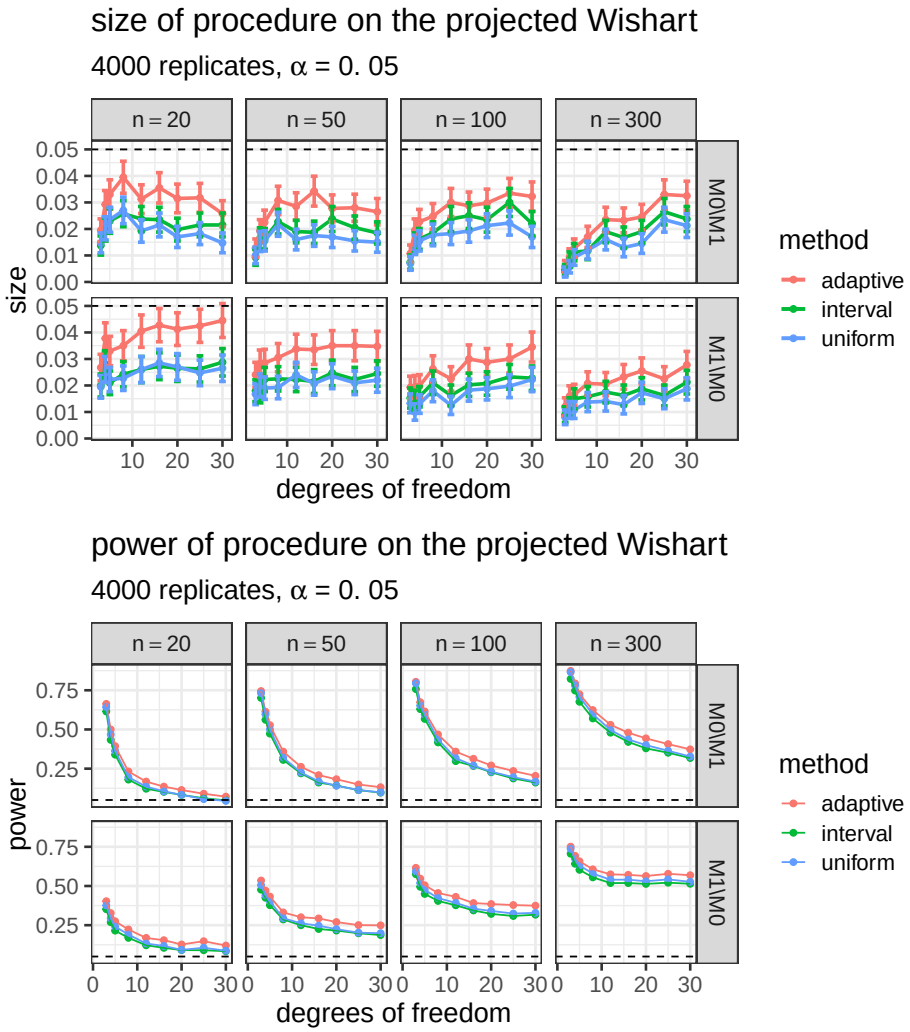


Figure 2.15: Size $\Pr(\phi = \mathcal{M}_{1-i}|\mathcal{M}_i)$ and power $\Pr(\phi = \mathcal{M}_i|\mathcal{M}_i)$ of the procedures on projected Wishart matrices (with 95% confidence intervals). $\alpha = 0.05$ is marked as dashed. The naive method makes large errors and is excluded.

where we use the previous projected Wishart to generate error covariance Σ under $\mathcal{M}_0$ and $\mathcal{M}_1$. We perform model selection by firstly regressing (Y_1, Y_2, Y_3) onto X with least squares and then apply the model selection procedures to the residual covariance. Covariates are randomly drawn from standard Gaussians and regression coefficients are generated from a t-distribution with 4 degrees of freedom. We fix $n = 1,000$ and vary the number of covariates p. The results are presented in Figure 2.16. We observe that the proposed procedure continues to maintain nominal size until p is relatively large compared to n. The power performance, on the other hand, does not seem to vary much as p grows.

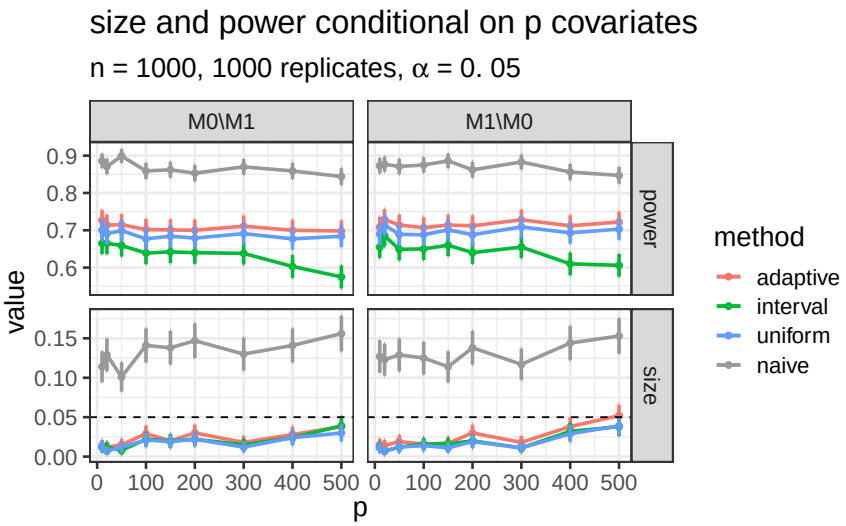


Figure 2.16: Size $\Pr(\phi = \mathcal{M}_{1-i}|\mathcal{M}_i)$ and power $\Pr(\phi = \mathcal{M}_i|\mathcal{M}_i)$ of the model selection procedures conditioned on p covariates (with 95% confidence intervals). Error covariances are generated from the projected Wishart. The procedures are applied to the least-squares residual covariance. $\alpha = 0.05$ is marked as dashed.

2.8 Example: American occupational structure

In this section we showcase an example of applying the method to edge orientation in learning a DAG. In studying the American occupational structure, Blau and Duncan (1967) measured

the following covariates on $n = 20,700$ subjects:

V: father's educational attainment,

X: father's occupational status,

U: child's educational attainment,

W: status of child's first job,

Y: status of child's occupation in 1962.

The data is summarized as the following correlation matrix of (V, X, U, W, Y)

$$S_n = \begin{pmatrix} 1.000 & 0.516 & 0.453 & 0.332 & 0.322 \\ 0.516 & 1.000 & 0.438 & 0.417 & 0.405 \\ 0.453 & 0.438 & 1.000 & 0.538 & 0.596 \\ 0.332 & 0.417 & 0.538 & 1.000 & 0.541 \\ 0.322 & 0.405 & 0.596 & 0.541 & 1.000 \end{pmatrix}.$$

At level $\alpha = 0.01$, the PC algorithm identifies the skeleton by d-separation, which only removes the edge between V and Y based on $Y \perp\!\!\!\perp V \mid U, X$. This is because the PC algorithm tests for conditional independence given smaller conditioning sets first. By a common-sense temporal ordering $\{V, X\} < U < \{W, Y\}$ among the variables, edges can be oriented except for $X - V$ and $W - Y$; see Fig. 2.17. The edge $V - X$ does not involve a collider and the orientation is statistically unidentifiable.

However, the orientation of $W - Y$ raises the interesting question of testing

$$\mathcal{M}_0\ (Y \to W) : V \perp\!\!\!\perp Y \mid U, X \quad \text{versus} \quad \mathcal{M}_1\ (Y \leftarrow W) : V \perp\!\!\!\perp Y \mid W, U, X.$$

We apply our method to the conditional correlation of (V, W, Y) given (U, X). We have $\lambda_n^{(0:1)} = 3.72$ and p-value $= 0.026$ under the envelope distribution $\bar{F}_{\hat{\rho}_n}$. Therefore, under $\alpha = 0.01$ the adaptive procedure would choose $\mathcal{M}_0 \cup \mathcal{M}_1$ and leave the orientation undetermined

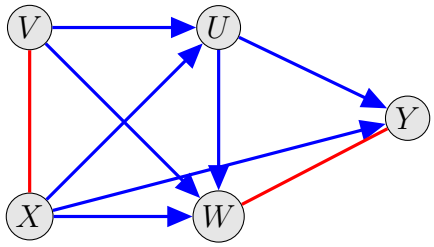


Figure 2.17: CPDAG inferred from Blau and Duncan (1967) dataset. The skeleton is inferred based on d-separation at level $\alpha = 0.01$ with the PC algorithm. Blue edges are oriented based on temporal ordering $\{V, X\} < U < \{W, Y\}$.

(the procedure would choose $Y \to W$ under $\alpha = 0.05$). This example illustrates the potential ambiguity in model selection even under a large sample size. The reader is referred to Spirtes et al. (2000, Section 5.8.4) for another discussion of the same example.

2.9 Discussion

We have considered choosing between marginal independence and conditional independence in a Gaussian graphical model, assuming we know at least one of them is true. The loglikelihood ratio statistic converges to a tight law under a sequence of truths converging to the intersection of the two models at a certain rate. The asymptotic distribution is shown to be non-uniform as it depends on *where* and *how* the sequence converges. We address this non-uniformity issue by introducing a family of envelope distributions that are well-behaved and bring back the continuity of asymptotic laws, as indexed by a parameter that can be consistently estimated. Contrary to the usual Neyman–Pearson hypothesis testing, we treat the two models symmetrically and develop model selection rules that choose both models when they are indistinguishable under a given sample size. Such rules can be designed according to the quantiles of the envelope distributions to uniformly control the type-I error below a desired level. As noted before we believe that "rate-free" asymptotic guarantees that are uniform are more useful in practice, since they do not rely upon untestable assumptions

regarding the sample size and the signal strength.

In this Chapter, we restricted ourselves to the Gaussian case. For testing conditional independence, some form of distributional assumption seems inevitable, since recent work of Shah and Peters (2020) shows that testing conditional independence without restricting the form of conditional independence is impossible in general.

Selection of non-nested models routinely relies on penalized scores based on loglikelihoods, such as the negated AIC and BIC. However, as we show, in the context of a weak signal relative to the sample size, simply choosing the model with the highest score can lead to considerable errors. To select models with "confidence", one should also look at the "gaps" between the top scores. We believe that the method developed here may be generalizable to a wider range of model selection problems.

Chapter 3

EFFICIENT LEAST SQUARES FOR LINEAR CAUSAL MODELS

Linear causal models, also known as linear structural equation models (SEMs), are widely used to postulate causal mechanisms underlying observational data. In these models, each variable equals a linear combination of a subset of the remaining variables plus an error term. When there is no unobserved confounding or selection bias, the error terms are assumed to be independent. We consider estimating a total causal effect in this setting. The causal structure is assumed to be known only up to a maximally oriented partially directed acyclic graph (MPDAG), a general class of graphs that can represent a Markov equivalence class of directed acyclic graphs (DAGs) with added background knowledge.

We propose a simple estimator based on recursive least squares, which can consistently estimate any identified total causal effect, under point or joint intervention. This estimator is derived as the MLE under Gaussian errors. Interestingly, we show that even when the errors are not necessarily Gaussian, this estimator still achieves excellent performance. In fact, it achieves efficiency bound among all regular estimators that are based on the sample covariance, which subsumes common estimators previously proposed in the literature.

This Chapter is organized as follows. In Section 3.2, we review related work on efficient estimation of total effects in over-identified settings. In Section 3.3, we introduce the preliminaries on linear SEMs, causal graphs and the identification of total effects. The concept of bucket decomposition is introduced. In Section 3.4, we introduce a block-recursive representation for the observational data and identify the total effect under such a representation. We first derive the proposed least squares estimator by finding the MLE under the assumption of Gaussian errors in Section 3.5. We then prove the optimal efficiency of our proposed

estimator under arbitrary error distributions in Section 3.6. Numerical comparisons are provided in Section 3.7. Additional preliminaries, proofs and numerical results can be found in Appendix B.

3.1 Introduction

A SEM specifies a causal mechanism underlying a set of variables (Bollen, 1989). Each variable equals a linear combination of a subset of the remaining variables plus an error term. A SEM is *associated* with a mixed graph, also known as a path diagram (Wright, 1921, 1934), which consists of both directed edges and bi-directed edges. A directed edge $i \to j$ represents that variable i appears as a covariate in the structural equation defining variable j. The equation for variable j takes the form

$$X_j = \sum_{i:i\to j} \gamma_{ij} X_i + \epsilon_j, \tag{3.1}$$

where ϵ_j is an error term. Often, the errors are assumed to follow a multivariate normal distribution, but it need not be the case. A bi-directed edge $i \leftrightarrow j$ indicates that errors ϵ_i and ϵ_j are dependent, which is assumed when there exists an unobserved (i.e., latent) confounder between i and j. The mixed graph is usually assumed to be *acyclic*, i.e., the graph does not contain cycles made of directed edges.

We focus on the setting when there is no unobserved confounder or selection bias, a condition also known as *causal sufficiency*; see Spirtes et al. (2000, Chap. 3) and Pearl (2009, Chap. 6). In this setting, all the error terms are assumed to be *mutually independent* and the mixed graph associated with the linear SEM is a *directed acyclic graph* (DAG), often called a *causal DAG*. Aside from being a statistical model for observational data, the linear SEM is also a causal model in the sense that it specifies the behavior of the system under interventions (see Section 3.3.2). Therefore, the total causal effect of one treatment variable (point intervention) or several treatment variables (joint intervention) on some outcome varibles can be defined.

The underlying causal DAG is usually unknown. In fact, linear SEMs associated with

different DAGs may define the same *observed* distribution (Drton et al., 2011). Without further assumptions on the error distributions, the underlying DAG can only be learned from observational data up to its Markov equivalence class, which can be uniquely represented by a completed partially directed acyclic graph (CPDAG) (Meek, 1995; Andersson et al., 1997). Additional background knowledge, such as knowledge of certain causal relationships (Meek, 1995; Fang and He, 2020) or partial orderings (Scheines et al., 1998), restrictions on the error distributions (Shimizu et al., 2006, 2011; Shimizu, 2014; Hoyer et al., 2008; Peters and Bühlmann, 2014), and other assumptions (Hauser and Bühlmann, 2012; Wang et al., 2017b; Rothenhäusler et al., 2018; Eigenmann et al., 2017) can be used to further refine the Markov equivalence class of DAGs, resulting in representing the causal structure as a maximally oriented PDAG (MPDAG), which is a rather general class of graphs containing directed and undirected edges that subsumes DAGs and CPDAGs (Meek, 1995). A given total causal effect is *identified* given a graph, if it can be expressed as a functional of the observed distribution, which is the same for every DAG in the equivalence class. Recently, a necessary and sufficient graphical criterion for identification given an MPDAG has been shown by Perković (2020). In general, there may be more than one identifying functional.

Naturally, the next step is to develop estimators for an identified total effect with desirable properties. When the effect is unidentified, the reader is referred to IDA-type (Maathuis et al., 2009; Nandy et al., 2017) or enumerative (Guo and Perković, 2021) approaches, which are beyond the scope of this Chapter. Among others, we consider the following desiderata.

Completeness. Can the estimator consistently estimate every identified effect, under either point or joint interventions?

Efficiency. Does the estimator achieve the smallest asymptotic (co)-variance compared to a reasonably large class of estimators?

To the best of our knowledge, no estimator proposed in the literature fulfills both desiderata. Indeed, the commonly used covariate adjustment estimators (Pearl, 1993; Shpitser et al., 2010; Maathuis and Colombo, 2015; Perković et al., 2015) do not exist for certain total

effects under joint interventions (Nandy et al., 2017; Perković et al., 2018; Perković, 2020). Furthermore, when they exist, even with an optimal adjustment set chosen to maximize efficiency (Henckel et al., 2019; Rotnitzky and Smucler, 2020; Witte et al., 2020), we will show in Section 3.7 that covariate adjustment can compare less favorably against a larger class of estimators.

We propose an estimator that is based on simple recursive least squares, that affirmatively fulfills both desiderata. In particular, our proposed estimator achieves the efficiency bound among all regular estimators that only depend on the sample covariance; see Section 3.6 for the precise definition of the class of estimators. Remarkably, our result holds regardless of the type of error distribution in the underlying linear SEM. Our method is implemented in the R (R Core Team, 2020) package `eff2` (`https://github.com/richardkwo/eff2`), which stands for "efficient effect" (estimate).

Admittedly, our estimator can be less efficient when compared to an even larger class of estimators, such as the class of all regular estimators considered in standard semiparametric theory. A semiparametric efficient estimator, relative to all regular estimators, can in principle be constructed by computing the efficient influence function and employing estimation strategies such as one-step correction or targeted maximum likelihood estimation (van der Laan and Rose, 2011). In fact, the semiparametric model we consider (see Eq. (3.21)) is a generalized, multivariate location-shift regression model with additional conditional independence constraints; see also Tsiatis (2006, §5.1) and Bickel et al. (1993, §4.3). While it is theoretically possible to construct a semiparametric efficient estimator by firstly estimating the error score and then solving the associated estimating equations (Bickel et al., 1993, §7.8), the resulting estimator tends to be too complicated and unstable for practical purposes unless the sample size is very large (Tsiatis, 2006, page 111). On the other hand, despite the potential loss of efficiency, our least squares estimator is easily computed and numerically stable. Hence, our proposal can be viewed as a deliberate trade-off between optimality and practicality.

3.2 Related work

The statistical performance of an estimator of a total causal effect, in *over-identified* settings, has recently received more attention; see, e.g, Kuroki and Miyakawa (2003); Henckel et al. (2019); Witte et al. (2020); Gupta et al. (2020); Rotnitzky and Smucler (2020); Smucler et al. (2020); Kuroki and Nanmo (2020). Here, "over-identified" (Koopmans and Reiersøl, 1950) refers to the fact that the total causal effect can be expressed as more than one functional of the (population) observed distribution, all of which coincide due to the additional conditional independence constraints obeyed by the observed distribution. For example, in the case where a total causal effect can be identified through covariate adjustment, usually there exists more than one valid adjustment set (Henckel et al., 2019). This is in contrast to the more traditional setting of causal inference, where the observed data distribution is nonparametric and is not expected to satisfy extra conditional independences.

Intuitively, the conditional independences in over-identified models can be exploited to maximize asymptotic efficiency; see, e.g., Sargan (1958); Hansen (1982) for early works in this direction. Under a linear SEM with independent errors, a total causal effect can be estimated via covariate adjustment as the least squares coefficient from the regression of the outcome on the treatment and adjustment variables. Henckel et al. (2019) recently showed that, under a linear SEM with independent errors, a valid adjustment set that minimizes asymptotic variance, also referred to as the *optimal adjustment set*, can be graphically characterized; see also Witte et al. (2020) for further properties of such an optimal set. This result was generalized by Rotnitzky and Smucler (2020) beyond linear SEMs: an optimal adjustment set is shown to always exist for point interventions, and a semiparametric efficient estimator is developed for this case. Note that, while valid adjustment sets (called "time-independent" adjustment sets by Rotnitzky and Smucler (2020)) exist for point interventions (Perković, 2020, Proposition 4.2), they may not exist for joint interventions (Nandy et al., 2017; Perković et al., 2018; Perković, 2020).

Less is known about how to efficiently estimate the total causal effect of a joint inter-

vention, at least in a generic fashion. For linear SEMs with independent errors, with the knowledge of the parents of the treatment variables in the underlying causal DAG, Nandy et al. (2017) considered two estimators for the joint-IDA algorithm, one based on recursive least squares and one based on a modified Cholesky decomposition. However, the efficiency properties of these estimators were not explored. In Section 3.7, numerical comparisons will show that our proposed estimator significantly outperforms these estimators.

Other results on the linear SEM include explicit calculations and comparisons for typical examples with either a particular structure or only a few variables; see, e.g., Kuroki and Cai (2004); Gupta et al. (2020). Gaussian errors are also assumed in these calculations.

3.3 Linear SEMs, causal graphs and effect identification

3.3.1 Linear SEMs under causal sufficiency

A linear SEM postulates a causal mechanism that generates data. Let X denote a vector of variables generated by a linear SEM, where X is indexed by V ($X = X_V$). Let $\mathcal{D}$ be the associated DAG on vertices V. For this $|V|$-dimensional random vector X, the model in Eq. (3.1) can be compactly rewritten as

$$X = \Gamma^\top X + \epsilon, \quad \Gamma = (\gamma_{ij}), \quad i \to j \text{ not in } \mathcal{D} \Rightarrow \gamma_{ij} = 0. \tag{3.2}$$

where $\Gamma \in \mathbb{R}^{|V|\times|V|}$ is a coefficient matrix, and $\epsilon = (\epsilon_i)$ is a $|V|$-dimensional random vector. DAG $\mathcal{D}$ is associated with the linear SEM in Eq. (3.1) in the sense that the non-zero entries of Γ correspond to the edges in $\mathcal{D}$.

Under causal sufficiency (no latent variables), we assume

$$\{\epsilon_i : i \in V\} \text{ are independent, } \mathbb{E}\,\epsilon = 0, \ \mathbb{E}\,\epsilon\epsilon^\top \succ \mathbf{0}, \tag{3.3}$$

where for a real symmetric matrix A, $A \succ \mathbf{0}$ means A is positive definite. The errors $\{\epsilon_i : i \in V\}$ are not necessarily Gaussian, nor identically distributed.

The law $P(X)$ is called the *observed distribution*. For a given $\mathcal{D}$, we will use $\mathcal{P}_\mathcal{D}$ to denote the set of possible laws of X, namely the collection of $P(X)$ as Γ and the error distribution

vary subject to Eqs. (3.2) and (3.3). The linear SEM poses certain restrictions on the set of laws $\mathcal{P}_{\mathcal{D}}$. Let $\mathrm{Pa}(i, \mathcal{D})$ denote the set of parents of vertex i, i.e., $\{j : j \to i \text{ is in } \mathcal{D}\}$. For any $P \in \mathcal{P}_{\mathcal{D}}$, among other constraints, (i) P factorizes according to $\mathcal{D}$, (ii) $\mathbb{E}[X_i \mid X_{\mathrm{Pa}(i,\mathcal{D})}]$ is linear in $X_{\mathrm{Pa}(i,\mathcal{D})}$ and (iii) $\mathrm{var}[X_i \mid X_{\mathrm{Pa}(i,\mathcal{D})}]$ is constant in $X_{\mathrm{Pa}(i,\mathcal{D})}$.

We observe n iid samples generated by the model above, namely $X^{(i)} = (I - \Gamma)^{-\top}\epsilon^{(i)}$ for $i = 1, \ldots, n$. Note that $(I - \Gamma)$ is invertible because Γ can be permuted into a lower-triangular matrix by a topological ordering (i.e., causal ordering) of vertices in $\mathcal{D}$.

3.3.2 Interventions and total causal effects

The assumed linear SEM also dictates the behavior of the system under interventions. Let $A \subseteq V$ be a set of vertices indexing treatment variables X_A. We use $\mathrm{do}(X_A = x_A)$ to denote *intervening* on variables X_A and forcing them to take values x_A (Pearl, 1995b). We call this a *point* intervention if A is a singleton, and a *joint* intervention if A consists of several vertices, which correspond to the case of multiple treatments. While X_A is fixed to x_A, the remaining variables are generated by their corresponding structural equations Eq. (3.1), with each X_i for $i \in A$ appearing in the equations replaced by the corresponding enforced value x_i (Strotz and Wold, 1960). This generating mechanism defines the *interventional distribution*, denoted by $P(X|\mathrm{do}(X_A = x_A))$, where the conditional probability notation is only conventional. More formally, the interventional distribution is expressed as

$$P(X|\mathrm{do}(X_A = x_A)) = \prod_{j \in A} \delta_{x_j}(X_j) \prod_{i \notin A} P\left(X_i | X_{\mathrm{Pa}(i,\mathcal{D})}\right), \tag{3.4}$$

where δ denotes a Dirac measure. Factor $P\left(X_i | X_{\mathrm{Pa}(i)}\right)$ is defined by the structural equation for X_i. Eq. (3.4) is known as the truncated factorization formula (Pearl, 2009), manipulated density formula (Spirtes et al., 2000) or the g-formula (Robins, 1986).

Definition 3.1 (Total causal effect, Pearl, 2009; Nandy et al., 2017)**.** *Let X_A be a vector of treatment variables and X_Y with $Y \in V \setminus A$ be an outcome variable. The total causal effect*

of X_A on X_Y is defined as the vector $\tau_{AY} \in \mathbb{R}^{|A|}$, where

$$(\tau_{AY})_i = \frac{\partial}{\partial x_{A_i}} \mathbb{E}[X_Y \mid \mathrm{do}(X_A = x_A)], \quad i = 1, \ldots, |A|.$$

That is, τ_{AY} is the gradient of the linear map $x_A \mapsto \mathbb{E}[Y|\mathrm{do}(X_A = x_A)]$. When multiple outcomes $Y = \{Y_1, \ldots, Y_k\}$, $k > 1$, are considered, the total causal effect of X_A on $X_{Y_1}, \ldots, X_{Y_k}$ can be defined by concatenating $\tau_{AY_1}, \ldots, \tau_{AY_k}$. Therefore, throughout, we assume the outcome variable is a singleton without loss of generality. Each coordinate of the total causal effect τ_{AY} can be expressed as a sum-product of the underlying linear SEM coefficients along certain causal paths from A to Y in $\mathcal{D}$, that is, certain paths of the form $A_1 \to \cdots \to Y_i$ for $A_1 \in A$; see also Wright (1934); Sullivant et al. (2010).

3.3.3 Causal graphs

Two different linear SEMs on the same set of variables can define the same observed distribution. For example, under Gaussian errors, linear SEMs associated with DAGs $A \to Y$ and $A \leftarrow Y$, define the same set of observed distributions, namely the set of centered bivariate Gaussian distributions. Without making additional assumptions on the error distribution, such as non-Gaussianity (Shimizu et al., 2006), partial non-Gaussianity (Hoyer et al., 2008), or equal variance of errors (Peters and Bühlmann, 2014; Chen et al., 2019), the underlying causal DAG can only be learned from the observed distribution up to its Markov equivalence class (Pearl and Verma, 1995; Chickering, 2002).

CPDAGs Two DAGs on the same set of vertices are Markov equivalent if they encode the same set of d-separation relations between the vertices. The d-separations between the vertices, prescribe conditional independences between the corresponding variables (known as the Markov condition (Lauritzen, 1996, §3.2.2)); see Appendix B.4 for the definition of d-separation and more background. This equivalence relation defines a Markov equivalence class, which consists of DAGs as elements. A Markov equivalence class can be uniquely represented by a *completed partially directed acyclic graph* (CPDAG), also known as an

essential graph (Meek, 1995; Andersson et al., 1997). A CPDAG $\mathcal{C}$ is a graph on the same set of vertices, that can contain both directed and undirected edges. We use $[\mathcal{C}]$ to denote the Markov equivalence class represented by CPDAG $\mathcal{C}$. A directed edge $i \to j$ in $\mathcal{C}$ implies $i \to j$ is in every $\mathcal{D} \in [\mathcal{C}]$, whereas an undirected edge $i - j$ in $\mathcal{C}$ implies there exist $\mathcal{D}_1, \mathcal{D}_2 \in [\mathcal{C}]$ such that $i \to j$ in $\mathcal{D}_1$ but $i \leftarrow j$ in $\mathcal{D}_2$. Given a DAG $\mathcal{D}$, the CPDAG $\mathcal{C}$ representing the Markov equivalence class of $\mathcal{D}$ can be drawn by keeping the skeleton of $\mathcal{D}$, adding all the unshielded colliders from $\mathcal{D}$ and completing the orientation rules R1–R3 of Meek (1995); see Fig. B.3 in Appendix B.4. For example, DAGs $A \to Y$ and $A \leftarrow Y$ are represented by CPDAG $A - Y$. To slightly abuse the notation, for a distribution Q, we write $Q \in [\mathcal{C}]$ if Q factorizes according to some DAG $\mathcal{D} \in [\mathcal{C}]$; see Lauritzen (1996, §3.2.2).

There are various *structure learning* algorithms that can be used to uncover CPDAG $\mathcal{C}$ from observational data. Some well-known examples are the PC algorithm (Spirtes et al., 2000) and the greedy equivalence search (Chickering, 2002). Choosing an appropriate algorithm for the dataset at hand is beyond the scope of this Chapter; the reader is referred to Drton and Maathuis (2017, §4) for a recent overview.

MPDAGs Certain background knowledge, if present, can be used to further orient some undirected edges in a CPDAG $\mathcal{C}$. Typically, knowledge of temporal orderings can inform the orientation of certain undirected edges; see Spirtes et al. (2000, §5.8.4) for an example. Adding these background-knowledge orientations and the additionally implied orientations based on the orientation rules of Meek (1995) to $\mathcal{C}$ results in a *maximally oriented partially directed graph* (MPDAG) $\mathcal{G}$. See Fig. B.3 and Algorithm 1 in Appendix B.4. MPDAGs are a rather general class of graphs that subsumes both DAGs and CPDAGs. An MPDAG $\mathcal{G}$ represents a *restricted* Markov equivalence class of DAGs, which we also denote by $[\mathcal{G}]$. Analogously to the case of a CPDAG, $i \to j$ in $\mathcal{G}$ implies $i \to j$ is in every $\mathcal{D} \in [\mathcal{G}]$, and $i - j$ in $\mathcal{G}$ implies there exist $\mathcal{D}_1, \mathcal{D}_2 \in [\mathcal{G}]$ such that $i \to j$ in $\mathcal{D}_1$ but $i \leftarrow j$ in $\mathcal{D}_2$.

For the rest of the Chapter, we will assume that we have access to an MPDAG $\mathcal{G}$ that

represents our structural knowledge about the underlying DAG $\mathcal{D}$. That is,

$$\text{causal DAG } \mathcal{D} \in [\mathcal{G}], \quad \mathcal{G} \text{ is an MPDAG}, \tag{3.5}$$

where $[\mathcal{G}]$ represents a collection of DAGs that are Markov equivalent, but can be strictly smaller than the corresponding Markov equivalence class due to background knowledge.

3.3.4 Causal effect identification

Throughout we will use the following notations. Given treatment variables X_A and an outcome variable X_Y such that $Y \notin A$, we are interested in learning the total causal effect τ_{AY}. We assume that we have access to an MPDAG $\mathcal{G}$, and to observational data that are generated as iid samples from a linear SEM defined by Eqs. (3.2) and (3.3), where the causal DAG $\mathcal{D}$ is in $[\mathcal{G}]$. Before estimation can be performed, we need to make sure that τ_{AY} can be identified from observational data. That is, we need to ensure that τ_{AY} can be expressed as a functional of the observed distribution that is the same for *every* DAG in $[\mathcal{G}]$. We have the following graphical criterion.

Theorem 3.1 (Perković, 2020)**.** *The total causal effect τ_{AY} of X_A on X_Y is identified given an MPDAG $\mathcal{G}$ if and only if there is no proper, possibly causal path from A to Y in $\mathcal{G}$ that starts with an undirected edge.*

Theorem 3.1 is Proposition 3.2 of Perković (2020), which holds for nonparametric causal graphical models. It does not require that the data is generated by a linear SEM. However, Perković (2020) proves that when the criterion fails, then two linear SEMs with Gaussian errors can be constructed such that their observed distributions coincide but their τ_{AY}'s are different. Hence, even if we restrict ourselves to linear SEMs, Theorem 3.1 still holds.

A few terms need some explanation. A *path* from A to Y in $\mathcal{G}$ is a sequence of distinct vertices $\langle v_1, \ldots, v_k \rangle$ for $k > 1$ with $v_1 \in A$ and $v_k = Y$, such that every pair of successive vertices are adjacent in $\mathcal{G}$. The path is *proper* when only its first vertex is in A. The path is *possibly causal* if no edge $v_l \leftarrow v_r$ is in $\mathcal{G}$ for $1 \leq l < r \leq k$. The reader is referred to

Appendix B.4 for more graphical preliminaries. When $\mathcal{G}$ satisfies Theorem 3.1 relative to vertex sets A and Y, the interventional distribution $P(X_Y|\text{do}(X_A = x_A))$, and hence the total effect, can be computed from the observed distribution $P(X)$. To express the identification formula, we require the following concepts.

Buckets and bucket decomposition

Let $\mathcal{G} = (V, E, U)$ be a partially directed graph, where V is the set of vertices, and E and U are sets of directed and undirected edges respectively. Let $B_1, \ldots, B_K$ be the *maximal connected components* of the undirected graph $\mathcal{G}_U := (V, \emptyset, U)$. Then $V = B_1 \dot\cup \ldots \dot\cup B_K$, where symbol $\dot\cup$ denotes disjoint union. Note that all the directed edges within each B_i are due to background knowledge. If we ignore the distinction between directed and undirected edges, then the subgraph induced by each B_i is chordal (Andersson et al., 1997, §4).

Suppose the connected components are ordered such that

$$i \to j \in E,\ i \in B_i,\ j \in B_j \quad \Rightarrow \quad i < j. \tag{3.6}$$

One can show that such a *partial causal ordering* always exists, though it may not be unique; see Algorithm 2 in Appendix B.4 to obtain such an ordering. Our result does not depend on the particular choice of partial causal ordering. We call $B_1, \ldots, B_K$ the *bucket decomposition* of V and call each B_k for $k = 1, \ldots, K$ a *bucket*; see Fig. 3.1(a) for an example. If it is clear which graph $\mathcal{G}$ is being referred to, we will shorten $\text{Pa}(j, \mathcal{G})$ as $\text{Pa}(j)$ to reduce clutter. For a set of vertices C in $\mathcal{G}$, we use $\text{Pa}(C) := \cup_{i \in C} \text{Pa}(i) \setminus C$ to denote the set of their *external parents*. Clearly, $\text{Pa}(B_k) \subseteq B_{[k-1]}$, where $B_{[k-1]} := B_1 \cup \cdots \cup B_{k-1}$.

Lemma 3.1. *Let i and j be two distinct vertices in MPDAG $\mathcal{G} = (V, E, U)$ such that $i \to j \in E$. Suppose that there is no undirected path from i to j in $\mathcal{G}$. If there is a vertex k, and an undirected path $j - \cdots - k$ in $\mathcal{G}$, then $i \to k \in E$.*

By definition of the parent set above we have that $\text{Pa}(B_k) = \cup_{i \in B_k} \text{Pa}(i) \setminus B_k$, $k = 1, \ldots, K$. However, since a bucket B_k is a maximal subset of V that is connected by undirected edges in $\mathcal{G}$, Lemma 3.1 implies the following important property.

Corollary 3.1 (Restrictive property). *Let $B_1, \ldots, B_K$ be the bucket decomposition of V in MPDAG $\mathcal{G} = (V, E, U)$. Then, all vertices in the same bucket have the same set of external parents, namely*

$$\mathrm{Pa}(B_k) = \mathrm{Pa}(i) \setminus B_k, \quad \textit{for any } i \in B_k,\ k = 1, \ldots, K.$$

The causal identification formula for $P(X_Y|\mathrm{do}(X_A = x_A))$ of Perković (2020) relies on a decomposition of certain *ancestors* of Y in MPDAG $\mathcal{G}$ according to the buckets. We call vertex i an ancestor of vertex j in $\mathcal{G}$ if there exists a directed path $i \to \cdots \to j$ in $\mathcal{G}$; we use the convention that j is an ancestor of itself. We denote the set of ancestors of j in $\mathcal{G}$ as $\mathrm{An}(j, \mathcal{G})$, or shortened as $\mathrm{An}(j)$.

Let $\mathcal{G}_{V\setminus A} = (V \setminus A, E', U')$ denote the subgraph of $\mathcal{G}$ *induced* by the vertices $V \setminus A$, where E' includes those edges in E that are between vertices in $V \setminus A$, and similarly for U'. Consider the set of ancestors of Y in $\mathcal{G}_{V\setminus A}$, denoted as

$$D := \mathrm{An}(Y, \mathcal{G}_{V\setminus A}). \tag{3.7}$$

The bucket decomposition $D_1, \ldots, D_K$ of D, induced by the bucket decomposition of V, is simply

$$D = \dot{\bigcup_{k=1}^{K}} D_k, \quad D_k = D \cap B_k, \quad i = 1, \ldots, K. \tag{3.8}$$

Lemma 3.2. *When the criterion in Theorem 3.1 is satisfied, we have* $\mathrm{Pa}(D_k, \mathcal{G}) = \mathrm{Pa}(B_k, \mathcal{G})$ *for every nonempty* D_k.

Proofs of Lemmas 3.1 and 3.2 are left to Appendix B.2.

Theorem 3.2 (Perković, 2020). *Suppose the criterion in Theorem 3.1 is satisfied for A, Y in MPDAG $\mathcal{G} = (V, E, U)$ such that $Y \notin A$. Let $P(X)$ be the observed distribution. Let $D = \mathrm{An}(Y, \mathcal{G}_{V\setminus A})$ and $D_1, \ldots D_K$ be the bucket decomposition of D as in Eq. (3.8). Then the interventional distribution $P(X_Y|\mathrm{do}(X_A = x_A))$ can be identified as*

$$P(X_Y|\mathrm{do}(X_A = x_A)) = \int \left\{ \prod_{k=1}^{K} P\left(X_{D_k}|X_{\mathrm{Pa}(D_k)}\right) \right\} \mathrm{d}X_{D\setminus Y} \tag{3.9}$$

for values $X_{\mathrm{Pa}(D_k)}$ in agreement with x_A, where $P\left(X_{D_k}|X_{\mathrm{Pa}(D_k)}\right) \equiv 1$ if $D_k = \emptyset$.

The expression in Eq. (3.9) above is a generalization of the truncated factorization Eq. (3.4) from DAGs to MPDAGs. Theorem 3.2 holds generally even when an underlying linear SEM is not assumed.

3.4 Block-recursive representation

In this section, we express the observed distribution $P(X)$ induced by a linear SEM compatible with MPDAG $\mathcal{G} = (V, E, U)$ in a block-recursive form. Each block corresponds to a bucket in the bucket decomposition of V. Such a reparameterization is necessitated by the fact that the causal ordering of $\mathcal{D}$ is unknown, whereas the buckets can be arranged into a valid partial causal ordering as in Eq. (3.6). We will use this representation to compute an estimator for the total causal effect.

Recall that $\mathcal{P}_{\mathcal{D}}$ denotes the family of laws of X arising from a linear SEM Eqs. (3.2) and (3.3) compatible with DAG $\mathcal{D}$. Let $\mathcal{P}_{\mathcal{G}} := \cup_{\mathcal{D}\in[\mathcal{G}]}\mathcal{P}_{\mathcal{D}}$, which denotes the family of laws of X arising from a linear SEM compatible with a DAG in $[\mathcal{G}]$.

Proposition 3.1 (Block-recursive form)**.** *Let $\mathcal{D}$ be the causal DAG associated with the linear SEM and $\mathcal{G}$ an MPDAG such that $\mathcal{D} \in [\mathcal{G}]$. Further, let $B_1, \ldots, B_K$ be the bucket decomposition of V in $\mathcal{G}$. Then the linear SEM Eqs. (3.2) and (3.3) can be rewritten as*

$$X = \Lambda^\top X + \varepsilon,$$

for some matrix of coefficients $\Lambda = (\lambda_{ij}) \in \mathbb{R}^{|V|\times|V|}$ and random vector $\varepsilon = (\varepsilon_i) \in \mathbb{R}^{|V|}$ such that

$$j \in B_l,\ i \notin \mathrm{Pa}(B_l, \mathcal{G}) \quad \Rightarrow \quad \lambda_{ij} = 0, \tag{3.10}$$

$$\mathbb{E}\,\varepsilon = 0, \quad \mathbb{E}\,\varepsilon_{B_k}\varepsilon_{B_k}^\top \succ \mathbf{0},\ (k = 1, \ldots, K), \quad \varepsilon_{B_1}, \ldots, \varepsilon_{B_K} \text{ are mutually independent}, \tag{3.11}$$

and

$$\text{law of } (\varepsilon_{B_k}) \in \mathcal{P}_{\mathcal{G}_{B_k}}, \quad k = 1, \ldots, K, \tag{3.12}$$

where $\mathcal{G}_{B_k}$ is the subgraph of $\mathcal{G}$ induced by B_k.

Note that in contrast to symbol ϵ used in Eqs. (3.2) and (3.3), symbol ε is used here to denote the errors in the block-recursive form. The coordinates within each ε_{B_k} may be *dependent*.

Proof. For $k = 2, \ldots, K$, by Eq. (3.2) and the restrictive property (Corollary 3.1), we have

$$X_{B_k} = \Gamma_{\mathrm{Pa}(B_k),B_k}^{\top} X_{\mathrm{Pa}(B_k)} + \Gamma_{B_k}^{\top} X_{B_k} + \epsilon_{B_k},$$

where $\mathrm{Pa}(B_k) = \mathrm{Pa}(B_k, \mathcal{G})$. The expression can be rewritten as

$$\begin{aligned} X_{B_k} &= (I - \Gamma_{B_k})^{-\top} \Gamma_{\mathrm{Pa}(B_k),B_k}^{\top} X_{\mathrm{Pa}(B_k)} + (I - \Gamma_{B_k})^{-\top} \epsilon_{B_k} \\ &= \Lambda_{\mathrm{Pa}(B_k),B_k}^{\top} X_{\mathrm{Pa}(B_k)} + \varepsilon_{B_k}, \end{aligned}$$

where $\varepsilon_{B_k} := (I - \Gamma_{B_k})^{-\top} \epsilon_{B_k}$ for $k = 1, \ldots, K$ (note that $X_{B_1} = \varepsilon_{B_1}$). Additionally, $\Lambda_{\mathrm{Pa}(B_k),B_k} = \Gamma_{\mathrm{Pa}(B_k),B_k} (I - \Gamma_{B_k})^{-1}$ for $k = 2, \ldots, K$.

Matrix $\Lambda \in \mathbb{R}^{|V| \times |V|}$ in the statement of the proposition is defined by blocks $\Lambda_{\mathrm{Pa}(B_k),B_k}$ for $k = 2, \ldots, K$ and zero entries otherwise. Therefore, $\lambda_{ij} = 0$ if $j \in B_l$ and $i \notin \mathrm{Pa}(B_l)$ for some $l = 1, \ldots K$. Hence, by putting the blocks together, the model can be written as $X = \Lambda^{\top} X + \varepsilon$.

The "new" errors ε satisfy

$$\varepsilon_{B_k} = \Gamma_{B_k}^{\top} \varepsilon_{B_k} + \epsilon_{B_k}, \quad k = 1, \ldots, K.$$

It then follows from Eqs. (3.2) and (3.3) that for every k,

$$\text{law of } \varepsilon_{B_k} \in \mathcal{P}_{\mathcal{D}_{B_k}} \subset \mathcal{P}_{\mathcal{G}_{B_k}},$$

since $D \in [\mathcal{G}]$. Moreover, for every k,

$$\mathbb{E}\, \varepsilon_{B_k} = 0, \quad \mathbb{E}\, \varepsilon_{B_k} \varepsilon_{B_k}^{\top} = (I - \Gamma_{B_k})^{-\top}\, \mathbb{E}\, \epsilon_{B_k} \epsilon_{B_k}^{\top} (I - \Gamma_{B_k})^{-1} \succ \mathbf{0},$$

where both $(I - \Gamma_{B_k})$ and $\mathbb{E}\, \epsilon_{B_k} \epsilon_{B_k}^{\top}$ are full rank, because Γ_{B_k} can be permuted into an upper-triangular matrix and $\mathbb{E}\, \epsilon\epsilon^{\top} \succ \mathbf{0}$ by Eq. (3.3). □

Corollary 3.2. *Under the same conditions as Proposition 3.1, it holds that*

$$\begin{aligned} X_{B_1} &= \varepsilon_{B_1}, \\ X_{B_k} &= \Lambda_{\mathrm{Pa}(B_k),B_k}^{\top} X_{\mathrm{Pa}(B_k)} + \varepsilon_{B_k}, \quad \varepsilon_{B_k} \perp\!\!\!\perp X_{\mathrm{Pa}(B_k)}, \quad k = 2, \ldots, K, \end{aligned} \tag{3.13}$$

where $\mathrm{Pa}(B_k) = \mathrm{Pa}(B_k, \mathcal{G})$.

