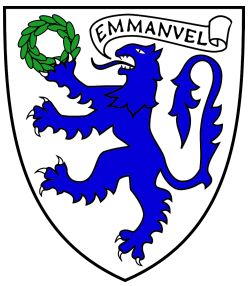
Semiparametric Methods for Two Problems in Causal Inference using Machine Learning



Harvey Carter Klyne

Statistical Laboratory
Department of Pure Mathematics and Mathematical Statistics
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Emmanuel College June 2023

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee. Chapters 2 and 3 are joint work with Rajen Shah (University of Cambridge). Chapter 2 is currently being submitted for publication as Klyne and Shah (2023). We hope to submit Chapter 3 for publication soon.

Harvey Carter Klyne
June 2023

Abstract

Semiparametric Methods for Two Problems in Causal Inference using Machine Learning

Harvey Carter Klyne

Scientific applications such as personalised (precision) medicine require statistical guarantees on causal mechanisms, however in many settings only observational data with complex underlying interactions are available. Recent advances in machine learning have made it possible to model such systems, but their inherent biases and black-box nature pose an inferential challenge. Semiparametric methods are able to nonetheless leverage these powerful nonparametric regression procedures to provide valid statistical analysis on interesting parametric components of the data generating process.

This thesis consists of three chapters. The first chapter summarises the semiparametric and causal inference literatures, paying particular attention to doubly-robust methods and conditional independence testing. In the second chapter, we explore the doubly-robust estimation of the average partial effect — a generalisation of the linear coefficient in a (partially) linear model and a local measure of causal effect. This framework involves two plug-in nuisance function estimates, and trades their errors off against each other. The first nuisance function is the conditional expectation function, whose estimate is required to be differentiable. We propose convolving an arbitrary plug-in machine learning regression — which need not be differentiable — with a Gaussian kernel, and demonstrate that for a range of kernel bandwidths we can achieve the semiparametric efficiency bound at no asymptotic cost to the regression mean-squared error. The second nuisance function is the derivative of the log-density of the predictors, termed the score function. This score function does not depend on the conditional distribution of the response given the predictors. Score estimation is only well-studied in the univariate case. We propose using a location-scale model to reduce the problem of multivariate score estimation to conditional mean and variance estimation plus univariate score estimation. This enables the use of an arbitrary machine learning regression. Simulations confirm the desirable properties of our

approaches, and code is made available in the R package `drape` (Doubly-Robust Average Partial Effects) available from https://github.com/harveyklyne/drape.

In the third chapter, we consider testing for conditional independence of two discrete random variables X and Y given a third continuous variable Z. Conditional independence testing forms the basis for constraint-based causal structure learning, but it has been shown that any test which controls size for all null distributions has no power against any alternative. For this reason it is necessary to restrict the null space, and it is convenient to do so in terms of the performance of machine learning methods. Previous works have additionally made strong structural assumptions on both X and Y. A doubly-robust approach which does not make such assumptions is to compute a generalised covariance measure using an arbitrary machine learning method, reducing the test for conditional correlation to testing whether an asymptotically Gaussian vector has mean zero. This vector is often high-dimensional and naive tests suffer from a lack of power. We propose greedily merging the labels of the underlying discrete variables so as to maximise the observed conditional correlation. By doing so we uncover additional structure in an adaptive fashion. Our test is calibrated using a novel double bootstrap. We demonstrate an algorithm to perform this procedure in a computationally efficient manner. Simulations confirm that we are able to improve power in high-dimensional settings with low-dimensional structure, whilst maintaining the desired size control. Code is made available in the R package `catci` (CATegorical Conditional Independence) available from https://github.com/harveyklyne/catci.

Acknowledgements

It is a great pleasure to be able to thank Rajen Shah for his generous and patient supervision. Throughout these last few years Rajen has provided me with a wealth of excellent ideas and conscientious feedback, and I am truly grateful to have had this opportunity. Throughout my PhD — and especially during COVID — I have been fortunate to have such wonderful friends to rely on, especially Elliot Klyne, Anton Rask Lundborg, Joakim Andersen, Elliot Young, Ben Stokell, Alex Chamolly, Mohit Dhiman, Elise French, Jamie Scott, Lukas Pin, Eric LeGresley, Liza Hadley, Eleanor Clifford, Hanna Martin, Jonathan Hoare, Aureliane Pierret, Louisa Snape, Elise Chang, Pieter Durman, Jordan Smith, and Miriam Hulley. Finally, I am grateful to my examiners Qingyuan Zhao and Oliver Dukes for their thoughtful suggestions which have considerably improved this work.

Contents

Chapter 1

Introduction

This thesis adds to the growing body of literature on semiparametric methods for causal inference. In this introductory chapter we summarise the semiparametric and causal inference literatures, paying particular attention to doubly-robust methods and conditional independence testing. In so doing we motivate the themes developed in Chapters 2 and 3.

The following, much celebrated, quote is attributed to George Box.

All models are wrong, but some are useful.

Real world processes do not follow standard parametric models. Practitioners doing inference based on such models must cross their fingers that the model is not so misspecified as to completely invalidate the conclusions. In many present-day applications this hope is hollow: the mechanisms concerned are very complex and standard parametric techniques entirely fail. This is especially true in statistical problems arising from the field of causal inference, where incorrect modelling assumptions can lead to heavy biases on the analysis.

Many scientific questions are — implicitly or explicitly — causal in nature. One may ask what effect certain observed variables had on one another, or how an intervention on one variable would have propagated to others. Such questions involve unobserved, "counterfactual" variables, and so some causal model for how these out-of-distribution variables operate is required. The underlying processes can be very complex, and non-linearities and variable interactions are often much better handled by modern machine learning methods such as tree ensembles than by classical statistical estimators. These nonparametric regression procedures are individually challenging to analyse ("black-box"), and — worse — the state-of-the-art moves so quickly that analytical insights for today's preferred methods are rendered useless tomorrow. Modern statistical frameworks should therefore make as few assumptions on the underlying regression procedures as possible. This is precisely the remit of semiparametric statistics — inference is done only on a finite-dimensional parameter of interest, and less interesting infinite-dimensional components of the model ("nuisance functions") may be estimated however we like.

Previous works have tended to propose specific estimators for the nuisance functions, particularly those based on kernel smoothing (Nadaraya, 1964; Watson, 1964) and sparse high-dimensional bases (Candes and Tao, 2007; Tibshirani, 1996). In this way the statistician provides the practitioner with a fully-formed estimation and inference procedure, and the statistician is able to take advantage of additional properties of their chosen methods, such as boundedness or smoothness properties. We take the view that the practitioner is the expert on their data, and should be free to plug-in whatever nuisance estimators best suit their needs. With this in mind, Chapters 2 and 3 are totally agnostic to the plug-in estimators, asking only that they satisfy certain (weak) convergence rates. In the rest of this introduction we further elucidate the development of semiparametric methods and their application to causal inference.

1.1 Semiparametric statistics

Semiparametic statistics is interested in estimating and doing inference on a finite-dimensional parameter of interest in the presence of nonparametric "nuisance" components. By focussing on a parametric target we are often able to derive root-n consistent and asymptotically normal estimators. The nonparametric component greatly enlarges the model class, reducing the risks from model misspecification and enabling the use of flexible machine learning procedures. In this section we have made use of work by van der Vaart (1998, 2002) and Kennedy (2016).

Suppose we observe an independent, identically distributed (i.i.d.) sample $(W_1, \ldots, W_n)$ taking values on some space $\mathcal{W}$ and distributed according to an unknown distribution P. A statistical model is a collection of distributions $\mathcal{P}$ on $\mathcal{W}$, which is assumed to contain P. A parametric model is a model which can be indexed by a finite-dimensional vector $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$, and we are usually interested in estimating a sub-vector of the true parameter $\psi \subseteq \theta$. Typically we can estimate $\psi \in \mathbb{R}^d$ at the root-n rate.

Semiparametric models, on the other hand, are also indexed by one or more infinite-dimensional parameters, called the nuisance functions. Even in this case, often we are only interested in estimating part of the parametric component of the model. In this case we can achieve the best of both worlds. Consider the following partially linear regression model $\mathcal{P}$ for a random triple $W = (Y, X, Z)$ taking values in $\mathcal{W} = \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^p$, which is the set of distributions P which satisfy

$$Y = \theta_P^T X + g_P(Z) + \varepsilon_P, \tag{1.1}$$

where $\theta_P \in \mathbb{R}^d$, $g_P : \mathbb{R}^p \to \mathbb{R}$ is continuous, $\mathbb{E}_P(\varepsilon_P \mid X, Z) = 0$, $\mathbb{E}_P\Big(\varepsilon_P^2 \;\Big|\; X, Z\Big) < \infty$. We are still able to estimate the parameter of interest θ_P at the root-n rate, despite the model

containing the nonparametric nuisance functions g_P and the distribution of ε_P. The price we pay is an enlarged multiplicative constant in the asymptotic variance of our estimator.

For a class of distributions $\mathcal{P}$ containing the data-generating distribution P and a functional $\psi : \mathcal{P} \to \mathbb{R}^d$, no unbiased estimator of $\psi(P)$ can achieve variance smaller than the so-called efficient lower bound. An estimator which achieves this bound is called efficient. The idea of efficiency dates back to Fisher (1922, 1925). When $\mathcal{P} = \{P_\theta \ : \ \theta \in \Theta\}$ is a parametric family, this bound is given by the famous Cramér–Rao theorem (Cramér, 1946). In an abuse of notation, we write $\psi(\theta) = \psi(P_\theta)$ and consider $\psi : \Theta \to \mathbb{R}^d$.

Theorem 1 (Cramér–Rao)**.** *If $\theta \mapsto P_\theta$ is differentiable at θ with likelihood p_θ and Fisher information $I_\theta := \mathrm{Var}_\theta\Big(\frac{\partial}{\partial\theta} \log p_\theta(W)\Big)$, and furthermore $T_n = T_n(W_1, \ldots, W_n)$ is an unbiased estimator of $\psi(\theta)$ for a differentiable function $\psi : \mathbb{R} \to \mathbb{R}$, then under regularity conditions $\mathrm{Var}_\theta(\sqrt{n}T_n) \geq \psi'(\theta)^2/I_\theta$.*

A semiparametric model contains a collection of parametric sub-models. The semiparametric efficiency bound is not smaller than the Cramér–Rao bound for any of these parametric sub-models, so taking the supremum over these sub-models' lower bounds gives us a (potentially loose) variance lower bound. In fact it is often possible to achieve a tight bound by considering straightforward classes of sub-models.

In Chapter 2 we consider the average partial effect (defined below), and we include here a derivation of the semiparametric efficiency bound for this parameter, first shown by Newey and Stoker (1993). What follows is not intended as a rigorous proof, but rather to demonstrate the key idea of applying the Cramér–Rao theorem over a collection of parametric sub-models.