Next, we show that if the total causal effect τ_{AY} is identifiable from MPDAG $\mathcal{G}$ (Theorem 3.1), then it can be calculated from Λ in the block-recursive representation of Proposition 3.1. Therefore, the distribution of ε is a nuisance relative to estimating τ_{AY}.

Proposition 3.2. *Suppose the criterion in Theorem 3.1 is satisfied for A, Y in MPDAG $\mathcal{G} = (V, E, U)$ such that $Y \notin A$. Let Λ be the block-recursive coefficient matrix given by Proposition 3.1. The total causal effect of X_A on X_Y is identified as*

$$\tau_{AY} = \Lambda_{A,D} \left[(I - \Lambda_{D,D})^{-1} \right]_{D,Y}, \tag{3.14}$$

where $D = \mathrm{An}(Y, \mathcal{G}_{V \setminus A})$ and the last subscript denotes the column corresponding to $Y \in D$.

Proof. We derive this result using Theorem 3.2. Recall that $D_1, \ldots, D_K$ is a partition of D induced by the bucket decomposition $B_1, \ldots, B_K$ of V in the sense that $D_k = D \cap B_k$ for $k = 1, \ldots, K$. When $D_k = \emptyset$, we use the convention that $P(X_{D_k} | X_{\mathrm{Pa}(D_k)}) \equiv 1$. By definition of $D = \mathrm{An}(Y, \mathcal{G}_{V \setminus A})$ and Eq. (3.6), observe that a vertex in $\mathrm{Pa}(D_k) = \mathrm{Pa}(D_k, \mathcal{G})$ is either in $D_1 \cup \cdots \cup D_{k-1}$ or in A. Let $F_k := A \cap \mathrm{Pa}(D_k)$. In Eq. (3.9), we note that the joint interventional distribution of X_D is given by

$$P(X_D | \mathrm{do}(X_A = x_A)) = \prod_{k=1}^{K} P(X_{D_k} | X_{\mathrm{Pa}(D_k)}) = \prod_{k=1}^{K} P(X_{D_k} | X_{\mathrm{Pa}(D_k) \setminus F_k}, X_{F_k} = x_{F_k}),$$

where x_{F_k} is fixed by the $\mathrm{do}(X_A = x_A)$ operation. Further, fix a factor $i \in \{1, \ldots, K\}$. By Lemma 3.2, $\mathrm{Pa}(D_i) = \mathrm{Pa}(B_i)$. By Eq. (3.13) and $\varepsilon_{D_i} \perp\!\!\!\perp X_{\mathrm{Pa}(B_i)}$, we have

$$\begin{aligned} X_{D_i} \mid \left\{ X_{\mathrm{Pa}(D_i) \setminus F_i}, X_{F_i} = x_{F_i} \right\} &=_d \Lambda_{\mathrm{Pa}(D_i) \setminus F_i, D_i}^{\top} X_{\mathrm{Pa}(D_i) \setminus F_i} + \Lambda_{F_i, D_i} x_{F_i} + \varepsilon_{D_i} \\ &= \Lambda_{\mathrm{Pa}(D_i) \cap D, D_i}^{\top} X_{\mathrm{Pa}(D_i) \cap D} + \Lambda_{\mathrm{Pa}(D_i) \cap A, D_i} x_{\mathrm{Pa}(D_i) \cap A} + \varepsilon_{D_i} \end{aligned}$$

The fact that the display above holds for every $i = 1, \ldots, K$ implies that the joint interventional distribution $P(X_D|\mathrm{do}(X_A = x_A))$ satisfies

$$X_D = \Lambda_{D,D}^T X_D + \Lambda_{A,D}^\top x_A + \varepsilon_D.$$

It follows that $X_D = (I - \Lambda_{D,D})^{-\top}(\Lambda_{A,D}^\top x_A + \varepsilon_D)$ and hence

$$\mathbb{E}[X_D \mid \mathrm{do}(X_A = x_A)] = (I - \Lambda_{D,D})^{-\top}\Lambda_{A,D}^\top x_A.$$

Since $Y \in D$, by Definition 3.1 we have

$$\tau_{AY} = \frac{\partial}{\partial x_A}\mathbb{E}[X_Y \mid \mathrm{do}(X_A = x_A)] = \Lambda_{A,D}\left[(I - \Lambda_{D,D})^{-1}\right]_{D,Y}.$$

□

We say vertex j is a possible descendant of i, denoted as $j \in \mathrm{PossDe}(i)$, if there exists a possibly causal path from i to j. For a set of vertices A, define $\mathrm{PossDe}(A) := \cup_{i\in A}\mathrm{PossDe}(i)$. See Appendix B.4 for more details.

Corollary 3.3. *If* $Y \notin \mathrm{PossDe}(A)$, *then* $\tau_{AY} = 0$.

Proof. Since $D = \mathrm{An}(Y, \mathcal{G}_{V\setminus A})$ and $Y \notin \mathrm{PossDe}(A)$, $\Lambda_{A,D} = \mathbf{0}$. □

3.5 Recursive least squares

Consider the *special case* when the errors in the linear SEM Eq. (3.1) are jointly Gaussian. In this case, by the standard maximum likelihood theory, the Cramér–Rao bound is achieved by the maximum likelihood estimator (MLE) of the total causal effect, which can be obtained by plugging in the MLE for Λ in the block-recursive form (Proposition 3.1) into the formula Eq. (3.14). We now compute the MLE for Λ given an MPDAG $\mathcal{G}$.

When ϵ is multivariate Gaussian, the block-recursive form in Proposition 3.1 is a linear Gaussian model parameterized by $\{(\Lambda_k)_{k=2}^K, (\Omega_k)_{k=1}^K\}$, where $\Lambda_k := \Lambda_{\mathrm{Pa}(B_k),B_k}$ and Ω_k is

the covariance for ε_{B_k}. Because ε are independent between blocks (Proposition 3.1), the likelihood factorizes as

$$\mathcal{L}((\Lambda_k)_k, (\Omega_k)_k) = \prod_{k=1}^{K} \mathcal{N}\left(X_{B_k} - \Lambda_k^\top X_{\mathrm{Pa}(B_k)}; \mathbf{0}, \Omega_k\right). \tag{3.15}$$

Denote the MLE of Λ by $\hat{\Lambda}^{\mathcal{G}}$, which consists of blocks $(\hat{\Lambda}_k^{\mathcal{G}})_{k=2}^K$ and zero values elsewhere, and the MLE of Ω by $\hat{\Omega}^{\mathcal{G}} = (\hat{\Omega}_k^{\mathcal{G}})_{k=1}^K$. The superscripts highlight the dependence on MPDAG $\mathcal{G}$. The MLE maximizes $\mathcal{L}((\Lambda_k)_k, (\Omega_k)_k)$ subject to Eq. (3.12), namely

$$\mathcal{N}(\mathbf{0}, \Omega_k) \in \mathcal{P}_{\mathcal{G}_{B_k}}, \quad k = 1, \ldots, K,$$

where $\mathcal{G}_{B_k}$ is the subgraph of $\mathcal{G}$ induced by B_k. This further translates to a set of algebraic constraints on $(\Omega_k)_{k=1}^K$, namely for $k = 1, \ldots, K$,

$$\det\left[(\Omega_k)_{\{i\}\cup C, \{j\}\cup C}\right] = 0, \text{ if } i \text{ and } j \text{ are d-separated by } C \text{ in } \mathcal{G}_{B_k}; \tag{3.16}$$

see, e.g., Drton et al. (2008, §3.1). Although the constraints Eq. (3.16) may seem daunting, we will show that they do not affect the MLE for Λ.

Let the sample covariance matrix be computed with respect to mean zero, i.e.,

$$\hat{\Sigma}^{(n)} := \frac{1}{n} \sum_{i=1}^{n} X^{(i)} X^{(i)\top}, \tag{3.17}$$

where n is the sample size, and the superscripts are reserved to index samples. To reduce clutter, for a set of indices C, we often abbreviate $\Sigma_{C,C}$ as Σ_C.

Lemma 3.3. *Suppose $X^{(i)} : i = 1, \ldots, n$ is generated iid from a linear SEM Eqs. (3.2) and (3.3) associated with an unknown causal DAG $\mathcal{D}$. Suppose the error ϵ is distributed as multivariate Gaussian. Suppose $\mathcal{D} \in [\mathcal{G}]$ for a known MPDAG $\mathcal{G}$. Let $\hat{\Sigma}^{(n)}$ be the sample covariance as defined in Eq. (3.17). The MLE for $\Lambda_k = \Lambda_{\mathrm{Pa}(B_k), B_k}$ in the block-recursive form is given by*

$$\hat{\Lambda}_k^{\mathcal{G}} = \left(\hat{\Sigma}^{(n)}_{\mathrm{Pa}(B_k)}\right)^{-1} \hat{\Sigma}^{(n)}_{\mathrm{Pa}(B_k), B_k}, \quad k = 2, \ldots, K. \tag{3.18}$$

Proof. By factorization in Eq. (3.15), MLE $(\hat{\Lambda}_k^{\mathcal{G}}, \hat{\Omega}_k^{\mathcal{G}})$ is the maximizer of log-likelihood

$$\begin{aligned}
&\ell_n(\Lambda_k, \Omega_k) \\
&= -\frac{1}{2}\sum_{i=1}^{n}\left(X_{B_k}^{(i)} - \Lambda_k^\top X_{\mathrm{Pa}(B_k)}^{(i)}\right)^\top \Omega_k^{-1}\left(X_{B_k}^{(i)} - \Lambda_k^\top X_{\mathrm{Pa}(B_k)}^{(i)}\right) - \frac{n}{2}\log\det(\Omega_k) \\
&= -\frac{1}{2}\operatorname{Tr}\left(\sum_{i=1}^{n}\Omega_k^{-1}(X_{B_k}^{(i)} - \Lambda_k^\top X_{\mathrm{Pa}(B_k)}^{(i)})(X_{B_k}^{(i)} - \Lambda_k^\top X_{\mathrm{Pa}(B_k)}^{(i)})^\top\right) - \frac{n}{2}\log\det(\Omega_k),
\end{aligned}$$

subject to Eq. (3.16). Taking a derivative with respect to $\Lambda_k \in \mathbb{R}^{|\,\mathrm{Pa}(B_k)|\times|B_k|}$, we have

$$\frac{\partial \ell_n(\Lambda_k, \Omega_k)}{\partial \Lambda_k} = -2\sum_{i=1}^{n} X_{\mathrm{Pa}(B_k)}^{(i)} X_{B_k}^{(i)\top}\Omega_k^{-1} + 2\sum_{i=1}^{n} X_{\mathrm{Pa}(B_k)}^{(i)} X_{\mathrm{Pa}(B_k)}^{(i)\top}\Lambda_k\Omega_k^{-1}.$$

For any positive definite Ω_k satisfying Eq. (3.16), setting the derivative $\ell_n(\Lambda_k, \Omega_k)/\partial\Lambda_k$ to zero yields the estimate

$$\hat{\Lambda}_k^{\mathcal{G}} = \left(\frac{1}{n}\sum_{i=1}^{n} X_{\mathrm{Pa}(B_k)}^{(i)} X_{\mathrm{Pa}(B_k)}^{(i)\top}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_{\mathrm{Pa}(B_k)}^{(i)} X_{B_k}^{(i)\top}\right) = \left(\hat{\Sigma}_{\mathrm{Pa}(B_k)}^{(n)}\right)^{-1}\hat{\Sigma}_{\mathrm{Pa}(B_k),B_k}^{(n)}.$$

□

Remark 3.1. Because of the restrictive property (Corollary 3.1), each $\hat{\Lambda}_k^{\mathcal{G}}$ is computed by optimizing over the space of $|\,\mathrm{Pa}(B_k)|\times|B_k|$ matrices and the resulting MLE takes the simple form as above; see also Anderson and Olkin (1985, §5) and Amemiya (1985, §6.4) for earlier discussions of this phenomenon.

However, such a simple form is unavailable in general, when the zero constraints on Λ do not obey the restrictive property, even if we ignore the algebraic constraints Eq. (3.16) on Ω. In fact, the likelihood function can be multimodal; see also Drton and Richardson (2004); Drton (2006b); Drton et al. (2009) on seemingly unrelated regressions.

Since $\hat{\Lambda}^{\mathcal{G}}$ is obtained by simply regressing each B_i onto $\mathrm{Pa}(B_i, \mathcal{G})$ using ordinary least squares, we call this specific recursive least squares *$\mathcal{G}$-regression*. The resulting MLE for an identified total causal effect is a plugin estimator using the formula in Proposition 3.2.

Definition 3.2 ($\mathcal{G}$-regression estimator)**.** *Suppose $X^{(i)} : i = 1, \ldots, n$ is generated iid from a linear SEM Eqs. (3.2) and (3.3) associated with an unknown causal DAG $\mathcal{D}$. Suppose $\mathcal{D} \in [\mathcal{G}]$ for a known MPDAG $\mathcal{G}$. Further, suppose for $A \subset V$, $Y \in V \setminus A$, τ_{AY} is identified under the criterion of Theorem 3.1. The $\mathcal{G}$-regression estimator for the total causal effect τ_{AY} is defined as*

$$\hat{\tau}^{\mathcal{G}}_{AY} = \hat{\Lambda}^{\mathcal{G}}_{A,D} \left[(I - \hat{\Lambda}^{\mathcal{G}}_{D,D})^{-1} \right]_{D,Y}, \tag{3.19}$$

where $\hat{\Lambda}^{\mathcal{G}}$ is given by Eq. (3.18).

3.6 Efficiency theory

In this section, we establish the asymptotic efficiency of our $\mathcal{G}$-regression estimator, when the errors in the generating linear SEM are *not* necessarily Gaussian, among a reasonably large class of estimators—all regular estimators that only depend on the sample covariance. This class of estimators, despite not covering all the estimators considered in the standard semiparametric efficiency theory, includes many in the literature, such as covariate adjustment (Henckel et al., 2019; Witte et al., 2020), recursive least squares (Gupta et al., 2020; Nandy et al., 2017), and modified Cholesky decomposition of the sample covariance (Nandy et al., 2017).

Definition 3.3. *Consider an estimator $\hat{\theta}_n$ of θ, $\theta \in \mathbb{R}^k$. We say that the asymptotic covariance of $\hat{\theta}_n$ is S, and write $\operatorname{acov} \hat{\theta}_n = S$, if $\sqrt{n}(\hat{\theta}_n - \theta) \to_d \mathcal{N}(\mathbf{0}, S)$. When $k = 1$, we write $\operatorname{avar} \hat{\theta}_n$ for asymptotic variance.*

For real symmetric matrices A and B, we say $A \succeq B$ if $A - B$ is positive semidefinite. We now state our main result.

Theorem 3.3 (Asymptotic efficiency of the $\mathcal{G}$-regression estimator)**.** *Suppose data is generated iid from a linear SEM Eqs. (3.2) and (3.3) associated with an unknown causal DAG $\mathcal{D}$. Suppose $\mathcal{D} \in [\mathcal{G}]$ for a known MPDAG $\mathcal{G}$. Further, suppose for $A \subset V$, $Y \in V \setminus A$, τ_{AY} is identified under the criterion of Theorem 3.1. Let $\hat{\tau}^{\mathcal{G}}_{AY}$ be the $\mathcal{G}$-regression estimator of τ_{AY}*

(Definition 3.2). Consider any consistent estimator $\hat{\tau}_{AY} = \hat{\tau}_{AY}(\hat{\Sigma}^{(n)})$ that is a differentiable function of the sample covariance. It holds that

$$\operatorname{acov}\left(\hat{\tau}_{AY}\right) \succeq \operatorname{acov}\left(\hat{\tau}^{\mathcal{G}}_{AY}\right).$$

It is clear from definitions that both $\hat{\tau}^{\mathcal{G}}_{AY}$ and $\hat{\tau}_{AY}$ are asymptotically linear. Therefore, their asymptotic covariances are well-defined. To prove Theorem 3.3, it suffices to show that for every $w \in \mathbb{R}^{|A|}$

$$\operatorname{avar}\left(w^\top \hat{\tau}_{AY}\right) \geq \operatorname{avar}\left(w^\top \hat{\tau}^{\mathcal{G}}_{AY}\right).$$

To this end, for any fixed $w \in \mathbb{R}^{|A|}$ we define τ_w as

$$\tau_w := w^\top \tau_{AY} = \tau_w(\Lambda), \tag{3.20}$$

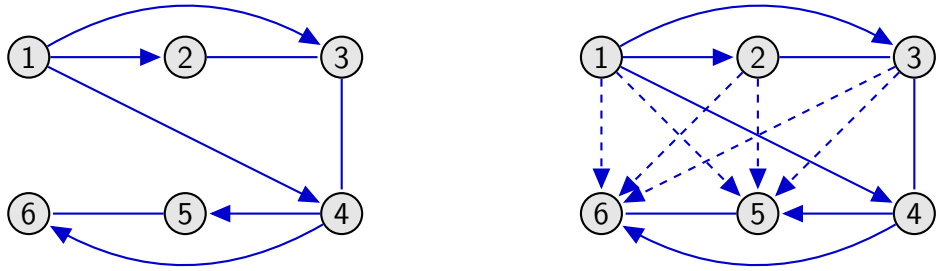
which is a smooth function of Λ. The corresponding $\mathcal{G}$-regression estimator $\hat{\tau}^{\mathcal{G}}_w := w^\top \hat{\tau}^{\mathcal{G}}_{AY} = \tau_w(\hat{\Lambda}^{\mathcal{G}})$ is still a plugin estimator (now of τ_w). Additionally, for a consistent estimator $\hat{\tau}_{AY}$ of τ_{AY}, the corresponding $\hat{\tau}_w := w^\top \hat{\tau}_{AY} = \hat{\tau}_w(\hat{\Sigma}^{(n)})$ is a consistent estimator of τ_w, in the form of a differentiable function of the sample covariance. It suffices to show $\operatorname{avar} \hat{\tau}_w \geq \operatorname{avar} \hat{\tau}^{\mathcal{G}}_w$ for every $w \in \mathbb{R}^{|A|}$.

The rest of this section is devoted to proving Theorem 3.3. First, we introduce graph $\bar{\mathcal{G}}$ as a saturated version of $\mathcal{G}$ (Proposition 3.3). In Section 3.6.1, we show that $\mathcal{G}$-regression with $\mathcal{G}$ replaced by $\bar{\mathcal{G}}$, aptly named $\bar{\mathcal{G}}$-regression, is a diffeomorphism between the space of covariance matrices and the space of parameters. In Section 3.6.2, we characterize the class of estimators relative to which $\mathcal{G}$-regression is optimal. To prove Theorem 3.3, we establish an efficiency bound for this class of estimators in Section 3.6.4 and verify that $\mathcal{G}$-regression achieves this bound in Section 3.6.5. Some of the proofs are left to Appendix B.1. See also Fig. B.1 in for an overview of the dependency structure of our results in this section.

3.6.1 $\bar{\mathcal{G}}$-regression as a diffeomorphism

Proposition 3.3 (Saturated MPDAG $\bar{\mathcal{G}}$)**.** *For MPDAG $\mathcal{G} = (V, E, U)$, an associated saturated MPDAG is $\bar{\mathcal{G}} = (V, \bar{E}, U)$, such that $\operatorname{Pa}(B_k, \bar{\mathcal{G}}) = B_{[k-1]}$ for $k = 2, \ldots, K$, where $(B_1, \ldots, B_K)$ is a bucket decomposition of V in both $\mathcal{G}$ and $\bar{\mathcal{G}}$.*

The proof can be found in Appendix B.2. In words, to create the saturated MPDAG $\bar{\mathcal{G}}$, we add all the possible directed edges between buckets $B_1, \ldots, B_K$ subject to the ordering $B_1, \ldots, B_K$. By construction, $\bar{\mathcal{G}}$ also satisfies the restrictive property in Corollary 3.1. See Fig. 3.1 for an example.



(a) (b)

Figure 3.1: (a) MPDAG $\mathcal{G} = (V, E, U)$ with buckets $B_1 = \{1\}$, $B_2 = \{2, 3, 4\}$ and $B_3 = \{5, 6\}$ and (b) its associated saturated MPDAG $\bar{\mathcal{G}} = (V, \bar{E}, U)$. The new edges in $\bar{E} \setminus E$ are drawn as dashed. Both $\mathcal{G}$ and $\bar{\mathcal{G}}$ satisfy the restrictive property in Corollary 3.1.

In the following, we introduce $\bar{\mathcal{G}}$-regression as a technical tool for establishing a diffeomorphism between the space of sample covariance matrices and the space of parameters in our semiparametric model. This link is the key to analyzing the efficiency of the estimators under consideration.

Recall that $\mathcal{P}_{\mathcal{G}}$ is the set of observed distributions generated by some linear SEM associated with a causal DAG $\mathcal{D} \in [\mathcal{G}]$, which is characterized by Proposition 3.1. More explicitly, let Q_k be the law of ε_{B_k} for $k = 1, \ldots, K$. The set of laws is explicitly prescribed as

$$\mathcal{P}_{\mathcal{G}} = \left\{ Q_1(X_{B_1}) \prod_{k=2}^{K} Q_k \left(X_{B_k} - \Lambda_{B_{[k-1]}, B_k}^{\top} X_{B_{[k-1]}} \right) : Q_k \in \mathcal{P}_{\mathcal{G}_{B_k}},\ i \to j \text{ not in } \mathcal{G} \Rightarrow \lambda_{ij} = 0 \right\} \tag{3.21}$$

where the law is indexed by $\Lambda = (\lambda_{ij})$ and $(Q_k)_{k=1}^K$. This is a *semiparametric* model and $(Q_k)_k$ is an infinite-dimensional nuisance parameter (van der Vaart, 2000, Chap. 25).

Consider the set of laws $\mathcal{P}_{\bar{\mathcal{G}}}$ associated with the saturated graph. Let $\Omega_k := \mathbb{E}_{Q_k} \varepsilon\varepsilon^\top$ be the covariance of Q_k for $k = 1, \ldots, K$. Let $\mathbb{R}^{n\times n}_{\text{PD}}$ denote the set of $n \times n$ symmetric, positive definite matrices. By our assumption, $\Omega_k \in \mathbb{R}^{|B_k|\times|B_k|}_{\text{PD}}$. Also, consider the coefficients $\Lambda = (\lambda_{ij})$ such that $\lambda_{ij} \neq 0$ only if $i \to j$ in $\bar{\mathcal{G}}$, or equivalently, $i \in B_l$ and $j \in B_m$ for $l < m$. Then, the covariance of X, denoted as Σ, under any $P \in \mathcal{P}_{\bar{\mathcal{G}}}$ is determined from $(\Omega_k)_k$ and Λ. Let us write this *covariance map* as

$$\Sigma = \phi_{\bar{\mathcal{G}}}\left((\Lambda_k)_{k=2}^K, (\Omega_k)_{k=1}^K\right),$$

where $\Lambda_k = \Lambda_{B_{[k-1]},B_k}$ is of dimension $(|B_1|+\cdots+|B_{k-1}|)\times|B_k|$. It follows from Corollary 3.2 that the covariance map $\phi_{\bar{\mathcal{G}}}$ is explicitly given by

$$\Sigma_{B_1} = \Omega_1, \quad \Sigma_{B_k} = \Lambda_k^\top \Sigma_{B_{[k-1]}} \Lambda_k + \Omega_k, \quad \Sigma_{B_{[k-1]},B_k} = \Sigma_{B_{[k-1]}} \Lambda_k, \quad k = 2, \ldots, K. \tag{3.22}$$

Further, the covariance map $\phi_{\bar{\mathcal{G}}}$ is a *diffeomorphism* between its domain and the set of $|V| \times |V|$ positive definite matrices.

Lemma 3.4. *Covariance map $\phi_{\bar{\mathcal{G}}}$ given by Eq. (3.22) is invertible. Further, $\left((\Lambda_k)_{k=2}^K, (\Omega_k)_{k=1}^K\right) \leftrightarrow \Sigma$ given by $\phi_{\bar{\mathcal{G}}}$ and its inverse $\phi_{\bar{\mathcal{G}}}^{-1}$ is a diffeomorphism between $\left(\times_{k=2}^K \mathbb{R}^{(|B_1|+\cdots+|B_{k-1}|)\times|B_k|}\right) \times \left(\times_{k=1}^K \mathbb{R}^{|B_k|\times|B_k|}_{PD}\right)$ and $\mathbb{R}^{|V|\times|V|}_{PD}$.*

Proof. By definition, covariance map $\phi_{\bar{\mathcal{G}}}$ is differentiable. To show diffeomorphism, we need to show that $\phi_{\bar{\mathcal{G}}}^{-1}(\Sigma)$ exists for every $\Sigma \in \mathbb{R}^{|V|\times|V|}_{\text{PD}}$ and that $\phi_{\bar{\mathcal{G}}}^{-1}$ is differentiable. For any positive definite Σ, the inverse covariance map $\phi_{\bar{\mathcal{G}}}^{-1}(\Sigma)$ is explicitly given by

$$\Lambda_k = \left(\Sigma_{B_{[k-1]}}\right)^{-1} \Sigma_{B_{[k-1]},B_k}, \quad k = 2, \ldots, K, \tag{3.23}$$

and

$$\Omega_k = \Sigma_{B_k \cdot B_{[k-1]}} = \Sigma_{B_k} - \Sigma^\top_{B_{[k-1]},B_k} \Sigma^{-1}_{B_{[k-1]}} \Sigma_{B_{[k-1]},B_k}, \quad k = 1, \ldots, K, \tag{3.24}$$

where $\Sigma_{B_k \cdot B_{[k-1]}}$ is the Schur complement of block B_k with respect to block $B_{[k-1]}$. Because Σ is positive definite, Schur complement Ω_k is also positive definite (Horn and Johnson, 2012, page 495). Clearly, the map $\phi_{\bar{\mathcal{G}}}^{-1}(\cdot)$ is differentiable. □

By Eqs. (3.23) and (3.24), Λ_k is the matrix of population least squares coefficients in a regression of X_{B_k} onto $X_{B_1 \cup \cdots \cup B_{k-1}}$ according to $\bar{\mathcal{G}}$, and Ω_k is the corresponding covariance of regression residuals. Hence, $\phi_{\bar{\mathcal{G}}}^{-1}(\Sigma)$ is called "$\bar{\mathcal{G}}$-regression".

Remark 3.2. In the special case when $\mathcal{G}$ is a DAG such that every bucket B_i is a singleton, Lemma 3.4 reduces to $(\Lambda, \omega) \leftrightarrow \Sigma$ given by $(\phi_{\bar{\mathcal{G}}}, \phi_{\bar{\mathcal{G}}}^{-1})$ being a diffeomorphism between

$$\left\{\Lambda \in \mathbb{R}^{|V| \times |V|} : \Lambda \text{ is upper-triangular}\right\} \times \left\{\omega \in \mathbb{R}^{|V|} : \omega_i > 0,\ i = 1, \ldots, |V|\right\} \longleftrightarrow \mathbb{R}_{\mathrm{PD}}^{|V| \times |V|}.$$

The covariance map is $\Sigma = \phi_{\bar{\mathcal{G}}}(\Lambda, \omega) = (I - \Lambda)^{-\top} \operatorname{diag}(\omega)(I - \Lambda)^{-1}$, and the inverse covariance map $\phi_{\bar{\mathcal{G}}}^{-1}$ is given by the unique LDL decomposition of Σ^{-1}. Lemma 3.4 is a generalization of Drton (2018, Theorem 7.2).

3.6.2 Covariance-based, consistent estimators

We now characterize the class of estimators relative to which the optimality of our estimator is established. Recall that under $P \in \mathcal{P}_{\mathcal{G}}$, $\hat{\Sigma} = \hat{\Sigma}^{(n)}$ is the sample covariance, Σ is the population covariance and $\tau_w = w^\top \tau_{AY}$. We assume that $n > \max_k\{|B_k| + |\operatorname{Pa}(B_k, \mathcal{G})|\}$ such that $\hat{\Sigma}^{(n)}$ is positive definite almost surely (Drton and Eichler, 2006, Sec. 3.1). For simplicity, the superscript (n) is often omitted.

Definition 3.4. *The class of estimators for τ_w under consideration is*

$$\mathcal{T}_w := \Bigg\{ \hat{\tau}_w\left(\hat{\Sigma}^{(n)}\right) : \mathbb{R}_{PD}^{|V| \times |V|} \to \mathbb{R} : \\ \hat{\tau}_w \text{ differentiable},\ \hat{\tau}_w(\hat{\Sigma}^{(n)}) \to_p \tau_w(P) \text{ as } n \to \infty \text{ under every } P \in \mathcal{P}_{\mathcal{G}} \Bigg\}. \quad (3.25)$$

By definition, in particular, $\mathcal{T}_w$ includes all regular estimators computable with least squares operations.

Characterizing $\mathcal{T}_w$ Let $(\hat{\Lambda}_k^{\bar{\mathcal{G}}})_{k=2}^K, (\hat{\Omega}_k^{\bar{\mathcal{G}}})_{k=1}^K$ be the image of $\hat{\Sigma}$ under $\phi_{\bar{\mathcal{G}}}^{-1}$. Recall that $(\Lambda_k)_{k=2}^K, (\Omega_k)_{k=1}^K$ is the image of Σ under $\phi_{\bar{\mathcal{G}}}^{-1}$. For a matrix C, let $\operatorname{vec} C$ denote vector-

izing C by concatenating its columns. Each $\operatorname{vec}\hat{\Lambda}_k^{\bar{\mathcal{G}}}$ can be split by *coordinates* into vectors

$$\hat{\Lambda}_{k,\mathcal{G}}^{\bar{\mathcal{G}}} = \left(\hat{\lambda}_{ij}^{\bar{\mathcal{G}}} : j \in B_k, i \in \mathrm{Pa}(B_k, \mathcal{G})\right), \quad \hat{\Lambda}_{k,\mathcal{G}^c}^{\bar{\mathcal{G}}} = \left(\hat{\lambda}_{ij}^{\bar{\mathcal{G}}} : j \in B_k, i \in \mathrm{Pa}(B_k, \bar{\mathcal{G}}) \setminus \mathrm{Pa}(B_k, \mathcal{G})\right), \tag{3.26}$$

where $\left(\hat{\Lambda}_{k,\mathcal{G}}^{\bar{\mathcal{G}}}\right)_k$ corresponds to between-bucket edges in $\mathcal{G}$ and $\left(\hat{\Lambda}_{k,\mathcal{G}^c}^{\bar{\mathcal{G}}}\right)_k$ corresponds to between-bucket edges in $\bar{\mathcal{G}}$ but not in $\mathcal{G}$. In the example of Fig. 3.1, we have $\hat{\Lambda}_{2,\mathcal{G}}^{\bar{\mathcal{G}}} = (\hat{\lambda}_{12}^{\bar{\mathcal{G}}}, \hat{\lambda}_{13}^{\bar{\mathcal{G}}}, \hat{\lambda}_{14}^{\bar{\mathcal{G}}})^\top$, $\hat{\Lambda}_{3,\mathcal{G}}^{\bar{\mathcal{G}}} = (\hat{\lambda}_{45}^{\bar{\mathcal{G}}}, \hat{\lambda}_{46}^{\bar{\mathcal{G}}})^\top$ and $\hat{\Lambda}_{2,\mathcal{G}^c}^{\bar{\mathcal{G}}} = \texttt{NULL}$, $\hat{\Lambda}_{3,\mathcal{G}^c}^{\bar{\mathcal{G}}} = (\hat{\lambda}_{15}^{\bar{\mathcal{G}}}, \hat{\lambda}_{16}^{\bar{\mathcal{G}}}, \hat{\lambda}_{25}^{\bar{\mathcal{G}}}, \hat{\lambda}_{26}^{\bar{\mathcal{G}}}, \hat{\lambda}_{35}^{\bar{\mathcal{G}}}, \hat{\lambda}_{36}^{\bar{\mathcal{G}}})^\top$. Similarly, $\operatorname{vec}\Lambda_k$ can be split into $\Lambda_{k,\mathcal{G}}$ and $\Lambda_{k,\mathcal{G}^c}$ for $k = 2, \ldots, K$.

The following lemma directly follows from Definition 3.4 and Lemma 3.4.

Lemma 3.5. *An estimator $\hat{\tau}_w \in \mathcal{T}_w$ can be written as*

$$\hat{\tau}_w\left(\hat{\Sigma}^{(n)}\right) = \hat{\tau}_w\left((\hat{\Lambda}_{k,\mathcal{G}}^{\bar{\mathcal{G}}})_{k=2}^K, (\hat{\Lambda}_{k,\mathcal{G}^c}^{\bar{\mathcal{G}}})_{k=2}^K, (\hat{\Omega}_k^{\bar{\mathcal{G}}})_{k=1}^K\right)$$

for function $\hat{\tau}_w\left((\hat{\Lambda}_{k,\mathcal{G}}^{\bar{\mathcal{G}}})_{k=2}^K, (\hat{\Lambda}_{k,\mathcal{G}^c}^{\bar{\mathcal{G}}})_{k=2}^K, (\hat{\Omega}_k^{\bar{\mathcal{G}}})_{k=1}^K\right)$ that is differentiable in its arguments.

The consistency of $\hat{\tau}_w$ implies the following two results.

Lemma 3.6. *For any $\hat{\tau}_w \in \mathcal{T}_w$, it holds that*

$$\hat{\tau}_w\left((\Lambda_{k,\mathcal{G}})_{k=2}^K, (\mathbf{0})_{k=2}^K, (\Omega_k)_{k=1}^K\right) \equiv \tau_w\left((\Lambda_{k,\mathcal{G}})_{k=2}^K\right) \tag{3.27}$$

for all $(\Lambda_{k,\mathcal{G}})_k$ and all positive definite $(\Omega_k)_k$.

Proof. Under any $P \in \mathcal{P}_{\mathcal{G}}$, since $\hat{\Sigma} \to_p \Sigma$ as $n \to \infty$ by the law of large numbers, by Lemma 3.4 and the continuous mapping theorem (van der Vaart, 2000, page 11), we have $\hat{\Lambda}_{k,\mathcal{G}}^{\bar{\mathcal{G}}} \to_p \Lambda_{k,\mathcal{G}}$, $\hat{\Lambda}_{k,\mathcal{G}^c}^{\bar{\mathcal{G}}} \to_p \mathbf{0}$ and $\hat{\Omega}_k^{\bar{\mathcal{G}}} \to_p \Omega_k$ for $k = 2, \ldots, K$. By Lemma 3.5 and continuous mapping again, $\hat{\tau}_w \to_p \hat{\tau}_w\left((\Lambda_{k,\mathcal{G}})_{k=2}^K, (\mathbf{0})_{k=2}^K, (\Omega_k)_{k=1}^K\right)$. The result then follows from the consistency of $\hat{\tau}_w$ under every $P \in \mathcal{P}_{\mathcal{G}}$. □

Corollary 3.4. *For $\hat{\tau}_w \in \mathcal{T}_w$, at any $((\Lambda_{k,\mathcal{G}})_{k=2}^K, (\mathbf{0})_{k=2}^K, (\Omega_k)_{k=1}^K)$, it holds that*

$$\frac{\partial\hat{\tau}_w}{\partial\Lambda_{k,\mathcal{G}}} = \frac{\partial\tau_w}{\partial\Lambda_{k,\mathcal{G}}} \ (k = 2, \ldots, K), \quad \frac{\partial\hat{\tau}_w}{\partial\Omega_k} = \mathbf{0} \ (k = 1, \ldots, K). \tag{3.28}$$

Proof. Let $\langle \cdot, \cdot \rangle$ denote the inner product. Since $\hat{\tau}_w$ is differentiable (Lemma 3.5), by a Taylor expansion at $((\Lambda_{k,\mathcal{G}})_{k=2}^K, (\mathbf{0})_{k=2}^K, (\Omega_k)_{k=1}^K)$, we have

$$\begin{aligned}
&\hat{\tau}_w\left((\Lambda_{k,\mathcal{G}} + \Delta\Lambda_{k,\mathcal{G}})_{k=2}^K, (\mathbf{0})_{k=2}^K, (\Omega_k + \Delta\Omega_k)_{k=1}^K\right) - \hat{\tau}_w\left((\Lambda_{k,\mathcal{G}})_{k=2}^K, (\mathbf{0})_{k=2}^K, (\Omega_k)_{k=1}^K\right) \\
&= \sum_{k=2}^K \left(\left\langle \frac{\partial \hat{\tau}_w}{\partial \Lambda_{k,\mathcal{G}}}, \Delta\Lambda_{k,\mathcal{G}} \right\rangle + o(\|\Delta\Lambda_{k,\mathcal{G}}\|) \right) + \sum_{k=1}^K \left(\left\langle \frac{\partial \hat{\tau}_w}{\partial \Omega_k}, \Delta\Omega_k \right\rangle + o(\|\Delta\Omega_k\|) \right),
\end{aligned}$$

which by Lemma 3.6 must equal $\tau_w((\Lambda_{k,\mathcal{G}} + \Delta\Lambda_{k,\mathcal{G}})_{k=2}^K) - \tau_w((\Lambda_{k,\mathcal{G}})_{k=2}^K)$. The result then follows from the differentiability of $\tau_w(\cdot)$ and the definition of derivatives. □

Note that Corollary 3.4 is similar to the conditions imposed on influence functions in standard semiparametric efficiency theory; see, e.g., Tsiatis (2006, Corollary 1, §3.1). However, the gradients $\partial\hat{\tau}_w / \partial\hat{\Lambda}^{\bar{\mathcal{G}}}_{k,\mathcal{G}^c}$ for $k = 2, \ldots, K$ are *free to vary* because $\hat{\Lambda}^{\bar{\mathcal{G}}}_{k,\mathcal{G}^c} \to_p \mathbf{0}$. That is, an estimator $\hat{\tau}_w \in \mathcal{T}_w$ can take arbitrary values as its second argument varies in the vicinity of zero, as long as differentiability is maintained.

3.6.3 Asymptotic covariance of least squares coefficients

We use this section to derive some asymptotic results that will be used to prove Theorem 3.3.

Consider a vertex $j \in B_k$ for $k \in \{2, \ldots, K\}$ and a set of vertices C such that $\mathrm{Pa}(B_k, \mathcal{G}) \subseteq C \subseteq \mathrm{Pa}(B_k, \bar{\mathcal{G}})$. Let $\hat{\lambda}^{(n)}_{C,j} \in \mathbb{R}^{|C|}$ be the least squares coefficients from regressing X_j onto X_C under sample size n. Let $\lambda_{C,j}$ be the corresponding true edge coefficient vector from Λ in Proposition 3.1. Then $\lambda_{C,j}$ has non-zero coordinates only for those indices in $\mathrm{Pa}(B_k, \mathcal{G})$. Because $X_j = \lambda_{C,j}^\top X_C + \varepsilon_j$ with $\varepsilon_j \perp\!\!\!\perp X_C$ by Corollary 3.2, we have $\hat{\lambda}^{(n)}_{C,j} \to_p \lambda_{C,j}$ under every $P \in \mathcal{P}_{\mathcal{G}}$. Moreover, we have the following asymptotic linear expansion.

Lemma 3.7. *Let j be a vertex in bucket B_k for $k \in \{2, \ldots, K\}$. Let C be a set of vertices such that $\mathrm{Pa}(B_k, \mathcal{G}) \subseteq C \subseteq \mathrm{Pa}(B_k, \bar{\mathcal{G}})$. Under any $P \in \mathcal{P}_{\mathcal{G}}$, it holds that*

$$\hat{\lambda}^{(n)}_{C,j} - \lambda_{C,j} = \frac{1}{n} \sum_{i=1}^n (\Sigma_C)^{-1} X_C^{(i)} \varepsilon_j^{(i)} + O_p(n^{-1}),$$

where $\Sigma = \mathbb{E}_P XX^\top$, $\hat{\lambda}^{(n)}_{C,j}$ is the vector of least squares coefficients from regressing X_j onto X_C under sample size n, and $\lambda_{C,j}$ is the vector of true coefficients in Proposition 3.1.

We now use Lemma 3.7 to obtain the covariance structure of $\bar{\mathcal{G}}$-regression coefficients $(\hat{\Lambda}_k^{\bar{\mathcal{G}}})_{k=2}^K$. Recall that $\hat{\Lambda}_k^{\bar{\mathcal{G}}} \in \mathbb{R}^{|B_{[k-1]}|\times|B_k|}$ with $B_{[k-1]} = B_1 \cup \cdots \cup B_{k-1}$ and

$$\left((\hat{\Lambda}_k^{\bar{\mathcal{G}}})_{k=2}^K, (\hat{\Omega}_k^{\bar{\mathcal{G}}})_{k=1}^K\right) = \phi_{\bar{\mathcal{G}}}^{-1}\left(\hat{\Sigma}^{(n)}\right),$$

as given by Eqs. (3.23) and (3.24). For matrices $A \in \mathbb{R}^{m\times n}, B \in \mathbb{R}^{p\times q}$, the Kronecker product $A \otimes B$ is an $mp \times nq$ matrix given by

$$A \otimes B = \begin{pmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{pmatrix}.$$

Lemma 3.8. *Let $(\hat{\Lambda}_k^{\bar{\mathcal{G}}})_{k=2}^K$ be the $\bar{\mathcal{G}}$-regression coefficients under sample size n. Under any $P \in \mathcal{P}_{\mathcal{G}}$, it holds that*

$$\sqrt{n}\begin{pmatrix} \mathrm{vec}(\hat{\Lambda}_2^{\bar{\mathcal{G}}} - \Lambda_2) \\ \vdots \\ \mathrm{vec}(\hat{\Lambda}_K^{\bar{\mathcal{G}}} - \Lambda_K) \end{pmatrix} \to_d \mathcal{N}\left(\mathbf{0}, \mathrm{diag}\left\{\Omega_2 \otimes \left(\Sigma_{B_{[1]}}\right)^{-1}, \ldots, \Omega_K \otimes \left(\Sigma_{B_{[K-1]}}\right)^{-1}\right\}\right).$$

Remark 3.3. $\sqrt{n}\,\mathrm{vec}(\hat{\Lambda}_k^{(n)} - \Lambda_k) \to_d \mathcal{N}\left(\mathbf{0}, \Omega_k \otimes \left(\Sigma_{B_{[k-1]}}\right)^{-1}\right)$ is equivalent to

$$\sqrt{n}(\hat{\Lambda}_k^{(n)} - \Lambda_k) \to_d \mathcal{MN}\left(\mathbf{0}, \left(\Sigma_{B_{[k-1]}}\right)^{-1}, \Omega_k\right),$$

where the RHS is a centered matrix normal distribution with row covariance $(\Sigma_{B_{[k-1]}})^{-1}$ and column covariance Ω_k; see Dawid (1981).