Consider a random pair $W = (Y, X)$ taking values in $\mathbb{R}^2$, and obeying the following nonparametric model $\mathcal{P}$. Under each $P \in \mathcal{P}$, the random variables (Y, X) follow a joint density p_P with respect to Lebesgue, the regression function

$$f_P(x) := \mathbb{E}_P[Y \mid X = x]$$

exists and is differentiable, and the marginal density of X is differentiable. The average partial effect is defined as

$$\theta_P := \mathbb{E}_P[f_P'(X)]. \tag{1.2}$$

The functions f_P and p_P satisfy the additional regularity conditions of Newey and Stoker (1993), which we do not restate here.

In the semiparametric notation, we are seeking to do inference on the functional $\psi : P \mapsto \mathbb{E}_P[f_P'(X)]$ for P in the model $\mathcal{P}$. Fix $P \in \mathcal{P}$, and consider the one-dimensional parametric sub-model $\{P_t\} \subset \mathcal{P}$ where P_t has joint density

$$p_t(y, x) = [1 + t\zeta(y, x)][1 + t\gamma(x)]p_P(y, x)$$

for some bounded, continuous functions ζ and γ which satisfy

$$\mathbb{E}_P[\zeta(Y,X) \mid X] = 0; \quad \mathbb{E}_P[\gamma(X)] = 0.$$

Note that $P_0 = P$. Write

$$\begin{aligned} p_t(y \mid x) &= \{1 + t\zeta(y,x)\} p_P(y \mid x); \\ p_t(x) &= \{1 + t\gamma(x)\} p_P(x). \end{aligned}$$

In order to apply Theorem 1 we need to know the Fisher information for the sub-model, which is

$$\mathbb{E}_P\Big[\Big\{\partial_t \log p_t(Y,X)|_{t=0}\Big\}^2\Big] = \mathbb{E}_P\Big[\{\zeta(Y,X) + \gamma(X)\}^2\Big].$$

Write $\xi_P = Y - f_P(X)$, and let $\mathbb{E}_t(\,\cdot \mid X = x)$ and $\mathbb{E}_t(\,\cdot\,)$ denote expectations with respect to the densities $p_t(\,\cdot \mid x)$ and $p_t(\,\cdot\,)$ respectively. Our target function is

$$\psi(t) = \mathbb{E}_t[\partial_x \mathbb{E}_t(Y \mid X)], \tag{1.3}$$

from which we need to compute $\psi'(0)$. By the definition of $p_t(\,\cdot \mid x)$, and using $\mathbb{E}_P[\xi_P \mid X] = \mathbb{E}_P[\zeta(Y,X) \mid X] = 0$, we have that

$$\begin{aligned} \mathbb{E}_t(Y \mid X = x) &= \mathbb{E}_P[Y\{1 + t\zeta(Y,X)\} \mid X = x] \\ &= f_P(x) + t\mathbb{E}_P[\xi_P\, \zeta(Y,X) \mid X = x]. \end{aligned} \tag{1.4}$$

Equation (1.3) suggests that we next need to differentiate (1.4) with respect to x, but this can be avoided using integration by parts in (1.3). Using the definition of $p_t(\,\cdot\,)$, and under appropriate regularity conditions,

$$\begin{aligned} \psi(P_t) &= \mathbb{E}_P[\{1 + t\gamma(X)\}\partial_x \mathbb{E}_t(Y \mid X)] \\ &= \int_{\mathbb{R}} p_P(x)\{1 + t\gamma(x)\}\partial_x \mathbb{E}_t(Y \mid X = x)\, dx \\ &= -\int_{\mathbb{R}} \partial_x\Big(p_P(x)\{1 + t\gamma(x)\}\Big)\mathbb{E}_t(Y \mid X = x)\, dx \\ &= -\mathbb{E}_P\left[\frac{\partial_x\Big(p_P(X)\{1 + t\gamma(X)\}\Big)}{p_P(X)}\mathbb{E}_t(Y \mid X)\right]. \end{aligned} \tag{1.5}$$

Plugging (1.4) in to (1.5), we have

$$\begin{aligned}\psi(P_t) &= -\mathbb{E}_P\left[\frac{\partial_x\Big(p_P(X)\{1+t\gamma(X)\}\Big)}{p_P(X)}\Big\{f_P(X) + t\mathbb{E}_P\Big(\xi_P\,\zeta(Y,X)\mid X\Big)\Big\}\right] \\ &= -\mathbb{E}_P\left[\frac{p_P'(X)}{p_P(X)}f_P(X)\right] \\ &\quad - t\mathbb{E}_P\left[\frac{\partial_x\{p_P(X)\gamma(X)\}}{p_P(X)}f_P(X) + \frac{p_P'(X)}{p_P(X)}\xi_P\,\zeta(Y,X)\right] \\ &\quad - t^2\mathbb{E}_P\left[\frac{\partial_x\{p_P(X)\gamma(X)\}}{p_P(X)}\xi_P\,\zeta(Y,X)\right]. \end{aligned} \tag{1.6}$$

Recall that we are seeking to compute the quantity $\psi'(0)$. Now equation (1.6) implies

$$\psi'(0) = -\mathbb{E}_P\left[\frac{\partial_x\{p_P(X)\gamma(X)\}}{p_P(X)}f_P(X) + \frac{p_P'(X)}{p_P(X)}\xi_P\,\zeta(Y,X)\right].$$

We do not wish to end up with terms involving γ', so we perform integration-by-parts once more on the first term. This yields

$$\psi'(0) = \mathbb{E}_P\left[f_P'(X) - \frac{p_P'(X)}{p_P(X)}\xi_P\,\zeta(Y,X)\right].$$

Using the definition of ξ_P and the conditions $\mathbb{E}_P[\xi_P \mid X] = \mathbb{E}_P[\zeta(Y,X) \mid X] = \mathbb{E}_P[\gamma(X)] = 0$, we may re-write this as

$$\psi'(0) = \mathbb{E}_P\left[\left\{f_P'(X) - \frac{p_P'(X)}{p_P(X)}\{Y - f_P(X)\} - C\right\}\{\zeta(Y,X) + \gamma(X)\}\right],$$

for any constant $C \in \mathbb{R}$. We will soon apply the Cauchy–Schwarz inequality to this expectation, which is tight when

$$f_P'(X) - \frac{p_P'(X)}{p_P(X)}\{Y - f_P(X)\} - C = \zeta(Y,X) + \gamma(X).$$

Since the right-hand side is mean-zero, we choose $C = \theta_P$.

By Theorem 1, the variance lower bound for an unbiased estimator of $\psi(P) = \theta_P$ is at least

$$\frac{\psi'(0)^2}{\mathbb{E}_P\Big[\Big\{\partial_t \log p_t(Y,X)|_{t=0}\Big\}^2\Big]}$$
$$= \frac{\Big(\mathbb{E}_P\Big[\Big\{f_P'(X) - \frac{p_P'(X)}{p_P(X)}\{Y - f_P(X)\} - \theta_P\Big\}\{\zeta(Y,X) + \gamma(X)\}\Big]\Big)^2}{\mathbb{E}_P\Big[\{\zeta(Y,X) + \gamma(X)\}^2\Big]}.$$

Since we could have chosen any suitable functions ζ and γ, we take a supremum to obtain the tightest bound. Indeed,

$$\sup_{\substack{\zeta\,:\,\mathbb{E}_P[\zeta(Y,X)|X=x]=0,\\ \gamma\,:\,\mathbb{E}_P[\gamma(X)]=0}} \frac{\Big(\mathbb{E}_P\Big[\Big\{f_P'(X) - \frac{p_P'(X)}{p_P(X)}\{Y - f_P(X)\} - \theta_P\Big\}\{\zeta(Y,X) + \gamma(X)\}\Big]\Big)^2}{\mathbb{E}_P\Big[\{\zeta(Y,X) + \gamma(X)\}^2\Big]}$$
$$= \mathbb{E}_P\left[\left\{f_P'(X) - \frac{p_P'(X)}{p_P(X)}\{Y - f_P(X)\} - \theta_P\right\}^2\right],$$

where we have applied a tight version of the Cauchy–Schwarz inequality, with $\zeta(y,x) = \frac{p_P'(x)}{p_P(x)}\{y - f_P(x)\}$ and $\gamma(x) = f_P'(x) - \theta_P$. This yields the variance lower bound

$$\mathbb{E}_P\left[\left\{f_P'(X) - \frac{p_P'(X)}{p_P(X)}\{Y - f_P(X)\} - \theta_P\right\}^2\right],$$

which is indeed that proved by Newey and Stoker (1993). The function $(y,x) \mapsto f_P'(x) - \frac{p_P'(x)}{p_P(x)}\{y - f_P(x)\} - \theta_P$ is called the efficient influence function (Kennedy, 2016; van der Vaart, 1998, 2002).

Many functionals of interest may be defined in terms of a moment equation involving the conditional mean function:

$$\psi(P) = \mathbb{E}_P\Big[m(f_P, Y, X)\Big],$$

where m is a fixed, known functional. A naive estimation approach would be to estimate f_P by some $\hat{f}$, and then take the empirical average $\frac{1}{n}\sum_{i=1}^n m(\hat{f}, Y_i, X_i)$ as an estimate of $\psi(P)$. Such approaches would typically suffer from plug-in bias and fail to attain the parametric rate of convergence. As well as poor estimation quality, this would also make inference, that is, performing hypothesis tests or forming confidence intervals, particularly problematic. More accurate results can be obtained by performing a one-step bias

correction, using additional information contained within the distribution of the predictor X. Such estimators were originally developed to account for missing data, and proved consistent if either a parametric model on the response mechanism or the missingness mechanism were correctly specified, hence "doubly-robust" (Robins and Rotnitzky, 2001; Robins et al., 2000; Scharfstein et al., 1999).

A modern perspective on doubly-robust procedures has become known as double machine learning. Rather than hoping to correctly specify one of two parametric models, users construct estimators based on two flexible machine learning methods and ask that they both converge at slower than root-n rates. See the review article Kennedy (2022) for further references.

1.2 Causal inference

Fisher (1958) famously argued that the observed association between smoking and cancer could be explained by external factors, such as certain genes causing an increased likelihood for both taking up smoking and getting cancer. He viewed the relationship between smoking and cancer to be a mere correlation, not causation. There are various frameworks for discussing causality, and each has their advocates and detractors.

So-called potential outcomes (or counterfactuals) represent hypothetical data points which can not be observed. We will never know what would have been if the smokers and non-smokers had each made different choices, or if they had been born with different genes, but we can still think of these as random variables. Potential outcomes date back to Neyman (1923) and were introduced for observational studies by Rubin (1974). The first stage of estimating a causal parameter posed in terms of potential outcomes is to check that it is well-defined in terms of the observed data alone. This identification step is the reduction of a causal problem to an ordinary statistical problem, and often suggests the form of estimators. Rothenhäusler and Yu (2020) give a causal interpretation of the average partial effect — which is the focus of Chapter 2 — as the causal effect of incrementing the predictors by an infinitesimal amount.

For systems of many variables it is often convenient to visualise causal effects as a directed graph. Wright (1921) discusses causation in terms of paths along such graphs. The counterfactual and graphical approaches to causal inference are unified via Single World Intervention Graphs (Richardson and Robins, 2013). Under additional probabilistic assumptions, one may attempt to learn some of the structure of an unknown causal graph from observational data (Pearl, 2009). The most well-known procedure is the PC Algorithm (Spirtes and Glymour, 1991; Spirtes et al., 1993), which works by testing for conditional independence between various subsets of the variables. Conditional independence testing

is also important for classical variable selection and significance testing, see for example Lundborg et al. (2022).

Conditional independence testing is a hard problem. Shah and Peters (2020, Thm. 2) prove that when Z has a continuous component, any test which controls the false rejection rate against all null hypotheses cannot have power against any alternative. Given a distribution for which there is conditional dependence between X and Y given Z and a test which reject the null with probability strictly greater than some $\alpha > 0$, one may construct a null distribution which is sufficiently close so that the test also rejects it with probability strictly greater than α. Since we cannot hope to control size over all null distributions and have any useful power, it is therefore necessary to restrict the class of null distributions under consideration. This is in contrast to unconditional independence testing, where one may calibrate any testing procedure

$$\phi\Big((X_1, Y_1), \ldots, (X_n, Y_n)\Big) \in \mathbb{R}$$

using a permutation test, i.e. the empirical distribution of

$$\Big\{\phi\Big((X_{\sigma(1)}, Y_1), \ldots, (X_{\sigma(n)}, Y_n)\Big) \ : \ \sigma \text{ a permutation of } \{1, \ldots, n\}\Big\}.$$

See Berrett and Samworth (2019, 2021); Hoeffding (1948) for examples. The continuity of Z is necessary for the argument of Shah and Peters (2020, Thm. 2). If Z takes only finitely many values, then one may partition the data based on the value of Z and perform unconditional independence tests on each subset. This sidesteps the hardness issue, controlling size against all null distributions and having non-trivial power against some alternatives. In Chapter 3 we consider conditional independence testing when X and Y are discrete variables, and Z is arbitrary.

Chapter 2

Average partial effect

2.1 Introduction

A common goal of practical data analysis is to quantify the effect that a particular predictor or set of predictors X has on a response Y, whilst accounting for the contribution of a vector of other predictors Z. Single-parameter summaries are often desirable for ease-of-interpretability, as demonstrated by the popularity of (partially) linear models. Such models, however, may not adequately capture the conditional mean of the response, potentially invalidating conclusions drawn. Indeed the successes of model-agnostic regression methods such as XGBoost (Chen and Guestrin, 2016), random forests (Breiman, 2001) and deep learning (Goodfellow et al., 2016) in machine learning competitions such as those hosted by Kaggle (Bojer and Meldgaard, 2021) suggest that such models fitting poorly is to be expected in many contemporary datasets of interest.

When $X \in \mathbb{R}$ is a continuous random variable and the conditional mean $f(x, z) := \mathbb{E}(Y \mid X = x, Z = z)$ is differentiable in the x-direction, a natural quantity of interest is the average slope with respect to x. This is known as the average partial effect (or average derivative), defined as

$$\theta := \mathbb{E}\left[\frac{\partial}{\partial x} f(X, Z)\right].$$

Historically, motivation for this estimand came from semiparametric single-index models, i.e., where $f(x, z) = G(\beta x + \gamma^T z)$; the coefficient β is then proportional to θ. The average partial effect also recovers the linear coefficient in a partially linear model

$$f(x, z) = \theta x + g(z).$$

Thus the average partial effect may be thought of as a generalisation of the coefficient in a partially linear model that appropriately measures the association of X and the response, while controlling for Z, even when there may be complex interactions between X and

Z present. Indeed, θ is also the average slope of the so-called partial dependence plot, popular in the field of interpretable machine learning and often used in conjunction with flexible regression methods that place no explicit restrictions on the form of regression functions to be estimated (Friedman, 2001; Molnar, 2022; Zhao and Hastie, 2021).

Rothenhäusler and Yu (2020) provide a causal interpretation of the average partial effect in the form of an average outcome change if the 'treatment' X of all subjects were changed by an arbitrarily small quantity. More precisely, let us denote by $Y(x)$ the potential outcome (Rubin, 1974) were X to be assigned the value x. Then under so-called weak ignorability, that is $\{Y(x) : x \in \mathbb{R}\} \perp\!\!\!\perp X \mid Z$, overlap, i.e., $p(x \mid z) > 0$ where $p(x \mid z)$ is the conditional density of X given Z, and mild regularity conditions,

$$\theta = \lim_{\delta \to 0} \frac{1}{\delta} \mathbb{E}\{Y(X + \delta) - Y(X)\}.$$

In this sense, θ may be thought of as a continuous analogue of the well-studied average treatment effect functional

$$\mathbb{E}(Y(1) - Y(0)) = \mathbb{E}\{\mathbb{E}(Y \mid Z, X = 1) - \mathbb{E}(Y \mid Z, X = 0)\} = \mathbb{E}\{f(1, Z) - f(0, Z)\}$$

in the case where X is discrete, only taking values 0 or 1, and analogous assumptions as above (Robins and Rotnitzky, 1995; Robins et al., 1994; Scharfstein et al., 1999).

While the average partial effect estimand is attractive from the perspective of interpretability, estimating the derivative of a conditional mean function is challenging. Regression estimators for f which have been trained to have good mean-squared prediction error can produce arbitrarily bad derivative estimates, if they are capable of returning these at all. For example, highly popular tree-based methods give piecewise constant estimated regression functions and so clearly provide unusable estimates for the derivative of f.

Moreover, even if the rate of convergence of the derivative estimates was comparable to the mean-squared prediction error when estimating f nonparametrically, an estimator of θ formed through their empirical average would typically suffer from plug-in bias and fail to attain the parametric rate of convergence. As well as poor estimation quality, this would also make inference, that is, performing hypothesis tests or forming confidence intervals, particularly problematic. The rich theory of semiparametric statistics (Bickel et al., 1993; Tsiatis, 2006) addresses the issue of such plug-in biases more generally, and supports the construction of debiased estimators based on (efficient) influence functions. This basic approach forms a cornerstone of what has become known as debiased machine learning: a collection of methodologies involving user-chosen machine learning methods to produce estimates of nuisance parameters that are used in the construction of estimators

of functionals that enjoy parametric rates of convergence (see for example the review article Kennedy (2022) and references therein).

Procedurally, this often involves modelling both the conditional expectation f and a function of the joint distribution of the predictors (X, Z), with the bias of the overall estimator controlled by a product of biases relating to each of these models (Rotnitzky et al., 2021). For the average partial effect, as shown by Newey and Stoker (1993); Powell et al. (1989), the predictor-based quantity to be estimated is the so-called score function (sometimes termed the negative score function)

$$\rho(x, z) := \frac{\frac{\partial}{\partial x} p(x \mid z)}{p(x \mid z)} = \frac{\partial}{\partial x} \log p(x \mid z),$$

where $p(x \mid z)$ is the (assumed differentiable) conditional density of $X \mid \{Z = z\}$. This has been studied in the unconditional setting (i.e. without any Z present) using estimators based on parametric families (Stoker, 1986), splines (Bera and Ng, 1995; Cox, 1985; Ng, 1994), and kernel smoothing methods (Härdle and Stoker, 1989; Li, 1996; Powell et al., 1989; Stoker, 1990). Nonparametric estimation of the score function in the multivariate setting however is particularly challenging owing to the complex nature of potential interactions. Direct estimation through plugging in a kernel density estimate of the joint density $p(x, z)$ can be plagued by stability issues where the estimated density is small. Recently Sriperumbudur et al. (2017) has considered an approach for multivariate score estimation based on infinite-dimensional exponential families parametrised by a reproducing kernel Hilbert space, and Chernozhukov et al. (2022b) has adapted deep learning architectures and tree splitting criteria to develop neural network and random forest-based approaches for estimating ρ.

One approach to tackling the challenges associated with estimating the derivative of the regression function f and the score function ρ is to assume that f, its derivative, and ρ are all sufficiently well-approximated by sparse linear combinations of basis functions (Chernozhukov et al., 2022d; Rothenhäusler and Yu, 2020). Similarly to the case with the debiased Lasso (Zhang and Zhang, 2014) where regression coefficients can be estimated without placing explicit sparsity assumptions relating to the conditional distribution of X given Z (see for example Shah and Bühlmann (2023)); in this case, fewer assumptions need to be placed on the estimator of ρ (Chernozhukov et al. (2022d) Remark 4.1). A related approach relies on ρ itself being well-approximated by a sparse linear combination of basis functions; see Chernozhukov et al. (2022a); Chernozhukov et al. (2021); Chernozhukov et al. (2022c,d, 2023b); Chernozhukov et al. (2020) for examples of both of these approaches. Hirshberg and Wager (2021) assume that the regression estimation error lies within some absolutely convex class of functions, and perform a convex optimisation to choose weights that minimise the worst-case mean-squared error over this class. In practice, the class

of functions may often be taken as sparse linear combinations of basis functions, and in general it may not always be clear how such basis functions may be chosen. Hirshberg and Wager (2020) and Wooldridge and Zhu (2020) consider parametric single index models for the conditional expectation f; this results in a helpful simplification of the problem in the high-dimensional setting these works consider, but may appear overly restrictive in the more moderate-dimensional settings we have in mind here. The difficulties of estimating θ have led Vansteelandt and Dukes (2022) and Hines et al. (2021) to propose interesting alternative estimands that aim to capture some notion of a conditional association of Y and X, given Z, but whose estimation avoids the challenges of nonparametric multivariate score estimation. Kennedy et al. (2017) instead estimate the whole curve $x \mapsto \mathbb{E}[f(x, Z)]$ using kernel smoothing, and Díaz and van der Laan (2012) consider a causal parameter defined in terms of stochastic interventions on X.

2.1.1 Our contributions and organisation of the chapter

In this chapter we take a different approach, and develop new approaches for addressing the two main challenges in estimating the average partial effect θ using a double machine learning framework as outlined above, namely estimation of the derivative of the conditional mean function f and the multivariate score ρ.

In Section 2.2 we first give a uniform asymptotic convergence result for such doubly robust estimators of θ requiring user-chosen estimators for f and ρ. We argue that uniform results as opposed to pointwise results are particularly important in nonparametric settings such as those considered here, otherwise for any sample size n there may exist candidate data generating mechanisms under which the asymptotic approximation error is not negligible. Indeed, considering the problem of testing for a non-zero partial effect, one can show that this is fundamentally hard: when Z is a continuous random variable, any test must have power at most its size. This comes as a consequence of noting that the null in question contains the null that $X \perp\!\!\!\perp Y \mid Z$, which is known to suffer from this form of impossibility (Shah and Peters, 2020, Thm. 2). This intrinsic hardness means that any non-trivial test must restrict the null further with the form of these additional conditions, which would be revealed in a uniform result but may be absent in a pointwise analysis, providing crucial guidance on the suitability of tests in different practical settings.