Similarly, we can compute the asymptotic covariance of the $\mathcal{G}$-regression coefficients. To obtain the result below, we rely on the restrictive property of $\mathcal{G}$ (Corollary 3.1).

Lemma 3.9. *Let $(\hat{\Lambda}_k^{\mathcal{G}})_{k=2}^K$ be the $\mathcal{G}$-regression coefficients as defined in Lemma 3.3 under sample size n. Under any $P \in \mathcal{P}_{\mathcal{G}}$, it holds that*

$$\sqrt{n}\begin{pmatrix} \mathrm{vec}(\hat{\Lambda}_2^{\mathcal{G}} - \Lambda_2) \\ \vdots \\ \mathrm{vec}(\hat{\Lambda}_K^{\mathcal{G}} - \Lambda_K) \end{pmatrix} \to_d \mathcal{N}\left(\mathbf{0}, \mathrm{diag}\left\{\Omega_2 \otimes \left(\Sigma_{\mathrm{Pa}(B_2,\mathcal{G})}\right)^{-1}, \ldots, \Omega_K \otimes \left(\Sigma_{\mathrm{Pa}(B_K,\mathcal{G})}\right)^{-1}\right\}\right).$$

3.6.4 Efficiency bound

We first notice a simple fact of the quadratic form and a property of the Kronecker product.

Lemma 3.10. *Let $S \in \mathbb{R}^{n\times n}_{PD}, x \in \mathbb{R}^n$ and suppose that (A, B) is a partition of the set $\{1, \dots, n\}$. For any fixed x_A, it holds that*

$$x^\top S x \geq x_A^\top (S_{A\cdot B}) x_A,$$

where $S_{A\cdot B} = S_{A,A} - S_{A,B}S_{B,B}^{-1}S_{B,A}$. The equality holds if and only if $x_B = -S_{B,B}^{-1}S_{B,A}x_A$.

Lemma 3.11 (Liu (1999, Theorem 1))**.** *Let $A \in \mathbb{R}^{m\times m}$ and $C \in \mathbb{R}^{n\times n}$ be non-singular. Suppose $\alpha \subset [m]$, $\beta \subset [n]$. Let α^c, β^c denote their respective complements. Let $\gamma^c = \{n(i-1)+j : i \in \alpha^c, j \in \beta^c\}$ and $\gamma = [mn] \setminus \gamma^c$. We have*

$$A_{\alpha^c\cdot\alpha} \otimes C_{\beta^c\cdot\beta} = (A \otimes C)_{\gamma^c\cdot\gamma}.$$

Lemma 3.12. *Suppose the assumptions of Theorem 3.3 hold. Fix $w \in \mathbb{R}^{|A|}$ and let $\tau_w = w^\top \tau_{AY} = \tau_w((\Lambda_{k,\mathcal{G}})_{k=2}^K)$ as in Eq. (3.20). Consider any estimator $\hat{\tau}_w \in \mathcal{T}_w$ given by Definition 3.4. Then under any $P \in \mathcal{P}_\mathcal{G}$, it holds that*

$$\operatorname{avar}(\hat{\tau}_w) \geq \sum_{k=2}^K h_k^\top \Omega_k \otimes (\Sigma_{\mathrm{Pa}(B_k,\mathcal{G})})^{-1} h_k, \tag{3.29}$$

where $(\Omega_k)_{k=2}^K$ and Σ are determined by P, and the gradient vectors $h_k = \partial\tau_w((\Lambda_{k,\mathcal{G}})_k)/\partial\Lambda_{k,\mathcal{G}}$ for $k = 2, \dots, K$ evaluated at $(\Lambda_{k,\mathcal{G}})_k$ are determined by $\tau_w(\cdot)$ and P.

Proof. By Lemma 3.5, estimator $\hat{\tau}_w \in \mathcal{T}_w$ can be written as

$$\hat{\tau}_w = \hat{\tau}_w\left((\hat{\Lambda}^{\bar{\mathcal{G}}}_{k,\mathcal{G}})_{k=2}^K,\ (\hat{\Lambda}^{\bar{\mathcal{G}}}_{k,\mathcal{G}^c})_{k=2}^K,\ (\hat{\Omega}^{\bar{\mathcal{G}}}_k)_{k=1}^K\right),$$

where the arguments correspond to the image of $\hat{\Sigma}$ under $\phi_{\bar{\mathcal{G}}}^{-1}$; see Eq. (3.26). Estimator $\hat{\tau}_w \in \mathcal{T}_w$ is asymptotically normal. By the delta method (Shorack, 2000, Sec 11.2), we have

$$\operatorname{avar}(\hat{\tau}_w) = \begin{pmatrix} \partial\hat{\tau}_w/\partial(\Lambda_{k,\mathcal{G}})_{k=2}^K \\ \partial\hat{\tau}_w/\partial(\Lambda_{k,\mathcal{G}^c})_{k=2}^K \\ \partial\hat{\tau}_w/\partial(\Omega_k)_{k=1}^K \end{pmatrix}^\top \operatorname{acov}\begin{Bmatrix} \operatorname{vec}(\hat{\Lambda}^{\bar{\mathcal{G}}}_{k,\mathcal{G}})_{k=2}^K \\ \operatorname{vec}(\hat{\Lambda}^{\bar{\mathcal{G}}}_{k,\mathcal{G}^c})_{k=2}^K \\ \operatorname{vec}(\hat{\Omega}^{\bar{\mathcal{G}}}_k)_{k=1}^K \end{Bmatrix} \begin{pmatrix} \partial\hat{\tau}_w/\partial(\Lambda_{k,\mathcal{G}})_{k=2}^K \\ \partial\hat{\tau}_w/\partial(\Lambda_{k,\mathcal{G}^c})_{k=2}^K \\ \partial\hat{\tau}_w/\partial(\Omega_k)_{k=1}^K \end{pmatrix},$$

where the partial derivatives of $\hat{\tau}_w(\cdot)$ are evaluated at $\left((\Lambda_{k,\mathcal{G}})_{k=2}^K, (\mathbf{0})_{k=2}^K, (\Omega_k)_{k=1}^K\right)$, the image of Σ under $\phi_{\bar{\mathcal{G}}}^{-1}$.

Using $\partial\hat{\tau}_w/\partial\Omega_k = \mathbf{0}$ for $k = 1, \ldots, K$ from Corollary 3.4, it follows that

$$\begin{aligned}\operatorname{avar}(\hat{\tau}_w) &= \begin{pmatrix} \partial\hat{\tau}_w/\partial(\Lambda_{k,\mathcal{G}})_{k=2}^K \\ \partial\hat{\tau}_w/\partial(\Lambda_{k,\mathcal{G}^c})_{k=2}^K \end{pmatrix}^\top \operatorname{acov}\left\{ \begin{matrix} \operatorname{vec}(\hat{\Lambda}_{k,\mathcal{G}}^{\bar{\mathcal{G}}})_{k=2}^K \\ \operatorname{vec}(\hat{\Lambda}_{k,\mathcal{G}^c}^{\bar{\mathcal{G}}})_{k=2}^K \end{matrix} \right\} \begin{pmatrix} \partial\hat{\tau}_w/\partial(\Lambda_{k,\mathcal{G}})_{k=2}^K \\ \partial\hat{\tau}_w/\partial(\Lambda_{k,\mathcal{G}^c})_{k=2}^K \end{pmatrix} \\ &= \sum_{k=2}^K \begin{pmatrix} \partial\hat{\tau}_w/\partial\Lambda_{k,\mathcal{G}} \\ \partial\hat{\tau}_w/\partial\Lambda_{k,\mathcal{G}^c} \end{pmatrix}^\top \operatorname{acov}\left\{ \begin{matrix} \hat{\Lambda}_{k,\mathcal{G}}^{\bar{\mathcal{G}}} \\ \hat{\Lambda}_{k,\mathcal{G}^c}^{\bar{\mathcal{G}}} \end{matrix} \right\} \begin{pmatrix} \partial\hat{\tau}_w/\partial\Lambda_{k,\mathcal{G}} \\ \partial\hat{\tau}_w/\partial\Lambda_{k,\mathcal{G}^c} \end{pmatrix},\end{aligned}$$

where we have used the block-diagonal structure of the asymptotic covariance from Lemma 3.8. Let

$$S^{(k)} := \operatorname{acov}\left\{ \begin{matrix} \hat{\Lambda}_{k,\mathcal{G}}^{\bar{\mathcal{G}}} \\ \hat{\Lambda}_{k,\mathcal{G}^c}^{\bar{\mathcal{G}}} \end{matrix} \right\}, \quad k = 2, \ldots, K,$$

which equals

$$S^{(k)} = \Omega_k \otimes \left(\Sigma_{B_{[k-1]}}\right)^{-1}, \quad k = 2, \ldots, K,$$

by Lemma 3.8. From Corollary 3.4, note that $\partial\hat{\tau}_w/\partial(\Lambda_{k,\mathcal{G}})_k \equiv h_k$ is *fixed* for $k = 2, \ldots, K$. Then, Lemma 3.10 yields the lower bound

$$\operatorname{avar}(\hat{\tau}_w) \geq \sum_{k=2}^K h_k^\top S_{\mathcal{G}\cdot\mathcal{G}^c}^{(k)} h_k,$$

where indices $\mathcal{G}$ and $\mathcal{G}^c$ correspond to the coordinates in $\hat{\Lambda}_{k,\mathcal{G}}^{\bar{\mathcal{G}}}$ and $\hat{\Lambda}_{k,\mathcal{G}^c}^{\bar{\mathcal{G}}}$ respectively. Indices $\mathcal{G}$ correspond to $\{(i,j) : j \in B_k, i \in \operatorname{Pa}(B_k, \mathcal{G})\}$; by construction of $\bar{\mathcal{G}}$, indices $\mathcal{G}^c$ correspond to $\{(i,j) : j \in B_k, i \in \operatorname{Pa}(B_k, \bar{\mathcal{G}}) \setminus \operatorname{Pa}(B_k, \mathcal{G})\}$. Now, to abuse the notation slightly, we apply Lemma 3.11 with

$$A = \Omega_k, \quad C = (\Sigma_{B_{[k-1]}})^{-1}, \quad \alpha = \emptyset, \quad \beta = \operatorname{Pa}(B_k, \bar{\mathcal{G}}) \setminus \operatorname{Pa}(B_k, \mathcal{G}),$$

such that

$$\alpha^c = \{1, \ldots, |B_k|\}, \quad \beta^c = \operatorname{Pa}(B_k, \mathcal{G}), \quad \gamma = \mathcal{G}^c, \quad \gamma^c = \mathcal{G}.$$

We obtain

$$S_{\mathcal{G}\cdot\mathcal{G}^c}^{(k)} = \Omega_k \otimes \left[(\Sigma_{B_{[k-1]}})^{-1}\right]_{\beta^c\cdot\beta} = \Omega_k \otimes \left(\Sigma_{\operatorname{Pa}(B_k,\mathcal{G})}\right)^{-1},$$

where the last step follows from $(H^{-1})_{\beta^c\cdot\beta} = (H_{\beta^c,\beta^c})^{-1}$ (Horn and Johnson, 2012, §0.8). □

3.6.5 *Efficiency of $\mathcal{G}$-regression estimator*

In Section 3.5, we have seen that when the errors are Gaussian, the $\mathcal{G}$-regression plugin is the MLE and hence achieves the efficiency bound. Here, we show that this is still true relative to the class of estimators we consider, even though the errors are not necessarily Gaussian. We verify that $\hat{\tau}_w^{\mathcal{G}} = w^\top \hat{\tau}_{AY}^{\mathcal{G}}$ achieves the efficiency bound above.

Lemma 3.13. *Let $\hat{\tau}_w^{\mathcal{G}} := w^\top \hat{\tau}_{AY}^{\mathcal{G}}$, where $\hat{\tau}_{AY}^{\mathcal{G}}$ is the $\mathcal{G}$-regression estimator (Definition 3.2). Under the same assumptions as Lemma 3.12, it holds that $\hat{\tau}_w^{\mathcal{G}} \in \mathcal{T}_w$ and $\hat{\tau}_w^{\mathcal{G}}$ achieves the efficiency bound in Eq. (3.29) under every $P \in \mathcal{P}_{\mathcal{G}}$.*

Proof. By Definition 3.2, $\hat{\tau}_w^{\mathcal{G}} \in \mathcal{T}_w$. Further, note that

$$\hat{\tau}_w^{\mathcal{G}} = \tau_w\left((\hat{\Lambda}_k^{\mathcal{G}})_{k=2}^K\right),$$

where $(\hat{\Lambda}_k^{\mathcal{G}})_{k=2}^K$ are the $\mathcal{G}$-regression coefficients in Eq. (3.18). Under any $P \in \mathcal{P}_{\mathcal{G}}$, we now verify that avar $\hat{\tau}_{\mathcal{G}}$ matches the RHS of Eq. (3.29). By the delta method (Shorack, 2000, Sec 11.2), we have

$$\begin{aligned}
\operatorname{avar} \hat{\tau}_w^{\mathcal{G}} &= \left(\partial \tau_w / \partial \operatorname{vec}(\Lambda_k)_{k=2}^K\right)^\top \operatorname{acov}\left\{\operatorname{vec}(\hat{\Lambda}_k^{\mathcal{G}})_{k=2}^K\right\}\left(\partial \tau_w / \partial \operatorname{vec}(\Lambda_k)_{k=2}^K\right) \\
&\overset{\text{(i)}}{=} \sum_{k=2}^K \left(\partial \tau_w / \partial \operatorname{vec} \Lambda_k\right)^\top \operatorname{acov}\left\{\operatorname{vec} \hat{\Lambda}_k^{\mathcal{G}}\right\}\left(\partial \tau_w / \partial \operatorname{vec} \Lambda_k\right) \\
&\overset{\text{(ii)}}{=} \sum_{k=2}^K \left(\partial \tau_w / \partial \operatorname{vec} \Lambda_k\right)^\top \Omega_k \otimes \left(\Sigma_{\mathrm{Pa}(B_k,\mathcal{G})}\right)^{-1}\left(\partial \tau_w / \partial \operatorname{vec} \Lambda_k\right),
\end{aligned}$$

which equals the RHS of Eq. (3.29). The partial derivatives of $\tau_w(\cdot)$ are evaluated at $(\Lambda_k)_{k=2}^K$. Step (i) follows from the block-diagonal structure of the asymptotic covariance of $\hat{\Lambda}_{\mathcal{G}}$ given by Lemma 3.9, and (ii) follows from the same lemma. □

Finally, we complete the proof of our main result.

Proof. of Theorem 3.3. Fix any $P \in \mathcal{P}_{\mathcal{G}}$. It suffices to show that for every $w \in \mathbb{R}^{|A|}$,

$$w^\top \operatorname{acov}(\hat{\tau}_{AY}) w \geq w^\top \operatorname{acov}(\hat{\tau}_{AY}^{\mathcal{G}}) w,$$

or equivalently

$$\operatorname{avar}\left(w^\top \hat{\tau}_{AY}\right) \geq \operatorname{avar}\left(w^\top \hat{\tau}^{\mathcal{G}}_{AY}\right).$$

This is true because for every $\hat{\tau}_{AY}$ in consideration, $\hat{\tau}_w := w^\top \hat{\tau}_{AY} \in \mathcal{T}_w$ and hence $\hat{\tau}_w$ is subject to the lower bound in Lemma 3.12. Meanwhile, by Lemma 3.13, such a lower bound is achieved by $\hat{\tau}^{\mathcal{G}}_w = w^\top \hat{\tau}^{\mathcal{G}}_{AY}$. The proof is complete because the choice of w is arbitrary. □

Remark 3.4. For Theorem 3.3 to hold, the independence error assumption Eq. (3.3) of the underlying linear SEM *cannot* be relaxed to uncorrelated errors. This comes from inspecting the proof of Lemma 3.8 in Appendix B.1. To show that the $\bar{\mathcal{G}}$-regression coefficients are asymptotically independent across buckets, the independence of errors is used to establish that for $2 \leq k < k' \leq K$, $j \in B_k$, $j' \in B_{k'}$, $\operatorname{cov}(\varepsilon_j X_{B_{[k-1]}}, \varepsilon_{j'} X_{B_{[k'-1]}}) = \mathbf{0}$.

Suppose for now $\{\epsilon_i : i \in V\}$ are only uncorrelated and hence $\{\varepsilon_{B_k} : k = 1, \ldots, K\}$ are only uncorrelated across buckets. Further, suppose $B_1 = \{1\}, B_2 = \{2\}, B_3 = \{3\}$ with $j = k = 2$ and $j' = k' = 3$. Then, we have

$$\begin{aligned}\operatorname{cov}\left(\varepsilon_j X_{B_{[k-1]}}, \varepsilon_{j'} X_{B_{[k'-1]}}\right) &= \operatorname{cov}\left(\varepsilon_2 \varepsilon_1, \varepsilon_3(\varepsilon_1, \gamma_{12}\varepsilon_1 + \varepsilon_2)^\top\right) \\ &= \mathbb{E}[\varepsilon_1 \varepsilon_2 (\varepsilon_1 \varepsilon_3, \gamma_{12}\varepsilon_1\varepsilon_3 + \varepsilon_2 \varepsilon_3)^\top],\end{aligned}$$

which may be non-zero.

3.7 Numerical Results

In this section, the finite-sample performance of $\mathcal{G}$-regression is evaluated against contending estimators. We use simulations and an *in silico* dataset for predicting expression levels in gene knockout experiments. All the numerical experiments were conducted with R v3.6, package `pcalg` v2.6 (Kalisch et al., 2012) and our package `eff2` v0.1.

3.7.1 Simulations

We compare the performance of $\mathcal{G}$-regression to several contending estimators under finite samples. We roughly follow the simulation setup of Henckel et al. (2019); Witte et al. (2020).

First, we draw a random undirected graph from the Erdős-Rényi model with average degree k, where k is drawn from $\{2, 3, 4, 5\}$ uniformly at random. The graph is converted to a DAG $\mathcal{D}$ with a random causal ordering and the corresponding CPDAG $\mathcal{G}$ is recorded. Then we fix a linear SEM by drawing γ_{ij} uniformly from $[-2, -0.1] \cup [0.1, 2]$ and choosing the error distribution randomly at random from the following:

1. $\epsilon_i \sim \mathcal{N}(0, v_i)$ with $v_i \sim \mathsf{Unif}(0.5, 6)$,
2. $\epsilon_i / \sqrt{v_i} \sim t_5$ with $v_i \sim \mathsf{Unif}(0.5, 1.5)$,
3. $\epsilon_i \sim \text{logistic}(0, s_i)$ with $s_i \sim \mathsf{Unif}(0.4, 0.7)$,
4. $\epsilon_i \sim \mathsf{Unif}(-a_i, a_i)$ with $a_i \sim \mathsf{Unif}(1.2, 2.1)$.

We generate n iid samples from the model. Treatments A of a fixed size are randomly selected from the set of vertices with non-empty descendants, and Y is selected randomly from their descendants; the drawing is repeated until τ_{AY} is identified from $\mathcal{G}$ according to the criterion of Theorem 3.1. Finally, the data and graph $\mathcal{G}$ are provided to each estimator of τ_{AY}.

We compare to the following three estimators:

- `adj.O`: optimal adjustment estimator (Henckel et al., 2019),
- `IDA.M`: joint-IDA estimator based on modifying Cholesky decompositions (Nandy et al., 2017),
- `IDA.R`: joint-IDA estimator based on recursive regressions (Nandy et al., 2017).

They are implemented in R package `pcalg`. The two joint-IDA estimators use the parents of treatment variables to estimate a causal effect. Both of them reduce to the IDA estimator of Maathuis et al. (2009) when $|A| = 1$. Admittedly, compared to $\mathcal{G}$-regression and `adj.O`, the joint-IDA estimators require less knowledge about the graph, namely only $\text{Pa}(i)$ for each $i \in A$.

For each estimator $\hat{\tau}_{AY}$, we compute its squared error $\|\hat{\tau}_{AY} - \tau_{AY}\|_2^2$. Dividing $\|\hat{\tau}_{AY} - \tau_{AY}\|_2^2$ by the squared error of $\mathcal{G}$-regression, we obtain the *relative squared error* of each contending

estimator. We consider $|A| \in \{1, 2, 3, 4\}$, $|V| \in \{20, 50, 100\}$ and $n \in \{100, 1000\}$; each configuration of $(|A|, |V|, n)$ is replicated 1,000 times.

Fig. 3.2 shows the distributions of relative squared errors. In Table 3.1, we summarize the relative errors with their geometric mean and median. Our estimator dominates all the contending estimators in all cases, and the improvement gets larger as $|A|$ gets bigger. Even though `adj.O` achieves the minimal asymptotic variance among all adjustment estimators, it can compare less favorably to our estimator by several folds. In general, the IDA estimators have very poor performances. Moreover, the results in Table 3.1 are computed only from the replications where a contending estimator exists. As mentioned in the Introduction, unlike $\mathcal{G}$-regression, none of the contending estimators is guaranteed to exist for every identified effect under joint intervention (`adj.O` always exists for point interventions); see Table 3.2 for the percentages of instances that are not estimable by contending estimators, even though the effect is identified by Theorem 3.1 and hence estimable by $\mathcal{G}$-regression.

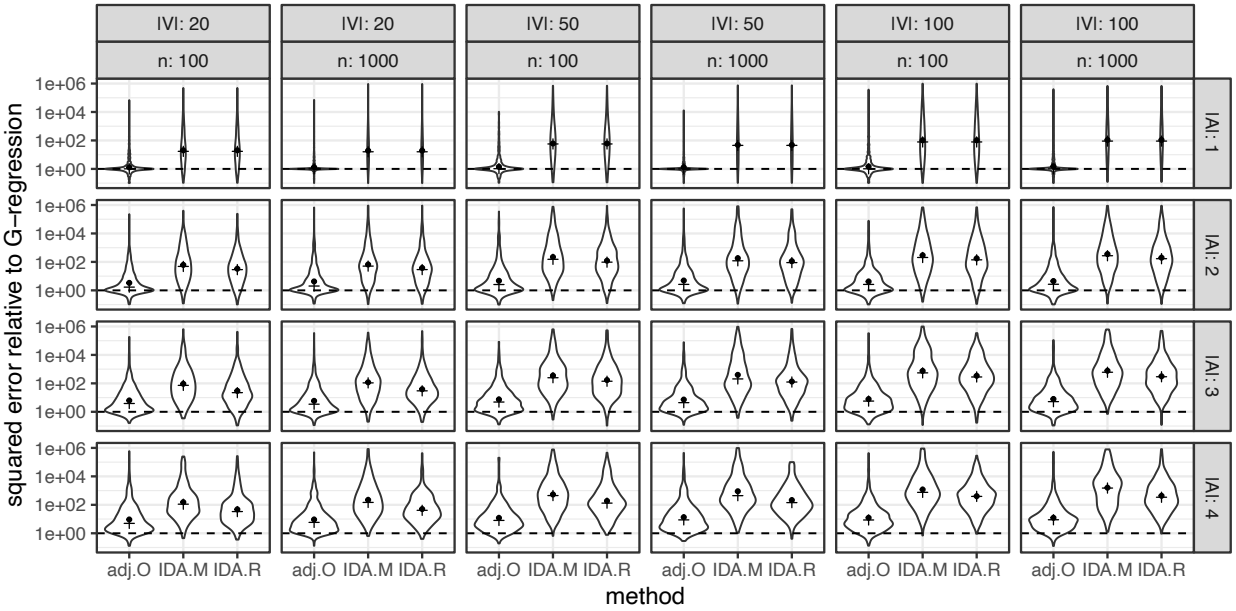


Figure 3.2: Violin plots for the relative squared errors of contending estimators ('·': geometric mean, '+': median).

Table 3.1: Geometric average (brackets: median) of relative squared errors compared to $\mathcal{G}$-regression

	$\|V\| = 20$				$\|V\| = 50$				$\|V\| = 100$			
$\|A\|$	$n = 100$		$n = 1000$		$n = 100$		$n = 1000$		$n = 100$		$n = 1000$	
`adj.O`												
1	1.3	(1.0)	1.3	(1.0)	1.4	(1.0)	1.3	(1.0)	1.5	(1.0)	1.5	(1.0)
2	3.4	(1.7)	4.2	(2.0)	4.7	(2.6)	4.9	(2.6)	4.2	(2.6)	4.5	(2.7)
3	6.3	(3.8)	5.9	(3.4)	7.4	(4.9)	7.2	(4.4)	7.8	(5.7)	8.0	(5.2)
4	9.3	(5.0)	9.3	(5.8)	12	(8.0)	14	(8.7)	12	(8.6)	12	(8.9)
`IDA.M`												
1	20	(17)	19	(16)	61	(57)	48	(45)	103	(78)	108	(90)
2	62	(48)	65	(51)	220	(153)	182	(120)	293	(205)	356	(272)
3	93	(72)	119	(108)	354	(249)	396	(205)	749	(547)	771	(604)
4	154	(111)	222	(147)	533	(448)	895	(440)	1188	(752)	1604	(1508)
`IDA.R`												
1	20	(17)	19	(16)	61	(57)	48	(45)	103	(78)	108	(90)
2	33	(29)	38	(29)	121	(96)	113	(89)	176	(140)	199	(168)
3	30	(22)	39	(30)	171	(141)	135	(125)	342	(281)	312	(292)
4	48	(34)	50	(41)	187	(132)	214	(143)	405	(391)	432	(342)

In Appendix B.3, we report additional simulation results where the CPDAG is estimated with the greedy equivalence search algorithm (Chickering, 2002) and provided to the estimators. The improvements are more modest but are still typically by several folds.

3.7.2 Predicting double knockouts in DREAM4 data

The DREAM4 *in silico* network challenge dataset (Marbach et al., 2009b) provides a benchmark for evaluating the reverse engineering of gene regulation networks. Here we use the 5th *Size10* dataset (Marbach et al., 2009a) as our example, which is a small network of 10 genes. Fig. 3.3 shows the true gene regulation network, which is constructed based on the networks of living organisms. A stochastic differential equation model was used to generate the data under wild type (steady state), perturbed steady state and knockout interventions.

Table 3.2: Percentage of identified instances not estimable using contending estimators (all estimable with $\mathcal{G}$-regression)

Estimator	$\lvert A\rvert$	$\lvert V\rvert = 20$	$\lvert V\rvert = 50$	$\lvert V\rvert = 100$
`adj.O`	1	0%	0%	0%
	2	17%	10%	5%
	3	30%	18%	15%
	4	36%	29%	22%
`IDA.M`	1	29%	32%	32%
	2	47%	51%	50%
	3	61%	59%	63%
	4	72%	69%	71%
`IDA.R`	1	29%	32%	32%
	2	47%	51%	50%
	3	61%	59%	63%
	4	72%	69%	71%

A task in the challenge is to use data under wild type and perturbed steady state (both are observational data) to predict the steady state expression levels under 5 different joint interventions, each of which knocks out a pair of genes. For our purpose, we also use the true network as input. However, the true network contains one cycle (other networks in DREAM4 contain more than one cycles). In the following, we remove one edge in the cycle and provide the resulting DAG to the estimators. Necessarily, the causal DAG is misspecified. Results are reported under 4 different edge removals.

Unfortunately, the wild type data only consists of one sample. To estimate the observational covariance, we use the perturbed steady state data, which consists of 5 segments of

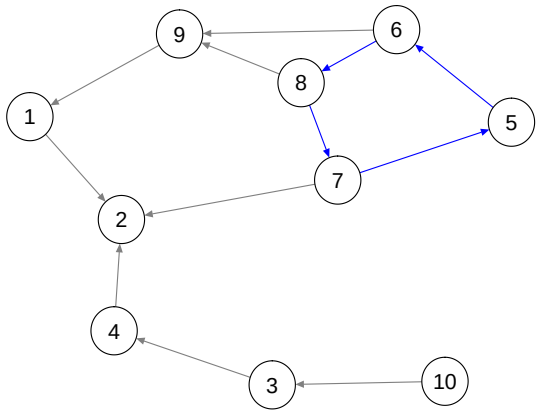


Figure 3.3: Gene regulation network from DREAM4 dataset, which contains a cycle (blue).

time series. A sample covariance is computed from each segment, and the final estimate is taken as their average. For a double knockout of genes (i, j), we use $\mathcal{G}$-regression to estimate the joint-intervention effect of $A = (i, j)$ on every other gene. The effect is identified because the DAG is given. For gene k, let s_k and $s_k^{(ij)}$ respectively denote its expression level under wild type and double knockout of genes (i, j). The expression level under double knockout is predicted as

$$\hat{s}_k^{(ij)} = \begin{cases} s_k - (s_i, s_j)^\top \hat{\tau}_{ij,k}, & k \notin \{i, j\} \\ 0, & k \in \{i, j\} \end{cases}.$$

The performance is evaluated with normalized squared error

$$\mathcal{E} = \frac{\sum_{(i,j)\in\mathcal{A}} \sum_{k=1}^{10} (\hat{s}_k^{(ij)} - s_k^{(ij)})^2}{\sum_{(i,j)\in\mathcal{A}} \sum_{k=1}^{10} (s_k^{(ij)})^2},$$

where $\mathcal{A} = \{(6, 8), (7, 8), (8, 10), (8, 5), (8, 9)\}$ consists of 5 double knockouts available in the dataset. For comparison, we also evaluate the performance of `adj.O` (optimal adjustment, Henckel et al. (2019)) and `IDA.R` (joint-IDA based on recursive regressions, Nandy et al. (2017)); `IDA.R` is chosen because it outperforms `IDA.M` according to Section 3.7.1. Unfortunately, `adj.O` is not able to estimate the effect on every k and a modified metric $\mathcal{E}^*$ is

computed by only summing over those estimable k's; the same metric $\mathcal{E}^*$ of $\mathcal{G}$-regression is also computed for comparison. As a baseline, we also compute $\mathcal{E}$ from naively estimating $s_k^{(ij)}$ with just s_k.

Table 3.3: Normalized squared errors of predicting gene double knockouts

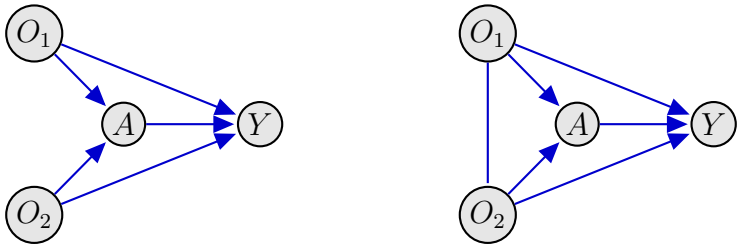
edge removed		$\mathcal{E}^*$		$\mathcal{E}$		
from cycle	$\nexists$ `adj.O`	`adj.O`	$\mathcal{G}$-`reg`	`IDA.R`	$\mathcal{G}$-`reg`	`baseline`
$5 \to 6$	36%	43%	**35%**	46%	**30%**	81%
$6 \to 8$	42%	**29%**	32%	33%	**26%**	81%
$8 \to 7$	60%	39%	**35%**	45%	**44%**	81%
$7 \to 5$	46%	40%	**33%**	45%	**34%**	81%

Table 3.3 reports the results, where the column '$\nexists$ `adj.O`' lists the percentage of effects not estimable by the adjustment estimator. In almost all the cases, $\mathcal{G}$-regression dominates all the contending estimators. In this example, even though both the causal graph and the linear SEM are misspecified, one can still witness some usefulness of our estimator.

3.8 Discussion

We have proposed $\mathcal{G}$-regression based on recursive least squares to estimate a total causal effect from observational data, under linearity and causal sufficiency assumuptions. $\mathcal{G}$-regression is applicable to estimating every identified total effect, under either point intervention or joint intervention. Further, via a new semiparametric efficiency theory, we have shown that the estimator achieves the efficiency bound within a restricted, yet reasonably large, class of estimators, including covariate adjustment and other regular estimators based on the sample covariance. Note that the restriction on the class of estimators is motivated by computational simplicity and numerical stability as mentioned in the Introduction. To

construct confidence intervals and conduct hypothesis tests, bootstrap can be easily applied to estimate the asymptotic covariance of our estimator. This is implemented in R package `eff2`.



(a) (b)

Figure 3.4: (a) and (b) lead to the same $\mathcal{G}$-regression estimator of τ_{AY}. The independence between O_1 and O_2 in (a) is dropped in (b).

We conclude this Chapter with a remark. We have seen that conditional independence constraints in Eq. (3.12) play no role for the restricted class of estimators considered — a feature that holds under the restrictive property (Corollary 3.1). For example, the marginal independence between O_1 and O_2 in Fig. 3.4(a) can be ignored without changing the $\mathcal{G}$-regression estimator of τ_{AY}. However, this is no longer true when linearity is dropped. For nonparametric causal graphical models, the asymptotic relative efficiency resulting from ignoring $O_1 \perp\!\!\!\perp O_2$ can be arbitrarily large; see Rotnitzky and Smucler (2020, Lemma 23).

Chapter 4

NON-ASYMPTOTIC BOUND OF MULTINOMIAL LIKELIHOOD RATIO

In this Chapter, we develop non-asymptotic tail bounds for the likelihood ratio statistic (for testing a simple null) under multinomial sampling with n samples and k categories. Data under this type of sampling is typically represented as a contingency table (Lauritzen, 1996, Chap. 4). Ignoring the usual factor of 2 and dividing by the sample size n, the statistic is also the relative entropy (Kullback-Leibler divergence) of the empirical probability vector with respect to the true probability vector.

Background and motivation are provided in Section 4.1. In Section 4.2, we generalize the technique of Agrawal (2020) and show that the moment generating function of the statistic is bounded by a polynomial of degree n on the unit interval, uniformly over all true probability vectors. We characterize the family of polynomials indexed by (k, n) and obtain explicit formulae. Asymptotic properties of these polynomials are studied. In Section 4.3, we present the resulting Chernoff bound and a closed-form approximation under large n. In Section 4.4, we show that our bound dominates the classic method-of-types bound and is competitive with the state of the art. Finally, in Section 4.5, we showcase a statistical application in estimating the proportion of unseen butterflies in Malay Peninsula. The bound defines a convex confidence region for the sampling probability vector, which will be used for statistical inference of discrete instrumental variable models in the next Chapter. Additional proofs can be found in Appendix C.

4.1 Introduction

Consider a multinomial experiment on an alphabet of size $k \geq 2$

$$(X_1, \ldots, X_k) \sim \mathsf{Mult}(n; (p_1, \ldots, p_k)), \tag{4.1}$$

where $(p_1, \ldots, p_k)$ belongs to the unit simplex Δ^{k-1}. The empirical measure is identified with the probability vector $(\hat{p}_1, \ldots, \hat{p}_k) = (X_1/n, \ldots, X_k/n)$. We are interested in its entropy relative to the true probability vector p, namely

$$\mathcal{D}(\hat{p}\|p) = \sum_{i=1}^{k} \hat{p}_i \log(\hat{p}_i/p_i), \tag{4.2}$$

where conventions $0 \cdot \log(0) = 0$ and $0 \cdot \log(0/0) = 0$ are adopted. The quantity $\mathcal{D}(\hat{p}\|p)$ is also known as the Kullback-Leibler divergence of p from $\hat{p}$. By the law of large numbers, $\mathcal{D}(\hat{p}\|p) \to 0$ as $n \to \infty$ almost surely.

Note that

$$n\,\mathcal{D}(\hat{p}\|p) = \sum_{i=1}^{k} X_i \log \frac{\hat{p}_i}{p_i} = \log \frac{\binom{n}{X_1,\ldots,X_k} \prod_{i=1}^{k} \hat{p}_i^{X_i}}{\binom{n}{X_1,\ldots,X_k} \prod_{i=1}^{k} p_i^{X_i}}$$

is also the log-likelihood ratio statistic (without the usual extra factor of 2). By standard asymptotic arguments (see, e.g., van der Vaart (2000, Example 16.1)), for fixed k and $n \to \infty$, it holds that

$$n\,\mathcal{D}(\hat{p}\|p) \;\to_d\; \chi^2_{k-1}/2 =_d \mathsf{Ga}((k-1)/2, 1), \tag{4.3}$$

which is a gamma distribution with shape $(k-1)/2$ and rate one.

Motivation We are interested in upper bounding the probability that $n\,\mathcal{D}(\hat{p}\|p)$ exceeds a given threshold. Tail bounds of this type are of interest to many problems in probability, statistics and machine learning, including Sanov's theorem in large deviations (Cover and Thomas, 2006, §11.4), goodness-of-fit tests (Cressie and Read, 1984; Jager and Wellner, 2007), construction of non-asymptotic confidence regions (Chafai and Concordet, 2009; Malloy et al., 2020) and the performance guarantees of various learning algorithms (Vinayak et al., 2019; Nowak and Tánczos, 2019).

The classic bound of this type is

$$P\left(n\,\mathcal{D}(\hat{p}\|p) > t\right) \leq \exp(-t)\binom{n+k-1}{k-1}, \quad (t > 0) \tag{4.4}$$

obtained by the "method of types" (Csiszár, 1998, Lemma II.1). For fixed k and t, this bound is asymptotically tight as $n \to \infty$, in the sense that the exponent $\exp(-t)$ matches the rate of the asymptotic gamma distribution in Eq. (4.3). Nevertheless, the bound above is far from optimal. There are recent developments in the literature that provide sharper results. In particular, Mardia et al. (2019) and Agrawal (2020) provide significant improvements over the method-of-types result by gaining tighter control for the binomial case ($k = 2$), and a reduction from multinomial ($k > 2$) to binomial, although their approaches are different. Additionally, bounds on the moments of $\mathcal{D}(\hat{p}\|p)$ have been studied; see Jiao et al. (2017); Mardia et al. (2019); Paninski (2003).

On a side note, by Pinsker's inequality, a tail bound on relative entropy implies a bound on the total variation. For bounds on the latter, see also van der Vaart and Wellner (1996, Appendix A.6) and Devroye (1983); Biau and Gyorfi (2005).

4.2 Bounding the moment generating function

In a vein similar to that of Agrawal (2020), we develop bounds with Chernoff's method, a classic workhorse for deriving exponential tail bounds; see, e.g., Vershynin (2018, §2.3). The key is to upper bound the moment generating function (MGF) of $n\,\mathcal{D}(\hat{p}\|p)$, which is defined as

$$\varphi_{k,n}(\lambda, p) := \mathbb{E}\exp\left(\lambda n\,\mathcal{D}(\hat{p}\|p)\right), \tag{4.5}$$

where the expectation is taken over $\mathsf{Mult}(n, p = (p_1, \ldots, p_k))$.

It follows that

$$\begin{aligned}
\varphi_{k,n}(\lambda, p) &= \sum_{X_1,\ldots,X_k} \binom{n}{X_1,\ldots,X_k} \prod_{i=1}^{k} p_i^{X_i} \left\{ \frac{\binom{n}{X_1,\ldots,X_k} \prod_{i=1}^{k} \hat{p}_i^{X_i}}{\binom{n}{X_1,\ldots,X_k} \prod_{i=1}^{k} p_i^{X_i}} \right\}^{\lambda} \\
&= \sum_{X_1,\ldots,X_k} \binom{n}{X_1,\ldots,X_k} \left\{ \prod_{i=1}^{k} \hat{p}_i^{X_i} \right\}^{\lambda} \left\{ \prod_{i=1}^{k} p_i^{X_i} \right\}^{1-\lambda},
\end{aligned} \tag{4.6}$$

where $X_1, \ldots, X_k$ are non-negative integers that sum to n.

Definition 4.1. *For $k \geq 1$, $n \geq 1$, $p \in \Delta^{k-1}$ and $\lambda \in [0,1]$, define*

$$G_{k,n}(\lambda, p) := \sum_{X_1,\ldots,X_k} \binom{n}{X_1, \ldots, X_k} \prod_{i=1}^{k} [\lambda X_i/n + (1-\lambda)p_i]^{X_i}, \tag{4.7}$$

where the summation is over non-negative integers that sum to n.

By definition, $G_{k,n}(\lambda, p)$ is a polynomial in λ of degree at most n. For the trivial case of $k = 1$, it is easy to see that $G_{1,n}(\lambda) \equiv 1$.

The multinomial probability in Eq. (4.6) is log-concave in $(p_1, \ldots, p_k)$. For $0 \leq \lambda \leq 1$, by Jensen's inequality, we have

$$\varphi_{k,n}(\lambda, p) \leq G_{k,n}(\lambda, p), \quad p \in \Delta^{k-1}.$$

The obvious obstacle here is to obtain a bound on the RHS that does not depend on the true probability vector p.

4.2.1 Family of $G_{k,n}(\lambda)$

First comes a surprising fact noticed by Agrawal (2020) in the $k = 2$ case.

Proposition 4.1. *$G_{k,n}(\lambda, p)$ does not depend on $p = (p_1, \ldots, p_k)$.*

Proof. This is true for $k = 1$. Fix any $k \geq 2$, we prove by induction on n that $G_{k,n}(\lambda, p)$ does not depend on p. For the base case,

$$G_{k,1}(\lambda, p) = \sum_{i=1}^{k} (\lambda + (1-\lambda)p_i) = k\lambda + 1 - \lambda,$$

which does not depend on p.

Suppose $G_{k,m}(\lambda, p) \equiv G_{k,m}(\lambda)$ for $m \leq n-1$. We now show that $G_{k,n}(\lambda, p)$ does not depend on p. Since $p_k = 1 - p_1 - \cdots - p_{k-1}$, it suffices to verify that $\partial G_{k,n}(\lambda, p)/\partial p_i \equiv 0$ for

$i = 1, \ldots, k-1$. Further, by symmetry, it suffices to show $\partial G_{k,n}(\lambda, p)/\partial p_1 \equiv 0$. Replacing p_k with $(1 - p_1 - \cdots - p_{k-1})$, we have

$$G_{k,n}(\lambda, p) = \sum_{X_1,\ldots,X_k} \binom{n}{X_1, \ldots, X_k} \prod_{j=2}^{k-1} \left[\lambda X_j/n + (1-\lambda)p_j\right]^{X_j} \times \left[\lambda X_1/n + (1-\lambda)p_1\right]^{X_1} \left[\lambda X_k/n + (1-\lambda)(1 - p_1 - \cdots - p_{k-1})\right]^{X_k},$$

and

$$\frac{\partial G_{k,n}(\lambda, p)}{\partial p_1} = \sum_{X_1,\ldots,X_k} \binom{n}{X_1, \ldots, X_k} \prod_{j=2}^{k-1} \left[\lambda X_j/n + (1-\lambda)p_j\right]^{X_j} \left\{(1-\lambda)X_1 \left[\lambda X_1/n + (1-\lambda)p_1\right]^{X_1-1} \left[\lambda X_k/n + (1-\lambda)p_k\right]^{X_k} - (1-\lambda)X_k \left[\lambda X_1/n + (1-\lambda)p_1\right]^{X_1} \left[\lambda X_k/n + (1-\lambda)p_k\right]^{X_k-1}\right\}.$$

Hence, it suffices to show

$$\sum_{X_1,\ldots,X_k} \binom{n}{X_1, \ldots, X_k} \prod_{j=2}^{k-1} \left[\lambda X_j/n + (1-\lambda)p_j\right]^{X_j} \times X_1 \left[\lambda X_1/n + (1-\lambda)p_1\right]^{X_1-1} \left[\lambda X_k/n + (1-\lambda)p_k\right]^{X_k} \equiv \sum_{X_1,\ldots,X_k} \binom{n}{X_1, \ldots, X_k} \times \prod_{j=2}^{k-1} \left[\lambda X_j/n + (1-\lambda)p_j\right]^{X_j} X_k \left[\lambda X_1/n + (1-\lambda)p_1\right]^{X_1} \left[\lambda X_k/n + (1-\lambda)p_k\right]^{X_k-1}.$$

We first simplify the LHS. Clearly, those summands with $X_1 = 0$ are zero and can be dropped. For $X_1 \geq 1$, $X_1\binom{n}{X_1,\ldots,X_k} = n\binom{n-1}{X_1-1,X_2,\ldots,X_k}$. Let $\lambda' := \lambda(n-1)/n$. For $j = 2, \ldots, k$, by setting $p'_j := \frac{1-\lambda}{1-\lambda'} p_j < p_j$, we have

$$\lambda X_j/n + (1-\lambda)p_j = \lambda' X_j/(n-1) + (1-\lambda')p'_j.$$

Further, letting $p'_1 := 1 - \sum_{j=2}^{k} p'_j$ it is easy to see that

$$\lambda' \frac{X_1 - 1}{n-1} + (1-\lambda')p'_1 = \lambda \frac{X_1}{n} + (1-\lambda)p_1.$$

Therefore, by introducing $X_1' = X_1 - 1$, we have

$$\begin{aligned}\text{LHS} &= n \sum_{X_1', X_2, \ldots, X_k} \binom{n-1}{X_1', X_2, \ldots, X_k} \left[\lambda' X_1'/(n-1) + (1-\lambda')p_1'\right]^{X_1'} \\ &\quad \times \prod_{j=2}^{k} \left[\lambda' X_j/(n-1) + (1-\lambda')p_j'\right]^{X_j} \\ &= nG_{k,n-1}(\lambda', p'),\end{aligned}$$

where the summation is over non-negative integers $X_1', X_2, \ldots, X_k$ summing to $n-1$. For the RHS, similarly, let $q_j' = \frac{1-\lambda}{1-\lambda'} p_j$ for $j = 1, \ldots, k-1$ and $q_k' = 1 - \sum_{j=1}^{k-1} q_j'$. With $X_k' = X_k - 1$, it follows that

$$\begin{aligned}\text{RHS} &= n \sum_{X_1, \ldots, X_{k-1}, X_k'} \binom{n-1}{X_1, \ldots, X_{k-1}, X_k'} \prod_{j=1}^{k-1} \left[\lambda' X_j/(n-1) + (1-\lambda')q_j'\right]^{X_j} \\ &\quad \times \left[\lambda' X_k'/(n-1) + (1-\lambda')q_k'\right]^{X_k'} \\ &= nG_{k,n-1}(\lambda', q').\end{aligned}$$

Finally, by the induction hypothesis,

$$\text{LHS} = nG_{k,n-1}(\lambda', p') = nG_{k,n-1}(\lambda', q') = \text{RHS}.$$

□

In view of this fact, we shall write $G_{k,n}(\lambda)$ in place of $G_{k,n}(\lambda, p)$. The set of polynomials $\{G_{k,n}(\lambda)\}$ are characterized by the following recurrence.

Proposition 4.2. *For $0 \le \lambda \le 1$, it holds that*

$$G_{k,n}(\lambda) = G_{k-1,n}(\lambda) + \lambda G_{k,n-1}\left(\frac{n-1}{n}\lambda\right), \quad k \ge 2,\ n \ge 1 \tag{4.8}$$

with $G_{1,n}(\lambda) \equiv 1$ and $G_{k,0}(\lambda) := 1$.

By Proposition 4.1, we have the freedom to choose p in the definition to evaluate $G_{k,n}(\lambda)$. In particular, by choosing $p_k = 0$ and $p_1 + \cdots + p_{k-1} = 1$, we can decompose $G_{k,n}(\lambda)$ into $G_{k-1,n}(\lambda)$ and a remainder. By a similar manipulation used in the previous proof, the remainder can be expressed in terms of $G_{k,n-1}$. We leave the detailed proof to the Appendix.

Theorem 4.1. *For $k \geq 2$, $n \geq 0$ and $0 \leq \lambda \leq 1$, it holds that*

$$G_{k,n}(\lambda) = \sum_{m=0}^{n} \frac{n!}{n^m (n-m)!} \binom{m+k-2}{k-2} \lambda^m. \tag{4.9}$$

Proof. We prove this by induction. For the base case, the formula gives $G_{k,0}(\lambda) \equiv 1$ for $k \geq 2$, which matches the value imposed by Proposition 4.2.

First, supposing the formula holds for $G_{2,n-1}$, we show that it also holds for $G_{2,n}$. By Proposition 4.2, it is easy to check that

$$\begin{aligned} G_{2,n}(\lambda) &= G_{1,n}(\lambda) + \lambda G_{2,n-1}(\lambda(n-1)/n) \\ &= 1 + \sum_{m=0}^{n-1} \frac{(n-1)!}{n^m (n-m-1)!} \lambda^{m+1} = \sum_{m=0}^{n} \frac{n!}{n^m (n-m)!} \lambda^m. \end{aligned}$$

Now, for any $k \geq 3$ and $n \geq 1$, suppose the formula holds for $G_{k-1,n}$ and $G_{k,n-1}$. We show that it also holds for $G_{k,n}$. By Proposition 4.2, we have

$$\begin{aligned} G_{k,n}(\lambda) &= G_{k-1,n}(\lambda) + \lambda G_{k,n-1}(\lambda(n-1)/n) \\ &= \sum_{m=0}^{n} \frac{n!}{n^m (n-m)!} \binom{m+k-3}{k-3} \lambda^m + \sum_{m=0}^{n-1} \frac{(n-1)!}{n^m (n-m-1)!} \binom{m+k-2}{k-2} \lambda^{m+1} \\ &= \sum_{m=0}^{n} \frac{n!}{n^m (n-m)!} \binom{m+k-3}{k-3} \lambda^m + \sum_{m=1}^{n} \frac{n!}{n^m (n-m)!} \binom{m+k-3}{k-2} \lambda^m \\ &= \sum_{m=0}^{n} \frac{n!}{n^m (n-m)!} \binom{m+k-2}{k-2} \lambda^m, \end{aligned}$$

where in the last step the addition formula $\binom{n}{l} = \binom{n-1}{l} + \binom{n-1}{l-1}$ is used (Graham et al., 1994, §5.1). □

Remark 4.1. For $k \geq 2$, $G_{k,n}(\lambda)$ is not a moment generating function of some distribution. Suppose $G_{k,n}(\lambda)$ is the MGF of random variable Y. Since $G_{k,n}(\lambda)$ is a polynomial of degree n, then $\mathbb{E}\, Y^{2n} = G_{k,n}^{(2n)}(0) = 0$, which implies Y is zero almost surely. However, the MGF of zero is identically one.

A few polynomials $G_{k,n}(\lambda)$ are listed in Table 4.1.

Table 4.1: Polynomials $G_{k,n}(\lambda)$

	$n=1$	$n=2$	$n=3$	$n=4$
$k=2$	$1+\lambda$	$1+\lambda+\frac{1}{2}\lambda^2$	$1+\lambda+\frac{2}{3}\lambda^2+\frac{2}{9}\lambda^3$	$1+\lambda+\frac{3}{4}\lambda^2+\frac{3}{8}\lambda^3+\frac{3}{32}\lambda^4$
$k=3$	$1+2\lambda$	$1+2\lambda+\frac{3}{2}\lambda^2$	$1+2\lambda+2\lambda^2+\frac{8}{9}\lambda^3$	$1+2\lambda+\frac{9}{4}\lambda^2+\frac{3}{2}\lambda^3+\frac{15}{32}\lambda^4$
$k=4$	$1+3\lambda$	$1+3\lambda+3\lambda^2$	$1+3\lambda+4\lambda^2+\frac{20}{9}\lambda^3$	$1+3\lambda+\frac{9}{2}\lambda^2+\frac{15}{4}\lambda^3+\frac{45}{32}\lambda^4$

4.2.2 Asymptotic properties

We consider the asymptotic behaviors of $G_{k,n}(\lambda)$, which can inform how well it captures the right dependence on k and n.

$n \to \infty$ under fixed k

Lemma 4.1. *For $k \geq 2$, $G_{k,n}(\lambda)$ increases in n.*

Proof. By Theorem 4.1, it suffices to show that

$$\frac{n!}{n^m(n-m)!} \geq \frac{(n-1)!}{(n-1)^m(n-m-1)!}$$

for $m = 0, \ldots, n$. By canceling factors from both sides, this is equivalent to $(1-\frac{1}{n})^m \geq 1-\frac{m}{n}$, which holds by Bernoulli's inequality. □

Proposition 4.3. *For $0 \leq \lambda < 1$ and any fixed $k \geq 2$, we have*

$$G_{k,n}(\lambda) \nearrow G_{k,\infty}(\lambda) := (1-\lambda)^{-(k-1)}, \quad \text{as } n \to \infty. \tag{4.10}$$

Proof. For $k = 2$ and $\lambda \in [0, 1)$,

$$G_{2,n}(\lambda) = \sum_{m=0}^{n} \frac{n!}{n^m(n-m)!}\lambda^m \leq \sum_{m=0}^{n} \lambda^m \to \frac{1}{1-\lambda},$$

where we used

$$\frac{n!}{n^m(n-m)!} = \frac{n \times (n-1) \times \ldots (n-m+1)}{n \times \cdots \times n} \leq 1.$$

Further, by Lemma 4.1, $G_{2,n}(\lambda)$ must converge as $n \to \infty$ for $\lambda \in [0,1)$. Suppose the limit is $G_{2,\infty}(\lambda)$. Clearly, $G_{2,\infty}(\lambda) = \lim_n G_{2,n}(\lambda) = \sup_n G_{2,n}(\lambda)$ is lower-semicontinuous. Taking limits on both sides of Eq. (4.8), we have

$$G_{2,\infty}(\lambda) = 1 + \lambda G_{2,\infty}(\lambda^-),$$

where we note $\frac{n-1}{n}\lambda \nearrow \lambda$. Meanwhile, by Theorem 4.1, $G_{2,n}(\lambda)$ is increasing in λ. Hence, we have $G_{2,\infty}(\lambda^-) = G_{2,\infty}(\lambda)$ by lower-semicontinuity and monotonicity of $G_{2,\infty}(\lambda)$. It follows that $G_{2,\infty} = (1-\lambda)^{-1}$. Applying the same reasoning to $k = 3$, we have

$$G_{3,\infty}(\lambda) = G_{2,\infty}(\lambda) + \lambda G_{3,\infty}(\lambda),$$

and hence $G_{3,\infty} = (1-\lambda)^{-2}$. Iterating this process, we get $G_{k,n}(\lambda) \nearrow (1-\lambda)^{-(k-1)}$ for $\lambda \in [0,1)$ and $k \geq 2$. □

Note that $G_{k,\infty}(\lambda) = (1-\lambda)^{-(k-1)}$ is the moment generating function of $\mathsf{Ga}(k-1,1)$. Further, $n\,\mathcal{D}(\hat{p}\|p) \to_d \mathsf{Ga}((k-1)/2, 1)$. This means, for fixed k and $n \to \infty$, $G_{k,n}(\lambda)$ is asymptotically tight in the exponent (rate parameter of gamma), but loose by a factor of 2 in the polynomial term (shape parameter of gamma).

$k \to \infty$ under fixed n

Proposition 4.4. *For fixed $0 < \lambda \leq 1$ and $n \geq 1$, as $k \to \infty$ we have*

$$\log G_{k,n}(\lambda) \asymp n \log k. \tag{4.11}$$

Proof. By Theorem 4.1, for fixed n and λ, the diverging term should be the largest term of $\{\binom{m+k-2}{k-2} : 0 \leq m \leq n\}$, which is when $m = n$. And $\log \binom{n+k-2}{k-2} \asymp n \log k$. □

The following shows that, as $k \to \infty$, the logarithmic dependence on k for an upper bound on the logarithm of MGF is also necessary.

Proposition 4.5. *Suppose $H_{k,n}(\lambda) \geq \varphi_{k,n}(\lambda; p)$ for all p and all $\lambda \in (0,1)$. For fixed $0 < \lambda \leq 1$ and $n \geq 1$, we have lower bound $\log H_{k,n}(\lambda) \gtrsim \lambda n \log k$ as $k \to \infty$.*

Proof. Let $p = (1/k, \ldots, 1/k)$. It follows from Eq. (4.6) that

$$\varphi_{k,n}(\lambda, p) = n^{-\lambda n} \sum_{X_1,\ldots,X_k} \binom{n}{X_1, \ldots, X_k} \prod_{i=1}^{k} \left(\frac{X_i^\lambda}{k^{1-\lambda}}\right)^{X_i}.$$

We claim that $\varphi_{k,n}(\lambda, p) \asymp k^{\lambda n}$. Consider the configurations of $(X_1, \ldots, X_k)$ such that n of them are one and the rest are zero. As $k \to \infty$, ignoring the factors that do not depend on k, the sum over these configurations becomes

$$n^{-\lambda n} \binom{k}{n} \left(\frac{1}{k^{1-\lambda}}\right)^n \asymp k^{\lambda n}.$$

We now show that the sum from all the other configurations is $O(k^{\lambda n - 1})$. Consider the contribution from those configurations with m non-zero categories. Their sum is

$$n^{-\lambda n} \binom{k}{m} \sum_{Y_1,\ldots,Y_m} \binom{n}{Y_1, \ldots, Y_m} \prod_{i=1}^{m} \left(\frac{Y_i^\lambda}{k^{1-\lambda}}\right)^{Y_i} =: \binom{k}{m} k^{-n(1-\lambda)} C_{n,m}(\lambda)$$

where $Y_1, \ldots, Y_m$ are positive integers that sum to n. It follows that

$$\sum_{m=1}^{n-1} \binom{k}{m} k^{-n(1-\lambda)} C_{n,m}(\lambda) < \left(\max_{1 \le m \le n-1} C_{n,m}(\lambda)\right) k^{-n(1-\lambda)} \sum_{m=1}^{n-1} k^m \asymp k^{\lambda n - 1}.$$

Hence, $\log H_{k,n}(\lambda) \gtrsim \lambda n \log k$. □

Remark 4.2. Agrawal (2020) uses the upper bound $G_{2,\infty}(\lambda)$ on $G_{2,n}(\lambda)$ to further bound $G_{k,n}(\lambda)$ for $k > 2$, by appealing to the chain rule of relative entropy (Cover and Thomas, 2006, §2.5). This leads to the following bound:

$$\varphi_{k,n}(\lambda) \le (1-\lambda)^{-(k-1)} = G_{k,\infty}(\lambda) \quad (0 \le \lambda < 1). \tag{4.12}$$

However, observe that for fixed n and large k, the logarithm of the above bound above grows *linearly* in k. In contrast, as we have shown via a direct approach, the bound $\log G_{k,n}(\lambda)$ has the right logarithmic dependence.

4.3 Chernoff bound

To highlight the dependence on (k, n), let $\hat{p}_{k,n}$ denote the empirical probability vector under k categories and n samples. For any $\lambda \in [0, 1]$, we have

$$P\left(n\,\mathcal{D}(\hat{p}_{k,n}\|p) > t\right) \leq \exp(-\lambda t) G_{k,n}(\lambda). \tag{4.13}$$

Minimizing over $\lambda \in [0, 1]$ yields the tightest bound.

Theorem 4.2. *For $k \geq 2$, $n \geq 1$, let $\hat{p}_{k,n}$ be the empirical probability vector from $\mathsf{Mult}(p, n)$ for $p \in \Delta^{k-1}$. For $t > 0$, it holds that*

$$P\left(n\,\mathcal{D}(\hat{p}_{k,n}\|p) > t\right) \leq \min_{\lambda \in [0,1]} \exp(-\lambda t) G_{k,n}(\lambda). \tag{4.14}$$

Proposition 4.6. *The bound in Theorem 4.2 is meaningful (RHS < 1) if $t > \min(\log G_{k,n}(1), k - 1)$.*

Proof. Let $f_{k,n}(\lambda, t) := \exp(-\lambda t) G_{k,n}(\lambda)$. Let $\psi_{k,n}(t) := \min_{\lambda \in [0,1]} f(\lambda, t)$ be the RHS of Eq. (4.14). First, suppose $t > \min(\log G_{k,n}(1), k - 1)$ and we show that $\psi_{k,n}(t) < 1$. Clearly, either $t > \log G_{k,n}(1)$ or $t > k - 1$. If $t > \log G_{k,n}(1)$, then $\psi_{k,n}(t) \leq f_{k,n}(1, t) = \exp(-t) G_{k,n}(1) < 1$. If $t > k - 1$, $\psi_{k,n}(t) \leq \psi_{k,\infty}(t)$ by Proposition 4.3. One can show that

$$\psi_{k,\infty}(t) = \begin{cases} 0, & t \leq k - 1 \\ \exp(k - 1 - t) \left(\frac{t}{k-1}\right)^{k-1}, & t > k - 1 \end{cases}.$$

Writing $t = k - 1 + \delta$ for $\delta > 0$, it follows that

$$\psi_{k,n}(t) \leq \psi_{k,\infty}(k - 1 + \delta) = \exp\left\{(k - 1)\log(1 + \frac{\delta}{k - 1}) - \delta\right\} < 1$$

by $\log(1 + x) < x$ for $x > 0$. □

We have verified that the converse also holds at least for $k \leq 500$.

Let $\lambda_{k,n}(t)$ be the minimizer in Theorem 4.2. Unfortunately, in general, $\lambda_{k,n}(t)$ does not permit a closed-form solution. In fact, finding $\lambda_{k,n}(t)$ is a non-convex problem and

$\exp(-\lambda t)G_{k,n}(\lambda)$ can have more than one local minima on the unit interval. In the following, we develop a simple closed-form approximation to $\lambda_{k,n}(t)$ that leads to a bound that is only slightly looser than Theorem 4.2, when n is relatively big compared to k.

Large n expansion of the minimizer By Proposition 4.3, when $n \to \infty$ we have

$$\exp(-\lambda t)G_{k,n}(\lambda) \to \exp(-\lambda t)(1-\lambda)^{-(k-1)} = e^{-\lambda t-(k-1)\log(1-\lambda)}.$$

Note that $\lambda \mapsto -\lambda t-(k-1)\log(1-\lambda)$ is convex. The previous display is uniquely minimized at

$$\lambda_{k,\infty}(t) = 1 - \frac{k-1}{t}, \quad \text{for } t > k-1. \tag{4.15}$$

Plugging in $\lambda_{k,\infty}(t)$ into Eq. (4.13) yields the following bound.

Corollary 4.1 (without correction)**.** *For $t > k-1$, it holds that*

$$P\left(n\,\mathcal{D}(\hat{p}_{k,n}\|p) > t\right) \le e^{-t}e^{k-1}G_{k,n}\left(1-\frac{k-1}{t}\right). \tag{4.16}$$

$\lambda_{k,\infty}(t)$ is the zeroth-order large n approximation to $\lambda_{k,n}(t)$. Yet, the bound can be significantly tightened by a further correction.

Proposition 4.7. *Suppose $k \ge 2$ and $t > k-1$. As $n \to \infty$, we have*

$$\lambda_{k,n}(t) = \lambda_{k,\infty}(t) + \frac{k}{k-1}\frac{t-k+1}{n} + o(n^{-1}). \tag{4.17}$$

Proof. Fix $k \ge 2$ and $t > k-1$. Let $f_{k,n} := \exp(-\lambda t)G_{k,n}(\lambda)$. First, we claim that there exists $N(k,t)$ such that $f'_{k,n}(\lambda_{k,n}) = 0$ for $n \ge N(k,t)$ at the minimizer $\lambda_{k,n}$. To see this, note that asymptotically $\lambda_{k,n}$ cannot be 0 or 1. In particular, (i) $\lambda_{k,n} = 0$ would imply RHS $= 1$ for Eq. (4.14), and (ii) $\lambda_{k,n} \to 1$ would imply RHS $\to \infty$ for Eq. (4.14) — both contradict Proposition 4.6. Given

$$f'_{k,n}(\lambda_{k,n}) = f'_{k,n}(\lambda_{k,\infty}) + f''_{k,n}(\lambda_{k,\infty})(\lambda_{k,n}-\lambda_{k,\infty}) + o(|\lambda_{k,n}-\lambda_{k,\infty}|),$$

it follows that

$$\lambda_{k,n} = \lambda_{k,\infty} - \frac{f'_{k,n}(\lambda_{k,\infty})}{f''_{k,n}(\lambda_{k,\infty})} + o(|\lambda_{k,n}-\lambda_{k,\infty}|).$$

Since $f_{k,n} \to f_{k,\infty} = \exp(-\lambda t) G_{k,\infty}(\lambda)$, it is easy to check that

$$\begin{aligned} f''_{k,n}(\lambda_{k,\infty}) &= f''_{k,\infty}(\lambda_{k,\infty}) + o(1) \\ &= (k-1) e^{-\lambda_{k,\infty} t} (1 - \lambda_{k,\infty})^{-(k+1)} + o(1), \end{aligned}$$

where the limit $(k-1) e^{-\lambda_{k,\infty} t} (1 - \lambda_{k,\infty})^{-(k+1)}$ is non-zero and finite. Meanwhile, we have

$$f'_{k,n}(\lambda_{k,\infty}) = e^{-\lambda_{k,\infty} t} \left(G'_{k,n}(\lambda_{k,\infty}) - t G_{k,n}(\lambda_{k,\infty}) \right).$$

Using $\lambda_{k,\infty} = 1 - (k-1)/t$, it follows that

$$\lambda_{k,n} = \lambda_{k,\infty} + \frac{e^{-(t-k+1)} \left[\frac{k-1}{1-\lambda_{k,\infty}} G_{k,n}(\lambda_{k,\infty}) - G'_{k,n}(\lambda_{k,\infty}) \right]}{(k-1) e^{-(t-k+1)} \left(\frac{k-1}{t} \right)^{-(k+1)} + o(1)} + o(|\lambda_{k,n} - \lambda_{k,\infty}|).$$

It is easy to check that the proof is complete given the following lemma. □

Lemma 4.2. *For $k \geq 2$ and $\lambda \in (0,1)$, it holds that*

$$n \left(\frac{k-1}{1-\lambda} G_{k,n}(\lambda) - G'_{k,n}(\lambda) \right) \to \frac{k(k-1)\lambda}{(1-\lambda)^{k+2}}, \quad \text{as } n \to \infty. \tag{4.18}$$

The proof relies on asymptotic expansions of the incomplete Gamma function and is left to the Appendix.

Remark 4.3. The correction in Proposition 4.7 can be viewed as a one-step Newton's iteration based on $\lambda_{k,\infty}(t)$.

In Fig. 4.1, we compare the correction term (the n^{-1} term) from Proposition 4.7 to the numerical values. The numerical value corresponding to a pair (t,k) is obtained by numerically finding $\lambda_{k,n}(t)$ for a sequence of n varying from 200 to 2×10^4, then fitting $\log(\lambda_{k,n} - \lambda_{k,\infty})$ against $-\log n$ in least squares, and finally taking the intercept and exponentiating.

Plugging the correction into Eq. (4.13) yields the following bound.

Corollary 4.2 (with correction)**.** *Let $\hat{\lambda}_{k,n} := \min\left\{1 - \frac{k-1}{t} + \frac{k}{k-1}\frac{t-k+1}{n}, 1\right\}$. For $n \geq 1$, $k \geq 2$ and $t > k-1$, it holds that*

$$P\left(n \, \mathcal{D}(\hat{p}_{k,n} \| p) > t\right) \leq \exp(-\hat{\lambda}_{k,n} t) G_{k,n}(\hat{\lambda}_{k,n}). \tag{4.19}$$

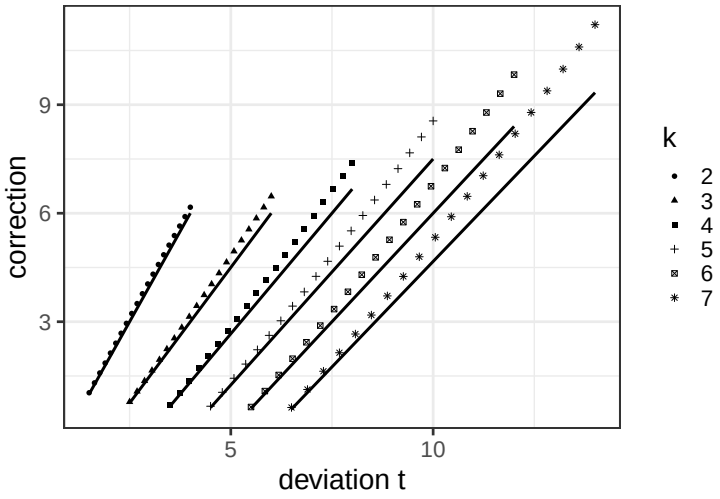


Figure 4.1: The ideal correction $\lim_n n(\lambda_{k,n}(t) - \lambda_{k,\infty}(t))$ (dots, fitted from numerical values) and the theoretical first-order correction $k(t-k+1)/(k-1)$ (lines), both plotted against the deviation t.

4.4 Discussion

In this section, we discuss the behavior of our bound and compare to bounds previously proposed in the literature.

4.4.1 Comparison

We briefly compare the bounds for several sample sizes under $k = 6$ in Figure 4.2; see also Fig. 4.3 for $k = 20$. First, our bound is always tighter than Agrawal (2020), since Agrawal (2020) uses Chernoff bound based on $G_{k,\infty}$, which upper-bounds $G_{k,n}$. Second, in the settings plotted, our bound is tighter than that of Mardia et al. (2019) for t smaller than some $T_{k,n}$ and vice versa for $t > T_{k,n}$ — an explanation for this phenomenon is provided in the following section. Third, the closed-form correction-based bound is significantly tighter than the bound without correction, and is in fact very close to the exact bound, with the difference between the two only noticeable when both n and t are small.

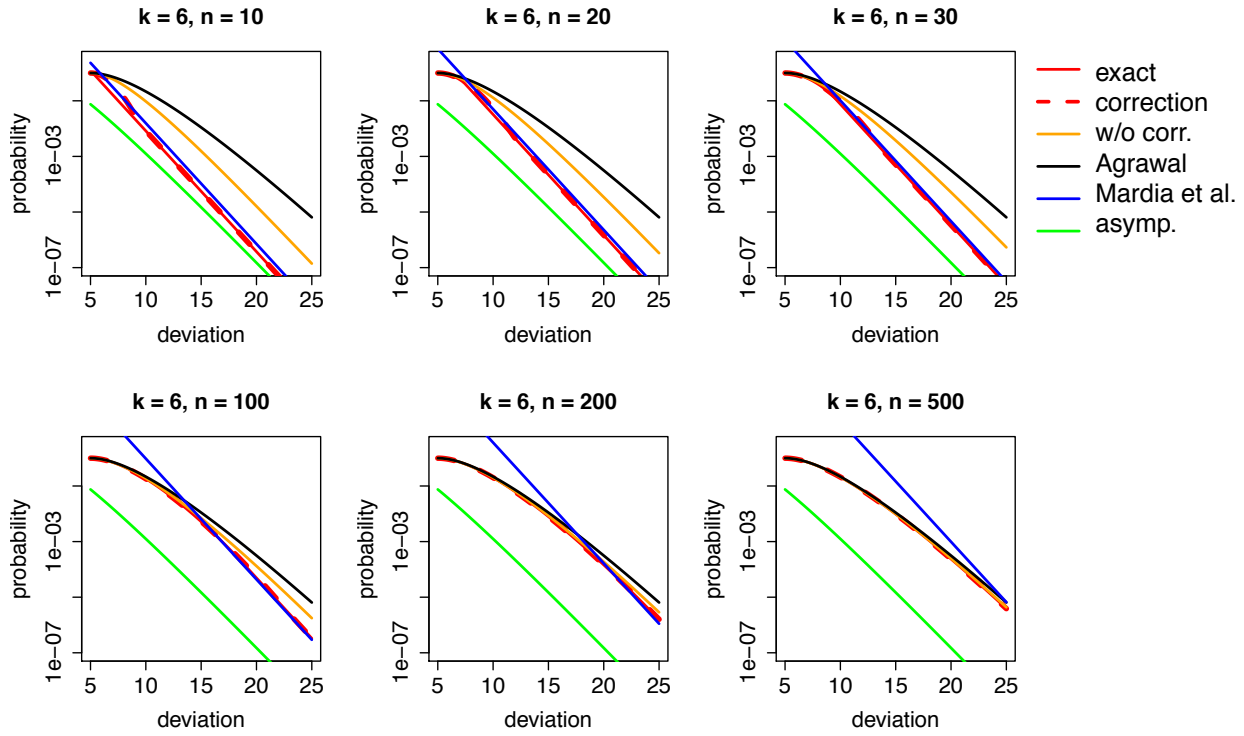


Figure 4.2: Comparison of probability bounds on $P(n\,\mathcal{D}(\hat{p}_{k,n}\|p) > t)$ for $k = 6$ and $t > \min(\log G_{k,n}(1), k-1)$. The y-axis is in logarithmic scale. The methods compared include: "exact" (Theorem 4.2 from numerical minimization), "correction" (Corollary 4.2), "w/o corr." (Corollary 4.1), Agrawal (2020, Theorem 1.2), Mardia et al. (2019, Theorem 3), and the asymptotic bound that is the exact probability when $n \to \infty$. Note that "asymp." might not be a valid bound and is for reference only.

4.4.2 Combinatorial scaling

Recently Mardia et al. (2019) considered a bound of the form

$$P\left(n\,\mathcal{D}(\hat{p}_{k,n}\|p) > t\right) \leq C(k,n)\exp(-t), \tag{4.20}$$

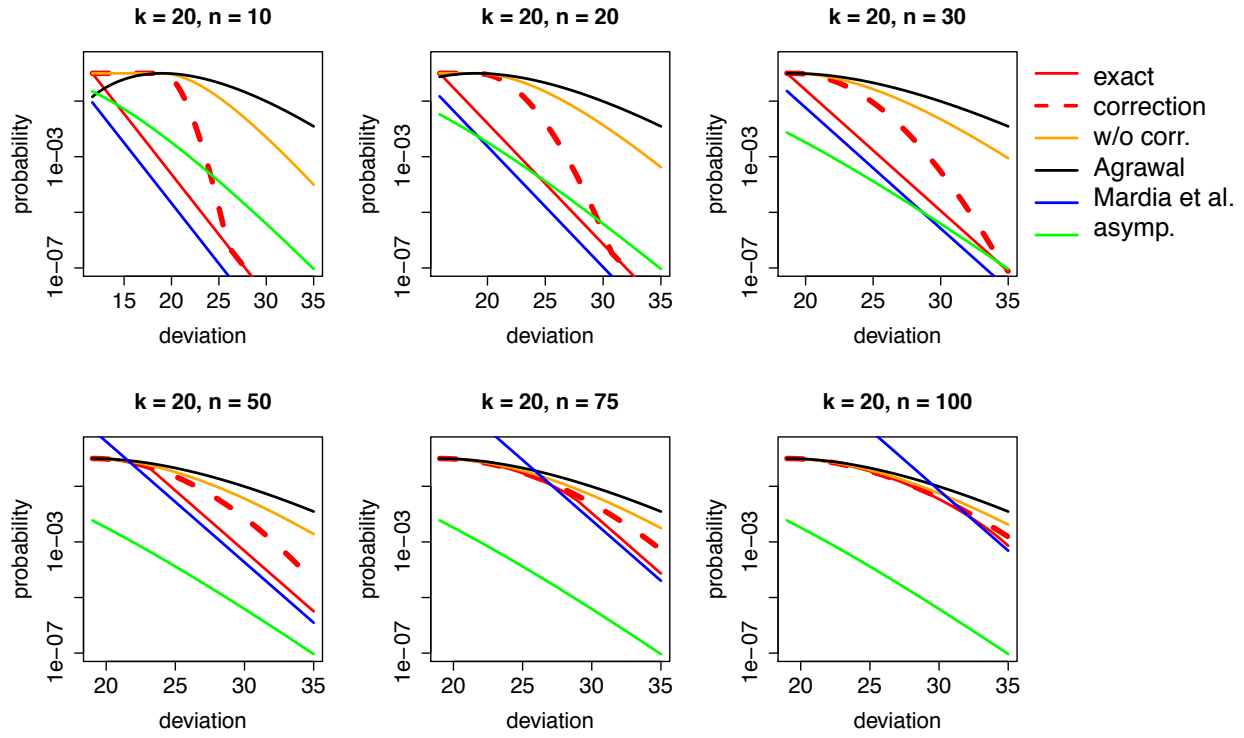


Figure 4.3: Comparison of probability bounds on $P(n\,\mathcal{D}(\hat{p}_{k,n}\|p) > t)$ for $k = 20$ and $t > \min(\log G_{k,n}(1), k-1)$. The y-axis is in logarithmic scale. The methods compared include: "exact" (Theorem 4.2 from numerical minimization), "correction" (Corollary 4.2), "w/o corr." (Corollary 4.1), Agrawal (2020, Theorem 1.2), Mardia et al. (2019, Theorem 3), and the asymptotic bound that is the exact probability when $n \to \infty$. Note that "asymp." might not be a valid bound and is for reference only.

where $C(k,n)$ captures the combinatorial dependence on k and n. This is motivated by the classic method-of-types inequality Eq. (4.4), which holds with

$$C_{\mathrm{T}}(k,n) = \binom{n+k-1}{k-1}.$$

Note that $C_{\mathrm{T}}(k,n)$ is the number of ways that $\{1,\ldots,n\}$ can be partitioned into k groups, and hence counts the "types" of possible empirical distributions. Mardia et al. (2019) showed

that $C_{\mathrm{T}}(k,n)$ can be improved to

$$C_{\mathrm{M}}(k,n) = \frac{12}{\pi}\sum_{i=0}^{k-2} K_i \left(\frac{e\sqrt{n}}{2\pi}\right)^i,$$

where

$$K_i = \begin{cases} \frac{\pi(2\pi)^{m/2}}{2\times 4\times\cdots\times m} & (m \text{ is even}) \\ \frac{(2\pi)^{(m+1)/2}}{1\times 3\times\cdots\times m} & (m \text{ is odd}) \end{cases}, \quad K_{-1} = 1$$

are constants. It can be shown that $C_{\mathrm{M}}(k,n)$ is smaller than $C_{\mathrm{T}}(k,n)$ for all $k,n \geq 2$.

Since the choice of λ that tightens our bound depends on t, the bounds presented in the previous section do not take the form of Eq. (4.20). For comparison, we use the following bound from setting $\lambda = 1$ in Eq. (4.13), which is not the tightest bound except for very large t.

Corollary 4.3. *For $n \geq 1$, $k \geq 2$ and $t > 0$, it holds that*

$$P\left(n\,\mathcal{D}(\hat{p}_{k,n}\|p) > t\right) \leq G_{k,n}(1)\exp(-t).$$

Like $C_{\mathrm{M}}(k,n)$ the resulting combinatorial factor $G_{k,n}(1)$ is also uniformly smaller than the method-of-types combinatorial factor $C_{\mathrm{T}}(k,n)$.

Proposition 4.8. *For $k \geq 2$, $n \geq 1$, $G_{k,n}(1) < C_{\mathrm{T}}(k,n)$.*

Proof. By Theorem 4.1,

$$\begin{aligned} G_{k,n}(1) &= \sum_{m=0}^{n} \frac{n\times(n-1)\times\cdots\times(n-m+1)}{n^m}\binom{m+k-2}{k-2} \\ &< \sum_{m=0}^{n}\binom{m+k-2}{k-2} = \binom{n+k-1}{k-1}, \end{aligned}$$

where the last equality follows from the "parallel summation" (Graham et al., 1994, Eq. (5.9)).

□

In fact, the improvement can be significant when n is large.

Proposition 4.9. *For fixed $k \geq 2$, as $n \to \infty$, $\frac{\log G_{k,n}(1)}{\log C_{\mathrm{T}}(k,n)} \to 1/2$.*

This basically says, in the regime of fixed k and large n, $G_{k,n}(1)$ is a square-root improvement over the method-of-types combinatorial factor. We leave its proof to Appendix C.3. In fact, $C_{\mathrm{M}}(k,n)$ achieves the same rate of improvement in the same regime; see Mardia et al. (2019, §1.2). For other regimes, we do not have an explicit comparison. Instead, in Fig. 4.4 we graphically compare the combinatorial factors for a few (k,n). We observe: (i) $\log G_{k,n}(1)$ and $\log C_{\mathrm{M}}(k,n)$ scale quite closely; (ii) for a fixed k, one can check that $G_{k,n}(1) < C_{\mathrm{M}}(k,n)$ for small n, and vice versa for large n. Note that (ii) explains why in Fig. 4.2 the bound of Mardia et al. (2019) becomes tighter than our bound for very large deviations when $n \in \{100, 200, 500\}$ — the tightening $\lambda_{k,n}(t) = 1$ for t large enough and the exact bound reduces to Corollary 4.3.

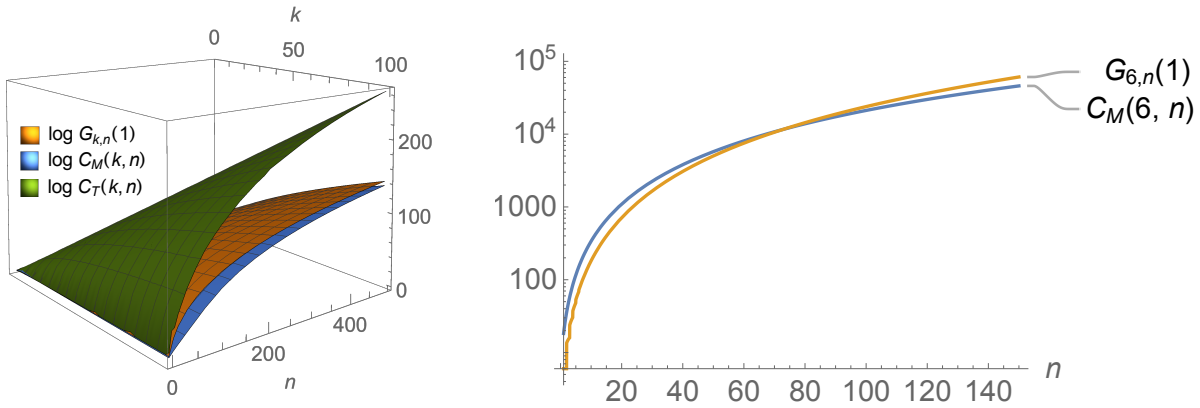


Figure 4.4: Comparison of combinatorial scaling factors $G_{k,n}(1)$ (ours), $C_{\mathrm{M}}(k,n)$ (Mardia et al., 2019) and $C_{\mathrm{T}}(k,n)$ (method of types).

Finally, we stress that the improved combinatorial factors are by no means optimal. To see this, note that as $n \to \infty$, $G_{k,n}(1) \to \infty$ for any fixed $k \geq 2$ and $C_{\mathrm{M}}(k,n) \to \infty$ for any fixed $k \geq 3$, which would render the bound in the form of Eq. (4.20) meaningless (for fixed k and t). However, by Proposition 4.3 because $G_{k,\infty}(\lambda)$ only diverges at $\lambda = 1$, our bounds stated in Theorem 4.1, Corollaries 4.1 and 4.2 do not suffer from this problem. Nevertheless,

we expect future improvements on $C(k, n)$ such that $C(k, \infty) < \infty$ for $k \geq 2$.

4.5 Application: unseen butterflies

The bound developed can be used to obtain a conservative critical value for the multinomial likelihood ratio. The bound in Theorem 4.2 can be determined numerically by searching for the minimizer over the unit interval, which is a non-convex but smooth, univariate optimization. Further given a level $\alpha \in (0, 1)$ (e.g., $\alpha = 0.05$), by a binary search, a critical value $t_{k,n}(\alpha)$ can be determined such that the bound at $t_{k,n}(\alpha)$ evaluates to α. The critical value on the likelihood ratio can be inverted to form a convex confidence region on p, which is guaranteed to contain p with probability at least $(1 - \alpha)$. This can be applied to the cases where k is comparable to n, and the standard large-sample χ^2 approximation is unlikely to be accurate (see Frydenberg and Jensen (1989)). We demonstrate with the following example.

Proportion of the unseen butterflies Table 4.2 shows the famous dataset (Orlitsky et al., 2016) that naturalist Corbet presented to Ronald Fisher in the 1940's. Corbet spent two years trapping butterflies in Malay Peninsula, and his intriguing question to Fisher was how many new species would he discover had he spent another two years on the islands. Corbet's original question led to the fruitful investigation of estimating the number of unseen species; see Fisher et al. (1943); Good and Toulmin (1956); Orlitsky et al. (2016).

Table 4.2: Butterflies recorded by Corbet

Frequency	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Species	118	74	44	24	29	22	20	19	20	15	12	14	6	12	6

However, here we pose a different question — what percentage of butterflies in Malay Peninsula belonged to the species that Corbet had not seen? That is, we want to estimate the proportion of butterflies from all the unseen species. Clearly, the MLE is zero based on

the sample. Instead, we ask for an upper bound with 95% confidence. Let $k = 435+1$, where 435 is the number of species observed by Corbet. Let $\hat{p} = (\hat{q}, 0)$, where $\hat{q}$ is the empirical distribution corresponding to Table 4.2. The sample size is $n = 2,029$ and the corresponding critical value is $t_{k,n}(\alpha) = 481.20$. The upper bound is given by the convex program

$$\max p_k \quad \text{s.t. } p \in \Delta^{k-1},\ n\,\mathcal{D}(\hat{p}\|p) \le t_{k,n}(\alpha),$$

which evaluates to 21.1%. See Appendix C.4 for the R code.

4.6 Conclusion

We have shown that for a multinomial experiment with alphabet size k and sample size n, the moment generating function of the entropy of the empirical distribution relative to the true distribution (scaled by n) can be uniformly bounded by a degree-n polynomial $G_{k,n}(\lambda)$ over the unit interval. We generalize Agrawal's (2020) result on $k = 2$ and characterize the family of $G_{k,n}(\lambda)$. The result gives rise to a one-sided Chernoff bound on the relative entropy for deviations $t > \min(\log G_{k,n}(1), k-1)$. The bound significantly improves the classic method-of-types bound and is competitive with the state of the art (Mardia et al., 2019). Further, since the tightest Chernoff bound does not permit a closed-form, we have developed a first-order large-n expansion of the minimizing λ, which provides a good approximation to the tightest bound in closed form. On a technical note, our approach directly constructs bounds for a generic k, in contrast to some other approaches (Mardia et al., 2019; Agrawal, 2020) that are based on a reduction from multinomial to binomial via the chain rule of relative entropy.