In our case, the conditions for our result involve required rates of convergence for estimation of the conditional mean f, the score ρ and also a condition on the quality of our implied estimate of the derivative of f. While estimation of conditional means is a task statisticians are familiar with tackling using machine learning methods, for example, the latter two remain challenging to achieve. In contrast to existing work, rather than relying on well-chosen basis function expansions or developing bespoke estimation tools we aim to leverage once again the predictive ability of modern machine learning methods,

which have a proven track record of success in practice. A general limitation of the type of semiparametric asymptotic theory we present in Section 2.2 is that the nuisance estimation problem may be challenging in moderate dimensions with finite samples (Robins and Ritov, 1997). This can be particularly problematic for certain Bayesian procedures in some settings (Ritov et al., 2014; Robins and Ritov, 1997), although we make use of a frequentist estimator here. In Section 2.5 we demonstrate that our proposed estimator achieves good finite sample results even in challenging settings.

For derivative estimation, we propose a post-hoc kernel smoothing procedure applied to the output of the chosen regression method for estimating f. In Section 2.3 we show that under mild conditions, our resmoothing method achieves consistent derivative estimation (in terms of mean-square error) at no asymptotic cost to estimation of f when comparing to the convergence rate enjoyed by the original regression method. Importantly, we do not require the use of a specific differentiable estimator $\hat{f}$ or any explicit assumptions on its complexity or stability properties. This contrasts in particular much of the literature on estimation of the derivative of a regression function; see for example Dai et al. (2016) and references therein, and also Da Rosa et al. (2008); Fonseca et al. (2018) for smoothing approaches using sigmoid functions specific to tree-based estimators.

Turning to score estimation, we seek to reduce the problem of multivariate score estimation to that of univariate score estimation, which as explained above, is better studied and more tractable. In Section 2.4 we advocate modelling the conditional distribution of $X \mid Z$ as a location-scale model (see for example Kennedy et al. (2017, Sec. 5) who work with this in an application requiring conditional density estimation),

$$X = m(Z) + \sigma(Z)\varepsilon,$$

where ε is mean-zero and independent of Z. Through estimating the conditional mean m and σ via some $\hat{m}$ and $\hat{\sigma}$, one can form scaled residuals $\{X - \hat{m}(Z)\}/\hat{\sigma}(Z)$ which may be fed to a univariate score estimator. Theoretically, we consider settings where ε is sub-Gaussian and σ is nonparametric, and also the case where ε is allowed to be heavy-tailed and $\sigma = 1$ (i.e. a location only model where the errors $X - m(Z) \perp\!\!\!\perp X$). We also demonstrate good numerical performance in heterogeneous, heavy-tailed settings. Given how even the univariate score involves a division by a density, one concern might be that any errors in estimating m and σ may propagate unfavourably to estimation of the score. We show however that the estimation error for the multivariate $\rho(x, z)$ may be bounded by the sum of the estimation errors for the conditional mean m, the conditional scale σ and the univariate score function for the residual ε alone. In this way we reduce the problem of multivariate score estimation to univariate score estimation, plus regression and heterogeneous scale estimation, all of which may be relatively more straightforward.

Our results rely on proving a sub-Gaussianity property of Lipschitz score functions, which may be of independent interest.

Numerical comparisons of our methodology to existing approaches are contained in Section 2.5, where we demonstrate in particular that the coverage properties of confidence intervals based on our estimator have favourable coverage over a range of settings, both where our theoretical assumptions are met and where they are not satisfied. We conclude with a discussion in Section 2.6. Proofs and additional results are relegated to later sections in this chapter. We provide an implementation of our methods in the R package `drape` (Doubly Robust Average Partial Effects) available from https://github.com/harveyklyne/drape.

2.1.2 Notation

Let (Y, X, Z) be a random triple taking values in $\mathbb{R} \times \mathbb{R}^d \times \mathcal{Z}$, where $\mathcal{Z}$ may be a mixture of discrete and continuous space. In order to present results that are uniform over a class of distributions P for (Y, X, Z), we will often subscript associated quantities by P. For example when $(Y, X, Z) \sim P$, we denote by $\mathbb{P}_P((Y, X, Z) \in A)$, the probability that (Y, X, Z) lies in a (measurable) set A, and write $f_P(x, z) := \mathbb{E}_P(Y \mid X = x, Z = z)$ for the conditional mean function.

Let $\mathcal{P}_0$ be the set of distributions P for (Y, X, Z) where both f_P and the conditional density of the predictors (with respect to the Lebesgue measure) $p_P(x \mid z)$ exist and are differentiable over all of $\mathbb{R}^d$, for almost every $z \in \mathcal{Z}$.

Write ∇ for the d-dimensional differentiation operator with respect to the x, which we replace with $'$ if we are enforcing the case $d = 1$. For each $P \in \mathcal{P}_0$, define the score function $\rho_P(x, z) := \nabla \log p_P(x \mid z)$, where $p_P(x \mid z)$ is the conditional probability density of $X \mid \{Z = z\}$ according to P. Note that ρ_P exists for each $P \in \mathcal{P}_0$ and almost every (X, Z), taking values in $\mathbb{R}^d$. Denote by Φ the standard d-dimensional normal cumulative distribution function (c.d.f.), and understand inequalities between vectors to apply elementwise.

We will sometimes introduce standard Gaussian random variables $W \sim N(0, 1)$ independent of (X, Z). Recall that a random variable $X \in \mathbb{R}$ is sub-Gaussian with parameter σ if it satisfies $\mathbb{E}[\exp(\lambda X)] \leq \exp(\lambda^2 \sigma^2 / 2)$ for all $\lambda \in \mathbb{R}$. A vector $V \in \mathbb{R}^d$ is sub-Gaussian with parameter σ if $u^T V$ is sub-Gaussian with parameter σ for any $u \in \mathbb{R}^d$ satisfying $\|u\|_2 = 1$.

As in Lundborg et al. (2022), given a family of sequences of real-valued random variables $(W_{P,n})_{P \in \mathcal{P}, n \in \mathbb{N}}$ taking values in a finite-dimensional vector space and whose distributions are determined by $P \in \mathcal{P}$, we write $W_{P,n} = o_{\mathcal{P}}(1)$ if $\sup_{P \in \mathcal{P}} \mathbb{P}_P(|W_{P,n}| > \epsilon) \to 0$ for every $\epsilon > 0$. Similarly, we write $W_{P,n} = O_{\mathcal{P}}(1)$ if, for any $\epsilon > 0$, there exist $M_\epsilon, N_\epsilon > 0$ such that $\sup_{n \geq N_\epsilon} \sup_{P \in \mathcal{P}} \mathbb{P}_P(|W_{P,n}| > M_\epsilon) < \epsilon$. Given a second family of

sequences of random variables $(V_{P,n})_{P\in\mathcal{P},n\in\mathbb{N}}$, we write $W_{P,n} = o_{\mathcal{P}}(V_{P,n})$ if there exists $R_{P,n}$ with $W_{P,n} = V_{P,n}R_{P,n}$ and $R_{P,n} = o_{\mathcal{P}}(1)$; likewise, we write $W_{P,n} = O_{\mathcal{P}}(V_{P,n})$ if $W_{P,n} = V_{P,n}R_{P,n}$ and $R_{P,n} = O_{\mathcal{P}}(1)$. If $W_{P,n}$ is vector or matrix-valued, we write $W_{P,n} = o_{\mathcal{P}}(1)$ if $\|W_{P,n}\| = o_{\mathcal{P}}(1)$ for some norm, and similarly $O_{\mathcal{P}}(1)$. By the equivalence of norms for finite-dimensional vector spaces, if this holds for some norm then it holds for all norms.

2.2 Doubly robust average partial effect estimator

We consider a nonparametric model

$$\mathbb{E}_P(Y \mid X, Z) =: f_P(X, Z),$$

for a response $Y \in \mathbb{R}$, continuous predictors of interest $X \in \mathbb{R}^d$, and additional predictors $Z \in \mathcal{Z}$ of arbitrary type. We assume that $(Y, X, Z) \sim P \in \mathcal{P}_0$, and so the conditional mean $f_P(x, z)$ and the conditional density $p_P(x \mid z)$ are differentiable with respect to x. Our goal is to do inference on the average partial effect

$$\theta_P := \mathbb{E}_P[\nabla f_P(X, Z)].$$

Recall that the score function ρ_P plays an important role when considering estimation of θ_P because it acts like the differentiation operator in the following sense; see also Newey and Stoker (1993); Stoker (1986).

Proposition 2. *Let the conditional density of the predictors $p_P(\cdot \mid z)$ exist and be differentiable in the jth coordinate x_j for every $x_{-j} \in \mathbb{R}^{d-1}$ and $z \in \mathcal{Z}$. Let $g : \mathbb{R}^d \times \mathcal{Z} \to \mathbb{R}$ similarly be differentiable with respect to x_j and satisfy*

$$\mathbb{E}_P|\nabla_j g(X, Z) + \rho_{P,j}(X, Z)g(X, Z)| < \infty. \tag{2.1}$$

Suppose for every (x_{-j}, z) there exist sequences $a_n \to -\infty$, $b_n \to \infty$ such that

$$\lim_{n\to\infty}\{g(b_n, x_{-j}, z)p_P(b_n, x_{-j} \mid z) - g(a_n, x_{-j}, z)p_P(a_n, x_{-j} \mid z)\} = 0;$$

here with some abuse of notation, we write for example $g(a, x_{-j}, z)$ for

$$g(x_1, \ldots, x_{j-1}, a, x_{j+1}, \ldots, x_d, z).$$

Then

$$\mathbb{E}_P[\nabla_j g(X, Z) + \rho_{P,j}(X, Z)g(X, Z)] = 0.$$

Proposition 2, which follows from integration-by-parts, allows one to combine estimates of f_P and ρ_P to produce a doubly-robust estimator as we now explain. Suppose we have some fixed function estimates $(\hat{f}, \hat{\rho})$, for example computed using some independent auxiliary data, and $g := f_P - \hat{f}$ obeys the conditions of Proposition 2. Then

$$\begin{aligned}\mathbb{E}_P\Big[\nabla\hat{f}(X,Z) - \hat{\rho}(X,Z)\Big\{Y - \hat{f}(X,Z)\Big\}\Big] - \theta_P \\ = \mathbb{E}_P\Big[\Big\{\rho_P(X,Z) - \hat{\rho}(X,Z)\Big\}\Big\{f_P(X,Z) - \hat{f}(X,Z)\Big\}\Big],\end{aligned} \tag{2.2}$$

which will be zero if either $\hat{f}$ or $\hat{\rho}$ equal f_P or ρ_P respectively. Given independent, identically distributed (i.i.d.) samples $(y_i, x_i, z_i) \sim P$ for $i = 1, \ldots, n$, this motivates an average partial effect estimator of the form

$$\frac{1}{n}\sum_{i=1}^{n} \nabla\hat{f}(x_i, z_i) - \hat{\rho}(x_i, z_i)\Big\{y_i - \hat{f}(x_i, z_i)\Big\}.$$

From (2.2) and using the Cauchy–Schwarz inequality, we see that the squared-bias of such an estimator is at worst the product of the mean-square error rates of the conditional mean and score function estimates. A consequence of this (see Theorem 3 below) is that the average partial effect estimate can achieve root-n consistency even when both conditional mean and score function estimators converge at a slower rate. Such an estimator is typically called doubly robust (Robins and Rotnitzky, 2001; Robins et al., 2000; Scharfstein et al., 1999).

In practice, the function estimates $\hat{f}$ and $\hat{\rho}$ would not be fixed and must be computed from the same data. For our theoretical analysis, it is helpful to have independence between the function estimates and the data points on which they are evaluated. For this reason we mimic the setting with auxiliary data by employing a sample-splitting scheme known as cross-fitting (Chernozhukov et al., 2018; Schick, 1986), which works as follows.

Given a sequence of i.i.d. data sets $\{(y_i, x_i, z_i) \ : \ i = 1, \ldots, n\}$, define a K-fold partition $\left(I^{(n,k)}\right)_{k=1,\ldots,K}$ of $\{1, \ldots, n\}$ for some K fixed (in all our numerical experiments we take $K = 5$). For simplicity of our exposition, we assume that n is a multiple of K and each subset is of equal size n/K. Let the pair of function estimates $\left(\hat{f}^{(n,k)}, \hat{\rho}^{(n,k)}\right)$ be estimated using data

$$D^{(n,k)} := \Big\{(y_i, x_i, z_i) \ : \ i \in \{1, \ldots, n\} \setminus I^{(n,k)}\Big\}.$$

The cross-fitted, doubly-robust estimator is

$$\hat{\theta}^{(n)} := \frac{1}{n}\sum_{k=1}^{K}\sum_{i \in I^{(n,k)}} \nabla\hat{f}^{(n,k)}(x_i, z_i) - \hat{\rho}^{(n,k)}(x_i, z_i)\Big\{y_i - \hat{f}^{(n,k)}(x_i, z_i)\Big\}, \tag{2.3}$$

with corresponding variance estimator

$$\hat{\Sigma}^{(n)} := \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I^{(n,k)}}\Big[\nabla\hat{f}^{(n,k)}(x_i,z_i)-\hat{\rho}^{(n,k)}(x_i,z_i)\Big\{y_i-\hat{f}^{(n,k)}(x_i,z_i)\Big\}-\hat{\theta}^{(n)}\Big] \\ \Big[\nabla\hat{f}^{(n,k)}(x_i,z_i)-\hat{\rho}^{(n,k)}(x_i,z_i)\Big\{y_i-\hat{f}^{(n,k)}(x_i,z_i)\Big\}-\hat{\theta}^{(n)}\Big]^T. \tag{2.4}$$

In the next section, we study the asymptotic behaviour of these estimators.

2.2.1 Uniform asymptotic properties

There has been a flurry of recent work (Chernozhukov et al., 2022a; Chernozhukov et al., 2021; Chernozhukov et al., 2022b,c,d, 2023b; Chernozhukov et al., 2020) on doubly-robust inference on a broad range of functionals of the conditional mean function, satisfying a moment equation of the form

$$\mathbb{E}_P[\Psi(X,Z;f_P)]=\beta_P,$$

for a known operator Ψ and unknown target parameter β_P. This encompasses estimation of θ_P by taking $\Psi(x,z;\Delta)=\partial_x\Delta(x,z)$. The general theory in this line of work however typically relies on a mean-squared continuity assumption of the form

$$\mathbb{E}_P\Big[\{\Psi(X,Z;\Delta)\}^2\Big]\leq C\Big\{\mathbb{E}_P\Big[\Delta^2(X,Z)\Big]\Big\}^q,$$

for some $q>0$ and all $\Delta\in\mathcal{F}$, where $\mathcal{F}$ contains the conditional mean estimation errors $f_P-\hat{f}$. This is a potentially strong assumption is our setting, which we wish to avoid. We therefore give below a uniform asymptotic result relating to estimators $\hat{\theta}^{(n)}$ (2.3) and $\hat{\Sigma}^{(n)}$ (2.4), not claiming any substantial novelty (see, for example, Chernozhukov et al. (2018, Thm. 5.1) for a similar theorem regarding the binary Average Treatment Effect, and Chernozhukov et al. (2022d, Cor. 4.1) and Rothenhäusler and Yu (2020, Lem. 5) for similar theorems with specific nuisance estimators), but so as to introduce the quantities $A_f^{(n)}, A_\rho^{(n)}, E_f^{(n)}, E_\rho^{(n)}$ which we will seek to bound in later sections. We stress that it is in achieving these bounds in a model-agnostic way that our major contributions lie.

Recall that the optimal variance bound is equal to the variance of the efficient influence function, which for the average partial effect in the nonparametric model $P\in\mathcal{P}_0$ is equal to

$$\psi_P(y,x,z):=\nabla f_P(x,z)-\rho_P(x,z)\{y-f_P(x,z)\}-\theta_P,$$

provided that $\Sigma_P:=\mathbb{E}_P\Big[\psi_P(Y,X,Z)\psi_P(Y,X,Z)^T\Big]$ exists and is non-singular (Newey and Stoker, 1993, Thm. 3.1). The theorem below shows that $\hat{\theta}$ achieves this variance bound asymptotically.

Theorem 3. *Define the following sequences of random variables:*

$$\begin{aligned}
A_f^{(n)} &:= \mathbb{E}_P\Big[\{f_P(X,Z) - \hat{f}^{(n,1)}(X,Z)\}^2 \;\Big|\; D^{(n,1)}\Big],\\
A_\rho^{(n)} &:= \max_{j=1,\dots,d} \mathbb{E}_P\Big[\{\rho_{P,j}(X,Z) - \hat{\rho}_j^{(n,1)}(X,Z)\}^2 \;\Big|\; D^{(n,1)}\Big],\\
E_f^{(n)} &:= \max_{j=1,\dots,d} \mathbb{E}_P\Big(\Big[\nabla_j f_P(X,Z) - \nabla_j \hat{f}^{(n,1)}(X,Z)\\
&\qquad\qquad + \rho_{P,j}(X,Z)\{f_P(X,Z) - \hat{f}^{(n,1)}(X,Z)\}\Big]^2 \;\Big|\; D^{(n,1)}\Big),\\
E_\rho^{(n)} &:= \max_{j=1,\dots,d} \mathbb{E}_P\Big[\{\rho_{P,j}(X,Z) - \hat{\rho}_j^{(n,1)}(X,Z)\}^2 \operatorname{Var}_P(Y \mid X,Z) \;\Big|\; D^{(n,1)}\Big];
\end{aligned}$$

note that we have suppressed P-dependence in the quantities defined above. Let $\mathcal{P} \subset \mathcal{P}_0$ and the chosen methods producing $\hat{f}^{(n,1)}$ and $\hat{\rho}^{(n,1)}$ be such that all of the following hold. The covariance matrix Σ_P exists for every $P \in \mathcal{P}$, with minimum eigenvalue at least $c_1 > 0$. Furthermore,

$$\sup_{P\in\mathcal{P}} \mathbb{E}_P\Big[\|\psi_P(Y,X,Z)\|_2^{2+\eta}\Big] \le c_2$$

for some $c_2, \eta > 0$. Finally, suppose the remainder terms defined above satisfy:

$$A_f^{(n)} = O_{\mathcal{P}}(1); \quad A_f^{(n)} A_\rho^{(n)} = o_{\mathcal{P}}(n^{-1}); \quad E_f^{(n)} = o_{\mathcal{P}}(1); \quad E_\rho^{(n)} = o_{\mathcal{P}}(1). \tag{2.5}$$

Then the doubly robust average partial effect estimator (2.3) *is root-n consistent, asymptotically Gaussian and efficient:*

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}} \sup_{t\in\mathbb{R}^d} \Big|\mathbb{P}_P\Big[\sqrt{n}(\Sigma_P)^{-1/2}(\hat{\theta}^{(n)} - \theta_P) \le t\Big] - \Phi(t)\Big| = 0,$$

and moreover the covariance estimate (2.4) *satisfies $\hat{\Sigma}^{(n)} = \Sigma_P + o_{\mathcal{P}}(1)$, and one may perform asymptotically valid inference (e.g. constructing confidence intervals) using*

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}} \sup_{t\in\mathbb{R}^d} \Big|\mathbb{P}_P\Big[\sqrt{n}(\hat{\Sigma}^{(n)})^{-1/2}(\hat{\theta}^{(n)} - \theta_P) \le t\Big] - \Phi(t)\Big| = 0.$$

The assumptions on $A_f^{(n)}$, $A_\rho^{(n)}$ and $E_\rho^{(n)}$ are relatively weak and standard; for example they are satisfied if the conditional variance $\operatorname{Var}_P(Y \mid X, Z)$ is bounded almost surely and each of $A_f^{(n)}$, $A_\rho^{(n)}$ converge at the nonparametric rate $o_{\mathcal{P}}(n^{-1/2})$; see Section 2.4 for our scheme on score estimation. For example, consider the case where $\mathcal{Z} = \mathbb{R}^p$ and f_P is $s > 0$ Hölder smooth, i.e., writing $m := \lceil s \rceil - 1$, for every $\alpha := (\alpha_1, \dots, \alpha_{d+p})$ with

$\alpha_1 + \cdots + \alpha_d = m$ and $\alpha_j \in \mathbb{Z}_{\geq 0}$, the partial derivatives (assumed to exist) satisfy

$$\left| \frac{\partial^\alpha f_P}{\partial^{\alpha_1} x_1 \cdots \partial^{\alpha_d} x_d \, \partial^{\alpha_{d+1}} z_1 \cdots \partial^{\alpha_{d+p}} z_p}(x,z) - \frac{\partial^\alpha f_P}{\partial^{\alpha_1} x_1 \cdots \partial^{\alpha_d} x_d \, \partial^{\alpha_{d+1}} z_1 \cdots \partial^{\alpha_{d+p}} z_p}(x',z') \right| \leq C \|(x,z) - (x',z')\|_2^{s-m}$$

for all $P \in \mathcal{P}$ and $(x,z), (x',z') \in \mathbb{R}^{d+p}$. Then we can expect that $A_f^{(n)} = O_{\mathcal{P}}(n^{-2s/(2s+d+p)})$ for appropriately chosen regression procedures; see for example Györfi et al. (2002). Then when $s > (d+p)/2$, this is $o_{\mathcal{P}}(n^{-1/2})$. Moreover, a faster rate for $A_f^{(n)}$ permits a slower rate for $A_\rho^{(n)}$ and vice versa.

However assuming $E_f^{(n)} = o_{\mathcal{P}}(1)$ needs justification. In particular, while the result aims to give guarantees for a version of $\hat{\theta}$ constructed using arbitrary user-chosen regression function and score estimators $\hat{f}^{(n,1)}$ and $\hat{\rho}^{(n,1)}$, in particular it requires $\hat{f}^{(n,1)}$ to be differentiable in the x coordinates, which for example is not the case for popular tree-based estimates of f_P. In the next section, we address this issue by proposing a resmoothing scheme to yield a suitable estimate of f_P that can satisfy the requirements on $A_f^{(n)}$ and $E_f^{(n)}$ simultaneously.