Chapter 5

CONVEX ANALYSIS OF DISCRETE INSTRUMENTAL VARIABLE MODELS

In this Chapter, we study instrumental variable (IV) models when the state space is finite. Such models can be viewed as a convex polytope in the space of counterfactual distributions, which maps to another convex polytope in the space of observed distributions. In Section 5.1, we review some background on IV and summarize common variants of IV assumptions. In Section 5.2, we present a framework for analyzing discrete IV models and illustrate with the case of Vietnam draft lottery data. Using the lottery number as an instrument, we study the effect of military service on annual earnings. We show that the difficulty in inference can be handled in an automated fashion with the aid of modern convex optimization and likelihood ratio bounds developed in the Chapter 4. Finally, in Section 5.3, we study partial identification of the average treatment effect. We show that, in a binary IV model, the Balke–Pearl bounds on the average treatment effect, which are typically derived under untestable cross-world assumptions, can be recovered under testable assumptions that hold naturally in a population single-world intervention graph (SWIG). The key ingredient is a set of inequalities that are implied by the latent confounder, which are analogues to the famous Bell-CHSH inequality in quantum mechanics. Polytope computation aids the proof of our result.

5.1 Background and assumptions

The instrumental variable approach is widely employed to infer a causal effect when there is latent confouding between treatment X and outcome Y. Intuitively, the instrument provides *exogeneity* for identifying the effect, an idea that dates back to Wright (1928). Roughly

speaking, variable Z is called an instrument if the following three conditions are met.

1. **Exclusion restriction**: Z has no direct effect on Y other than through X.

2. **Exogeneity**: Z is independent of the latent confounder between X and Y.

3. **Relevance**: Z is not independent of X; or, Z has a non-zero effect on X.

Many designed experiments and natural experiments can be viewed and analyzed from this perspective. Examples include clinical trials with non-compliance (Imbens and Rubin, 1997a), encouragement design, Mendelian randomization (Gray and Wheatley, 1991), environment factors (Miguel et al., 2004), etc; see Angrist and Krueger (2001, Table 1) for more examples.

5.1.1 IV model in econometrics

Traditionally, IV models are analyzed in the framework of linear structural equation models (SEM). Due to the focus of this Chapter on nonparametric IV models, we only give a brief overview of linear IV models; see Angrist and Pischke (2008, Chap. 4) for more details.

2SLS Linear IV model postulates structural equations

$$\begin{aligned} \text{1st-stage}: &\quad X = \pi Z + \alpha U + \eta^X, \\ \text{2nd-stage}: &\quad Y = \beta X + \delta U + \eta^Y, \end{aligned} \tag{5.1}$$

where U is an unobserved latent confounder. The parameter of interest is β, which measures the effect of X on Y. We assume all variables have zero mean and finite variance. The three IV assumptions are fulfilled as (1) Z does not enter the 2nd stage equation (exclusion), (2) $Z \perp\!\!\!\perp U, \eta^X, \eta^Y$ (exogeneity), and (3) $\pi \neq 0$ (relevance). By substituting the 2nd-stage equation into the 1st-stage equation, we get

$$\text{reduced form}: Y = \beta\pi Z + \beta(\alpha U + \eta^X) + \eta^Y.$$

Note that $\beta(\alpha U + \eta^X) + \eta^Y \perp\!\!\!\perp Z$ in the reduced form and $\alpha U + \eta^X \perp\!\!\!\perp Z$ in the 1st stage. Hence, by regressing Y on Z and regressing X on Z, we can identify $\beta\pi$ and π respectively, leading to

$$\hat{\beta} = \frac{\widehat{\beta\pi}}{\hat{\pi}} = \frac{\operatorname{cov}(Y, Z)/\operatorname{var} Z}{\operatorname{cov}(X, Z)/\operatorname{var} Z} = \frac{\operatorname{cov}(Y, Z)}{\operatorname{cov}(X, Z)}.$$

This estimator can be rephrased as the two-stage least squares (2SLS):

$$\begin{aligned} \text{1st stage}: &\quad \hat{\pi} \text{ from } X \sim Z \\ \text{2nd stage}: &\quad \hat{\beta} \text{ from } Y \sim \hat{\pi} Z, \end{aligned}$$

where $\hat{\pi}Z$ is the fitted value of X from the 1st-stage regression. When Z is binary, this is also known as Wald's estimator

$$\hat{\beta} = \frac{\operatorname{cov}(Y, Z)}{\operatorname{cov}(X, Z)} = \frac{\mathbb{E}[Y|Z=1] - \mathbb{E}[Y|Z=0]}{\mathbb{E}[X|Z=1] - \mathbb{E}[X|Z=0]}. \tag{5.2}$$

When X is multivariate, 2SLS can be generalized to cases when $\dim Z \geq \dim X$. Because 2SLS takes the form of a division, its finite-sample behavior can be erratic when π is close to zero, in which case Z is called a *weak instrument*; see Stock et al. (2002) for a survey of tools to handle this case.

Heterogeneity and LATE Assumptions posed by Eq. (5.1) that lead to the point identification of β may seem rather strong: for every individual, Z has a constant effect on X and X has a constant effect on Y. Suppose both Z and X are binary and let us drop Eq. (5.1). Intuitively, one is unable to identify the effect for those whose treatment status is unchanged by the instrument. Imbens and Angrist (1994) shows that under a monotonicity assumption that $X(z=1) \geq X(z=0)$ for every individual, the effect of X on Y can be identified for the subgroup of *compliers*, namely the set of individuals with $X(z=1) = 1$ and $X(z=0) = 0$, or in other words, those who are treated if and only if $Z = 1$. This effect is called the local average treatment effect (LATE) and can be estimated with Eq. (5.2). We will describe a generalization of this idea in Section 5.2.

5.1.2 Discrete IV models

For the rest of the this Chapter, we will focus on another line of research that revolves around (nonparametric) discrete IV models, where the state spaces for X, Y and Z, denoted by $\mathcal{X}$, $\mathcal{Y}$ and $\mathcal{Z}$, are finite. A primary case is when $\mathcal{X} = \mathcal{Y} = \mathcal{Z} = \{0, 1\}$, such as clinical trials with non-compliance (Z: assigned treatment, X: actual treatment, Y: recovery status). These models can also arise from discretizing an IV dataset with continuous variables.

Studies of discrete IV models are primarily concerned with two tasks: *partial identification* and *falsification*. Without additional assumptions, the counterfactual distributions of Y (or the treatment effect) are not uniquely identified from the observed distribution of X, Y and Z. However, non-trivial bounds on these parameters of interest are often implied from the observed distribution; such as the bounds obtained by Balke and Pearl (1997); Robins (1989); Manski (1990) for binary IV models. Similarly, non-trivial inequalities on the observed distribution (Pearl, 1995a; Balke and Pearl, 1997; Bonet, 2001; Kédagni and Mourifié, 2020) are typically implied by IV assumptions, which can be exploited for falsification tests. As one can imagine, the purpose of both tasks is to characterize *sharp* bounds or implications that cannot be improved. However, the sharp characterization can vary according to the exact form of IV assumptions postulated; see Swanson et al. (2018) for a survey of partial identification results under different assumptions. Next, we review some common variants of exclusion restriction and exogeneity assumption suitable for discrete IV models; stronger assumptions are listed first. Note that relevance, as a form of faithfulness assumption, is not closed under weak convergence and is not useful for the purpose of partial identification and falsification.

Exclusion restriction We have the following variants.

Assumption 5.1 (individual exclusion)**.** $Y(x, z) = Y(x, z') = Y(x)$ *for all* $x \in \mathcal{X}$, $z, z' \in \mathcal{Z}$.

Assumption 5.2 (latent-variable stochastic exclusion)**.** *There exists a latent variable* U *such that* $P(Y(x, z) = y \mid U) = P(Y(x, z') = y \mid U)$ *for all* $x \in \mathcal{X}$, $y \in \mathcal{Y}$ *and* $z, z' \in \mathcal{Z}$

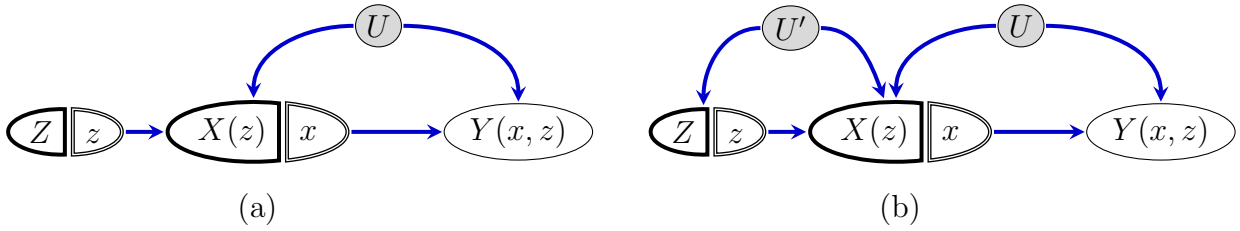


Figure 5.1: SWIGs for IV models, where Z is randomly assigned in (a) but not in (b). U and U' are latent variables.

almost surely.

Assumption 5.3 (stochastic exclusion)**.** $P(Y(x,z) = y) = P(Y(x,z') = y)$ *for all* $x \in \mathcal{X}$, $y \in \mathcal{Y}$ *and* $z, z' \in \mathcal{Z}$.

Clearly,

$$\text{individual exclusion} \Rightarrow \text{latent-variable stochastic exclusion} \Rightarrow \text{stochastic exclusion},$$

where the second "⇒" follows by marginalizing. Stochastic exclusion can be interpreted as the average controlled directed effect of Z on Y holding X fixed being zero. While individual exclusion is not subject to any empirical test, stochastic exclusion, being a statement on equality between groups stratified by Z, can be tested in a future experiment that randomizes both X and Z. Individual exclusion can be read off from SWIGs in Fig. 5.1 by noting $z \not\to Y(x,z)$. Assumption 5.2 and Assumption 5.3 can be read off from a population interpretation of SWIGs (Richardson and Robins, 2013, §7); see (Shpitser et al., 2020, Rule 3 of §3).

Exogeneity The assumption depends on whether Z is randomly assigned; compare the two SWIGs in Fig. 5.1.

Assumption 5.4 (random assignment)**.** $Z \perp\!\!\!\perp \{X(z), Y(x,z) : x \in \mathcal{X}, z \in \mathcal{Z}\}$.

Assumption 5.5 (joint exogeneity)**.** $Z \perp\!\!\!\perp \{Y(x,z) : x \in \mathcal{X}, z \in \mathcal{Z}\}$.

Assumption 5.6 (latent-variable marginal exogeneity)**.** *There exists a latent variable* U *such that* $Z \perp\!\!\!\perp U$ *and* $Y(x,z) \perp\!\!\!\perp Z, X(z) \mid U$ *for each* $x \in \mathcal{X}$ *and* $z \in \mathcal{Z}$*.*

Assumption 5.7 (marginal exogeneity)**.** $Z \perp\!\!\!\perp Y(x,z)$ *for every* $x \in \mathcal{X}$ *and* $z \in \mathcal{Z}$*.*

Exogeneity is also referred to as (statistical) independence or exchangeability. We have

$$\text{random assignment} \Rightarrow \text{joint exogeneity} \Rightarrow \text{latent-variable} \Rightarrow \text{marginal exogeneity},$$

where the second "⇒" follows from identifying U with $(Y(x,z) : x \in \mathcal{X}, z \in \mathcal{Z})$, and the third follows from contraction axiom of conditional independence. Latent-variable exogeneity can be read off from the SWIGs in Fig. 5.1. Contrary to the first two, marginal exogeneity does not make cross-world assumptions and can be (hypothetically) tested if the natural value of Z is observed immediately before X and Z are intervened on.

5.2 Inference via convex programming: Vietnam draft lottery

In this Section, we use the convex programming technique developed to analyze the effect of Vietnam war veteran status on annual earning. We use the CPS (Current Population Survey) extract dataset[1] prepared for Angrist and Krueger (1992b, 1995).

5.2.1 Background and motivation

Each year between 1970 and 1973, the priority for military service in Vietnam for draft-age men was determined by lotteries. In each lottery, all 366 days of a year are randomly permuted and each birthdate is assigned with a Random Sequence Number (RSN) between 1 and 366. Men are drafted for service by the order of their RSNs, with people with lower numbers called first. At some point during the year, a ceiling RSN is announced such that only those with RSNs below the ceiling are officially draft-eligible. However, of course, the RSN does not completely determine the veteran status. Many men, especially those with a lower RSN, would voluntarily enlist to improve their service conditions (Angrist, 1991).

[1] https://economics.mit.edu/faculty/angrist/data1/data/angkru95

In addition, there was great uncertainty associated with the exact ceiling RSN, which was generally unknown untill later in the year. For example, the 1971 ceiling RSN (125) was announced in October, 5 months after the number was called for service. Therefore, even for those with a relatively high RSN, there was incentive to seek draft deferment (e.g., by remaining in school).

Men were drafted by lottery from specific birth cohorts. The 1970 lottery drafted men born between 1944 and 1950; each lottery between 1971 and 1973 only drafted men born in a specific year, namely 1951-1953 accordingly. For the purpose of this Section, we focus on a subsample consisting of 14,464 white men (after removing individuals with missing data), born between 1949 and 1953. The dataset records the annual earnings in the surveyed year and the wages are converted to 1978 dollars using the CPI (Angrist and Krueger, 1992b). To protect anonymity, the dataset was curated such that RSNs are recoded into 14 groups: each of the first 13 groups contains 25 consecutive numbers (RSN 1-324), while the last group contains all the remaining numbers.

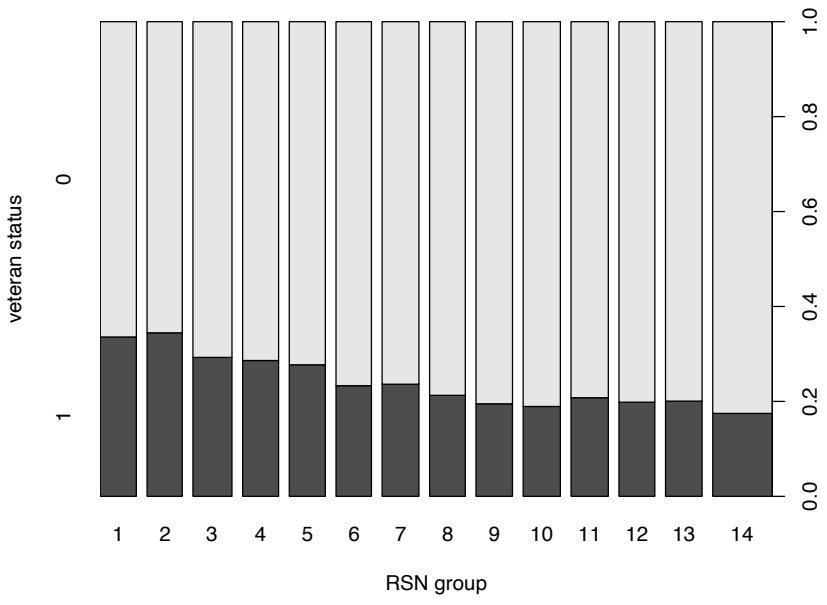


Figure 5.2: Propensity of veteran status by lottery RSN groups.

Instrument and monotonicity Let Z denote the RSN group, X denote military service (X=1 for served) and Y denote the annual earning. Since (1) Z is randomly assigned (exogenous), (2) Z has no effect on Y except through X, (3) Z is associated with X, Z is an instrument; see Angrist (1990) for more discussions. Fig. 5.2 shows the empirical propensity of service for each level of Z. Let $X(z)$ denote the potential outcome of enlistment had the person been assigned with RSN level z. Suppose Z takes levels $1, \ldots, |\mathcal{Z}|$. Consider the following monotonicity assumption.

Assumption 5.8 (Monotonicity)**.** $X(z) \geq X(z+1)$ *almost surely for* $z = 1, \ldots, |\mathcal{Z}| - 1$.

In words, this says that if a person is enlisted, then he must still be enlisted had he been assigned to a lower RSN group.

LATE analysis of counterfactual distributions The dataset was analyzed by Abadie (2002) in terms of the counterfactual distributions of the annual earning of the treated (military service) versus the untreated (no military service). By assuming monotonicity (Assumption 5.8), the counterfactual distributions of $Y(0)$ and $Y(1)$ among the compliers can be identified (Imbens and Rubin, 1997b). More specifically, let $Z' = \mathbb{I}_{Z>4}$ be a dichotomized instrument. Note that under Assumption 5.8, $X(z'=0) \geq X(z'=1)$ almost surely. Under monotonicity, joint exogeneity and positivity, the distribution functions of $Y(0)$ and $Y(1)$ among the compliers can be identified as

$$\begin{aligned} P(Y(1) \leq y \mid X(0)=1, X(1)=0) &= \frac{\mathbb{E}[I_{Y\leq y}X|Z'=0] - \mathbb{E}[I_{Y\leq y}X|Z'=1]}{\mathbb{E}[X|Z'=0] - \mathbb{E}[X|Z'=1]}, \\ P(Y(0) \leq y \mid X(0)=1, X(1)=0) &= \frac{\mathbb{E}[I_{Y\leq y}(1-X)|Z'=0] - \mathbb{E}[I_{Y\leq y}(1-X)|Z'=1]}{\mathbb{E}[1-X|Z'=0] - \mathbb{E}[1-X|Z'=1]}, \end{aligned} \tag{5.3}$$

where individuals with $X(0) = 1$ and $X(1) = 0$ are called compliers. Replacing the conditional expectations above with their empirical estimates, the two counterfactual distribution functions can be estimated from data; see Fig. 5.3. It can be seen from the figure that military service seems to reduce the lower quantiles of the earnings, while leaving the upper quantiles unaffected (Abadie, 2002).

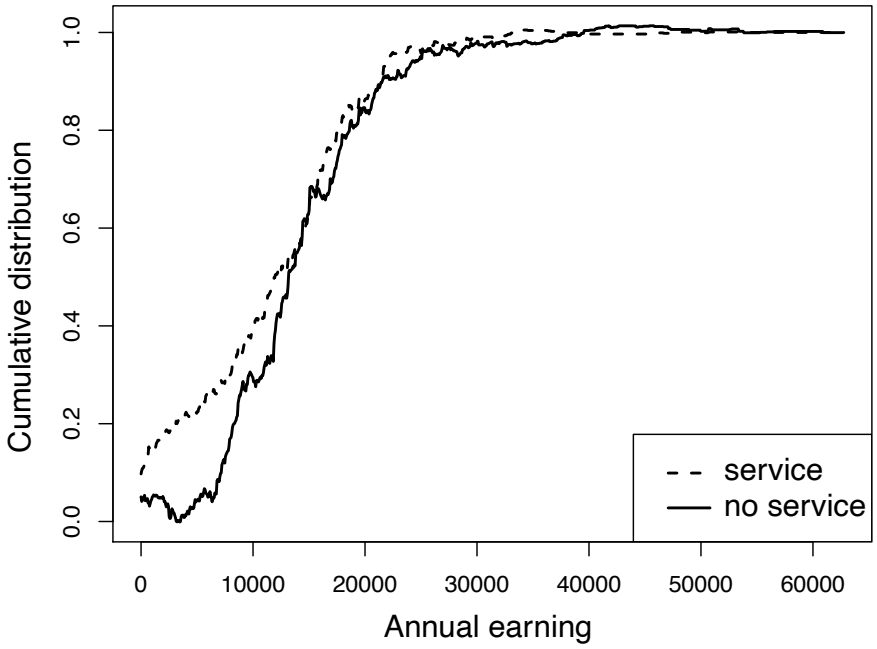


Figure 5.3: Counterfactual CDFs of annual earning among the compliers estimated by the method of Abadie (2002); the estimates seem to suggest that the untreated distribution stochastically dominates the treated distribution. The instrument employed is the dichotomized $Z' = \mathbb{I}_{Z>4}$.

Hence, in other words, did the military service make the poor even poorer? This observation is far from being conclusive. There are a few issues.

1. It is unclear whether the difference in the lower quantiles is statistically significant. Abadie (2002) conducted a bootstrap test for the hypothesis of the first-order stochastic dominance of $Y(0)$ over $Y(1)$ among compliers, namely

$$P(Y(0) \leq y \mid X(0) = 1, X(1) = 0) \leq P(Y(1) \leq y \mid X(0) = 1, X(1) = 0), \quad \forall y,$$

   with a modified Kolmogorov-Smirnov statistic (McFadden, 1989), and found the hypothesis not rejected. However, this by no means *proves* the hypothesis.

2. The analysis is restricted to the group of compliers relative to the binary instrument $Z' = \mathbb{I}_{Z>4}$. It is unclear whether the same observation can be made for compliers relative to other thresholds, or more generally, for the whole population. This is concerning given that most of those who served in Vietnam were likely volunteers who would have enlisted regardless of the assigned lottery number (Angrist and Krueger, 2001).

3. The empirical plugin estimate for Eq. (5.3) does not account for the restrictions on the observed distribution imposed by the counterfactual assumptions. There are two further points.

   (a) The plugin estimates are inefficient and may result in large biases. For the same reason, inverse probability weighting (IPW) estimators are inefficient and can be augmented to maximize efficiency (Tsiatis, 2006, Chap. 9)

   (b) The analysis does not test for falsification implications of the assumed counterfactual model. Ideally, one should test for such implications, and if the model is not rejected, proceed with further analysis. However, this is usually a challenging post-selection inference problem.

   There are various methods (Pearl, 1995a; Wang et al., 2017a; Kédagni and Mourifié, 2020) for falsification tests of IVs. Meanwhile, there is also a growing literature on the inference of "intersection bounds", bounds defined as supremum or infimum of conditional moments; see Chernozhukov et al. (2013); Andrews and Shi (2013); Fan and Park (2014), to name a few. However, most methods do not formulate this as a post-(model)-selection inference problem. For an exception, see Ramsahai and Lauritzen (2011) for an approach to binary IV based on specialized bootstrap (Andrews, 2000); see also Bi et al. (2019) for a related issue on weak instruments.

To address these issues, in what follows, we present a non-asymptotic analysis of the same data using a discrete IV model, conducted via convex programming in a unified fashion.

5.2.2 *Counterfactual model*

Recall that Z is the RSN group with $|\mathcal{Z}|$ levels, and $X(z) \in \{0, 1\}$ is the veteran status potential outcome. We discretize the earning Y by binning into brackets. Suppose the resulting Y takes $|\mathcal{Y}|$ levels. Let $Y(x, z)$ be its potential outcome. In addition to Assumption 5.8, we make counterfactual assumptions of individual exclusion (Assumption 5.1) and joint exogeneity (Assumption 5.5). By Assumption 5.1, we suppose that the RSN group has no effect on an individual's earning except through military service. In other words, there is only one pathway that RSN can affect earning. That being said, it is hypothesized that RSN can also affect earning through schooling; see Angrist and Krueger (1992a). Assumption 5.5 is implied by the random assignment of Z.

5.2.3 *Monotonicity, complier types and identification*

Assumption 5.8 states that $X(z)$ is a non-increasing function of ordinal variable z. This assumption can be interpreted as dividing the population into $|\mathcal{Z}| + 1$ *monotone types*:

- always taker (AT): $X(1) = \cdots = X(|\mathcal{Z}|) = 1$
- never taker (NT): $X(1) = \cdots = X(|\mathcal{Z}|) = 0$
- complier type k (CP_k): $X(1) = \cdots = X(k) = 1$, $X(k+1) = \cdots = X(|\mathcal{Z}|) = 0$, for $k = 1, \ldots, |\mathcal{Z}| - 1$.

For simplicity, in this subsection, we illustrate with $|\mathcal{Z}| = 3$. See Fig. 5.4 for the four types.

Lemma 5.1. *Under Assumptions 5.5 and 5.8, the probability of each monotone type is identified.*

Proof. From Fig. 5.4, it is easy to see that

$$P(\text{AT}) = P(X(3) = 1) = P(X(3) = 1|Z = 3) = P(X = 1|Z = 3),$$

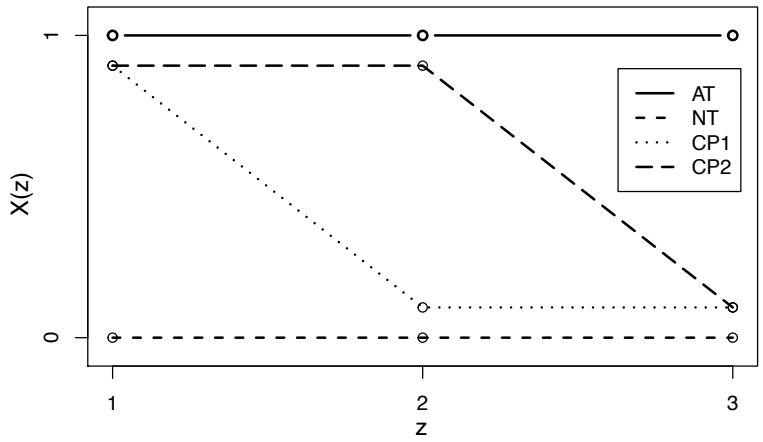


Figure 5.4: Four types of individuals when $|\mathcal{Z}| = 3$: always taker (AT), never taker (NT), complier 1 (CP_1) and complier 2 (CP_2).

where we used Assumption 5.5. Similarly, $P(\text{NT}) = P(X = 0|Z = 1)$. Now, note from the graph that $P(X = 1|Z = 2)$ consist of both AT and CP_2. Hence, we have

$$P(\text{CP}_2) = P(X = 1|Z = 2) - P(X = 1|Z = 3).$$

Similarly, it holds that

$$\begin{aligned} P(\text{CP}_1) = P(X = 0|Z = 2) - P(\text{NT}) &= P(X = 0|Z = 2) - P(X = 0|Z = 1) \\ &= P(X = 1|Z = 1) - P(X = 1|Z = 2). \end{aligned} \tag{5.4}$$

The case for other levels of Z follows similarly. □

Corollary 5.1. *Under Assumptions 5.5 and 5.8, the observed $\mathbb{E}[X|Z = z]$ is non-increasing in z. Further, it is strictly decreasing if the probability of each monotone type is positive.*

This prediction seems to be compatible with Fig. 5.2.

Lemma 5.2. *Suppose Assumptions 5.1, 5.5 and 5.8 hold. Also, suppose the probability of each monotone type is positive. Then, the marginal distribution of $Y(0)$ for NT and the marginal distribution of $Y(1)$ for AT are identified. Further, for each $k = 2, \ldots, |\mathcal{Z}| - 1$,*

the marginal distribution of $Y(0)$ *and the marginal distribution of* $Y(1)$ *among each* CP_k *are identified.*

Proof. Note distribution of $Y(0)$ among NT is identified from

$$P(Y(0)=y|\text{NT}) = P(Y(0)=y|X(1)=0) = P(Y(0)=y|X(1)=0,Z=1)$$
$$= P(Y=y|X=0,Z=1),$$

where we used Assumption 5.5 and consistency. Similarly, distribution of $Y(1)$ of AT can be identified from $P(Y|X=1,Z=3)$.

Now we identify $Y(0)$ and $Y(1)$'s marginal distributions for CP_2. Note

$$\begin{aligned}
P(Y=y,X=0|Z=3) &= P(Y(0)=y,X(3)=0|Z=3)\\
&= P(Y(0)=y,X(3)=0)\\
&= P(Y(0)=y,X(2)=1,X(3)=0) + P(Y(0)=y,X(2)=0,X(3)=0)\\
&= P(Y(0)=y|\text{CP}_2)P(\text{CP}_2) + P(Y(0)=y,X(2)=0)\\
&= P(Y(0)=y|\text{CP}_2)P(\text{CP}_2) + P(Y(0)=y,X(2)=0|Z=2)\\
&= P(Y(0)=y|\text{CP}_2)P(\text{CP}_2) + P(Y=y,X=0|Z=2),
\end{aligned}$$

where the 4th equality uses monotonicity $X(2)=0 \Rightarrow X(3)=0$. Using Eq. (5.4), we get

$$P(Y(0)=y \mid \text{CP}_2) = \frac{P(Y=y,X=0|Z=3) - P(Y=y,X=0|Z=2)}{P(X=0|Z=3) - P(X=0|Z=2)}.$$

Similarly, by decomposing $P(Y=y,X=1|Z=2)$, we get

$$\begin{aligned}
P(Y=y,X=1|Z=2) &= P(Y(1)=y,X(2)=1|Z=2)\\
&= P(Y(1)=y,X(3)=0,X(2)=1) + P(Y(1)=y,X(3)=1,X(2)=1)\\
&= P(Y(1)=y|\text{CP}_2)P(\text{CP}_2) + P(Y(1)=y,X(3)=1)\\
&= P(Y(1)=y|\text{CP}_2)P(\text{CP}_2) + P(Y(1)=y,X(3)=1|Z=3)\\
&= P(Y(1)=y|\text{CP}_2)P(\text{CP}_2) + P(Y=y,X=1|Z=3).
\end{aligned}$$

Dividing through by the formula for $P(\text{CP}_2)$, we have

$$P(Y(1)=y|\text{CP}_2) = \frac{P(Y=y,X=1|Z=2) - P(Y=y,X=1|Z=3)}{P(X=1|Z=2) - P(X=1|Z=3)}.$$

The identification among CP_1, or any other CP_k, follows in the same way. □

It then follows that, the *distributional treatment effect* among each CP_k

$$F_1(y|\mathrm{CP}_k) - F_0(y|\mathrm{CP}_k) := P(Y(1) \leq y|\mathrm{CP}_k) - P(Y(0) \leq y|\mathrm{CP}_k) \tag{5.5}$$

is identified. However, the distributional treatment effect over the whole population

$$F_1(y) - F_0(y) := P(Y(1) \leq y) - P(Y(0) \leq y)$$

is unidentified because $P(Y(0)|\mathrm{AT})$ and $P(Y(1)|\mathrm{NT})$ are unidentified. However, we will show that non-trivial bounds can still be obtained for the overall effect.

5.2.4 Convex program

We now describe the convex program for our analysis. The programming variable is the *counterfactual probability table*

$$P(X(1), \ldots, X(|\mathcal{Z}|), Y(0), Y(1)),$$

which is an element in probability simplex $\Delta^{2^{|\mathcal{Z}|}|\mathcal{Y}|^2-1}$. The counterfactual probability table is subject to the following constraints.

1. **Assumption 5.4**: For all $z \in \mathcal{Z}$,

$$P(X(1), \ldots, X(|\mathcal{Z}|), Y(0), Y(1)) \equiv P(X(1), \ldots, X(|\mathcal{Z}|), Y(0), Y(1)|Z = z). \tag{5.6}$$

2. **Assumption 5.8**: All the non-monotone types have probability zero. That is,

$$P(X(1) = x_1, \ldots, X(|\mathcal{Z}|) = x_{|\mathcal{Z}|}) = 0,$$

for all $(x_1, \ldots, x_{|\mathcal{Z}|})$ such that $x_i < x_{i+1}$ for some i. The equation is linear in P because marginalization is linear. There are $(2^{|\mathcal{Z}|} - |\mathcal{Z}| - 1)$ such equalities.

3. **Consistency**: Describe the observed distribution in terms of *conditional probability tables*

$$Q(X, Y|Z = z), \quad z = 1, \ldots, |\mathcal{Z}|.$$

These tables are linear functions of P through consistency conditions

$$Q(X = x, Y = y|Z = z) = P(X(z) = x, Y(x) = y), \quad x \in \{0, 1\}, y \in \mathcal{Y}, z \in \mathcal{Z}, \quad (5.7)$$

where Eq. (5.6) is implicitly used. There are $2|\mathcal{Y}||\mathcal{Z}|$ such equalities. These linear maps automatically ensure that $Q(X, Y|Z = z)$ is a probability table.

Finally, the observed distribution Q is linked to the data via a LRT concentration bound developed in the previous Chapter.

5. **LRT concentration**: Let $\mathbb{Q}_n$ denote the empirical distribution. We have

$$\sum_{z \in \mathcal{Z}} n_z \mathcal{D}\left(\mathbb{Q}_n(X, Y|Z = z) \| Q(X, Y|Z = z)\right) \leq C_\alpha, \quad (5.8)$$

where C_α is a non-asymptotic α-level critical value that can be computed by tightening the Chernoff bound. Note that the Kullback-Leibler divergence is convex in Q (Cover and Thomas, 2006, Theorem 2.7.2), and therefore the LHS above is also convex in P.

Remark 5.1. While there are other methods that construct smaller (in Lebesgue measure) confidence regions for the multinomial probability vector, such as Chafai and Concordet (2009); Malloy et al. (2020) in particular, these typically produce non-convex (even disconnected) confidence regions that are not suitable for the type of analysis presented here.

Confidence region Under the assumed causal model, the feasible region of this convex program is a confidence region that is guaranteed to contain the underlying the counterfactual distribution with at least $(1 - \alpha)$ probability. More explicitly, let $\mathcal{P}$ denote the space of counterfactual model P. Let $\mathcal{Q}$ denote the induced space of observed distributions, i.e., $\mathcal{Q} = \{Q(P) : P \in \mathcal{P}\}$, where $Q(P)$ is the linear map Eq. (5.7). Let $\mathcal{Q}_n$ be the set of

observed distribution allowed by Eq. (5.8). Then, the confidence region for the counterfactual distribution is given by

$$\mathcal{P}_n := Q^{-1}(\mathcal{Q} \cap \mathcal{Q}_n) = \{P \in \mathcal{P} : Q(P) \in \mathcal{Q}_n\}.$$

See Fig. 5.5 for an illustration. If this feasible region is empty, then the assumed causal model is rejected at level α.

If the feasible is non-empty, bounds of a functional $f(P)$ can be computed. In particular, the lower bound of a convex f, the upper bound of a concave f, and both bounds of a linear f can be easily obtained by optimizing f over $\mathcal{P}_n$. Suppose f is a linear functional of P that is not point-identified from observed $Q(P)$. Interval $[\min_{P\in\mathcal{P}_n} f(P), \max_{P\in\mathcal{P}_n} f(P)]$ computed in such a way is guaranteed, with at least $(1-\alpha)$ probability, to contain the population-level interval $[\min_{P\in Q^{-1}(Q)} f(P), \max_{P\in Q^{-1}(Q)} f(P)]$, which further contains the true $f(P)$. The reported interval is called an *ignorance region* (Vansteelandt et al., 2006), which can often be slightly shortened for the purpose of just covering $f(P)$ instead of the entire population-level interval (Imbens and Manski, 2004) — an idea we do not pursue here. By definition of confidence region, given a collection of functionals $f_1, f_2, \ldots$, the resulting intervals are guaranteed to cover the corresponding functionals simultaneously with probability at least $1-\alpha$. The guarantee is uniform; no multiplicity correction is required.

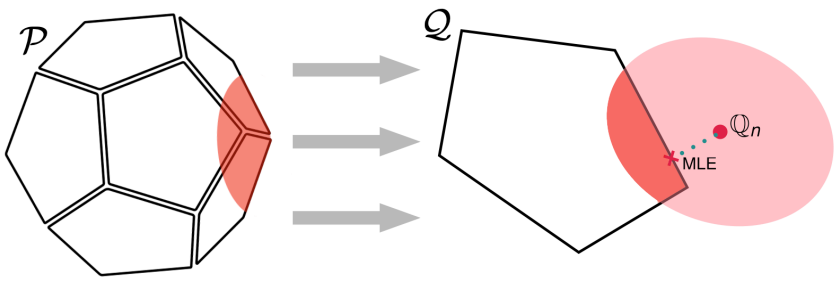


Figure 5.5: The confidence region $\mathcal{P}_n$ is the inverse image of the intersection of observed model and LRT lower level set.

5.2.5 Results

We now present our analysis of the CPS extract data. Since the size of the program is exponential in $|\mathcal{Z}|$, to fit into memory, we have to reduce the levels of Z to 5 by regrouping

$$\{1,2\} \mapsto 1, \quad \{3,4,5\} \mapsto 2, \quad \{6,7,8\} \mapsto 3, \quad \{9,10,11\} \mapsto 4, \quad \{12,13,14\} \mapsto 5.$$

Outcome Y is binned into 12 levels using breaks

$$0.1,\ 4000,\ 7279,\ 9500,\ 11486,\ 13451,\ 15540,\ 17931,\ 21620,\ 30000,\ 50000,$$

which are based on empirical deciles; the first bin is used to signify zero income. We choose significance level $\alpha = 0.05$ and the corresponding critical value $C_\alpha = 143.13$, which is computed with R package[2] `multChernoff`. The convex program is specified with R package CVXR (Fu et al., 2020) and solved with MOSEK (Andersen and Andersen, 2000). The feasible region is non-empty so the assumed model is not rejected.

Table 5.1: Proportions of monotone types

	NT	AT	CP_1	CP_2	CP_3	CP_4
overall (MLE)	0.652	0.183	0.061	0.060	0.028	0.016
overall (empirical)	0.660	0.188	0.054	0.058	0.030	0.009
among veterans (MLE)	0	0.768	0.034	0.085	0.063	0.051

Monotone types Table 5.1 shows the MLE for the six monotone types. Compliers CP_1–CP_4 only make up 16% of the population. As discussed in Angrist and Krueger (2001), most of the veterans indeed seem to be volunteers (AT).

[2] https://github.com/richardkwo/multChernoff

Remark 5.2. It is perhaps interesting to notice from Table 5.1 that the MLEs for monotone types are different from their empirical estimates. Due to the convex polytope described by the causal model, the information on Y is used to estimate the proportions of types. To understand this phenomenon, consider estimating a 2×2 table $\left(\begin{smallmatrix} p & r \\ r & q \end{smallmatrix}\right)$ whose minor diagonal entries must be identical, with empirical frequencies $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$. Although either margin of the model is unconstrained (the projection of the polytope is the unit interval), the MLE for the margin is not the corresponding empirical margin, because otherwise $p+r = a+c$ contradicts $p + r = a + d$. In fact, the MLE is $\left(\begin{smallmatrix} a & (b+c)/2 \\ (b+c)/2 & d \end{smallmatrix}\right)$ with margin $(a + (b + c)/2, d + (b + c)/2)$.

Distributional treatment effect within each complier type By definition, the distributional treatment effect Eq. (5.5) among CP_k is given by the ratio

$$\begin{aligned} &F_1(y|\mathrm{CP}_k) - F_0(y|\mathrm{CP}_k) \\ &= \frac{P(Y(1) \le y, X(k) = 1, X(k+1) = 0) - P(Y(0) \le y, X(k) = 1, X(k+1) = 0)}{P(X(k) = 1, X(k+1) = 0)}, \end{aligned} \tag{5.9}$$

which is not an affine function of the counterfactual distribution P. Since $Y(0)$ stochastically dominates $Y(1)$ if and only if $F_1(y) - F_0(y) \ge 0$, the sign of the effect is determined simply by the sign of the numerator, which is linear in P. Hence, we compute the confidence bands for the numerator from the convex program, which are confidence bands on the distributional effect up to an unknown positive constant. For ease of presentation, dividing the bands by the MLE of the denominator (bounds beyond $[-1, +1]$ are capped to $[-1, +1]$), we get the results shown in the first column of Fig. 5.6. There is hardly any evidence from data to determine the direction of the effect.

To get an idea on the sample size required to reach the conclusion, the 2nd and 3rd columns show the results if the counts in the contingency table were multiplied by 5 or 10. Because these CP_k effects are identified (Lemma 5.2), the width of the bands decreases at $n^{-1/2}$ rate. Still, no conclusion can be reached even at $\times 10$ sample size.

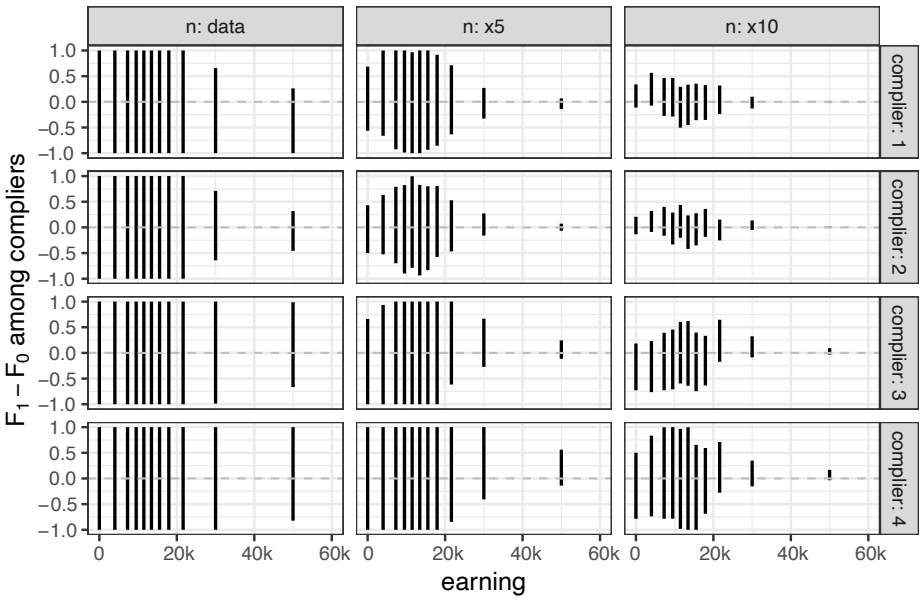


Figure 5.6: 95% confidence bands on distributional treatment effect $F_1(y|\mathrm{CP}_k) - F_0(y|\mathrm{CP}_k)$ within each complier subgroup (rows: complier types CP_1 through CP_4; columns: data, $\times 5$ sample size, $\times 10$ sample size). The bounds drawn are computed from dividing the bounds on the numerator of Eq. (5.9) by the MLE of $P(\mathrm{CP}_k)$.

Overall distributional treatment effect Fig. 5.7 shows the confidence bands for the distributional treatment effect over the whole population. Still, no conclusion can be reached. Note that, because $F_1 - F_0$ is unidentified, the width of bands does not diminish at $n^{-1/2}$ rate.

Falsifiability The assumed counterfactual model would be rejected (empty feasible region) if the sample size is multiplied by 12 with empirical frequencies unchanged. This is to say, hypothetically, at $\times 12$ sample size, the empirical distribution significantly violates the the implication on the observed distribution $P(X, Y, Z)$. In particular, the violation does not stem from the $P(X, Z)$ margin because the empirical estimates of monotone types are positive (Table 5.1), which is compatible with the monotonicity of $\mathbb{E}(X|Z = z)$ (Corollary 5.1). What

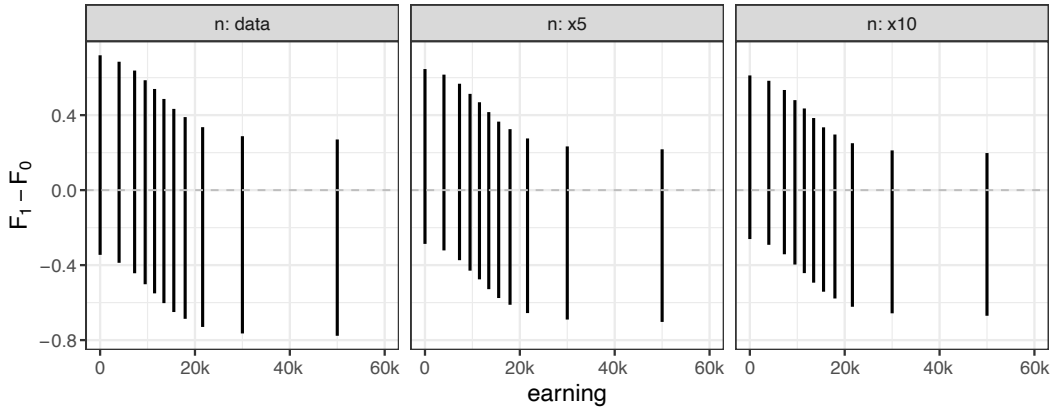


Figure 5.7: 95% confidence bands on the overall distributional treatment effect $F_1(y) - F_0(y)$ (columns: data, $\times 5$ sample size, $\times 10$ sample size).

is violated is the implications on $P(Y|X, Z)$; see Richardson et al. (2011, §4) for implications in the binary case.