2.3 Resmoothing

In this section we propose doing derivative estimation via a kernel convolution applied to an arbitrary initial regression function estimate. This is inspired by similar approaches for edge detection in image analysis (Canny, 1986). We do this operation separately for each dimension ($x_1, \ldots, x_d$, d fixed) of interest, so without loss of generality in this section we take the dimension of x to be $d = 1$ (so $\nabla(\cdot) = (\cdot)'$). Motivated by Theorem 3, we seek a class of differentiable regression procedures so that the errors

$$A_f^{(n)} := \mathbb{E}_P\Big[\{f_P(X,Z) - \hat{f}^{(n,1)}(X,Z)\}^2 \;\Big|\; D^{(n,1)}\Big],$$
$$E_f^{(n)} := \mathbb{E}_P\bigg(\Big[f_P'(X,Z) - (\hat{f}^{(n,1)})'(X,Z) + \rho_P(X,Z)\{f_P(X,Z) - \hat{f}^{(n,1)}(X,Z)\}\Big]^2 \;\Big|\; D^{(n,1)}\bigg),$$

satisfy

$$A_f^{(n)} = O_{\mathcal{P}}(n^{-\alpha}); \quad E_f^{(n)} = o_{\mathcal{P}}(1),$$

for $\alpha > 0$ as large as possible.

Consider regressing Y on (X,Z) using some favoured machine learning method, whatever that may be. By training on $D^{(n,k)}$ we get a sequence of estimators $\tilde{f}^{(n,k)}$ of the conditional mean function $f_P(x,z) := \mathbb{E}(Y \mid X = x, Z = z)$, which we expect to have a good mean-squared error convergence rate but that are not necessarily differentiable (or that their derivatives are hard to compute, or numerically unstable). Additional

smoothness may be achieved by convolving $\tilde{f}^{(n,k)}$ with a kernel, yielding readily computable derivatives. Let $K : \mathbb{R} \to \mathbb{R}$ be a differentiable kernel function. The convolution yields a new regression estimator

$$\hat{f}^{(n,k)}(x,z) = \{\tilde{f}^{(n,k)}(\cdot,z) * K_{h_n}\}(x) = \int_{\mathbb{R}} \tilde{f}^{(n,k)}(u,z) K_{h_n}(x-u)\, du,$$

where $K_{h_n}(t) = h_n^{-1} K(h_n^{-1} t)$ for a sequence of bandwidths $h_n > 0$. Here we will use a standard Gaussian kernel,

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right),$$

but we do not expect the choice of kernel to be critical. In kernel smoothing, other popular choices include the box kernel $K(u) = \frac{1}{2}\mathbb{1}\{|u| \leq 1\}$, the Epanechnikov kernel $K(u) = \frac{3}{4}(1-u^2)\mathbb{1}\{|u| \leq 1\}$, and the tricube kernel $K(u) = \frac{70}{81}(1-|u|^3)^3\mathbb{1}\{|u| \leq 1\}$ (Wasserman, 2006), although these are not everywhere differentiable. The Gaussian kernel is positive, symmetric, and satisfies $K'(u) = -uK(u)$, which makes it convenient for a theoretical analysis.

We provide visual examples of resmoothing in Figure 2.1. In the left-hand plot we apply our procedure to a smoothing spline regression $\tilde{f}$, which is already capable of returning derivative estimates. In the right-hand plot we instead use a random forest as $\tilde{f}$, which is piecewise constant and so not appropriate for derivative estimation. In both cases our resmoothed estimator seems to capture the underlying smoothness of the true regression function.

2.3.1 Theoretical results

Our goal in this section is to demonstrate that for some sequence of bandwidths h_n and a class of distributions $\mathcal{P} \subset \mathcal{P}_0$ that will encode any additional assumptions we need to make, we have relationships akin to

$$A_f^{(n)} = O_{\mathcal{P}}\left(\mathbb{E}_P\left[\left\{f_P(X,Z) - \tilde{f}^{(n,1)}(X,Z)\right\}^2 \,\middle|\, D^{(n,1)}\right]\right) \quad \text{and} \quad E_f^{(n)} = o_{\mathcal{P}}(1). \tag{2.6}$$

This means that we can preserve the mean squared error properties of the original $\tilde{f}^{(n,1)}$ but also achieve the required converge to zero of the term $E_f^{(n)}$; see Theorem 3. A result of this flavour is given by the following theorem.

Theorem 4. *Define the following random quantities*

$$\tilde{A}_f^{(n)} := \mathbb{E}_P\left[\left\{f_P(X,Z) - \tilde{f}^{(n,1)}(X,Z)\right\}^2 \,\middle|\, D^{(n,1)}\right];$$
$$\tilde{B}_f^{(n)} := \left(\mathbb{E}_P\left[\left|f_P(X,Z) - \tilde{f}^{(n,1)}(X,Z)\right|^{2+\eta} \,\middle|\, D^{(n,1)}\right]\right)^{\frac{2}{2+\eta}},$$

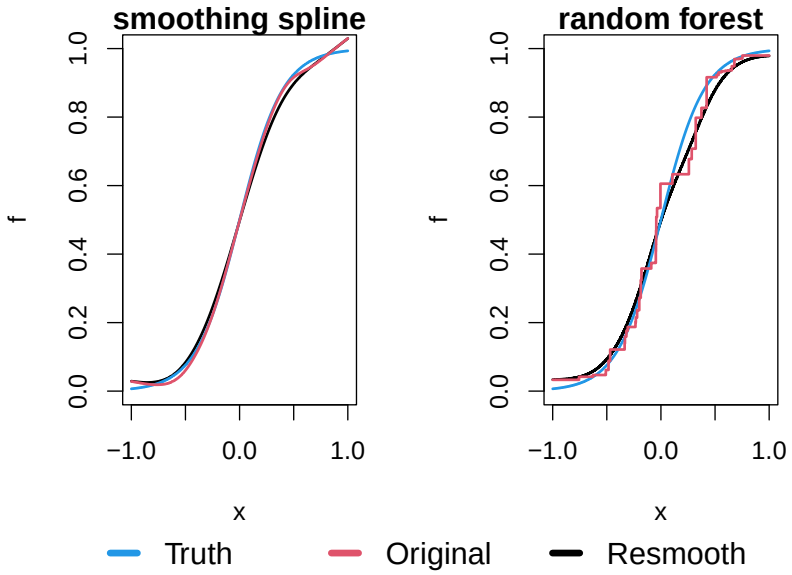


Figure 2.1 Visual example of our resmoothing procedure applied to two orginal regression functions (red lines), fitted to data from a smooth, univariate regression model (blue lines). The resmoothed estimators (black lines) use bandwidth chosen by our data-driven selection procedure (Algorithm 1). In the left-hand plot we have used a smoothing spline for the original regression (`stats` package (R Core Team, 2021)), and in the right-hand plot a random forest (`grf` package (Athey et al., 2019)) with 10 trees.

where $\eta > 0$ is a uniform constant, and the randomness is over the training data set $D^{(n,1)} \sim P$. Let $\mathcal{P} \subset \mathcal{P}_0$ and the chosen regression method producing the fitted regression function $\tilde{f}^{(n,1)}$ be such that all of the following hold. The regression error of $\tilde{f}^{(n,1)}$ is bounded with high probability:

$$\sup_{x,z} \left| f_P(x,z) - \tilde{f}^{(n,1)}(x,z) \right| = O_{\mathcal{P}}(1).$$

For each $P \in \mathcal{P}$ and almost every $z \in \mathcal{Z}$ the conditional density $p_P(\cdot \mid z)$ is twice continuously differentiable and

$$\sup_{P \in \mathcal{P}} \sup_{x,z} \left| \rho'_P(x,z) \right| < \infty. \tag{2.7}$$

There exists a class of functions $C_P : \mathbb{R} \to \mathbb{R}$, $P \in \mathcal{P}$ such that $\sup_{P \in \mathcal{P}} \mathbb{E}_P \left[C_P^2(Z) \right] < \infty$ and

$$\sup_{x} \left| f''_P(x,z) \right| \leq C_P(z) \tag{2.8}$$

for almost every $z \in \mathcal{Z}$, for each $P \in \mathcal{P}$.

If $\tilde{A}_f^{(n)} = O_{\mathcal{P}}\left(n^{-\alpha}\right)$ and $\tilde{B}_f^{(n)} = O_{\mathcal{P}}\left(n^{-\beta}\right)$ for $\alpha, \beta > 0$, then the choice $h_n = cn^{-\gamma}$ for any $c > 0$ and

$$\gamma \in \left[\max\left(\frac{\alpha}{4}, \frac{\alpha - \beta}{2} \right), \frac{\alpha}{2} \right),$$

achieves $A_f^{(n)} = O_{\mathcal{P}}\left(n^{-\alpha}\right)$ and $E_f^{(n)} = O_{\mathcal{P}}\left(n^{2\gamma - \alpha}\right) = o_{\mathcal{P}}(1)$.

For an $\eta > 0$ that is small, one might expect that the squared $L_{2+\eta}$ error of the initial estimator $\tilde{f}^{(n,1)}$ given by $\tilde{B}_f^{(n)}$ is fairly close to the corresponding squared L_2 error $\tilde{A}_f^{(n)}$; in this case we might expect the range of permissible γ to be the generous interval $[\alpha/4, \alpha/2)$, indicating that the particular choice of resmoothing bandwidth is not critical asymptotically. That is, we have a large range of possible bandwidth sequences, whose decay to zero varies in orders of magnitude, for which we have the desirable conclusion that the associated smoothed estimator $\hat{f}^{(n,1)}$ enjoys $A_f^{(n)} = O_{\mathcal{P}}\left(n^{-\alpha}\right)$ as in the case of the original $\tilde{f}^{(n,1)}$, but crucially also $E_f^{(n)} = o_{\mathcal{P}}(1)$ as required by Theorem 3. The availability of theoretical guarantees on modern machine learning methods is limited. When f_P is in a $d-$dimensional Hölder class with smoothness $s > 0$, various regression procedures can achieve $\tilde{A}_f^{(n)} = O_{\mathcal{P}}(n^{-2s/(2s+d)})$ (Györfi et al., 2002) and deep neural networks can achieve $\tilde{A}_f^{(n)} = O_{\mathcal{P}}(n^{-s/(s+d)})$ (Farrell et al., 2021), up to logarithmic terms in n.

The additional assumptions required by the result are a Lipschitz property of the score function (2.7) and a particular bound on the second derivative of the regression function f_P (2.8), both of which we consider to be relatively mild. As shown in Theorem 5, the

former condition implies a sub-Gaussianity property of the score function. The latter condition satisfied if f''_P is uniformly bounded in that $\sup_{P\in\mathcal{P}} \sup_{x,z} \left|f''_P(x,z)\right| < \infty$.

Comparing Theorem 4 to classical results on the Nadaraya–Watson kernel smoothing estimator (Nadaraya, 1964; Watson, 1964), these typically require $\gamma = 1/5$ for optimal mean-squared error (Wasserman (2006, Thm. 5.28)). We are able to allow a range of bandwidths because we are plugging in regression estimators $\tilde{f}^{(n,k)}$, rather than just using the noisy observations Y. This means that we do not get the usual bias–variance tradeoff in the bandwidth h_n, with the leading terms in our mean-squared error being $\tilde{A}_f^{(n)} + h_n^2 \tilde{B}_f^{(n)} + h_n^4$, as opposed to $(nh_n)^{-1} + h_n^4$. The derivative of the Nadaraya–Watson estimator can provide consistent estimates for $\gamma \in (0, 1/3)$ (Collomb, 1979; Schuster and Yakowitz, 1979), with kernel K being a Lipschitz probability density function. In Theorem 4 we fix K to be the Gaussian kernel for analytical convenience, and leave generalisations to future work.

With this result on the insensitivity of the bandwidth choice with respect to the adherence to the conditions of Theorem 3 in hand, we now discuss a simple practical scheme for choosing an appropriate bandwidth.

2.3.2 Practical implementation

There are two practical issues that require consideration. First, we must decide how to compute the convolutions. Second, we discuss bandwidth selection for the kernel for which we suggest a data-driven selection procedure, which picks the largest resmoothing bandwidth achieving a cross-validation score within some specified tolerance of the original regression.

Now for $W \sim N(0,1)$ independent of (X, Z), we have

$$\hat{f}^{(n,k)}(x,z) = \mathbb{E}\Big[\tilde{f}^{(n,k)}(x + hW, z) \;\Big|\; D^{(n,k)}\Big];$$
$$\hat{f}^{(n,k)\prime}(x,z) = \frac{1}{h}\mathbb{E}\Big[W\tilde{f}^{(n,k)}(x + hW, z) \;\Big|\; D^{(n,k)}\Big],$$

the latter expression following from differentiating under the integral sign; see Lemma 13 for a derivation. While this indicates it is possible to compute Gaussian expectations to any degree of accuracy by Monte Carlo, the regression function $\tilde{f}^{(n,k)}$ may be expensive to evaluate too many times, and we have found that derivative estimates are sensitive to sample moments deviating from their population values. These issues can be alleviated by using antithetic variates, however we have found the simpler solution of grid-based numerical integration (as is common in image processing (Canny, 1986)) to be very effective. We require a deterministic set of pairs $\{(w_j, q_j) \,:\, j = 1, \ldots, J\}$ such that, for

functions g,

$$\mathbb{E}[g(W)] \approx \sum_{j=1}^{J} g(w_j) q_j.$$

We suggest taking the $\{w_j\}$ to be an odd number of equally spaced grid points covering the first few standard deviations of W, and $\{q_j\}$ to be proportional to the corresponding densities, such that $\sum_{j=1}^{J} q_j = 1$. This ensures that the odd sample moments are exactly zero and the leading even moments are close to their population versions. In particular we suggest computing the resmoothed regression and derivative estimates as

$$\hat{f}^{(n,k)}(x, z) = \sum_{j=1}^{J} \tilde{f}^{(n,k)}(x + h w_j, z) q_j;$$

$$\hat{f}^{(n,k)\prime}(x, z) = \frac{1}{h} \sum_{j=1}^{J} w_j \tilde{f}^{(n,k)}(x + h w_j, z) q_j.$$

See Section 2.11.2 for further details. We do not expect the use of the Gaussian kernel to be essential to resmoothing in general, however the derivative formulas here are based on the property $K'(u) = -uK(u)$.

Recall that the goal of resmoothing is to yield a differentiable regression estimate without sacrificing the good prediction properties of the first-stage regression. With this intuition, we suggest choosing the largest bandwidth such that quality of the regression estimate in terms of squared error, as measured by cross-validation score, does not substantially decrease.

Specifically, the user first specifies a non-negative tolerance and a collection of positive trial bandwidths, for instance an exponentially spaced grid up to the empirical standard deviation of X. Next, we find the bandwidth $h_{\min}$ minimising the cross-validation error across the given set of bandwidths including bandwidth 0 (corresponding to the original regression function). Then, for each positive bandwidth at least as large as $h_{\min}$, we find the largest bandwidth h such that the corresponding cross-validation score $\mathrm{CV}(h)$ exceeds $\mathrm{CV}(h_{\min})$ by no more than some tolerance times as estimate of the standard deviation of the difference $\mathrm{CV}(h) - \mathrm{CV}(h_{\min})$; if no such h exists, we pick the minimum positive bandwidth. Given a sufficiently small minimum bandwidth, this latter case should typically not occur.

The procedure is summarised in Algorithm 1 below. We suggest computing all the required evaluations of $\tilde{f}^{(n,k)}$ at once, since this only requires loading a model once per fold. In all of our numerical experiments presented in Section 2.5 we used $K = 5$ and set the tolerance to be $2 \approx \Phi^{-1}(0.975)$, though the results were largely unchanged for a wide range of tolerances.

Input: Data set $D^{(n)}$, number of folds $K \in \mathbb{N}$, set of L positive potential bandwidths $\mathcal{H} := \{h_1, \ldots, h_L\}$, tolerance tol ≥ 0 controlling the permissible increase in regression error.
Output: Bandwidth $\hat{h} \geq 0$.
Partition $D^{(n)}$ into K folds.
for *each fold* $k = 1, \ldots, K$ **do**
  Train $\tilde{f}^{(n,k)}$ on the out-of-fold data $D^{(n,k)}$.
  For each $i \in I^{(n,k)}$, set $\mathrm{err}_i(0) := \{Y_i - \tilde{f}^{(n,k)}(X_i, Z_i)\}^2$.
  for *each trial bandwidth* $h \in \{h_1, \ldots, h_L\}$ **do**
    for *each in-fold data point* (X_i, Z_i), $i \in I^{(n,k)}$ **do**
      Compute $\hat{f}^{(n,k)}(X_i, Z_i) = \sum_{j=1}^{J} \tilde{f}^{(n,k)}(X_i + hw_j, Z_i) q_j$.
      Set $\mathrm{err}_i(h) := \{Y_i - \hat{f}^{(n,k)}(X_i, Z_i)\}^2$
    end
  end
end
Writing $h_0 := 0$, for each $l = 1, \ldots, L$, set $\mathrm{CV}(h_l)$ to be the mean of the $\{\mathrm{err}_i(h_l)\}_{i=1}^n$.
Set $h_{\min} := \mathrm{argmin}_h \mathrm{CV}(h)$.
For each $h \in \mathcal{H}$ such that $h \geq h_{\min}$, set $\mathrm{se}(h)$ to be the empirical standard deviation of $\{\mathrm{err}_i(h_{\min}) - \mathrm{err}_i(h)\}_{i=1}^n$ divided by $\sqrt{n}$.
Set $\hat{h}$ to be the largest $h \in \mathcal{H}$ with $h \geq h_{\min}$ such that $\mathrm{CV}(h) \leq \mathrm{CV}(h_{\min}) + \mathrm{tol} \times \mathrm{se}(h)$, or set $\hat{h} := \min \mathcal{H}$ if no such h exists.

Algorithm 1: Cross validation selection procedure for the resmoothing bandwidth.

2.4 Score estimation

In this section we consider the problem of constructing an estimator of the score function ρ_P of a random variable X conditional on Z as required in the estimator $\hat{\theta}^{(n)}$ (2.3) of the average partial effect θ_P; however, score function estimation is also of independent interest more broadly, for example in distributional testing, particularly for assessing tail behaviour (Bera and Ng, 1995).

The multivariate score estimation problem that we seek to address has received less attention than the simpler problem of score estimation on a single univariate random variable; the latter may equivalently be expressed as the problem of estimating the ratio of the derivative of a univariate density and the density itself. In Section 2.4.1, we propose a location–scale model that then reduces our original problem to the latter, and in 2.4.2 by strengthening our modelling assumption to a location-only model, we weaken requirements on the tail behaviour of the errors.

Note that

$$\begin{aligned}\rho_{P,j}(x,z) &= \nabla_j \log p_P(x \mid z) \\ &= \nabla_j \log\{p_P(x_j \mid x_{-j}, z)\, p_P(x_{-j} \mid z)\} \\ &= \nabla_j \log p_P(x_j \mid x_{-j}, z).\end{aligned}$$

Therefore each component of ρ_P may be estimated separately using the conditional distribution of X_j given (X_{-j}, Z). This means that we can consider each variable separately, so for the rest of this section we assume that $d = 1$ without loss of generality.

Before we discuss location–scale families, we first present a theorem on the sub-Gaussianity of Lipschitz score functions that is key to the results to follow and may be of independent interest.

An interesting property of score functions is that their tail behaviour is "nicer" for heavy-tailed random variables. If a distribution has Guassian tails, its score has linear tails. If a distribution has exponential tails, the score function has constant tails. If a distribution has polynomial tails, the score function tends to zero. This trade-off has a useful consequence: that $\rho_P(X, Z)$ can be sub-Gaussian even when $X \mid Z$ is not. One straightforward implication of this is that the moments of the score are bounded. More importantly however, this shows for example that the expectation of the exponential of the score is finite, this quantity being particularly useful for bounding the ratio of a density and a version shifted by a given amount: see Lemma 26. As such, this shows that while score estimation may appear to be highly delicate given that it involves the derivative and inverse of a density, it does in fact enjoy a certain robustness. Estimation of the score based on data corrupted by a perturbation, which in our case here would be our estimates of the errors in the location–scale model, can still yield estimates whose quality is somewhat comparable with those obtained using the original uncorrupted data, as Theorems 6, 7 and 8 to follow indicate.

The result below, which is proved using repeated integration by parts, is stated for a univariate (unconditional) score. However we note that when the conditional distribution of X given Z satisfies the conditions of Theorem 5, the same conclusions hold conditionally on Z.

Theorem 5. *Let X be a univariate random variable with density p twice differentiable on $\mathbb{R}$ and score function ρ satisfying $\sup_{x\in\mathbb{R}} |\rho'(x)| \leq C < \infty$. Then for all positive integers k,*

$$\mathbb{E}\big[\rho^{2k}(X)\big] \leq C^k (2k-1)!!,$$

where $m!!$ denotes the double factorial of m, that is the product of all positive integers up to m that have the same parity as m. Furthermore, the random variable $\rho(X)$ is

sub-Gaussian with parameter $\sqrt{2C}$*. If additionally* X *is symmetrically distributed, then the sub-Gaussian parameter may be reduced to* $\sqrt{C}$*.*

The moment bound in Theorem 5 is tight when X is Gaussian. For X following a logistic distribution, the true moment is $\mathbb{E}[\rho^{2k}(X)] = C^k \frac{2^k}{2k+1}$.

2.4.1 Estimation for location–scale families

In this section, we consider a location–scale model for X on Z. Our goal is to reduce the conditions of Theorem 3 to conditions based on regression, scale estimation and univariate score estimation alone. The former two tasks are more familiar to analysts and amenable to the full variety of flexible regression methods that are available.

We assume that we have access to an i.i.d. dataset $D^{(n)} := \{(x_i, z_i) \,:\, i = 1, \ldots, n\}$ of size n, with which to estimate the score. Let us write $\mathcal{P}_{\text{ls}}$ for the class of location–scale models of the form

$$X = m_P(Z) + \sigma_P(Z)\varepsilon_P, \tag{2.9}$$

where ε_P is mean-zero and independent of Z, both $m_P(Z)$ and $\sigma_P(Z)$ are square-integrable, and ε_P has a differentiable density on $\mathbb{R}$. This enables us to reduce the problem of estimating the score of $X \mid Z$ to that of estimating the score function of the univariate variable ε_P alone. Note that we have not assumed that ε_P has unit variance here, though in practice a finite variance may be required in order to obtain a sufficiently good estimate of σ_P.