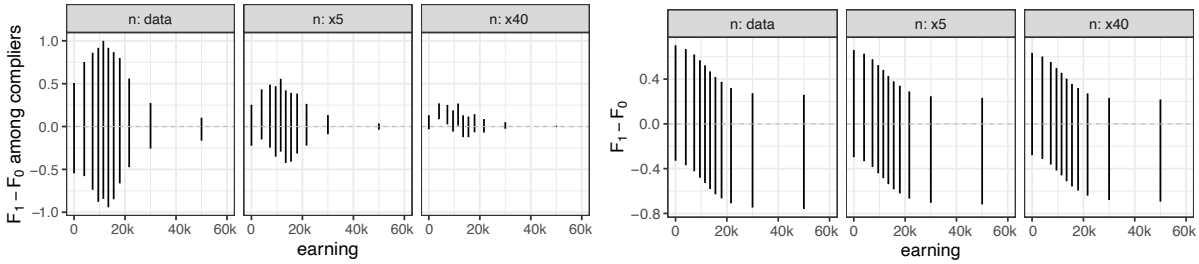


Figure 5.8: 95% confidence bands on the distributional treatment effect $F_1(y) - F_0(y)$ using the dichotomized instrument $Z' = \mathbb{I}_{Z>4}$. (left: among compliers, right: the whole population). Columns: data, $\times 5$ sample size, $\times 40$ sample size.

Comparison to the analysis of Abadie (2002) From our results, there is no evidence to conclude that the military service lowered annual earnings, with respect to each complier type or the whole population. Fig. 5.8 additionally shows the results from using the same

instrument $Z' = \mathbb{I}_{Z>4}$ as Abadie (2002) — significance from distributional treatment effect among compliers would require about $\times 40$ sample size.

5.3 Balke–Pearl bounds, CHSH inequality and SWIG independences

We conclude this Chapter by studying partial identification of the average treatment effect (ATE) in terms of IV model on $\mathcal{X} = \mathcal{Y} = \mathcal{Z} = \{0, 1\}$; no monotonicity is assumed. We resolve a long-standing gap between two set of bounds on the ATE — the stronger Balke–Pearl/"sharp IV" bounds and the weaker Robins-Manski/"natural" bounds. In the literature, the Balke–Pearl bounds are typically derived (Balke and Pearl, 1997; Richardson and Robins, 2014) under stronger assumptions, i.e., either individual exclusion or joint exogeneity, which are untestable cross-world statements, while the natural bounds only require testable assumptions. In this Section, we derive Balke–Pearl bounds under the latent-variable assumptions of marginal exogeneity (Assumption 5.6) and stochastic exclusion (Assumption 5.2); further, these bounds are shown to be sharp. The "secret sauce" that closes this gap is a set of CHSH-type inequalities (Clauser et al., 1969), a generalized form of Bell's (1964) inequality, implied by the latent variable.

5.3.1 Background

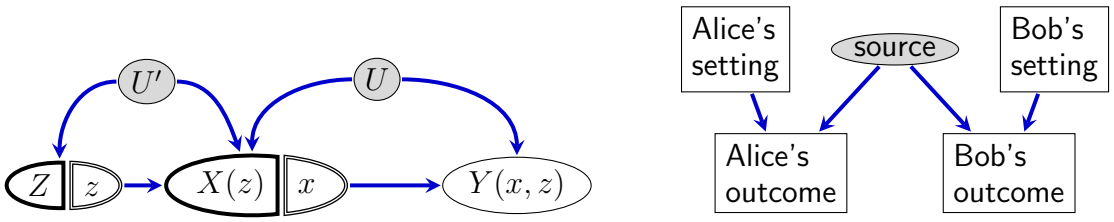
The latent-variable formulation of stochastic exclusion and marginal exogeneity

$$\text{Assumption 5.2}: \quad P(Y(x,0) = 1 \mid U) = P(Y(x,1) = 1 \mid U), \quad x \in \{0,1\} \tag{5.10}$$

$$\text{Assumption 5.6}: \quad Z \perp\!\!\!\perp U, \quad Y(x,z) \perp\!\!\!\perp Z, X(z) \mid U, \quad x, z \in \{0,1\} \tag{5.11}$$

can be read off from the population SWIG (Richardson and Robins, 2013, §7) drawn in Fig. 5.1(b), reproduced below in Fig. 5.9(a). In particular, Eq. (5.10) can be read off from the fact that the fixed node z is d-separated from $Y(x,z)$ by U; see Shpitser et al. (2020, Rule 3 of §3).

Marginalizing over U, these latent-variable assumptions *imply* stochastic exclusion and



(a) (b)

Figure 5.9: (a) SWIG for IV model (b) DAG for a Bell-CHSH experiment, where the potential outcomes are A (Alice's outcome when her setting is $\boldsymbol{a}$), A' (Alice's outcome when her setting is $\boldsymbol{a}'$), B (Bob's outcome when his setting is $\boldsymbol{b}$) and B' (Bob's outcome when his setting is $\boldsymbol{b}'$). Latent variables are shaded. Note the correspondence: $U-$ source, $X(z)-$ Alice's outcome, $z-$ Alice's setting, $Y(x,z)-$ Bob's outcome and $x-$ Bob's setting.

marginal exogeneity:

$$\text{Assumption 5.3}: \quad P(Y(x,z)=1) = P(Y(x,z')=1), \quad x,z,z' \in \{0,1\} \tag{5.12}$$

$$\text{Assumption 5.7}: \quad Z \perp\!\!\!\perp Y(x,z), \quad x,z \in \{0,1\}, \tag{5.13}$$

which leads to ATE $\tau := \mathbb{E}[Y(1,0)] - \mathbb{E}[Y(0,0)]$ being partially identified as

$$\begin{aligned} \max \left\{ \begin{array}{c} -P(X=1,Y=0 \mid Z=0) - P(X=0,Y=1 \mid Z=0) \\ -P(X=1,Y=0 \mid Z=1) - P(X=0,Y=1 \mid Z=1) \\ P(X=1,Y=1 \mid Z=0) + P(X=0,Y=0 \mid Z=1) - 1 \\ P(X=1,Y=1 \mid Z=1) + P(X=0,Y=0 \mid Z=0) - 1 \end{array} \right\} &\leq \tau \\ \leq \min \left\{ \begin{array}{c} -P(X=1,Y=1 \mid Z=0) - P(X=0,Y=0 \mid Z=0) \\ -P(X=1,Y=1 \mid Z=1) - P(X=0,Y=0 \mid Z=1) \\ 1 - P(X=1,Y=0 \mid Z=0) + P(X=0,Y=1 \mid Z=1) \\ 1 - P(X=1,Y=0 \mid Z=1) + P(X=0,Y=1 \mid Z=0) \end{array} \right\}, \end{aligned} \tag{5.14}$$

known as the natural bounds (Robins, 1989; Manski, 1990). These bounds are sharp under Eqs. (5.12) and (5.13).

Suppose that we strengthened the assumptions to individual exclusion and joint exogene-

ity:

$$\text{Assumption 5.1}: \quad Y(0,0) = Y(0,1) =: Y(0), \quad Y(1,0) = Y(1,1) =: Y(1) \tag{5.15}$$

$$\text{Assumption 5.5}: \quad Z \perp\!\!\!\perp Y(0), Y(1). \tag{5.16}$$

Note that both assumptions are cross-world statements, for which no consistent test exists even by a hypothetical randomized experiment. Under these assumptions, the interval can be narrowed to

$$\begin{aligned}
\max \left\{ \begin{array}{c}
-P(X=1,Y=0 \mid Z=0) - P(X=0,Y=1 \mid Z=0) \\
-P(X=1,Y=0 \mid Z=1) - P(X=0,Y=1 \mid Z=1) \\
P(X=1,Y=1 \mid Z=0) + P(X=0,Y=0 \mid Z=1) - 1 \\
P(X=1,Y=1 \mid Z=1) + P(X=0,Y=0 \mid Z=0) - 1 \\
P(X=1,Y=1 \mid Z=0) - P(Y=1 \mid Z=1) - P(X=1,Y=0 \mid Z=0) - P(X=0,Y=1 \mid Z=0) \\
P(X=1,Y=1 \mid Z=1) - P(Y=1 \mid Z=0) - P(X=1,Y=0 \mid Z=1) - P(X=0,Y=1 \mid Z=1) \\
P(X=0,Y=0 \mid Z=1) - P(Y=0 \mid Z=0) - P(X=1,Y=0 \mid Z=1) - P(X=0,Y=1 \mid Z=1) \\
P(X=0,Y=0 \mid Z=0) - P(Y=0 \mid Z=1) - P(X=1,Y=0 \mid Z=0) - P(X=0,Y=1 \mid Z=0)
\end{array} \right\} \leq \tau \\
\leq \min \left\{ \begin{array}{c}
-P(X=1,Y=1 \mid Z=0) - P(X=0,Y=0 \mid Z=0) \\
-P(X=1,Y=1 \mid Z=1) - P(X=0,Y=0 \mid Z=1) \\
1 - P(X=1,Y=0 \mid Z=0) + P(X=0,Y=1 \mid Z=1) \\
1 - P(X=1,Y=0 \mid Z=1) + P(X=0,Y=1 \mid Z=0) \\
-P(X=1,Y=0 \mid Z=0) + P(Y=0 \mid Z=1) + P(X=1,Y=1 \mid Z=0) + P(X=0,Y=0 \mid Z=0) \\
-P(X=1,Y=0 \mid Z=1) + P(Y=0 \mid Z=0) + P(X=1,Y=1 \mid Z=1) + P(X=0,Y=0 \mid Z=1) \\
-P(X=0,Y=1 \mid Z=1) + P(Y=1 \mid Z=0) + P(X=1,Y=1 \mid Z=1) + P(X=0,Y=0 \mid Z=1) \\
-P(X=0,Y=1 \mid Z=0) + P(Y=1 \mid Z=1) + P(X=1,Y=1 \mid Z=0) + P(X=0,Y=0 \mid Z=0)
\end{array} \right\},
\end{aligned} \tag{5.17}$$

which is known as the Balke–Pearl bounds (Balke and Pearl, 1997). The Balke–Pearl bounds are sharp under Eqs. (5.15) and (5.16). Hence, the obvious gap here is to understand if Balke–Pearl bounds can hold under weaker, falsifiable assumptions in the latent variable formulation.

5.3.2 CHSH inequality

Fig. 5.9(b) shows the DAG that depicts the Bell-CHSH experiment in quantum mechanics. Two particles are generated at the source and then travel away from each other; en route, they are measured by Alice and Bob. Alice and Bob measure the spin of a particle along particular

directions, denoted by Alice's setting and Bob's setting. Alice's setting is randomly chosen from $\boldsymbol{a}$ and $\boldsymbol{a}'$; Bob's setting is randomly chosen from $\boldsymbol{b}$ and $\boldsymbol{b}'$. The spin measurement is either $+1$ or -1. So, depending on the setting, Alice observes either $A = \pm 1$ or $A' = \pm 1$, and Bob observes either $B = \pm 1$ or $B' = \pm 1$. The CHSH inequality describes constraints obeyed by correlations $\mathbb{E}\, AB$, $\mathbb{E}\, A'B$, $\mathbb{E}\, AB'$ and $\mathbb{E}\, A'B'$, predicted by the local hidden variable theory; see Gill (2014) for a statistician's derivation.

Here we derive CHSH-type inequalities for Fig. 5.9(a), which describes inequalities implied by the latent variable U. See also Richardson et al. (2017, Example 57) for a related inequality. We start with a simple fact.

Lemma 5.3. *For $a, b, u, v \in [0, 1]$, we have*

$$0 \leq (1-a)(1-b) + av + ub - uv \ \leq 1.$$

Proof. The expression is equivalent to

$$0 \leq (1-a)(1-b) + ab - (u-a)(v-b) \ \leq 1.$$

This can be proved by first noting that

$$\min\{-ab, -(1-a)(1-b)\} \leq -(u-a)(v-b) \leq \max\{a(1-b), b(1-a)\}, \tag{5.18}$$

from which the conclusion then follows.

The claim (5.18) can be seen by noting that $f(u, v) = -(u-a)(v-b)$ is a hyperbola with saddlepoint (a, b), hence for $(u, v) \in [0, 1]^2$, maxima will occur at either $(0, 1)$ or $(1, 0)$ and minima will occur at $(0, 0)$ or $(1, 1)$. □

Proposition 5.1 (CHSH)**.** *Suppose X, Y, Z are all binary. Under Assumptions 5.2 and 5.6, we have*

$$0 \leq P(X=1, Y(x,z)=1 \mid Z=z) + P(X=0, Y(1-x,z)=0 \mid Z=z)$$
$$+P(X=0, Y(x,1-z)=0 \mid Z=1-z) - P(X=0, Y(1-x,1-z)=0 \mid Z=1-z) \leq 1, \quad x, z \in \{0,1\}.$$

Proof. By marginalization, it suffices to prove for $x, z \in \{0, 1\}$,

$$0 \leq P(X = 1, Y(x, z) = 1 \mid Z = z, U) + P(X = 0, Y(1 - x, z) = 0 \mid Z = z, U)$$
$$+P(X = 0, Y(x, 1-z) = 0 \mid Z = 1-z, U) - P(X = 0, Y(1-x, 1-z) = 0 \mid Z = 1-z, U) \leq 1.$$

The first term can be rewritten as

$$\begin{aligned}
&P(X = 1, Y(x, z) = 1 \mid Z = z, U) \\
&= P(X(z) = 1, Y(x, z) = 1 \mid Z = z, U) \\
&= P(X(z) = 1 \mid Z = z, U) P(Y(x, z) = 1 \mid X(z) = 1, Z = z, U) \\
&= P(X = 1 \mid Z = z, U) P(Y(x, z) = 1 \mid U).
\end{aligned}$$

The first step uses consistency. The last step uses Eq. (5.11) and consistency again. Applying a similar argument to the other three terms, we get

$$\begin{aligned}
&P(X = 1, Y(x, z) = 1 \mid Z = z, U) + P(X = 0, Y(1 - x, z) = 0 \mid Z = z, U) \\
&\quad + P(X = 0, Y(x, 1 - z) = 0 \mid Z = 1 - z, U) - P(X = 0, Y(1 - x, 1 - z) = 0 \mid Z = 1 - z, U) \\
&= P(X = 1 \mid Z = z, U) P(Y(x, z) = 1 \mid U) + P(X = 0 \mid Z = z, U) P(Y(1 - x, z) = 0 \mid U) \\
&\quad + P(X = 0 \mid Z = 1 - z, U) P(Y(x, 1 - z) = 0 \mid U) \\
&\quad - P(X = 0 \mid Z = 1 - z, U) P(Y(1 - x, 1 - z) = 0 \mid U) \\
&= \{1 - P(X = 0 \mid Z = z, U)\} \{1 - P(Y(x, z) = 0 \mid U)\} \\
&\quad + P(X = 0 \mid Z = z, U) P(Y(1 - x, z) = 0 \mid U) \\
&\quad + P(X = 0 \mid Z = 1 - z, U) P(Y(x, 1 - z) = 0 \mid U) \\
&\quad - P(X = 0 \mid Z = 1 - z, U) P(Y(1 - x, 1 - z) = 0 \mid U) \\
&= \{1 - P(X = 0 \mid Z = z, U)\} \{1 - P(Y(x, z) = 0 \mid U)\} \\
&\quad + P(X = 0 \mid Z = z, U) P(Y(1 - x, z) = 0 \mid U) \\
&\quad + P(X = 0 \mid Z = 1 - z, U) P(Y(x, z) = 0 \mid U) \\
&\quad - P(X = 0 \mid Z = 1 - z, U) P(Y(1 - x, z) = 0 \mid U),
\end{aligned}$$

where the last step uses Eq. (5.10). The result then follows from applying Lemma 5.3 with $a = P(X = 0 \mid Z = z, U)$, $b = P(Y(x, z) = 0 \mid U)$, $u = P(X = 0 \mid Z = 1 - z, U)$ and $v = P(Y(1 - x, z) = 0 \mid U)$. □

5.3.3 Balke–Pearl bounds

The natural bounds can be "boosted" to Balke-Pearl by CHSH inequalities.

Theorem 5.1. *Suppose X, Y, Z are all binary. Under Assumptions 5.2 and 5.6, the Balke-Pearl bounds in Eq. (5.17) hold and are sharp.*

Proof. Suppose the bounds hold. For sharpness, note that by identifying

$$U = (Y(0, 1), Y(1, 0), Y(0, 0), Y(1, 1)),$$

model defined by Eqs. (5.15) and (5.16) is a submodel and the same bounds are achieved under the submodel; see Richardson and Robins (2014, Theorem 2).

Now we prove Balke–Pearl bounds by polytope computation. Parametrize variables $P(X, Y \mid Z = z)$ and $P(X, Y(0, 0), Y(0, 1), Y(1, 0), Y(1, 1) \mid Z = z)$ for $z = 0, 1$. The variables are subject to (1) simplex constraints, (2) consistency, (3) stochastic exclusion Eq. (5.12) and (4) marginal exogeneity Eq. (5.13). Also, importantly, they obey (5) the CHSH inequalities in Proposition 5.1. Note that all of these constraints are affine in the variables. Additionally, introduce affine function $\tau = \mathbb{E}\, Y(1, 0) - \mathbb{E}\, Y(0, 0)$ as a programming variable. Under the constraints, the variables $P(X, Y \mid Z = z)$, $P(X, Y(0, 0), Y(0, 1), Y(1, 0), Y(1, 1) \mid Z = z)$ for $z = 0, 1$ along with τ define a degenerate convex polytope. We obtain its projection on coordinates $P(X, Y \mid Z = 0)$, $P(X, Y \mid Z = 1)$ and τ by Fourier–Motzkin elimination. The computation is done with Julia packages `JuMP.jl` (Dunning et al., 2017), `Polyhedra.jl` (Legat et al., 2021) and `CDDLib.jl` (Legat et al., 2020; Fukuda, 2007); see Appendix D for the code.

The projection is a polytope characterized by 2 supporting hyperplanes and 28 supporting half-spaces. There are 16 half-spaces that involve τ, which exactly give the Balke–Pearl bounds:

$P(0,0\|0)$	$P(1,0\|0)$	$P(0,1\|0)$	$P(0,0\|1)$	$P(1,0\|1)$	$P(0,1\|1)$	τ	$\leq$
0	1	1	0	0	0	1	1
0	1	0	0	0	1	1	1
0	2	1	-1	-1	0	1	1
0	-1	-1	0	0	0	-1	0
1	0	0	-1	-1	-1	-1	0
1	-1	-1	-1	-1	0	-1	0
0	0	1	0	1	0	1	1
0	1	2	1	1	0	1	2
-1	-1	-1	1	0	0	-1	0
-1	-2	-2	1	1	0	-1	0
0	0	0	0	1	1	1	1
-1	-1	0	0	2	1	1	1
0	0	0	0	-1	-1	-1	0
-1	-1	0	1	-1	-1	-1	0
1	1	0	-1	-2	-2	-1	0
1	1	0	0	1	2	1	2

□

BIBLIOGRAPHY

Alberto Abadie. Bootstrap tests for distributional treatment effects in instrumental variable models. *Journal of the American statistical Association*, 97(457):284–292, 2002.

Milton Abramowitz and Irene A Stegun. *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables*. Number 55. Courier Dover Publications, 1972.

Rohit Agrawal. Finite-sample concentration of the multinomial in relative entropy. *IEEE Transactions on Information Theory*, 66(10):6297–6302, 2020.

Takeshi Amemiya. *Advanced Econometrics*. Harvard University Press, 1985.

Erling D. Andersen and Knud D. Andersen. The MOSEK interior point optimizer for linear programming: an implementation of the homogeneous algorithm. In *High performance optimization*, pages 197–232. Springer, 2000.

Theodore W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley New York, 2nd edition, 1984.

Theodore Wilbur Anderson and Ingram Olkin. Maximum-likelihood estimation of the parameters of a multivariate normal distribution. *Linear algebra and its applications*, 70: 147–171, 1985.

Steen A. Andersson, David Madigan, and Michael D. Perlman. A characterization of Markov equivalence classes for acyclic digraphs. *The Annals of Statistics*, 25:505–541, 1997.

Donald W. K. Andrews. Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica*, pages 399–405, 2000.

Donald W. K. Andrews. Testing when a parameter is on the boundary of the maintained hypothesis. *Econometrica*, 69(3):683–734, 2001.

Donald W. K. Andrews and Xiaoxia Shi. Inference based on conditional moment inequalities. *Econometrica*, 81(2):609–666, 2013.

Joshua D. Angrist. Lifetime earnings and the Vietnam era draft lottery: evidence from social security administrative records. *The American Economic Review*, pages 313–336, 1990.

Joshua D. Angrist. The draft lottery and voluntary enlistment in the Vietnam era. *Journal of the American statistical Association*, 86(415):584–595, 1991.

Joshua D. Angrist and Alan B. Krueger. The effect of age at school entry on educational attainment: an application of instrumental variables with moments from two samples. *Journal of the American statistical Association*, 87(418):328–336, 1992a.

Joshua D. Angrist and Alan B. Krueger. Estimating the payoff to schooling using the Vietnam-era draft lottery. *NBER working paper*, (w4067), 1992b.

Joshua D. Angrist and Alan B. Krueger. Split-sample instrumental variables estimates of the return to schooling. *Journal of Business & Economic Statistics*, 13(2):225–235, 1995.

Joshua D. Angrist and Alan B. Krueger. Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives*, 15(4):69–85, 2001.

Joshua D Angrist and Jörn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, 2008.

Alexander Balke and Judea Pearl. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1171–1176, 1997.

Ole Barndorff-Nielsen. *Information and Exponential Families: in Statistical Theory*. John Wiley & Sons, 2014.

John S. Bell. On the Einstein Podolsky Rosen paradox. *Physics Physique Fizika*, 1:195–200, Nov 1964.

Roger L. Berger and Dennis D. Boos. P values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association*, 89(427):1012–1016, 1994.

Dimitri P. Bertsekas, Angelia Nedi, and Asuman E. Ozdaglar. *Convex Analysis and Optimization*. Athena Scientific, 2003.

B. C. Bhattacharyya. The use of McKay's Bessel function curves for graduating frequency distributions. *Sankhyā: The Indian Journal of Statistics*, pages 175–182, 1942.

Nan Bi, Hyunseung Kang, and Jonathan Taylor. Inference after selecting plausibly valid instruments with application to mendelian randomization. *arXiv preprint arXiv:1911.03985*, 2019.

Gérard Biau and Laszlo Gyorfi. On the asymptotic properties of a nonparametric l_1-test statistic of homogeneity. *IEEE Transactions on Information Theory*, 51(11):3965–3973, 2005.

Peter J. Bickel, Chris A. J. Klaassen, Ya'acov Ritov, and Jon A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*, volume 4. Johns Hopkins University Press, Baltimore, 1993.

Peter M. Blau and Otis Dudley Duncan. *The American Occupational Structure*. Wiley New York, 1967.

Kenneth A. Bollen. *Structural Equations with Latent Variables*. Wiley, New York, 1989.

Blai Bonet. Instrumentality tests revisited. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann, 2001.

Djalil Chafai and Didier Concordet. Confidence regions for the multinomial parameter with small sample size. *Journal of the American Statistical Association*, 104(487):1071–1079, 2009.

Gary Chamberlain. Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of econometrics*, 34(3):305–334, 1987.

Sanjay Chaudhuri, Mathias Drton, and Thomas S. Richardson. Estimation of a covariance matrix with zeros. *Biometrika*, 94(1):199–216, 2007.

Wenyu Chen, Mathias Drton, and Y. Samuel Wang. On causal discovery with an equal-variance assumption. *Biometrika*, 106(4):973–980, 2019.

Victor Chernozhukov, Sokbae Lee, and Adam M. Rosen. Intersection bounds: estimation and inference. *Econometrica*, 81(2):667–737, 2013.

David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

John F. Clauser, Michael A. Horne, Abner Shimony, and Richard A. Holt. Proposed experiment to test local hidden-variable theories. *Phys. Rev. Lett.*, 23:880–884, Oct 1969.

Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. John Wiley and Sons, 2nd edition, 2006.

Noel Cressie and Timothy R. C. Read. Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 46(3):440–464, 1984.

Imre Csiszár. The method of types. *IEEE Transactions on Information Theory*, 44(6): 2505–2523, 1998.

A. Philip Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, pages 1–31, 1979.

A. Philip Dawid. Some matrix-variate distribution theory: notational considerations and a Bayesian application. *Biometrika*, 68(1):265–274, 1981.

Luc Devroye. The equivalence of weak, strong and complete convergence in ℓ_1 for kernel density estimates. *The Annals of Statistics*, 11(3):896–904, 1983.

DLMF. *NIST Digital Library of Mathematical Functions*. Release 1.0.25 of 2019-12-15. URL http://dlmf.nist.gov/. F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, B. V. Saunders, H. S. Cohl, and M. A. McClain, eds.

Mathias Drton. Algebraic techniques for Gaussian models. In M. Hušková and M. Janžura, editors, *Prague Stochastics*. Matfyzpress, Charles Univ., 2006a.

Mathias Drton. Computing all roots of the likelihood equations of seemingly unrelated regressions. *Journal of Symbolic Computation*, 41(2):245–254, 2006b.

Mathias Drton. Discrete chain graph models. *Bernoulli*, 15(3):736–753, 2009a.

Mathias Drton. Likelihood ratio tests and singularities. *The Annals of Statistics*, 37(2): 979–1012, 2009b.

Mathias Drton. Algebraic problems in structural equation modeling. In *The 50th Anniversary of Gröbner Bases*, pages 35–86. Mathematical Society of Japan, 2018.

Mathias Drton and Michael Eichler. Maximum likelihood estimation in Gaussian chain graph models under the alternative Markov property. *Scandinavian Journal of Statistics*, 33(2): 247–257, 2006.

Mathias Drton and Marloes H. Maathuis. Structure learning in graphical modeling. *Annual Review of Statistics and Its Application*, 4:365–393, 2017.

Mathias Drton and Michael D. Perlman. Model selection for Gaussian concentration graphs. *Biometrika*, 91(3):591–602, 2004.

Mathias Drton and Thomas S. Richardson. Multimodality of the likelihood in the bivariate seemingly unrelated regressions model. *Biometrika*, 91(2):383–392, 2004.

Mathias Drton and Seth Sullivant. Algebraic statistical models. *Statistica Sinica*, pages 1273–1297, 2007.

Mathias Drton, Bernd Sturmfels, and Seth Sullivant. *Lectures on algebraic statistics*, volume 39. Springer Science & Business Media, 2008.

Mathias Drton, Michael Eichler, and Thomas S. Richardson. Computing maximum likelihood estimates in recursive linear models with correlated errors. *Journal of Machine Learning Research*, 10(81):2329–2348, 2009.

Mathias Drton, Rina Foygel, and Seth Sullivant. Global identifiability of linear structural equation models. *The Annals of Statistics*, 39(2):865–886, 2011.

Iain Dunning, Joey Huchette, and Miles Lubin. JuMP: A modeling language for mathematical optimization. *SIAM Review*, 59(2):295–320, 2017.

Marco Eigenmann, Preetam Nandy, and Marloes H. Maathuis. Structure learning of linear Gaussian structural equation models with weak edges. In *Proceedings of the 33rd Annual Conference on Uncertainty in Artificial Intelligence (UAI-17)*, 2017.

Robin J. Evans. Model selection and local geometry. *The Annals of Statistics*, 48(6):3513 – 3544, 2020.

Robin J. Evans and Thomas S. Richardson. Markovian acyclic directed mixed graphs for discrete data. *The Annals of Statistics*, pages 1452–1482, 2014.

Yanqin Fan and Sang Soo Park. Nonparametric inference for counterfactual means: Bias-correction, confidence sets, and weak IV. *Journal of Econometrics*, 178:45–56, 2014.

Zhuangyan Fang and Yangbo He. IDA with background knowledge. In *Proceedings of the 36th Annual Conference on Uncertainty in Artificial Intelligence (UAI-20)*, 2020.

Ronald A. Fisher. The distribution of the partial correlation coefficient. *Metron*, 3:329–332, 1924.

Ronald A. Fisher, A. Steven Corbet, and Carrington B. Williams. The relation between the number of species and the number of individuals in a random sample of an animal population. *The Journal of Animal Ecology*, pages 42–58, 1943.

Gerald B. Folland. *Real Analysis: Modern Techniques and Their Applications*. Wiley & Sons, 2nd edition, 1999.

Morten Frydenberg and Jens Ledet Jensen. Is the 'improved likelihood ratio statistic' really improved in the discrete case? *Biometrika*, 76(4):655–661, 1989.

Anqi Fu, Balasubramanian Narasimhan, and Stephen Boyd. CVXR: An R package for disciplined convex optimization. *Journal of Statistical Software*, 94(14):1–34, 2020.

Komei Fukuda. cddlib reference manual. Technical report, McGill University, Montréal, Quebec, Canada, 2007. URL http://search.r-project.org/library/rcdd/doc/cddlibman.pdf.

Charles J. Geyer. On the asymptotics of constrained M-estimation. *The Annals of Statistics*, 22(4):1993–2010, 1994.

Richard D. Gill. Statistics, causality and Bell's theorem. *Statistical Science*, 29(4):512–528, 2014.

Irving J. Good and George H. Toulmin. The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika*, 43(1-2):45–63, 1956.

Ronald L Graham, Donald E. Knuth, and Oren Patashnik. *Concrete Mathematics. A Foundation for Computer Science*. Addison-Wesley, Reading, Mass. USA, 1994.

Richard Gray and Keith Wheatley. How to avoid bias when comparing bone marrow transplantation with chemotherapy. *Bone Marrow Transplantation*, 7:9–12, 1991.

Florian F. Gunsilius. Nontestability of instrument validity under continuous treatments. *Biometrika*, 12 2020.

F. Richard Guo and Emilija Perković. Minimal enumeration of all possible total effects in a markov equivalence class. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 2395–2403. PMLR, 13–15 Apr 2021.

F. Richard Guo and Thomas S. Richardson. On testing marginal versus conditional independence. *Biometrika*, 107(4):771–790, 07 2020.

F. Richard Guo and Thomas S. Richardson. Chernoff-type concentration of empirical probabilities in relative entropy. *IEEE Transactions on Information Theory*, 67(1):549–558, 2021.

Shantanu Gupta, Zachary C. Lipton, and David Childers. Estimating treatment effects with observed confounders and mediators. *arXiv preprint arXiv:2003.11991*, 2020.

Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, pages 1029–1054, 1982.

Alan Hauser and Peter Bühlmann. Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *Journal of Maching Learning Research*, 13:2409–2464, 2012.

Leonard Henckel, Emilija Perković, and Marloes H. Maathuis. Graphical criteria for efficient total effect estimation via adjustment in causal linear models. *arXiv preprint arXiv:1907.02435*, 2019.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2nd edition, 2012.

Harold Hotelling. New light on the correlation coefficient and its transforms. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 15(2):193–232, 1953.

Patrik O. Hoyer, Aapo Hyvärinen, Richard Scheines, Peter L. Spirtes, Joseph Ramsey, Gustavo Lacerda, and Shohei Shimizu. Causal discovery of linear acyclic models with arbitrary distributions. In *Proceedings of the 24th Annual Conference on Uncertainty in Artificial Intelligence (UAI-08)*, pages 282–289, 2008.

Guido W. Imbens and Joshua D. Angrist. Identification and estimation of local average treatment effects. *Econometrica*, pages 467–475, 1994.

Guido W. Imbens and Charles F. Manski. Confidence intervals for partially identified parameters. *Econometrica*, 72(6):1845–1857, 2004.

Guido W. Imbens and Donald B. Rubin. Bayesian inference for causal effects in randomized experiments with noncompliance. *The Annals of Statistics*, pages 305–327, 1997a.

Guido W. Imbens and Donald B. Rubin. Estimating outcome distributions for compliers in instrumental variables models. *The Review of Economic Studies*, 64(4):555–574, 1997b.

Leah Jager and Jon A. Wellner. Goodness-of-fit tests via phi-divergences. *The Annals of Statistics*, 35(5):2018–2053, 2007.

Jiantao Jiao, Kartik Venkat, Yanjun Han, and Tsachy Weissman. Maximum likelihood estimation of functionals of discrete distributions. *IEEE Transactions on Information Theory*, 63(10):6774–6798, 2017.

Norman L. Johnson, Samuel Kotz, and N. Balakrishnan. *Continuous Univariate Distributions*, volume 1 of *Wiley series in probability and mathematical statistics: Applied probability and statistics*. Wiley & Sons, 1995.

Markus Kalisch, Martin Mächler, Diego Colombo, Marloes H. Maathuis, and Peter

Bühlmann. Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software*, 47(11):1–26, 2012.

Désiré Kédagni and Ismael Mourifié. Generalized instrumental inequalities: testing the instrumental variable independence assumption. *Biometrika*, 107(3):661–675, 2020.

Daphne Koller, Nir Friedman, and Francis Bach. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

Tjalling C. Koopmans and Olav Reiersøl. The identification of structural characteristics. *The Annals of Mathematical Statistics*, 21(2):165–181, 1950.

Manabu Kuroki and Zhihong Cai. Selection of identifiability criteria for total effects by using path diagrams. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence (UAI-04)*, pages 333–340, 2004.

Manabu Kuroki and Masami Miyakawa. Covariate selection for estimating the causal effect of control plans by using causal diagrams. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1):209–222, 2003.

Manabu Kuroki and Hisayoshi Nanmo. Variance formulas for estimated mean response and predicted response with external intervention based on the back-door criterion in linear structural equation models. *AStA Advances in Statistical Analysis*, pages 1–19, 2020.

Steffen L. Lauritzen. *Graphical Models*. Oxford University Press, New York, 1996.

Benoît Legat, Robin Deits, Daisuke Oyama, Marcelo Forets, Mathieu Besançon, Sascha Timme, Elliot Saba, Julia TagBot, Sebastián Guadalupe, and François Pacaud. Juliapolyhedra/cddlib.jl: v0.6.3, November 2020. URL https://doi.org/10.5281/zenodo.4264518.

Benoît Legat, Robin Deits, Marcelo Forets, Oliver Evans, Gustavo Goretkin, Daisuke Oyama, Joey Huchette, Twan Koolen, Chase Coleman, Elliot Saba, Henrique Ferrolho, Robert

Schwarz, and Guillaume Berger. Juliapolyhedra/polyhedra.jl: v0.6.13, February 2021. URL https://doi.org/10.5281/zenodo.4498311.

Erich L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses*. Springer-Verlag New York, 2006.

Jing Lei. Classification with confidence. *Biometrika*, 101(4):755–769, 2014.

Jianzhou Liu. Some Löwner partial orders of Schur complements and Kronecker products of matrices. *Linear Algebra and its Applications*, 291(1-3):143–149, 1999.

Marloes H. Maathuis and Diego Colombo. A generalized back-door criterion. *The Annals of Statistics*, 43(3):1060–1088, 2015.

Marloes H. Maathuis, Markus Kalisch, and Peter Bühlmann. Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, 37(6A):3133–3164, 2009.

Matthew L Malloy, Ardhendu Tripathy, and Robert D. Nowak. Optimal confidence regions for the multinomial parameter. *arXiv preprint arXiv:2002.01044*, 2020.

Charles F. Manski. Nonparametric bounds on treatment effects. *The American Economic Review*, 80(2):319–323, 1990.

Daniel Marbach, Thomas Schaffter, Dario Floreano, Robert J. Prill, and Gustavo Stolovitzky. The DREAM4 in-silico network challenge. Draft, version 0.3 http://gnw.sourceforge.net/resources/DREAM4%20in%20silico%20challenge.pdf, 2009a.

Daniel Marbach, Thomas Schaffter, Claudio Mattiussi, and Dario Floreano. Generating realistic in silico gene networks for performance assessment of reverse engineering methods. *Journal of Computational Biology*, 16(2):229–239, 2009b.

Jay Mardia, Jiantao Jiao, Ervin Tánczos, Robert D. Nowak, and Tsachy Weissman. Concentration inequalities for the empirical distribution of discrete distributions: beyond the

method of types. *Information and Inference: A Journal of the IMA*, 11 2019. ISSN 2049-8772.

Daniel McFadden. Testing for stochastic dominance. In *Studies in the Economics of Uncertainty*, pages 113–134. Springer, 1989.

Christopher Meek. Causal inference and causal explanation with background knowledge. In *Proceedings of the 11th Annual Conference on Uncertainty in Artificial Intelligence (UAI-95)*, pages 403–410, 1995.

Edward Miguel, Shanker Satyanath, and Ernest Sergenti. Economic shocks and civil conflict: an instrumental variables approach. *Journal of political Economy*, 112(4):725–753, 2004.

Preetam Nandy, Marloes H. Maathuis, and Thomas S. Richardson. Estimating the effect of joint interventions from observational data in sparse high-dimensional settings. *The Annals of Statistics*, 45(2):647–674, 2017.

Robert Nowak and Ervin Tánczos. Tighter confidence intervals for rating systems. *arXiv preprint arXiv:1912.03528*, 2019.

Alon Orlitsky, Ananda Theertha Suresh, and Yihong Wu. Optimal prediction of the number of unseen species. *Proceedings of the National Academy of Sciences*, 113(47):13283–13288, 2016.

Liam Paninski. Estimation of entropy and mutual information. *Neural computation*, 15(6): 1191–1253, 2003.

Judea Pearl. Comment: graphical models, causality and intervention. *Statistical Science*, 8 (3):266–269, 1993.

Judea Pearl. On the testability of causal models with latent and instrumental variables. In P. Besnard and S. Hanks, editors, *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann, 1995a.

Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995b.

Judea Pearl. *Causality*. Cambridge University Press, Cambridge, 2nd edition, 2009.

Judea Pearl and Thomas S. Verma. A theory of inferred causation. In *Studies in Logic and the Foundations of Mathematics*, volume 134, pages 789–811. Elsevier, 1995.

Emilija Perković. Identifying causal effects in maximally oriented partially directed acyclic graphs. In *Proceedings of the 36th Annual Conference on Uncertainty in Artificial Intelligence (UAI-20)*, 2020.

Emilija Perković, Johannes Textor, Markus Kalisch, and Marloes H. Maathuis. A complete generalized adjustment criterion. In *Proceedings of the 31st Annual Conference on Uncertainty in Artificial Intelligence (UAI-15)*, 2015.

Emilija Perković, Markus Kalisch, and Marloes H. Maathuis. Interpreting and using CPDAGs with background knowledge. In *Proceedings of the 33rd Annual Conference on Uncertainty in Artificial Intelligence (UAI-17)*, 2017.

Emilija Perković, Johannes Textor, Markus Kalisch, and Marloes H. Maathuis. Complete graphical characterization and construction of adjustment sets in Markov equivalence classes of ancestral graphs. *Journal of Machine Learning Research*, 18(220):1–62, 2018.

Michael D. Perlman and Lang Wu. The emperor's new tests. *Statistical Science*, 14(4): 355–369, 1999.

Jonas Peters and Peter Bühlmann. Identifiability of Gaussian structural equation models with equal error variances. *Biometrika*, 101(1):219–228, 2014.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL https://www.R-project.org/.

Roland R. Ramsahai and Steffen L. Lauritzen. Likelihood analysis of the binary instrumental variable model. *Biometrika*, 98(4):987–994, 2011.

Hans Reichenbach. *The Direction of Time*. Dover Publications, 1956.

Thomas S. Richardson and James M. Robins. Single world intervention graphs (SWIGs): a unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, 128, 2013.

Thomas S. Richardson and James M. Robins. ACE Bounds; SEMs with Equilibrium Conditions. *Statistical Science*, 29(3):363 – 366, 2014.

Thomas S. Richardson and Peter Spirtes. Ancestral graph Markov models. *The Annals of Statistics*, 30(4):962 – 1030, 2002.

Thomas S. Richardson, Robin J. Evans, and James M. Robins. Transparent parameterizations of models for potential outcomes. *Bayesian Statistics*, 9:569–610, 2011.

Thomas S. Richardson, Robin J. Evans, James M. Robins, and Ilya Shpitser. Nested Markov properties for acyclic directed mixed graphs. *arXiv preprint arXiv:1701.06686*, 2017.

James M. Robins. A new approach to causal inference in mortality studies with a sustained exposure period-application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7:1393–1512, 1986.

James M. Robins. The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. *Health Service Research Methodology: a Focus on AIDS*, pages 113–159, 1989.

James M Robins, Richard Scheines, Peter Spirtes, and Larry Wasserman. Uniform consistency in causal inference. *Biometrika*, 90(3):491–515, 2003.

Dominik Rothenhäusler, Jan Ernest, and Peter Bühlmann. Causal inference in partially linear structural equation models. *The Annals of Statistics*, 46(6A):2904 – 2938, 2018.

Andrea Rotnitzky and Ezequiel Smucler. Efficient adjustment sets for population average causal treatment effect estimation in graphical models. *Journal of Machine Learning Research*, 21(188):1–86, 2020.

John D. Sargan. The estimation of economic relationships using instrumental variables. *Econometrica: Journal of the Econometric Society*, pages 393–415, 1958.

Richard Scheines, Peter Spirtes, Clark Glymour, Christopher Meek, and Thomas Richardson. The TETRAD project: constraint based aids to causal model specification. *Multivariate Behavioral Research*, 33(1):65–117, 1998.

Rajen D. Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514–1538, 06 2020.

Shohei Shimizu. LiNGAM: Non-Gaussian methods for estimating causal structures. *Behaviormetrika*, 41(1):65–98, 2014.

Shohei Shimizu, Patrik O. Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7 (Oct):2003–2030, 2006.

Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvärinen, Yoshinobu Kawahara, Takashi Washio, Patrik O. Hoyer, and Kenneth Bollen. DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. *Journal of Machine Learning Research*, 12:1225–1248, 2011.

Galen R. Shorack. *Probability for Statisticians*. Springer, 2000.

Ilya Shpitser, Tyler VanderWeele, and James M. Robins. On the validity of covariate adjustment for estimating causal effects. In *Proceedings of the 26th Annual Conference on Uncertainty in Artificial Intelligence (UAI-10)*, pages 527–536, 2010.

Ilya Shpitser, Thomas S. Richardson, and James M. Robins. Multivariate counterfactual systems and causal graphical models. *arXiv preprint arXiv:2008.06017*, 2020.

Marvin K. Simon. *Probability Distributions involving Gaussian Random Variables: A Handbook for Engineers and Scientists*. Springer Science & Business Media, 2007.

Ezequiel Smucler, Facundo Sapienza, and Andrea Rotnitzky. Efficient adjustment sets in causal graphical models with hidden variables. *arXiv preprint arXiv:2004.10521*, 2020.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, Cambridge, MA, 2nd edition, 2000.

James H. Stock, Jonathan H. Wright, and Motohiro Yogo. A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics*, 20(4):518–529, 2002.

Robert H. Strotz and Herman O. A. Wold. Recursive vs. nonrecursive systems: An attempt at synthesis (part I of a triptych on causal chain systems). *Econometrica*, 28(2):417–427, 1960.

Seth Sullivant, Kelli Talaska, and Jan Draisma. Trek separation for Gaussian graphical models. *The Annals of Statistics*, 38(3):1665–1685, 2010.

Sonja A. Swanson, Miguel A. Hernán, Matthew Miller, James M. Robins, and Thomas S. Richardson. Partial identification of the average treatment effect using instrumental variables: review of methods for binary instruments, treatments, and outcomes. *Journal of the American Statistical Association*, 113(522):933–947, 2018.

Anastasios Tsiatis. *Semiparametric Theory and Missing Data*. Springer, New York, 2006.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer-Verlag, 2009.

Mark J. van der Laan and Sherri Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, New York, 2011.

Aad van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 2000.

Aad van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes: with Applications to Statistics*. Springer, 1996.

Stijn Vansteelandt, Els Goetghebeur, Michael G. Kenward, and Geert Molenberghs. Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. *Statistica Sinica*, pages 953–979, 2006.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. doi: 10.1017/9781108231596.

Ramya Korlakai Vinayak, Weihao Kong, Gregory Valiant, and Sham Kakade. Maximum likelihood estimation for learning populations of parameters. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6448–6457, Long Beach, California, USA, 2019.

Linbo Wang, James M. Robins, and Thomas S. Richardson. On falsification of the binary instrumental variable model. *Biometrika*, 104(1):229–236, 2017a.

Yuhao Wang, Liam Solus, Karren Dai Yang, and Caroline Uhler. Permutation-based causal inference algorithms with interventions. In *Advances in Neural Information Processing Systems 30*, pages 5822–5831. 2017b.

Janine Witte, Leonard Henckel, Marloes H. Maathuis, and Vanessa Didelez. On efficient adjustment in causal graphs. *Journal of Machine Learning Research*, 21(246):1–45, 2020.

Philip G. Wright. *Tariff on animal and vegetable oils*. Macmillan Company, New York, 1928.

Sewall Wright. Correlation and causation. *Journal of Agricultural Research*, 20:557–585, 1921.

Sewall Wright. The method of path coefficients. *The Annals of Mathematical Statistics*, 5 (3):161–215, 1934.

Appendix A

APPENDIX TO CHAPTER 2

A.1 Asymptotic distribution of LRT

Proof of Proposition 2.3. For convenience, we assume the form of (sub)sequences of $\rho_{13,n}$ and $\rho_{23,n}$ as

$$\rho_{13,n} = \eta n^{-a}, \quad \rho_{23,n} = \tau n^{-(1/2-a)}$$

for $a \in (0, 1/2)$ and $\eta\tau = \delta$. We perform a manual change of measure by relating the law under $P_{\Sigma_n^{(i)}}$ to that under P_I, which is independent sampling of $\mathcal{N}(0, I)$. Under sample size n, suppose Ω_n is the sample covariance under $\mathcal{N}(0, I)$. Now suppose $S_n^{(i)}$ is the sample covariance under $P_{\Sigma_n^{(i)}}$ for $i = 0, 1$. Then it holds that

$$S_n^{(i)} =_d L_n^{(i)} \Omega_n L_n^{(i)\top}, \tag{A.1}$$

for some $L_n^{(i)}$ such that $\Sigma_n^{(i)} = L_n^{(i)} L_n^{(i)\top}$. Here we choose them as the Cholesky decompositions

$$L_n^{(0)} = \begin{pmatrix} \sigma_{11}^{1/2} & 0 & 0 \\ 0 & \sigma_{22}^{1/2} & 0 \\ n^{-a}\eta\sigma_{33}^{1/2} & n^{a-\frac{1}{2}}\tau\sigma_{33}^{1/2} & (1 - \eta^2 n^{-2a} - \tau^2 n^{2a-1})^{1/2}\, \sigma_{33}^{1/2} \end{pmatrix}$$

and

$$L_n^{(1)} = \begin{pmatrix} \sigma_{11}^{1/2} & 0 & 0 \\ \eta\tau n^{-1/2}\sigma_{22}^{1/2} & (1 - \eta^2\tau^2/n)^{1/2}\, \sigma_{22}^{1/2} & 0 \\ n^{-a}\eta\sigma_{33}^{1/2} & (n^a - \gamma^2 n^{-a})\, \tau (n - \eta^2\tau^2)^{-1/2}\sigma_{33}^{1/2} & (1 - \eta^2 n^{-2a})^{1/2}\, (n^{2a}\tau^2 - n)^{1/2}\, (\eta^2\tau^2 - n)^{-1/2}\sigma_{33}^{1/2} \end{pmatrix}.$$

By the central limit theorem, we have

$$n^{1/2}(\Omega_n - I) \to_d W \tag{A.2}$$

for W a 3×3 matrix of joint Gaussian variables whose covariance is determined by the Isserlis matrix. The asymptotic distribution of $\lambda_n^{(0:1)}$ can be obtained by substituting

$$S_n^{(i)} = L_n^{(i)} \left\{ I + n^{-1/2} W + o_p(n^{-1/2}) \right\} L_n^{(i)\top} \tag{A.3}$$

into the closed-form expression of Eq. (2.13) in the main text and simplifying. We have under $\Sigma_n^{(0)}$

$$\lambda_n^{(0:1)} = \gamma\tau(\gamma\tau - 2w_{12}) + o_p(1), \tag{A.4}$$

and under $\Sigma_n^{(1)}$

$$\lambda_n^{(0:1)} = -\gamma\tau(\gamma\tau + 2w_{12}) + o_p(1). \tag{A.5}$$

The result is immediate from $w_{12} \sim \mathcal{N}(0,1)$ and $\gamma\tau = \delta$. □

Proof of Proposition 2.6. Under $P^n_{\Sigma^* + n^{-1/2}\mathsf{G}\,h}$, by van der Vaart (2000, Theorem 16.7) $\lambda_n^{(0:1)}$ is asymptotically distributed as the loglikelihood ratio statistic for testing H_0 and H_1 based on a single sample from $\mathcal{N}(h, I_{\Sigma^*}^{-1})$. The theorem still applies to our case even though H_0 and H_1 are non-nested, as its proof does not require the two models to be nested. That is, given $X \sim \mathcal{N}(m = 0, I_{\Sigma^*}^{-1})$, we have

$$\begin{aligned} \lambda_n^{(0:1)} &\to_d \| I_{\Sigma^*}^{1/2}(X + h) - I_{\Sigma^*}^{1/2} H_0 \|^2 - \| I_{\Sigma^*}^{1/2}(X + h) - I_{\Sigma^*}^{1/2} H_1 \|^2 \\ &=_d \| I_{\Sigma^*}^{1/2} X - I_{\Sigma^*}^{1/2}(H_0 - h) \|^2 - \| I_{\Sigma^*}^{1/2} X - I_{\Sigma^*}^{1/2}(H_1 - h) \|^2, \end{aligned} \tag{A.6}$$

which is equivalent to testing $m \in H_0 - h$ versus $m \in H_1 - h$ from X. Given $I_{\Sigma^*}^{-1} = LL^\top$, by rewriting $X =_d LZ$ for $Z \sim \mathcal{N}(\mu = \mathbf{0}, I_2)$, the testing problem is mapped to that from Z by L^{-1}. Hence, this is further equivalent to testing

$$\mu \in L^{-1}(H_0 - h) \quad \text{versus} \quad \mu \in L^{-1}(H_1 - h)$$

from Z. Note that $H_i - h_i = H_i$ since H_i is affine. □

Proof of Proposition 2.2. Since the limit experiments are of the same type, we only derive for local alternatives $\Sigma_n^{(1)} \in \mathcal{M}_1 \setminus \mathcal{M}_0$. We set the coordinate system as in Fig. A.1, where

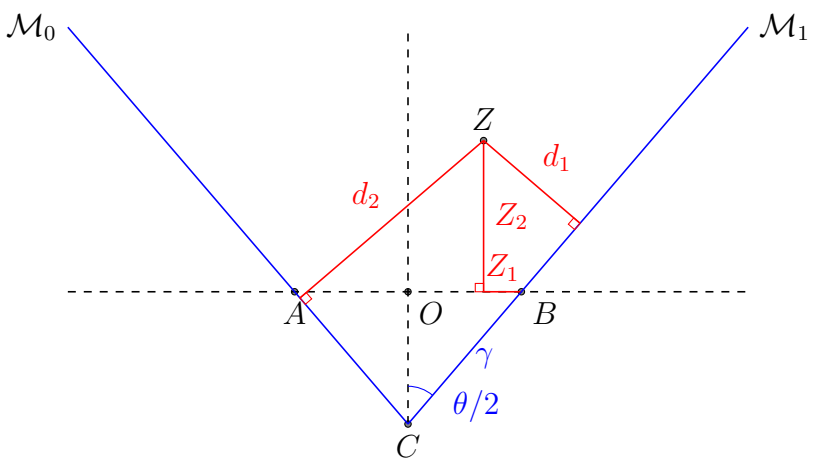


Figure A.1: Derivation of the asymptotic distribution Eq. (2.24) from the limit experiment of $\mathcal{M}_1 \setminus \mathcal{M}_0$ under the weak-strong regime (the middle panel of Fig. 2.6).

the bisector of angle $\angle BCA = \theta = \arcsin\rho$ is the y-axis. The standard Gaussian vector centered at B is represented as $Z = (x, y) = (\gamma\sin(\theta/2) - Z_1, Z_2)$. By the limit experiment, we have $\lambda_n^{(0:1)} \to_d d_1^2 - d_2^2$. $\mathcal{M}_0$ and $\mathcal{M}_1$ are respectively represented by lines $y = \pm kx + a$ for $k = \cot(\theta/2)$ and $a = -\gamma\cos(\theta/2)$. We have

$$\begin{aligned} d_1^2 - d_2^2 &= \frac{(a + kx - y)^2}{1 + k^2} - \frac{(a - kx - y)^2}{1 + k^2} \\ &= 2\rho(Z_1 - \gamma\sin(\theta/2))(Z_2 + \gamma\cos(\theta/2)), \end{aligned}$$

where we used

$$\frac{2k}{1 + k^2} = \frac{2\cot(\theta/2)}{1 + \cot^2(\theta/2)} = \sin\theta = \rho.$$

By a change of variables $(Z_1, Z_2) =_d ((U + V)/\sqrt{2}, (U - V)/\sqrt{2})$ for another pair of independent standard normals and using the fact

$$\sqrt{1 + \sqrt{1 - \rho^2}} - \sqrt{1 - \sqrt{1 - \rho^2}} = \sqrt{2(1 - \rho)},$$

upon simplifying we have

$$\lambda_n^{(0:1)} \to_d d_1^2 - d_2^2 =_d \rho \left[\left(U + \gamma \sqrt{\frac{1-\rho}{2}} \right)^2 - \left(V + \gamma \sqrt{\frac{1+\rho}{2}} \right)^2 \right].$$

□

A.2 Proofs of envelope distributions

Proof of Proposition 2.11. Firstly, we note that $\bar{F}_\rho(0) = \bar{G}(0) = 1/2$ and by Proposition 2.9 the non-negative part of $\bar{F}_\rho$ also converges to that of $\bar{G}$ as $\rho \to 0$, namely a point mass at zero. It remains to be shown that the negative part of $\bar{F}_\rho$ converges in law to the negative part of $\bar{G}$. It suffices to show for any $x \leq 0$

$$\sup_\gamma \Pr\left(\rho \left[\left\{ Z_1 + \gamma \left(\frac{1+\rho}{2} \right)^{1/2} \right\}^2 - \left\{ Z_2 + \gamma \left(\frac{1-\rho}{2} \right)^{1/2} \right\}^2 \right] \leq x \right) \to \Pr(-Z^2 \leq x)/2$$

as $\rho \to 0$. Given $\rho > 0$, the maximized probability can be rewritten as

$$\begin{aligned}
&\sup_\gamma \Pr\left(\rho \left[\left\{ Z_1 + \gamma \left(\frac{1+\rho}{2} \right)^{1/2} \right\}^2 - \left\{ Z_2 + \gamma \left(\frac{1-\rho}{2} \right)^{1/2} \right\}^2 \right] \leq x \right) \\
&= \sup_\gamma \Pr\left((\gamma\rho)^2 + 2\gamma\rho \left\{ \left(\frac{1+\rho}{2} \right)^{1/2} Z_1 - \left(\frac{1-\rho}{2} \right)^{1/2} Z_2 \right\} \leq x - \rho(Z_1^2 - Z_2^2) \right) \\
&= \sup_\delta \Pr\left(\delta^2 + 2\delta \left\{ \left(\frac{1+\rho}{2} \right)^{1/2} Z_1 - \left(\frac{1-\rho}{2} \right)^{1/2} Z_2 \right\} + \rho(Z_1^2 - Z_2^2) \leq x \right) \\
&= \sup_\delta \Pr\left\{ X_\rho(\delta) \leq x \right\},
\end{aligned}$$

where we define

$$X_\rho(\delta) = \delta^2 + 2\delta \left\{ \left(\frac{1+\rho}{2} \right)^{1/2} Z_1 - \left(\frac{1-\rho}{2} \right)^{1/2} Z_2 \right\} + \rho(Z_1^2 - Z_2^2)$$

for $\rho \in [0, 1)$ and $\delta \in \mathbb{R}$. Note that $\sup_\delta \Pr(X_0(\delta) \leq x) = \sup_\delta \Pr(\delta^2 + 2\delta Z \leq x) = \Pr(-Z^2 \leq x)/2$ for $Z \sim \mathcal{N}(0, 1)$ by Proposition 2.7. We are left to show $\sup_\delta \Pr\left(X_\rho(\delta) \leq x \right) \to$

$\sup_\delta \Pr\left(X_0(\delta) \leq x\right)$ as $\rho \to 0$. Choose $x < M < \infty$ and define

$$Y_\rho(\delta) := \begin{cases} X_\rho(\delta), & X_\rho(\delta) \leq M \\ M, & X_\rho(\delta) > M \end{cases}.$$

We observe that

$$\begin{aligned} &\left|\sup_\delta \Pr\left(X_\rho(\delta) \leq x\right) - \sup_\delta \Pr\left(X_0(\delta) \leq x\right)\right| \\ &= \left|\sup_\delta \Pr\left(Y_\rho(\delta) \leq x\right) - \sup_\delta \Pr\left(Y_0(\delta) \leq x\right)\right| \\ &\leq \sup_\delta \left|\Pr\left(Y_\rho(\delta) \leq x\right) - \Pr\left(Y_0(\delta) \leq x\right)\right| \\ &= \sup_\delta \left|\mathbb{E}\left(\mathbb{I}_{Y_\rho(\delta)\leq x} - \mathbb{I}_{Y_0(\delta)\leq x}\right)\right| \to 0, \end{aligned}$$

where the last step follows from weak convergence $\{Y_\rho(\delta) : \delta \in \mathbb{R}\} \rightsquigarrow \{Y_0(\delta) : \delta \in \mathbb{R}\}$ in $\ell^\infty(\mathbb{R})$ as $\rho \to 0$ for a bounded stochastic process; see van der Vaart (2000, Chap. 18). □

Proof of Proposition 2.7. It suffices to consider $\delta \geq 0$. Given any $x < 0$,

$$\sup_\delta \Pr(\delta^2 + 2\delta Z \leq x) = \sup_{\delta>0} \Phi\left(\frac{x-\delta^2}{2\delta}\right) = \sup_{\delta>0} \Phi\left(-\left(\frac{-x}{2\delta} + \frac{\delta}{2}\right)\right) = \Phi(-(-x)^{1/2}),$$

where $\delta^* = (-x)^{1/2}$ is the maximizer; Given any $x \geq 0$, $\delta = 0$ maximizes the probability to one. Hence, the envelope CDF is

$$\bar{G}(x) = \begin{cases} \Phi(-(-x)^{1/2}), & x < 0 \\ 1, & x \geq 0 \end{cases},$$

from which it follows that

$$\bar{g}(x) = \bar{G}'(x) = \frac{1}{2} f_{\chi^2_1}(-x)\mathbb{I}_{x<0} + \frac{1}{2}\delta_0(x).$$

The envelope for $\mathcal{M}_1 \setminus \mathcal{M}_0$ follows from symmetry. □

Proof of Proposition 2.8. Since $\bar{F}_\rho = \bar{F}_{-\rho}$, it suffices to consider $\rho \in (0, 1]$. First consider $\varphi_{\rho,\gamma}(x)$, the density function for $X^2 - Y^2$ with $X \sim \mathcal{N}\left(\mu_1 = \gamma\left(\frac{1-\rho}{2}\right)^{1/2}, 1\right)$ and

$Y \sim \mathcal{N}\left(\mu_2 = \gamma\left(\frac{1+\rho}{2}\right)^{1/2}, 1\right)$ for $\gamma \in \mathbb{R}$ and $\rho \in (0,1]$. Since $p(X^2 - Y^2 = v^2, Y^2 = t) = p(Y^2 = t)p(X^2 = t + v^2)$, the density $\varphi_{\rho,\gamma}$ has the following integral representation from marginalization

$$\begin{aligned}\varphi_{\rho,\gamma}(v^2) &= \int_0^\infty \chi_1^2(t;\mu_2^2)\chi_1^2(v^2+t;\mu_1^2)\,\mathrm{d}t \\ &= \frac{1}{2\pi}\exp\left(-v^2/2 - \gamma^2/2\right)\int_0^\infty \frac{\exp(-t)\cosh\left(\gamma\left\{\frac{(1+\rho)t}{2}\right\}^{1/2}\right)\cosh\left(\gamma\left\{\frac{(1-\rho)(t+v^2)}{2}\right\}^{1/2}\right)}{\{t(t+v^2)\}^{1/2}}\,\mathrm{d}t.\end{aligned}$$

Recall that $\lesssim$ allows for a positive multiplicative constant. Using $\cosh(x) < \exp(x)$ for $x > 0$, we have

$$\begin{aligned}\varphi_{\rho,\gamma}(v^2) &\lesssim \exp(-v^2/2 - \gamma^2/2)\int_0^\infty \frac{e^{-t}\cosh\left(\gamma\left\{\frac{(1+\rho)t}{2}\right\}^{1/2}\right)\cosh\left(\gamma\left\{\frac{(1-\rho)(t+v^2)}{2}\right\}^{1/2}\right)}{\{t(t+v^2)\}^{1/2}}\,\mathrm{d}t \\ &< \exp(-v^2/2)\int_0^\infty \frac{\exp\left(-t - \gamma^2/2 + \gamma\left\{\frac{(1+\rho)t}{2}\right\}^{1/2} + \gamma\left\{\frac{(1-\rho)(t+v^2)}{2}\right\}^{1/2}\right)}{\{t(t+v^2)\}^{1/2}}\,\mathrm{d}t.\end{aligned}$$

We note that

$$-\gamma^2/2 + \gamma\left\{\frac{(1+\rho)t}{2}\right\}^{1/2} + \gamma\left\{\frac{(1-\rho)(t+v^2)}{2}\right\}^{1/2} \le \frac{1}{2}\left[\left\{\frac{(1+\rho)t}{2}\right\}^{1/2} + \left\{\frac{(1-\rho)(t+v^2)}{2}\right\}^{1/2}\right]^2$$

by completing the square in γ. It then follows that

$$\begin{aligned}\varphi_{\rho,\gamma}(v^2) &< \exp(-v^2/2)\int_0^\infty \frac{\exp\left(-t + \frac{1}{2}\left[t + \frac{1-\rho}{2}v^2 + \{(1-\rho^2)t(t+v^2)\}^{1/2}\right]\right)}{\{t(t+v^2)\}^{1/2}}\,\mathrm{d}t \\ &= \exp\left(-\frac{1+\rho}{4}v^2\right)\int_0^\infty \frac{\exp\left(-t/2 + \{(1-\rho^2)t(t+v^2)\}^{1/2}/2\right)}{\{t(t+v^2)\}^{1/2}}\,\mathrm{d}t \\ &\le \exp\left(-\frac{1+\rho}{4}v^2\right)\int_0^\infty \frac{\exp\left(-t/2 + (1-\rho^2)^{1/2}(t+v^2/2)/2\right)}{\{t(t+v^2)\}^{1/2}}\,\mathrm{d}t \\ &= \exp\left(-v^2\{1+\rho-(1-\rho^2)^{1/2}\}/4\right)\int_0^\infty \frac{\exp\left(-\{1-(1-\rho^2)^{1/2}\}t/2\right)}{\{t(t+v^2)\}^{1/2}}\,\mathrm{d}t \\ &= \exp\left(-\rho v^2/4\right)K_0\left(\left\{1-(1-\rho^2)^{1/2}\right\}v^2/4\right),\end{aligned}$$

where we used $2\{t(t+v^2)\}^{1/2} \le 2t + v^2$ in the third line. $K_\nu(\cdot)$ is the modified Bessel function of the second kind, and has the following asymptotic expansion for $z > 0$ (Abramowitz and

Stegun, 1972, Page 378)

$$K_\nu(z) = \left(\frac{\pi}{2z}\right)^{1/2} \exp(-z)\left\{1 + \frac{4\nu^2 - 1}{8z} + O(z^{-2})\right\}.$$

Hence for large v^2, we have

$$\varphi_{\rho,\gamma}(v^2) \lesssim \frac{\exp\left[-v^2\{1+\rho-(1-\rho^2)^{1/2}\}/4\right]}{\{1-(1-\rho^2)^{1/2}\}v^2}.$$

Recall that $\{F_{\rho,\gamma} : \gamma \in \mathbb{R}\}$ is the family of distributions in Eq. (2.23). With $\gamma' = (1-\rho^2)^{1/2}\gamma$,

$$\rho\left[\left\{Z_1 + \gamma\left(\frac{1+\rho}{2}\right)^{1/2}\right\}^2 - \left\{Z_2 + \gamma\left(\frac{1-\rho}{2}\right)^{1/2}\right\}^2\right] \sim F_{\rho,\gamma'}.$$

It follows that the density function

$$f_{\rho,\gamma'}(-v^2) = \varphi_{\rho,\gamma}(v^2/\rho) \lesssim \frac{\rho \exp\left[-v^2\{1+\rho-(1-\rho^2)^{1/2}\}/(4\rho)\right]}{\{1-(1-\rho^2)^{1/2}\}v^2}, \tag{A.7}$$

where the exponent $\left\{1+\rho-(1-\rho^2)^{1/2}\right\}/(4\rho) \in (1/4, 1/2]$. By Definition 2.1, we have

$$\begin{aligned}\bar{F}^*_\rho(-v^2) &= \sup_{\gamma'\in\mathbb{R}} F_{\rho,\gamma'}(-v^2) \\ &= \sup_{\gamma'\in\mathbb{R}} \int_{v^2}^\infty f_{\rho,\gamma'}(-u)\,\mathrm{d}u \lesssim \int_{v^2}^\infty \frac{\rho \exp\left[-u\{1+\rho-(1-\rho^2)^{1/2}\}/(4\rho)\right]}{\{1-(1-\rho^2)^{1/2}\}u}\,\mathrm{d}u < \infty,\end{aligned}$$

and hence $\bar{F}^*_\rho(-v^2) \to 0$ as $v \to \infty$. By Lemma 2.4, $\bar{F}_\rho$ is a distribution function for every $\rho \in (0, 1]$. □

Proof of Proposition 2.9. Fix $\rho \in (0, 1]$ and $v^2 \geq 0$, with $\gamma' = \gamma(1-\rho^2)^{1/2}$ it follows from Proposition 2.2 that

$$1 - F_{\rho,\gamma'}(v^2) = \Pr\left\{(Z_1+\mu_1)^2 - (Z_2+\mu_2)^2 \geq v^2/\rho\right\}, \tag{A.8}$$

where $\mu_1 = \gamma\{(1+\rho)/2\}^{1/2}$, $\mu_2 = \gamma\{(1-\rho)/2\}^{1/2}$. Since $F_{\rho,\gamma'}$ is symmetric in γ, we show $\gamma = 0$ maximizes $F_{\rho,\gamma'}(v^2)$ by showing that the probability on the right hand side of Eq. (A.8) increases in $\gamma \in (0, \infty)$. The probability can be interpreted as the standard Gaussian measure of the hyperbolic set $\{(x, y) : x^2 - y^2 \geq v^2/\rho\}$ with the Gaussian centered

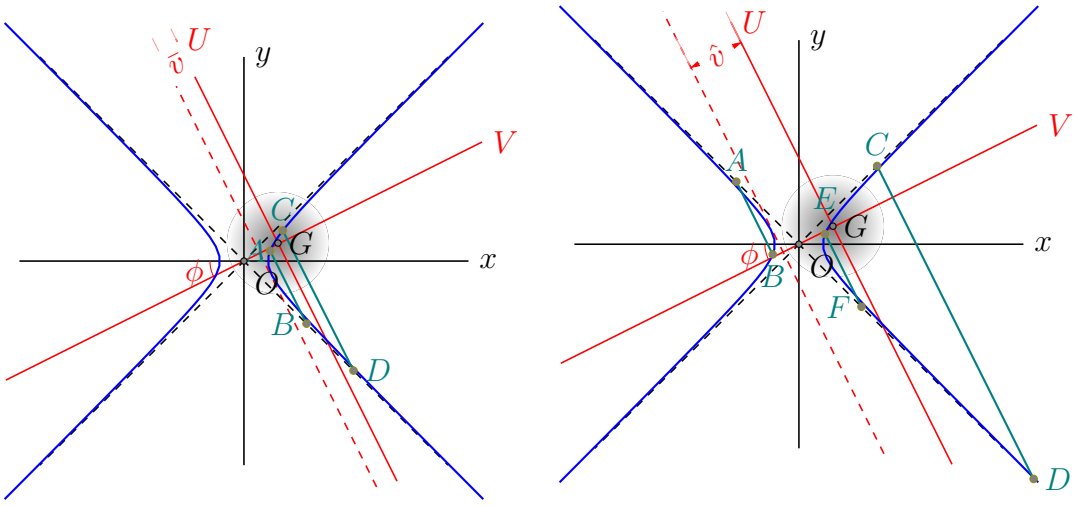


Figure A.2: The distribution function $F_{\rho,\gamma}(\cdot)$ at can be interpreted as the probability of a hyperbolic set (inside the two branches of blue curves) as measured by a standard normal centered $|\gamma|$ away from the origin, lying on the line V with slope $\tan\phi = \{(1-\rho)/(1+\rho)\}^{1/2}$. The asymptotes of the hyperbola are $y = \pm x$.

at $G = (\mu_1, \mu_2) = \gamma(\{(1+\rho)/2\}^{1/2}, \{(1-\rho)/2\}^{1/2})$. This is visualized in Fig. A.2, where $\gamma = \overline{OG}$, $\tan\phi = \{(1-\rho)/(1+\rho)\}^{1/2}$ and the hyperbolic set consists of the area inside the two branches. As γ increases from zero, the center moves away from the origin along the V line. Let U be the line perpendicular to V. The Gaussian measure has two independent standard normal projections (U, V), which is a rotation of (Z_1, Z_2). Now we show that for every $v > 0$, by conditioning on $|V| = v$, the conditional probability of U in the appropriate "section" of the hyperbolic set, denoted by probability $q(v)$, increases with γ.

Let $[A, B]$ and $[C, D]$ be the line segments that $V = -v$ and $V = v$ intersect the hyperbola respectively. By independence of U and V, we have $q(v) = \Pr(U \in [A, B]) + \Pr(U \in [C, D])$. Let $\hat{v}$ and $\bar{v}$ be the distance from G to the tangent to the left and right branch of the hyperbola respectively, parallel to line U; see Fig. A.2. There are three cases. (i) When $v \le \bar{v}$ (the first panel of Fig. A.2), as γ increases, both $[A, B]$ and $[C, D]$ become bigger, and thus $q(v)$ increases. (ii) When $\bar{v} < v \le \hat{v}$, $[A, B]$ is empty but $[C, D]$ becomes bigger, so $q(v)$ increases.

(iii) When $v > \hat{v}$, as γ increases (the second panel of Fig. A.2), $[C, D]$ increases but $[A, B]$ decreases. Let $[E, F]$ be the segment symmetric to $[A, B]$ about the origin. We observe that, as γ increases by an infinitesimal $\Delta\gamma$, the amount that $\Pr(U \in [A, B])$ decreases equals the amount that $\Pr(U \in [E, F])$ increases, which is smaller than the amount that $\Pr(U \in [C, D])$ increases. Hence, $q(v)$ still increases.

By the monotonicity for every value of $|V|$, we conclude that the total probability on the right hand side of Eq. (A.8) increases in γ. Hence, $F_{\gamma,\rho}(v^2)$ is maximized at $\gamma = 0$ for every v, namely $\bar{F}_\rho = F_{\rho,\gamma=0}$. It follows that for $X \sim \bar{F}_\gamma$, $(X)_+ =_d \rho(Z_1^2 - Z_2^2)_+$ for two independent standard normal variables Z_1, Z_2. □

Proof of Proposition 2.10. Under $\rho = 1$, the CDF is

$$F_\gamma(x) = \Pr\left\{(Z_1 + \gamma)^2 - Z_2^2 \le x\right\} = \mathbb{E}_{Z_2}\left[\Pr\left\{(Z_1 + \gamma)^2 \le x + Z_2^2 \mid Z_2\right\}\right].$$

Since the conditional probability is non-negative, it suffices to show that given any $x \in \mathbb{R}$, $\gamma = 0$ maximizes $\Pr\{(Z_1 + \gamma)^2 \le x + z_2^2 \mid Z_2 = z_2\} = \Pr\{(Z_1 + \gamma)^2 \le x + z_2^2\}$ for all $z_2 \in \mathbb{R}$. When $x + z_2^2 \le 0$, the conditional probability is zero and $\gamma = 0$ is trivially a maximizer. When $x + z_2^2 > 0$, then $\Pr\{(Z_1 + \gamma)^2 \le x + z_2^2\} = \Phi((x + z_2^2)^{1/2} - \gamma) - \Phi(-(x + z_2^2)^{1/2} - \gamma)$. Setting the derivative with respect to γ to zero requires $\phi(-(x + z_2^2)^{1/2} - \gamma) = \phi((x + z_2^2)^{1/2} - \gamma)$, to which $\gamma = 0$ is the unique solution. Therefore, $\gamma = 0$ is the unique maximizer of $F_\gamma(x)$ for all x. □

Appendix B

APPENDIX TO CHAPTER 3

B.1 Proofs of asymptotic efficiency

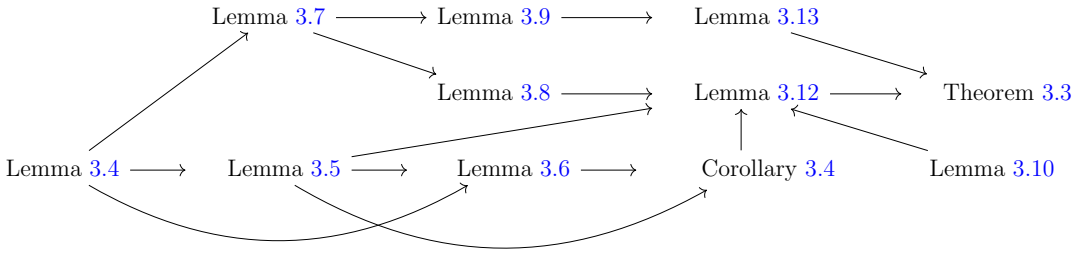


Figure B.1: Dependency structure of proofs in Section 3.6 of main text.

Proof of Lemma 3.7. For simplicity, we drop the superscripts in $\hat{\Sigma}^{(n)}$ and $\hat{\lambda}^{(n)}$. Since $j \in B_k$ and $\mathrm{Pa}(B_k, \mathcal{G}) \subseteq C \subseteq \mathrm{Pa}(B_k, \bar{\mathcal{G}})$, we have

$$\begin{aligned}\hat{\lambda}_{C,j} - \lambda_{C,j} &= (\hat{\Sigma}_C)^{-1}\hat{\Sigma}_{C,j} - (\Sigma_C)^{-1}\Sigma_{C,j} \\ &= \left(\Sigma_C + \hat{\Sigma}_C - \Sigma_C\right)^{-1}\left(\hat{\Sigma}_{C,j} - \Sigma_{C,j}\right) + \left((\hat{\Sigma}_C)^{-1} - (\Sigma_C)^{-1}\right)\Sigma_{C,j}.\end{aligned}$$

We compute the two terms separately. The first term becomes

$$\begin{aligned}\left(\Sigma_C + \hat{\Sigma}_C - \Sigma_C\right)^{-1}\left(\hat{\Sigma}_{C,j} - \Sigma_{C,j}\right) &= \left(\Sigma_C + O_p(n^{-1/2})\right)^{-1}\left(\hat{\Sigma}_{C,j} - \Sigma_{C,j}\right) \\ &= (\Sigma_C)^{-1}\left(\hat{\Sigma}_{C,j} - \Sigma_{C,j}\right) + O_p(n^{-1}),\end{aligned}$$

where we used the fact that Σ_C is positive definite (Lemma 3.4) and $\|\hat{\Sigma}_{C,j} - \Sigma_{C,j}\| = O_p(n^{-1/2})$, $\|\hat{\Sigma}_C - \Sigma_C\|_2 = O_p(n^{-1/2})$ by the central limit theorem.

In the second term,

$$\begin{aligned}(\hat{\Sigma}_C)^{-1} - (\Sigma_C)^{-1} &= \left(\Sigma_C + \hat{\Sigma}_C - \Sigma_C\right)^{-1} - (\Sigma_C)^{-1} \\ &= \left[I - \left(I - (\Sigma_C)^{-1}\hat{\Sigma}_C\right)\right]^{-1}(\Sigma_C)^{-1} - (\Sigma_C)^{-1}.\end{aligned}$$

Since $\|I - (\Sigma_C)^{-1}\hat{\Sigma}_C\|_2 = O_p(n^{-1/2})$, using Neumann series $(I-H)^{-1} = I + H + H^2 + \ldots$ for $H = I - (\Sigma_C)^{-1}\hat{\Sigma}_C$ with $\|H\|_2 \to_p 0 < 1$, we have

$$\begin{aligned}(\hat{\Sigma}_C)^{-1} - (\Sigma_C)^{-1} &= \left[I + H + O_p(n^{-1})\right](\Sigma_C)^{-1} - (\Sigma_C)^{-1} \\ &= H(\Sigma_C)^{-1} + O_p(n^{-1}) \\ &= \left[I - (\Sigma_C)^{-1}\hat{\Sigma}_C\right](\Sigma_C)^{-1} + O_p(n^{-1}).\end{aligned}$$

Combining the two terms, we obtain

$$\begin{aligned}\hat{\lambda}_{C,j} - \lambda_{C,j} &= (\Sigma_C)^{-1}\left(\hat{\Sigma}_{C,j} - \Sigma_{C,j}\right) + \left[I - (\Sigma_C)^{-1}\hat{\Sigma}_C\right](\Sigma_C)^{-1}\Sigma_{C,j} + O_p(n^{-1}) \\ &= (\Sigma_C)^{-1}\hat{\Sigma}_{C,j} - (\Sigma_C)^{-1}\Sigma_{C,j} + (\Sigma_C)^{-1}\Sigma_{C,j} - (\Sigma_C)^{-1}\hat{\Sigma}_C(\Sigma_C)^{-1}\Sigma_{C,j} + O_p(n^{-1}) \\ &\overset{\text{(i)}}{=} (\Sigma_C)^{-1}\left(\hat{\Sigma}_{C,j} - \hat{\Sigma}_C\lambda_{C,j}\right) + O_p(n^{-1}) \\ &= \frac{1}{n}\sum_{i=1}^{n}(\Sigma_C)^{-1}\left[X_j^{(i)}X_C^{(i)} - X_C^{(i)}X_C^{(i)\top}\lambda_{C,j}\right] + O_p(n^{-1}) \\ &= \frac{1}{n}\sum_{i=1}^{n}(\Sigma_C)^{-1}X_C^{(i)}\left(X_j^{(i)} - \lambda_{C,j}^{\top}X_C^{(i)}\right) + O_p(n^{-1}) \\ &\overset{\text{(ii)}}{=} \frac{1}{n}\sum_{i=1}^{n}(\Sigma_C)^{-1}X_C^{(i)}\left(X_j^{(i)} - \lambda_{\mathrm{Pa}(B_k,\mathcal{G}),j}^{\top}X_{\mathrm{Pa}(B_k,\mathcal{G})}^{(i)}\right) + O_p(n^{-1}) \\ &= \frac{1}{n}\sum_{i=1}^{n}(\Sigma_C)^{-1}X_C^{(i)}\varepsilon_j^{(i)} + O_p(n^{-1}),\end{aligned}$$

where (i) uses $\lambda_{C,j} = (\Sigma_C)^{-1}\Sigma_{C,j}$ and (ii) follows from Proposition 3.1 and $\mathrm{Pa}(B_k,\mathcal{G}) \subseteq C \subseteq \mathrm{Pa}(B_k,\bar{\mathcal{G}})$. □

Proof of Lemma 3.8. For each $k = 2,\ldots,K$, note that for $C = \mathrm{Pa}(B_k,\bar{\mathcal{G}}) = B_{[k-1]}$, $\mathrm{vec}\,\hat{\Lambda}_k^{\bar{\mathcal{G}}} = (\hat{\lambda}_{C,j}^{(n)})_{j\in B_k}$ by concatenation. By Lemma 3.7, we have the following asymptotic linear expansion

$$\hat{\lambda}_{B_{[k-1]},j}^{(n)} - \lambda_{B_{[k-1]},j} = \frac{1}{n}\sum_{i=1}^{n}\left(\Sigma_{B_{[k-1]}}\right)^{-1}X_{B_{[k-1]}}^{(i)}\varepsilon_j^{(i)} + O_p(n^{-1}). \tag{B.1}$$

By the central limit theorem,

$$\sqrt{n}\begin{pmatrix} \operatorname{vec}(\hat{\Lambda}_2^{\bar{\mathcal{G}}} - \Lambda_2) \\ \vdots \\ \operatorname{vec}(\hat{\Lambda}_K^{\bar{\mathcal{G}}} - \Lambda_K) \end{pmatrix}$$

converges to a centered multivariate normal distribution. Further, we claim that the asymptotic covariance must be block-diagonal according to $k = 2, \ldots, K$. To see this, take $k < k'$, $j \in B_k$, $j' \in B_{k'}$ and let $C = B_{[k-1]}$, $C' = B_{[k'-1]}$. Using Eq. (B.1), we have

$$\begin{aligned} &\lim_{n\to\infty} n^{-1}\,\mathbb{E}\left(\hat{\lambda}_{C,j}^{(n)} - \lambda_{C,j}\right)\left(\hat{\lambda}_{C',j}^{(n)} - \lambda_{C',j}\right)^\top \\ &= (\Sigma_C)^{-1}\operatorname{cov}(\varepsilon_j X_C, \varepsilon_{j'} X_{C'})(\Sigma_{C'})^{-1} \\ &= (\Sigma_C)^{-1}\left\{\mathbb{E}\left[\varepsilon_j\varepsilon_{j'} X_C X_{C'}^\top\right] - \mathbb{E}\left[\varepsilon_j X_C\right]\mathbb{E}\left[\varepsilon_{j'} X_{C'}^\top\right]\right\}(\Sigma_{C'})^{-1}. \end{aligned}$$

In the expression above, because $\varepsilon_{B_k} \perp\!\!\!\perp X_{B_{[k-1]}}$ and $\varepsilon_{B_{k'}} \perp\!\!\!\perp X_{B_{[k'-1]}}$ by Corollary 3.2 and $j \in B_k$, $j' \in B_{k'}$ for $k < k'$, we have $\mathbb{E}\left[\varepsilon_j\varepsilon_{j'} X_C X_{C'}^\top\right] = \mathbb{E}\,\varepsilon_{j'}\,\mathbb{E}\left[\varepsilon_j X_C X_{C'}^\top\right] = \mathbf{0}$, $\mathbb{E}\,\varepsilon_j X_C = \mathbf{0}$ and $\mathbb{E}\,\varepsilon_{j'} X_{C'} = \mathbf{0}$. It follows that the display above evaluates to $\mathbf{0}$ and hence the asymptotic covariance matrix is block-diagonal.

It remains to be shown that $\operatorname{acov}\operatorname{vec}(\hat{\Lambda}_k^{\bar{\mathcal{G}}} - \Lambda_k) = \Omega_k \otimes (\Sigma_{B_{[k-1]}})^{-1}$ for $k = 2, \ldots, K$. Fix k, take any two distinct $j, j' \in B_k$ and let $C = B_{[k-1]}$. Again using Eq. (B.1), we have

$$\operatorname{acov}\begin{pmatrix} \hat{\lambda}_{C,j}^{(n)} \\ \hat{\lambda}_{C,j'}^{(n)} \end{pmatrix} = \begin{pmatrix} H & F \\ F^\top & D \end{pmatrix},$$

where

$$\begin{aligned} H &= (\Sigma_C)^{-1}\operatorname{cov}(\varepsilon_j X_C, \varepsilon_j X_C)(\Sigma_C)^{-1} = \operatorname{var}(\varepsilon_j)(\Sigma_{B_{[k-1]}})^{-1}, \\ F &= (\Sigma_C)^{-1}\operatorname{cov}(\varepsilon_j X_C, \varepsilon_{j'} X_C)(\Sigma_C)^{-1} = \operatorname{cov}(\varepsilon_j, \varepsilon_{j'})(\Sigma_{B_{[k-1]}})^{-1}, \\ D &= (\Sigma_C)^{-1}\operatorname{cov}(\varepsilon_{j'} X_C, \varepsilon_{j'} X_C)(\Sigma_C)^{-1} = \operatorname{var}(\varepsilon_j')(\Sigma_{B_{[k-1]}})^{-1}. \end{aligned}$$

Noting that $\Omega_k = \operatorname{cov}(\varepsilon_{B_k})$ and $\operatorname{vec}\hat{\Lambda}_k^{\bar{\mathcal{G}}} = (\hat{\lambda}_{C,j}^{(n)})_{j\in B_k}$, the result then follows from comparing the expressions above to the definition of Kronecker product for every pair $j, j' \in B_k$. □

Proof of Lemma 3.9. Note that by the restrictive property of $\mathcal{G}$ (Corollary 3.1), we have $\operatorname{vec}\hat{\Lambda}_k^{\mathcal{G}} = \left(\hat{\lambda}_{\operatorname{Pa}(B_k,\mathcal{G}),j}^{(n)}\right)_{j\in B_k}$ for $k = 2, \ldots, K$. Using Lemma 3.7 with $C = \operatorname{Pa}(B_k, \mathcal{G})$, we have

the following asymptotic linear expansion

$$\hat{\lambda}^{(n)}_{\mathrm{Pa}(B_k,\mathcal{G}),j} - \lambda_{\mathrm{Pa}(B_k,\mathcal{G}),j} = \frac{1}{n}\sum_{i=1}^{n} \left(\Sigma_{\mathrm{Pa}(B_k,\mathcal{G})}\right)^{-1} X^{(i)}_{\mathrm{Pa}(B_k,\mathcal{G})}\varepsilon^{(i)}_j + O_p(n^{-1}). \tag{B.2}$$

The rest of computation follows similarly to the proof of Lemma 3.8. □

Proof of Lemma 3.10. Since $S \in \mathbb{R}^{n\times n}_{\mathrm{PD}}$, by completing the square, we have

$$\begin{aligned}
x^\top S x &= x_A^\top S_{A,A} x_A + x_B^\top S_{B,A} x_A + x_A^\top S_{A,B} x_B + x_B^\top S_{B,B} x_B \\
&\quad - x_A^\top S_{A,B} S_{B,B}^{-1} S_{B,A} x_A + x_A^\top S_{A,B} S_{B,B}^{-1} S_{B,A} x_A \\
&= x_A^\top (S_{A,A} - S_{A,B} S_{B,B}^{-1} S_{B,A}) x_A + (x_B + S_{B,B}^{-1} S_{B,A} x_A)^\top S_{B,B} (x_B + S_{B,B}^{-1} S_{B,A} x_A) \\
&\geq x_A^\top (S_{A,A} - S_{A,B} S_{B,B}^{-1} S_{B,A}) x_A = x_A^\top S_{A\cdot B} x_A,
\end{aligned}$$

where the equality holds if and only if $x_B = -S_{B,B}^{-1} S_{B,A} x_A$. □

B.2 Proofs of graphical results

Proof of Lemma 3.1. Let the undirected path between j and k be $p = \langle j = V_1, \ldots, V_l = k\rangle$ with $l > 1$. First note that i is not on p because there is no undirected path between i and j in $\mathcal{G}$.

Further, since $i \to j - V_2$ is in $\mathcal{G}$, by Meek rules R1 and R1 (Fig. B.3 in Appendix B.4), $i - V_2$ or $i \to V_2$ is in $\mathcal{G}$. Since, by assumption, there is no undirected path from i to j in $\mathcal{G}$, $i - V_2 \notin U$. Hence, $i \to V_2 \in E$ and if $l = 2$, the statement of the lemma holds. If $l > 2$, we can apply the above reasoning iteratively until we obtain $i \to V_l \in E$. □

Proof of Lemma 3.2. Let $l \in D_k$. Since $D_k \subseteq B_k$, $l \in B_k$. Then by Corollary 3.1, we have that $\mathrm{Pa}(B_k) = \mathrm{Pa}(l) \setminus B_k$. Therefore, $\mathrm{Pa}(B_k) \subseteq \cup_{j\in D_k} \mathrm{Pa}(j) \setminus B_k$ and furthermore, $\mathrm{Pa}(B_k) \subseteq \cup_{j\in D_k} \mathrm{Pa}(j) \setminus D_k = \mathrm{Pa}(D_k)$. Hence, it suffices to show $\mathrm{Pa}(D_k) \subseteq \mathrm{Pa}(B_k)$.

We prove $\mathrm{Pa}(D_k) \subseteq \mathrm{Pa}(B_k)$ by contradiction. Suppose there exists $j \in \mathrm{Pa}(D_k) \setminus \mathrm{Pa}(B_k)$. By definition $D = \mathrm{An}(Y, \mathcal{G}_{V\setminus A})$ and $D = \cup_{r=1}^{K} D_r$. Therefore, if $k = 1$, then $j \in A$; if $k > 1$, j must be contained in $\cup_{r=1}^{k-1} D_r$ or in A. If $j \in A$, this leads to a contradiction with Lemma B.1 in Appendix B.4. Suppose $k > 1$ and $j \in \cup_{r=1}^{k-1} D_r$. Because $\cup_{r=1}^{k-1} D_r \subseteq \cup_{r=1}^{k-1} B_r$

and buckets $\{B_1, \ldots, B_K\}$ are disjoint, we have $(\cup_{r=1}^{k-1} D_r) \cap B_k = \emptyset$. However, this contradicts that $j \in B_k$. □

Proof of Proposition 3.3. By construction, the undirected component of $\bar{\mathcal{G}}$ remains the same as that of $\mathcal{G}$. Hence, $\bar{\mathcal{G}}$ has the same bucket decomposition as $\mathcal{G}$. We only need to show that $\bar{\mathcal{G}}$ is an MPDAG. It is enough to show that the edge orientations in $\bar{\mathcal{G}}$ are closed under rules R1–R4 of Meek (1995) that are displayed in Fig. B.3 of Appendix B.4. Note that since $\mathcal{G}$ an MPDAG it is closed under R1–R4. So if any of the left-hand-side graphs in Figure B.3 are induced subgraphs of $\bar{\mathcal{G}}$, then at least one of the directed edges in these induced subgraphs must have been added in the construction of $\bar{\mathcal{G}}$.

Since the construction of $\bar{\mathcal{G}}$ does not involve adding directed edges within a bucket, the left-hand-side of rules R3 and R4 in Figure B.3 cannot appear as induced subgraphs of $\bar{\mathcal{G}}$. Hence, edge orientations in $\bar{\mathcal{G}}$ are complete under rules R3 and R4.

Consider the left-hand-side of rule R1 in Figure B.3, $A \to B - C$, for some $A, B, C \in V$. For $A \to B - C$ to be an induced subgraph of $\bar{\mathcal{G}}$, $A \to B$ must have been added in the construction of $\bar{\mathcal{G}}$ from $\mathcal{G}$. Hence, A and B would need to be in different buckets in V in $\mathcal{G}$. Since B and C are in the same bucket because of edge $B - C$, $A \to C$ would also be added to $\mathcal{G}$ in the construction of $\bar{\mathcal{G}}$. Hence, $A \to B - C$ will also not appear as an induced subgraph of $\bar{\mathcal{G}}$ and edge orientations in $\bar{\mathcal{G}}$ are also closed under R1.

Consider the left-hand-side of R2 in Figure B.3, and suppose for a contradiction that $A \to B \to C$ and $A - C$ is an induced subgraph of $\bar{\mathcal{G}}$ for some $A, B, C \in V$. Then $A \to B$, $B \to C$, or both $A \to B$ and $B \to C$, were added to $\mathcal{G}$ in the construction of $\bar{\mathcal{G}}$. Because of $A - C$, suppose A and C are in the same bucket B_i for some $i \in \{1, \ldots, K\}$ in $\mathcal{G}$. Also, suppose $B \in B_j$. Because only directed edges between buckets are added, $i \neq j$.

Now, $A \to B$ and $B \to C$ cannot be both added to $\mathcal{G}$ to construct $\bar{\mathcal{G}}$, because that would imply that $i < j$ and $j < i$. By R1, $B \to C - A$ cannot be an induced subgraph of MPDAG $\mathcal{G}$, so $A \to B$ alone also could not be added to $\mathcal{G}$. Therefore, $B \to C$ alone was added to $\mathcal{G}$. But $C - A \to B$ is an induced subgraph of $\mathcal{G}$, so $i < j$, which contradicts the direction of

$B \to C$. □

B.3 Additional simulation results

In this section, we report additional simulation results. The setup is the same as Section 2.7 of main text, but we replace the true CPDAG with the CPDAG estimated with the greedy equivalence search algorithm (Chickering, 2002) based on the same sample. The relative squared errors of the contending estimators are shown in Fig. B.2 and are summarized in Table B.1. Compared to the results with the true CPDAG, the performance improvement of $\mathcal{G}$-regression is more modest but still matters in practice. The reduced improvement is due to the error in estimating the graph, which diminishes as n increases.

Table B.1: Geometric average (brackets: median) of relative squared errors compared to $\mathcal{G}$-regression when CPDAGs are estimated

	$\lvert V\rvert = 20$				$\lvert V\rvert = 50$				$\lvert V\rvert = 100$			
$\lvert A\rvert$	$n = 100$		$n = 1000$		$n = 100$		$n = 1000$		$n = 100$		$n = 1000$	
`adj.O`												
1	1.0	(1.0)	1.0	(1.0)	1.2	(1.0)	1.3	(1.0)	1.8	(1.1)	1.6	(1.0)
2	2.0	(1.1)	3.1	(1.2)	2.4	(1.3)	3.1	(1.4)	3.2	(1.9)	3.7	(2.0)
3	3.3	(1.7)	5.2	(2.7)	4.0	(2.4)	5.9	(2.8)	4.7	(2.5)	5.5	(2.8)
4	4.6	(2.2)	7.9	(4.2)	5.0	(2.1)	9.0	(5.7)	10	(5.9)	8.9	(5.6)
`IDA.M`												
5	2.9	(1.4)	4.1	(1.4)	4.5	(2.7)	10	(5.7)	7.3	(4.5)	18	(11)
6	4.2	(2.0)	6.6	(2.1)	7.3	(4.8)	14	(7.2)	13	(7.9)	22	(14)
7	6.2	(3.1)	6.8	(2.5)	12	(7.1)	16	(8.3)	15	(10)	28	(18)
8	9.5	(5.6)	9.0	(3.1)	13	(10)	20	(12)	19	(14)	37	(26)
`IDA.R`												
9	2.9	(1.4)	4.1	(1.4)	4.5	(2.7)	10	(5.7)	7.3	(4.5)	18	(11)
10	2.7	(1.3)	4.6	(1.2)	4.5	(2.3)	9.6	(4.0)	8.5	(5.9)	15	(9.5)
11	3.1	(1.5)	4.1	(1.2)	5.8	(3.0)	7.8	(2.5)	7.6	(5.2)	14	(8.9)
12	3.6	(1.6)	4.2	(1.3)	4.9	(2.8)	8.2	(3.6)	8.1	(5.4)	15	(10)

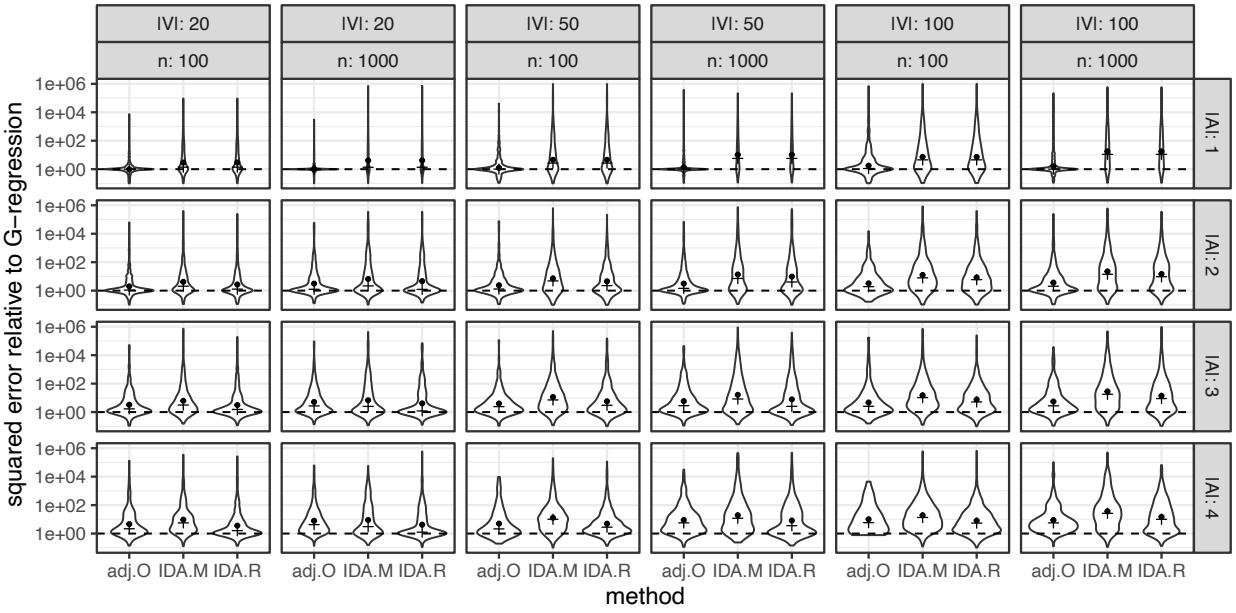


Figure B.2: Violin plots for the relative squared errors of contending estimators ('·': geometric mean, '+': median). The estimated CPDAGs are provided to the estimators.

B.4 Graphical preliminaries

Graphs, vertices, edges A graph $\mathcal{G} = (V, F)$ consists of a set of vertices (variables) V and a set of edges F. The graphs we consider are allowed to contain directed ($\rightarrow$) and undirected ($-$) edges and at most one edge between any two vertices. We can thus partition the set of edges F into a set of directed edges E and undirected edges U and denote graph $\mathcal{G} = (V, F)$ as $\mathcal{G} = (V, E, U)$. The corresponding undirected graph is simply $\mathcal{G}_U = (V, \emptyset, U)$.

Subgraphs and skeleton An *induced subgraph* $\mathcal{G}_{V'} = (V', F')$ of $\mathcal{G} = (V, F)$ consists of $V' \subseteq V$ and $F' \subseteq F$ where F' are all edges in F between vertices in V'. A *skeleton* of a graph $\mathcal{G} = (V, F)$ is an undirected graph $\mathcal{G} = (V, F')$, such that F' are undirected versions of all edges in F.

Paths. Directed, undirected, causal, non-causal, proper paths A *path* p from i to j in $\mathcal{G}$ is a sequence of distinct vertices $p = \langle i, \ldots, j \rangle$ in which every pair of successive vertices are adjacent. A path consisting of undirected edges is an *undirected* path. A *directed path* from i to j is a path from i to j in which all edges are directed towards j, that is, $i \to \cdots \to j$. We will use *causal path* instead of *directed path* when talking about causal graphs. Let $p = \langle v_1, \ldots, v_k \rangle$, $k > 1$ be a path in $\mathcal{G}$, p is a *possibly directed path* (*possibly causal path*) if no edge $v_i \leftarrow v_j, 1 \leq i < j \leq k$ is in $\mathcal{G}$. Otherwise, p is a *non-causal path* in $\mathcal{G}$ (see Definition 3.1 and Lemma 3.2 of Perković et al., 2017). A path from A to Y is *proper* (w.r.t. A) if only its first vertex is in A.

Directed cycles A directed path from i to j and the edge $j \to i$ form a *directed cycle*.

Colliders, shields and definite status paths If a path p contains $i \to j \leftarrow k$ as a subpath, then j is a *collider* on p. A path $\langle i, j, k \rangle$ is an *(un)shielded triple* if i and k are (not) adjacent. A path is *unshielded* if all successive triples on the path are unshielded. A node v_j is a *definite non-collider* on a path p if there is at least one edge out of v_j on p, or if $v_{j-1} - v_j - v_{j+1}$ is a subpath of p and $\langle v_{j-1}, v_j, v_{j+1} \rangle$ is an unshielded triple. A node is of *definite status* on a path if it is a collider, a definite non-collider or an endpoint on the path. A path p is of definite status if every node on p is of definite status.

Subsequences and subpaths A *subsequence* of a path p is obtained by deleting some nodes from p without changing the order of the remaining nodes. A subsequence of a path is not necessarily a path. For a path $p = \langle v_1, v_2, \ldots, v_m \rangle$, the *subpath* from v_i to v_k $(1 \leq i \leq k \leq m)$ is the path $p(v_i, v_k) = \langle v_i, v_{i+1}, \ldots, a_k \rangle$.

Ancestral relations If $i \to j$, then i is a *parent* of j, and j is a *child* of i. If there is a causal path from k to l, then k is an *ancestor* of l, and l is a *descendant* of k. If there is a possibly causal path from k to l, then k is a *possible ancestor* of l, and l is a *possible descendant* of k. We use the convention that every vertex is a descendant, ancestor, possible ancestor and

possible descendant of itself. The sets of parents, ancestors, descendants and possible descendants of i in $\mathcal{G}$ are denoted by $\text{Pa}(i, \mathcal{G})$, $\text{An}(i, \mathcal{G})$, $\text{De}(i, \mathcal{G})$ and $\text{PossDe}(i, \mathcal{G})$ respectively. For a set of vertices A, we let $\text{Pa}(A, \mathcal{G}) = (\cup_{i \in A} \text{Pa}(i, \mathcal{G})) \setminus A$, whereas, $\text{An}(A, \mathcal{G}) = \cup_{i \in A} \text{An}(i, \mathcal{G})$, $\text{De}(A, \mathcal{G}) = \cup_{i \in A} \text{De}(i, \mathcal{G})$ and $\text{PossDe}(A, \mathcal{G}) = \cup_{i \in A} \text{PossDe}(i, \mathcal{G})$

DAGs, PDAGs A *directed graph* contains only directed edges. A *partially directed graph* may contain both directed and undirected edges. A directed graph without directed cycles is a *directed acyclic graph* (DAG). A *partially directed acyclic graph* (PDAG) is a partially directed graph without directed cycles.

Blocking and d-separation (See Definition 1.2.3 of Pearl (2009) and Lemma C.1 of Henckel et al. (2019)). Let Z be a set of vertices in an PDAG $\mathcal{G} = (V, E, U)$. A definite status path p is *blocked* by Z if (i) p contains a non-collider that is in Z, or (ii) p contains a collider C such that no descendant of C is in Z. A definite status path that is not blocked by Z is *open* given Z. If A, B and Z are three pairwise disjoint sets of nodes in a PDAG $\mathcal{G} = (V, E, U)$, then Z *d-separates* A from B in $\mathcal{G}$ if Z blocks every definite status path between any node in A and any node in B in $\mathcal{G}$.

CPDAGs, MPDAGs Several DAGs can encode the same d-separation relationships. Such DAGs form a *Markov equivalence class* which is uniquely represented by a *completed partially directed acyclic graph* (CPDAG) (Meek, 1995; Andersson et al., 1997). A PDAG $\mathcal{G} = (V, E, U)$ is a *maximally oriented* PDAG (MPDAG) if it is closed under orientation rules R1-R4 of (Meek, 1995), presented in Figure B.3. The MPDAG can then be alternatively defined as any PDAG that does not contain graphs on the left-hand side of each orientation rule as induced subgraphs. Both DAGs and CPDAGs are types of MPDAGs (Meek, 1995).

Background knowledge and constructing MPDAGs A PDAG $\mathcal{G}'$ is *represented* by another PDAG $\mathcal{G}$ (equivalently $\mathcal{G}$ represents $\mathcal{G}'$) if $\mathcal{G}'$ and $\mathcal{G}$ have the same adjacencies and unshielded colliders and every directed edge $i \to j$ in $\mathcal{G}$ is also in $\mathcal{G}'$. Let R be a set of

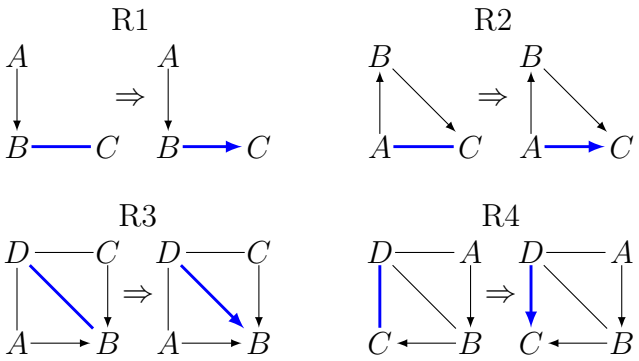


Figure B.3: The orientation rules from Meek (1995). If the graph on the left-hand side of a rule is an induced subgraph of a PDAG $\mathcal{G}$, then *orient* the blue undirected edge ($-$) as shown on the right-hand side of the rule. Hence, the graphs on the left-hand side of each rule are not allowed to be induced subgraphs of an MPDAG.

directed edges representing background knowledge. Algorithm 1 of Meek (1995) describes how to incorporate background knowledge R in an MPDAG $\mathcal{G}$. If Algorithm 1 does not return a FAIL, then it returns a new MPDAG $\mathcal{G}'$ that is represented by $\mathcal{G}$. Background knowledge R is *consistent* with MPDAG $\mathcal{G}$ if and only if Algorithm 1 does not return a FAIL (Meek, 1995).

Remark B.1. The MPDAG output by `ConstructMPDAG`($\mathcal{G}$, R) is the same independent of the ordering of edges in R. This stems from the fact that the orientation rules of Meek (1995) are necessary and sufficient for the construction of an MPDAG given a set of adjacencies and unshielded colliders.

$\mathcal{G}$ and $[\mathcal{G}]$ If $\mathcal{G}$ is a MPDAG, then $[\mathcal{G}]$ denotes every DAG represented by $\mathcal{G}$.

Causal and partial causal ordering of vertices A total ordering, $<$, of vertices $V' \subseteq V$ is *consistent* with a DAG $\mathcal{D} = (V, E, \emptyset)$ and called a *causal ordering* of V' if for every $i, j \in V'$, such that $i < j$ and such that i and j are adjacent in $\mathcal{D}$, $i \to j$ is in $\mathcal{D}$. There can be more than one causal ordering of V' in a DAG $\mathcal{D} = (V, E, \emptyset)$. For example, in DAG $i \leftarrow j \to k$

Algorithm 1: `ConstructMPDAG`, (Meek, 1995; Perković et al., 2017)

Data: MPDAG $\mathcal{G}$, background knowledge R

Result: MPDAG $\mathcal{G}'$ or FAIL

Let $\mathcal{G}' = \mathcal{G}$;
while $R \neq \emptyset$ **do**
  Choose an edge $\{X \to Y\}$ in R ;
  $R = R \setminus \{X \to Y\}$;
  if $\{X - Y\}$ *or* $\{X \to Y\}$ *is in* $\mathcal{G}'$ **then**
    Orient $\{X \to Y\}$ in $\mathcal{G}'$;
    Close the edge orientations under the rules in Figure B.3 in $\mathcal{G}'$;
  else
    FAIL;
  end
end

both orderings $j < i < k$ and $j < k < i$ are consistent.

Since an MPDAG may contain undirected edges, there is generally no unique causal ordering of vertices in an MPDAG. Instead, we define a *partial causal ordering,* $<$, of a vertex set V', $V' \subset V$ in an MPDAG $\mathcal{G} = (V, E, U)$ as a total ordering of pairwise disjoint vertex sets $A_1, \ldots, A_k$, $k \geq 1$, $\cup_{i=1}^{k} A_i = V'$, that satisfy the following: if $A_i < A_j$ and there is an edge between $i \in A_i$ and $j \in A_j$ in $\mathcal{G}$, then $i \to j$ is in $\mathcal{G}$.

Buckets and bucket decomposition Algorithm 2 describes how to obtain an ordered bucket decomposition for a set of vertices V in an MPDAG $\mathcal{G} = (V, E, U)$. By Perković (2020, Lemma 1), the ordered list of buckets output by Algorithm 2 is a partial causal ordering of V in $\mathcal{G}$.

Lemma B.1. *(see Lemma D.1 (i) of Perković, 2020) Let A and Y be disjoint node sets in*

Algorithm 2: Partial causal ordering (Perković, 2020)

input : vertex set V in MPDAG $\mathcal{G}$=(V, E, U) and MPDAG $\mathcal{G}$.

output: An ordered list B=$(B_1, \ldots, B_k), k \geq 1$, of the bucket decomposition of V in $\mathcal{G}$.

Let $\mathcal{G}_U$ denote the undirected subgraph of $\mathcal{G}$;
Let $ConComp$ be the bucket decomposition (i.e., maximal connected components) of V in $\mathcal{G}_U$;
Let B be an empty list;
while $ConComp \neq \emptyset$ **do**
    Let C be any element from $ConComp$;
    Let $\overline{C}$ be the set of vertices in $ConComp$ that are not in C;
    if *all edges between* C *and* $\overline{C}$ *are into* C *in* $\mathcal{G}$ **then**
        Remove C from $ConComp$;
        Add C to the beginning of B;
    end
end
return B;

MPDAG $\mathcal{G} = (V, E, U)$. *Suppose that there is no proper possibly causal path from* A *to* Y *that starts with an undirected edge in* $\mathcal{G}$, *that is, suppose that the criterion in Theorem 3.1 is satisfied. Further, let* $D = \mathrm{An}(Y, \mathcal{G}_{V \setminus A})$ *and* $D = \dot{\bigcup}_{i=1}^{K} D_i$ *for* $D_i = D \cap B_i$, $i = 1, \ldots, K$, *where* $B_1, \ldots B_K$ *is the bucket decomposition of* V. *Then for all* $i \in \{1, \ldots, K\}$, *there is no proper possibly causal path from* A *to* B_i *that starts with an undirected edge in* $\mathcal{G}$.

Appendix C

APPENDIX TO CHAPTER 4

C.1 Proof of Proposition 4.2

Proof. By Proposition 4.1, $G_{k,n}(\lambda) = G_{k,n}(\lambda, p)$ for $p_k = 0$ and $p_1 + \cdots + p_{k-1} = 1$. By Eq. (4.7), we split $G_{k,n}(\lambda) = A + B$, where A sums over those X with $X_k = 0$, and B sums over those with $X_k \geq 1$. Clearly,

$$A = \sum_{X_1,\ldots,X_{k-1}} \binom{n}{X_1, \ldots, X_{k-1}} \prod_{j=1}^{k-1} [\lambda X_j/n + (1-\lambda)p_j]^{X_j},$$

where the summation is over non-negative integers $X_1, \ldots, X_{k-1}$ such that they sum to n. Further, $(p_1, \ldots, p_{k-1})$ forms a probability vector. Hence, $A = G_{k-1,n}(\lambda)$.

Now we evaluate

$$B = \sum_{X_k=1}^{n} \sum_{X_1+\cdots+X_{k-1}=n-X_k} \binom{n}{X_1, \ldots, X_k} \prod_{j=1}^{k} [\lambda X_j/n + (1-\lambda)p_j]^{X_j}.$$

Using the fact that $\binom{n}{X_1,\ldots,X_k} = \left(\frac{n}{X_k}\right)\binom{n-1}{X_1,\ldots,X_{k-1},X_k-1}$ and $p_k = 0$, we have

$$\begin{aligned} B = &\sum_{X_1,\ldots,X_{k-1},X'_k} \frac{n}{X'_k+1} \binom{n-1}{X_1, \ldots, X_{k-1}, X'_k} \left(\frac{\lambda(X'_k+1)}{n}\right)^{X'_k+1} \\ &\times \prod_{j=1}^{k-1} [\lambda X_j/n + (1-\lambda)p_j]^{X_j} \qquad \text{(C.1)} \\ = \lambda &\sum_{X_1,\ldots,X_{k-1},X'_k} \binom{n-1}{X_1, \ldots, X_{k-1}, X'_k} \left(\frac{\lambda(X'_k+1)}{n}\right)^{X'_k} \prod_{j=1}^{k-1} [\lambda X_j/n + (1-\lambda)p_j]^{X_j}, \end{aligned}$$

where $X'_k := X_k - 1 \in \{0, \ldots, n-1\}$ and the summation is over $(X_1, \ldots, X_{k-1}, X'_k)$ such that they sum to $n-1$. Let $\lambda' := \frac{n-1}{n}\lambda$ and

$$p'_j := \frac{1-\lambda}{1-\lambda'} p_j \quad (j = 1, \ldots, k-1), \quad p'_k := \frac{\lambda/n}{1-\lambda'}$$

such that $\sum_{j=1}^{k} p'_j = \frac{1-\lambda}{1-\lambda'} + \frac{\lambda/n}{1-\lambda'} = 1$. Then we have

$$\frac{\lambda(X'_k+1)}{n} = \lambda' \frac{X'_k}{n-1} + (1-\lambda')p'_k,$$
$$\lambda X_j/n + (1-\lambda)p_j = \lambda' \frac{X_j}{n-1} + (1-\lambda')p'_j \quad (j = 1, \ldots, k-1).$$

Hence, by Eq. (4.7) and Proposition 4.1, Eq. (C.1) becomes

$$\begin{aligned} B &= \lambda \sum_{X_1,\ldots,X_{k-1},X'_k} \binom{n-1}{X_1, \ldots, X_{k-1}, X'_k} \prod_{j=1}^{k} \left[\lambda' X_j/n - 1 + (1-\lambda')p'_j\right]^{X_j} \\ &= \lambda G_{k,n-1}(\lambda') = \lambda G_{k,n-1}\left(\frac{n-1}{n}\lambda\right). \end{aligned}$$

Putting A and B together, we have $G_{k,n}(\lambda) = G_{k-1,n}(\lambda) + \lambda G_{k,n-1}\left(\frac{n-1}{n}\lambda\right)$. □

C.2 Proof of Lemma 4.2

We will use the following two properties of the incomplete Gamma function

$$\Gamma(a,z) := \int_z^\infty t^{a-1}e^{-t}\,\mathrm{d}t.$$

Lemma C.1 (DLMF, §8.8)**.** *It holds that*

$$\Gamma(a+1,z) = a\Gamma(a,z) + z^a e^{-z}, \tag{C.2}$$

and

$$\Gamma(a,z) = \frac{\Gamma(a)}{\Gamma(a-n)}\Gamma(a-n,z) + z^{a-1}e^{-z}\sum_{k=0}^{n-1}\frac{\Gamma(a)}{\Gamma(a-k)}z^{-k}, \tag{C.3}$$

where n is a non-negative integer.

Lemma C.2 (DLMF, §8.11(iii))**.** *For fixed $\gamma > 1$, as $a \to \infty$, it holds that*

$$\Gamma(a,\gamma a) = z^a e^{-z}\left\{\sum_{k=0}^{n}\frac{(-1)^k b_k(\gamma)}{(\gamma-1)^{2k+1}}a^{-k-1} + o(|a|^{-n-1})\right\}, \tag{C.4}$$

where $b_0(\gamma) = 1$, $b_1(\gamma) = \gamma$, $b_2(\gamma) = \gamma(2\gamma+1)$, and for $k = 1, 2, \ldots,$

$$b_k(\gamma) = \gamma(1-\gamma)b'_{k-1}(\gamma) + (2k-1)\gamma b_{k-1}(\gamma). \tag{C.5}$$

Proof. We first express $G_{k,n}(\lambda)$ in terms of the incomplete Gamma function. For the case of $k = 2$, we have

$$\begin{aligned} G_{2,n}(\lambda) &= \sum_{m=0}^{n} \frac{n!}{n^m(n-m)!}\lambda^m \\ &= \lambda^n \sum_{m=0}^{n} \frac{n!}{n^m(n-m)!}\lambda^{-(n-m)} \\ &= n^{-n}\lambda^n n! \sum_{m=0}^{n} \frac{(n/\lambda)^m}{m!} = n^{-n}\lambda^n e^{n/\lambda}\Gamma(n+1, n/\lambda), \end{aligned} \tag{C.6}$$

where we used the fact (DLMF, Eq. 8.4.8) that

$$\Gamma(n+1, z) = n!e^{-z}\sum_{k=0}^{n}\frac{z^k}{k!}, \quad n = 0, 1, 2, \ldots.$$

Similarly, we have

$$G'_{2,n}(\lambda) = n^{1-n}\lambda^{n-2}\left\{\left(\frac{n}{\lambda}\right)^n + (\lambda-1)e^{n/\lambda}\Gamma(n+1, n/\lambda)\right\}.$$

The LHS of Eq. (4.18) with $k = 2$ can be expressed as

$$n\left(\frac{1}{1-\lambda}G_{2,n}(\lambda) - G'_{2,n}(\lambda)\right) = \frac{B}{(1-\lambda)\lambda^2},$$

where

$$B = e^{n/\lambda}n^2\left(n\frac{(1-\lambda)^2}{\lambda} + \lambda\right)\left(\frac{n}{\lambda}\right)^{-n-1}\Gamma(n+1, n/\lambda) - (1-\lambda)n^2.$$

Using Lemma C.1, B can be expressed in terms of $\Gamma(n, n/\lambda)$ as

$$B = \lambda^2 n - \lambda(1-\lambda)n^2 + e^{n/\lambda}n^3\left(n\frac{(1-\lambda)^2}{\lambda} + \lambda\right)\left(\frac{n}{\lambda}\right)^{-n-1}\Gamma(n, n/\lambda).$$

By Lemma C.2, plugging in

$$\Gamma(n, n/\lambda) = \left(\frac{n}{\lambda}\right)^n e^{-n/\lambda}\left\{\sum_{k=0}^{2}\frac{(-1)^k b_k(\lambda^{-1})}{(\lambda^{-1}-1)^{2k+1}}(n/\lambda)^{-k-1} + o(n^{-3})\right\}$$

into the previous display and simplifying, we get

$$\begin{aligned} B &= \lambda^2 n - \lambda(1-\lambda)n^2 + (1-\lambda)^2\left[\frac{\lambda}{1-\lambda}n^2 - \frac{\lambda^2}{(1-\lambda)^3}n + \frac{\lambda^{-1}(2/\lambda+1)}{(\lambda^{-1}-1)^5} + o(1)\right] \\ &\quad + \lambda^2\left[\frac{\lambda}{1-\lambda}n - \frac{\lambda^2}{(1-\lambda)^3} + o(1)\right] \\ &= \frac{2\lambda^3}{(1-\lambda)^3} + o(1). \end{aligned}$$

And therefore,

$$n\left(\frac{1}{1-\lambda}G_{2,n}(\lambda) - G'_{2,n}(\lambda)\right) = \frac{2\lambda}{(1-\lambda)^4} + o(1).$$

By a similar computation for $k = 3, 4, \ldots$, one can show that

$$n\left(\frac{k-1}{1-\lambda}G_{k,n}(\lambda) - G'_{k,n}(\lambda)\right) = \frac{k(k-1)\lambda}{(1-\lambda)^{k+2}} + o(1).$$

□

C.3 Proof of Proposition 4.9

Lemma C.3 (DLMF, §8.11(v)). *As $z \to \infty$, it holds that*

$$\Gamma(z, z) = z^{z-1}e^{-z}\left(\sqrt{\frac{\pi}{2}}z^{1/2} + O(1)\right).$$

Proof of Proposition 4.9. For $k = 2$, we have

$$\begin{aligned} G_{2,n}(1) &\overset{\text{(i)}}{=} (e/n)^n \Gamma(n+1, n) \\ &\overset{\text{(ii)}}{=} (e/n)^n n\Gamma(n, n) + 1 \\ &\overset{\text{(iii)}}{=} \sqrt{\frac{\pi}{2}}n^{1/2} + O(1), \end{aligned}$$

where (i) follows from Eq. (C.6), (ii) from Lemma C.1 and (iii) from Lemma C.3. And hence,

$$\lim_{n\to\infty} \frac{\log G_{2,n}(1)}{\log \binom{n+1}{1}} = \lim_{n\to\infty} \frac{\log n^{1/2}}{\log n} = 1/2.$$

By a similar computation for $k = 3, 4, \ldots$, one can show that

$$\lim_{n\to\infty} \frac{\log G_{k,n}(1)}{\log \binom{n+k-1}{k-1}} = \lim_{n\to\infty} \frac{\log n^{(k-1)/2}}{\log n^{k-1}} = 1/2.$$

□

C.4 R code for unseen butterflies

The following R code is used to compute a 95% confidence upper bound on the total probability of unseen butterflies.

```
library(multChernoff)  # https://github.com/richardkwo/multChernoff
library(CVXR)
library(plyr)

# Corbet butterfly data
# https://en.wikipedia.org/wiki/Unseen_species_problem
corbet.butterfly <- data.frame(j=1:15, n_j=c(118,74,44,24,29,22,20,19,20,15,12,14,6,12,6))
n.butterfly <- Reduce(c, alply(corbet.butterfly, 1, function(.df) rep(.df$j, .df$n_j)))
alpha <- 0.05
n.observed <- c(n.butterfly, 0)    # the last one is the unseen
n <- sum(n.observed)
k <- length(n.observed)
p.observed <- n.observed / n

# critical value
t.alpha <- criticalValue(k, n, p=alpha, verbose = TRUE)
cat(sprintf("critical value = %f\n", t.alpha))

# convex program
p <- Variable(k)
obj <- p[k]
constr <- list(p>=0,
               sum(p) == 1,
               2 * n * sum(p.observed * (log(p.observed) - log(p))) <= t.alpha)
prob <- Problem(Maximize(obj), constr)
result <- solve(prob, verbose=TRUE)

# result
print(result$status)
unseen <- result$value
p.maximizer <- c(result$getValue(p))
cat(sprintf("unseen <= %f\n", unseen))
```

The convex program is specified with R package CVXR (Fu et al., 2020) and solved with

MOSEK (Andersen and Andersen, 2000). The session information is printed out as below.

```
> sessionInfo()
R version 4.0.2 (2020-06-22)
Platform: x86_64-apple-darwin17.0 (64-bit)
Running under: macOS  10.16

Matrix products: default
LAPACK: /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRlapack.dylib

locale:
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] plyr_1.8.6              CVXR_1.0-8              multChernoff_0.0.0.9000

loaded via a namespace (and not attached):
 [1] bit_4.0.4       compiler_4.0.2  R6_2.4.1        Matrix_1.2-18   tools_4.0.2     gmp_0.6-1
 [7] Rmosek_9.1.0    Rcpp_1.0.5      bit64_4.0.5     grid_4.0.2      Rmpfr_0.8-2     lattice_0.20-41
```

Appendix D

APPENDIX TO CHAPTER 5

D.1 Code for polytope computation

The following Julia v1.5 code is used for proving Theorem 5.1.

```
using JuMP                  # https://github.com/jump-dev/JuMP.jl
using Polyhedra             # https://github.com/JuliaPolyhedra/Polyhedra.jl
using CDDLib                # https://github.com/JuliaPolyhedra/CDDLib.jl
model = Model()
# counterfactuals P(X, Y(x=0,z=0), Y(0,1), Y(1,0), Y(1,1) | Z=z) for z=0,1
@variable(model, P[1:2, 1:2, 1:2, 1:2, 1:2, 1:2] >= 0)
# observed Q(X, Y | Z=z) for z=0,1
@variable(model, Q[1:2, 1:2, 1:2] >= 0)
# simplex constraints
@constraint(model, sum(P[:,:,:,:,:,1])==1)
@constraint(model, sum(P[:,:,:,:,:,2])==1)
@constraint(model, sum(Q[:,:,1])==1)
@constraint(model, sum(Q[:,:,2])==1)
# ME: marginal exogeneity
@constraint(model, sum(P[:,1,:,:,:,1]) == sum(P[:,1,:,:,:,2]))
@constraint(model, sum(P[:,:,1,:,:,1]) == sum(P[:,:,1,:,:,2]))
@constraint(model, sum(P[:,:,:,1,:,1]) == sum(P[:,:,:,1,:,2]))
@constraint(model, sum(P[:,:,:,:,1,1]) == sum(P[:,:,:,:,1,2]))
# stochastic exclusion
@constraint(model, sum(P[:,1,:,:,:,1]) == sum(P[:,:,1,:,:,1]))
@constraint(model, sum(P[:,:,:,1,:,1]) == sum(P[:,:,:,:,1,1]))
# consistency Q(X=x, Y=y | Z=z) = P(X=x, Y(x,z)=y | Z=z)
# x=0, y=0, z=0
@constraint(model, Q[1,1,1]==sum(P[1,1,:,:,:,1]))
```

```
# x=0, y=0, z=1
@constraint(model, Q[1,1,2]==sum(P[1,:,1,:,:,2]))
# x=0, y=1, z=0
@constraint(model, Q[1,2,1]==sum(P[1,2,:,:,:,1]))
# x=0, y=1, z=1
@constraint(model, Q[1,2,2]==sum(P[1,:,2,:,:,2]))
# x=1, y=0, z=0
@constraint(model, Q[2,1,1]==sum(P[2,:,:,1,:,1]))
# x=1, y=0, z=1
@constraint(model, Q[2,1,2]==sum(P[2,:,:,:,1,2]))
# x=1, y=1, z=0
@constraint(model, Q[2,2,1]==sum(P[2,:,:,2,:,1]))
# x=1, y=1, z=1
@constraint(model, Q[2,2,2]==sum(P[2,:,:,:,2,2]))
# CHSH
# 0 <= P(x0,y(x0,z0)=0 | z0) + P(x1,y(x0,z1)=1 | z1)
+ P(x0,y(x1,z1)=0 | z1) - P(x0,y(x1,z0)=0 | z0) <= 1
@constraint(model, 0 <= sum(P[1,1,:,:,:,1]) + sum(P[2,:,2,:,:,2])
+ sum(P[1,:,:,:,1,2]) - sum(P[1,:,:,1,:,1]))
@constraint(model,      sum(P[1,1,:,:,:,1]) + sum(P[2,:,2,:,:,2])
+ sum(P[1,:,:,:,1,2]) - sum(P[1,:,:,1,:,1]) <= 1)
# 0 <= P(x0,y(x1,z0)=0 | z0) + P(x1,y(x1,z1)=1 | z1)
+ P(x0,y(x0,z1)=0 | z1) - P(x0,y(x0,z0)=0 | z0) <= 1
@constraint(model, 0 <= sum(P[1,:,:,1,:,1]) + sum(P[2,:,:,:,2,2])
+ sum(P[1,:,1,:,:,2]) - sum(P[1,1,:,:,:,1]))
@constraint(model,      sum(P[1,:,:,1,:,1]) + sum(P[2,:,:,:,2,2])
+ sum(P[1,:,1,:,:,2]) - sum(P[1,1,:,:,:,1]) <= 1)
# 0 <= P(x0,y(x0,z1)=0 | z1) + P(x1,y(x0,z0)=1 | z0)
+ P(x0,y(x1,z0)=0 | z0) - P(x0,y(x1,z1)=0 | z1) <= 1
@constraint(model, 0 <= sum(P[1,:,1,:,:,2]) + sum(P[2,2,:,:,:,1])
+ sum(P[1,:,:,1,:,1]) - sum(P[1,:,:,:,1,2]))
@constraint(model,      sum(P[1,:,1,:,:,2]) + sum(P[2,2,:,:,:,1])
+ sum(P[1,:,:,1,:,1]) - sum(P[1,:,:,:,1,2]) <= 1)
```

```
# 0 <= P(x0,y(x1,z1)=0 | z1) + P(x1,y(x1,z0)=1 | z0)
+ P(x0,y(x0,z0)=0 | z0) - P(x0,y(x0,z1)=0 | z1) <= 1
@constraint(model, 0 <= sum(P[1,:,:,:,1,2]) + sum(P[2,:,:,2,:,1])
+ sum(P[1,1,:,:,:,1]) - sum(P[1,:,1,:,:,2]))
@constraint(model,      sum(P[1,:,:,:,1,2]) + sum(P[2,:,:,2,:,1])
+ sum(P[1,1,:,:,:,1]) - sum(P[1,:,1,:,:,2]) <= 1)
# ATE = P(Y(1,0)=1) - P(Y(0,0)=1)
@variable(model, ATE)
@constraint(model, ATE == sum(P[:,:,:,2,:,1]) - sum(P[:,2,:,:,:,1]))
# CDD
lib = CDDLib.Library(:exact)
_hrep = hrep(model)
var_names = dimension_names(_hrep)
# CCD H representation
poly = polyhedron(_hrep, lib)
removehredundancy!(poly)
proj_index = collect(65:73)
proj_names = var_names[proj_index]
proj = project(poly, proj_index)
removehredundancy!(proj)
vrep(proj)
print(proj)
```

VITA

F. Richard Guo was born in Sichuan, China, in April 1991. He received his B.Eng. from University of Electronic Science and Technology of China in 2013, followed by a M.Sc. from Duke University in 2016, both in computer science. He received his Ph.D. in Statistics from University of Washington, Seattle in 2021. He will join the Statistical Laboratory at University of Cambridge as a postdoctoral associate in September 2021.