We denote the density and score function (under P) of the residual ε_P by p_ε and ρ_ε respectively. Using these we may write the conditional density and score function as

$$\begin{aligned} p_P(x \mid z) &= p_\varepsilon\left(\frac{x - m_P(z)}{\sigma_P(z)}\right) \\ \rho_P(x, z) &= \frac{1}{\sigma_P(z)}\, \rho_\varepsilon\left(\frac{x - m_P(z)}{\sigma_P(z)}\right). \end{aligned}$$

Given conditional mean and (non-negative) scale estimates $\hat{m}^{(n)}$ and $\hat{\sigma}^{(n)}$, trained on $D^{(n)}$, define estimated residuals

$$\begin{aligned} \hat{\varepsilon}^{(n)} &:= \frac{X - \hat{m}^{(n)}(Z)}{\hat{\sigma}^{(n)}(Z)} \\ &= \frac{\sigma_P(Z)\ \varepsilon_P + m_P(Z) - \hat{m}^{(n)}(Z)}{\hat{\sigma}^{(n)}(Z)}. \end{aligned}$$

We will use the estimated residuals $\hat{\varepsilon}^{(n)}$ to construct a (univariate) residual score estimator $\hat{\rho}_{\hat{\varepsilon}}^{(n)}$, also trained on $D^{(n)}$, which we combine into a final estimate of $\rho_P(x,z)$:

$$\hat{\rho}^{(n)}(x,z) = \frac{1}{\hat{\sigma}^{(n)}(z)}\,\hat{\rho}_{\hat{\varepsilon}}^{(n)}\left(\frac{x-\hat{m}^{(n)}(z)}{\hat{\sigma}^{(n)}(z)}\right). \tag{2.10}$$

While there are a variety of univariate score estimators available (see Section 2.1), these will naturally have been studied in settings when supplied with i.i.d. data from the distribution whose associated score we wish to estimate. Our setting here is rather different in that we wish to apply such a technique to an estimated set of residuals. In order to study how existing performance guarantees for score estimation may be translated to our setting, we let $p_{\hat{\varepsilon}}$ be the density of the distribution of $\hat{\varepsilon}^{(n)}$, conditional on $D^{(n)}$, and write $\rho_{\hat{\varepsilon}}(\epsilon) := p'_{\hat{\varepsilon}}/p_{\hat{\varepsilon}}$ for the associated score function. With this, let us define the following quantities, which are random over the sampling of $D^{(n)} \sim P$.

$$\begin{aligned}
A_\rho^{(n)} &:= \mathbb{E}_P\left[\left\{\rho_P(X,Z) - \hat{\rho}^{(n)}(X,Z)\right\}^2 \;\middle|\; D^{(n)}\right],\\
A_m^{(n)} &:= \mathbb{E}_P\left[\left\{\frac{m_P(Z)-\hat{m}^{(n)}(Z)}{\sigma_P(Z)}\right\}^2 \;\middle|\; D^{(n)}\right],\\
A_\sigma^{(n)} &:= \mathbb{E}_P\left[\left\{\frac{\sigma_P(Z)-\hat{\sigma}^{(n)}(Z)}{\sigma_P(Z)}\right\}^2 \;\middle|\; D^{(n)}\right],\\
A_{\hat{\varepsilon}}^{(n)} &:= \mathbb{E}_P\left[\left\{\rho_{\hat{\varepsilon}}(\hat{\varepsilon}^{(n)}) - \hat{\rho}_{\hat{\varepsilon}}^{(n)}\left(\hat{\varepsilon}^{(n)}\right)\right\}^2 \;\middle|\; D^{(n)}\right].
\end{aligned}$$

The first quantity $A_\rho^{(n)}$ is what we ultimately seek to bound to satisfy the requirements of Theorem 3. The final quantity $A_{\hat{\varepsilon}}^{(n)}$ is the sort of mean squared error we might expect to have guarantees on: note that this is evaluated with respect to the distribution of $\hat{\varepsilon}^{(n)}$, from which we can access samples.

We will assume throughout that $\sigma_P(z)$ is bounded away from zero for all $z \in \mathcal{Z}$, so $A_m^{(n)}$ and $A_\sigma^{(n)}$ may be bounded above by multiples of their counterparts unscaled by $\sigma_P(Z)$. The former versions however allow for the estimation of m_P and σ_P to be poorer in regions where σ_P is large. We also introduce the quantities

$$u_\sigma^{(n)}(z) := \frac{\hat{\sigma}^{(n)}(z)-\sigma_P(z)}{\sigma_P(z)}; \quad u_m^{(n)}(z) := \frac{\hat{m}^{(n)}(z)-m_P(z)}{\sigma_P(z)},$$

which will feature in our conditions in the results to follow. Before considering the case where ρ_ε may be estimated nonparametrically, we first consider the case where the distribution of ε_P is known.

Known family

The simplest setting is where the distribution of ε_P, and hence the function ρ_ε, is known.

Theorem 6. *Let $\mathcal{P} \subset \mathcal{P}_{ls}$ be such that all of the following hold. Under each $P \in \mathcal{P}$, the location–scale model (2.9) holds with $\varepsilon_P \overset{d}{=} \varepsilon$ fixed such that $\|\rho_\varepsilon\|_{Lip} < \infty$ and $\mathbb{E}[\rho_\varepsilon^2(\varepsilon)] < \infty$. The scale parameter σ_P is bounded away from zero,*

$$\inf_{P\in\mathcal{P}} \inf_{z\in\mathcal{Z}} \sigma_P(z) > 0,$$

and the ratio $\sigma_P/\hat{\sigma}^{(n)}$ and the regression error $u_m^{(n)}$ are bounded with high probability:

$$\sup_{z\in\mathcal{Z}} \frac{\sigma_P(z)}{\hat{\sigma}^{(n)}(z)} = O_{\mathcal{P}}(1); \quad \sup_{z\in\mathcal{Z}} \left|u_m^{(n)}(z)\right| = O_{\mathcal{P}}(1).$$

Set

$$\hat{\rho}^{(n)}(x,z) = \frac{1}{\hat{\sigma}^{(n)}(z)}\, \rho_\varepsilon\left(\frac{x - \hat{m}^{(n)}(z)}{\hat{\sigma}^{(n)}(z)}\right).$$

Then

$$A_\rho^{(n)} = O_{\mathcal{P}}\Big(A_m^{(n)} + A_\sigma^{(n)}\Big).$$

We see that in this case, the error $A_\rho^{(n)}$ we seek to control is bounded by mean squared errors in estimating m_P and σ_P.

Sub-Gaussian family

We now consider the case where ε_P follows some unknown sub-Gaussian distribution and ρ_ε is Lipschitz. This for example encompasses the case where ε_P has a Gaussian mixture distribution.

Theorem 7. *Let $\mathcal{P} \subset \mathcal{P}_{ls}$ and uniform constants $C_\varepsilon, C_\rho, C_\sigma, > 0$ be such that all of the following hold. Under each $P \in \mathcal{P}$, the location–scale model (2.9) holds where ε_P is sub-Gaussian with parameter at most C_ε. Furthermore the density p_ε of ε_P is twice differentiable on $\mathbb{R}$, with*

$$\sup_{\epsilon\in\mathbb{R}} |\partial_\epsilon^2 \log p_\varepsilon(\epsilon)| = \sup_{\epsilon\in\mathbb{R}} |\rho_\varepsilon'(\epsilon)| \le C_\rho,$$

and p_ε' and p_ε'' are both bounded. The scale parameter σ_P is bounded away from zero,

$$\inf_{P\in\mathcal{P}} \inf_{z\in\mathcal{Z}} \sigma_P(z) > 0,$$

and with high probability $D^{(n)}$ is such that the regression error $u_m^{(n)}$ and the scale error $u_\sigma^{(n)}$ are bounded:

$$\sup_{z\in\mathcal{Z}}\left|u_m^{(n)}(z)\right| = O_{\mathcal{P}}(1); \quad \lim_{n\to\infty}\sup_{P\in\mathcal{P}}\mathbb{P}_P\Big(\sup_{z\in\mathcal{Z}}\left|u_\sigma^{(n)}(z)\right| > C_\sigma\Big) = 0,$$

for some

$$C_\sigma < \min\left(1, \frac{1}{18\sqrt{C_\rho}C_\varepsilon}\right).$$

Then

$$A_\rho^{(n)} = O_{\mathcal{P}}\Big(A_m^{(n)} + A_\sigma^{(n)} + A_{\hat{\varepsilon}}^{(n)}\Big).$$

We see that in addition to requiring that $A_m^{(n)}$ and $A_\sigma^{(n)}$ are well-controlled, the mean squared error associated with the univariate score estimation problem $A_{\hat{\varepsilon}}^{(n)}$ also features in the upper bound. We also have a condition on $\sup_z\left|u_\sigma^{(n)}(z)\right|$ in Theorem 7 that is stronger than $\sup_z\left|u_\sigma^{(n)}(z)\right| = O_{\mathcal{P}}(1)$, but weaker than $\sup_z\left|u_\sigma^{(n)}(z)\right| = o_{\mathcal{P}}(1)$.

2.4.2 Estimation for location families

Theorem 7 assumes that ε_P is sub-Gaussian, which we use to deal with the scale estimation error $u_\sigma^{(n)}$. If X only depends on Z through its location (i.e. σ_P is constant) then the same proof approach works for heavy-tailed ε_P. Consider the location only model

$$X = m_P(Z) + \varepsilon_P, \tag{2.11}$$

where ε_P is independent of Z, $m_P(Z)$ is square-integrable, and ε_P has a differentiable density on $\mathbb{R}$. Compared to Section 2.4.1, we have assumed σ_P does not depend on Z, and have relabelled $\sigma_P\varepsilon_P \mapsto \varepsilon_P$. We fix $\hat{\sigma}^{(n)}(z) = 1$.

Theorem 8. *Let $\mathcal{P} \subset \mathcal{P}_{ls}$ and uniform constant $C_\rho > 0$ be such that all of the following hold. Under each $P \in \mathcal{P}$, the location model* (2.11) *holds. Furthermore the density p_ε of ε_P is twice differentiable on $\mathbb{R}$, with*

$$\sup_{\epsilon\in\mathbb{R}}|\partial_\epsilon^2 \log p_\varepsilon(\epsilon)| = \sup_{\epsilon\in\mathbb{R}}|\rho_\varepsilon'(\epsilon)| \le C_\rho,$$

and p_ε' and p_ε'' are both bounded. With high probability $D^{(n)}$ is such that the regression error $u_m^{(n)}$ is bounded

$$\sup_z\left|u_m^{(n)}(z)\right| = O_{\mathcal{P}}(1).$$

Then

$$A_\rho^{(n)} = O_{\mathcal{P}}\Big(A_m^{(n)} + A_{\hat{\varepsilon}}^{(n)}\Big).$$

As to be expected, compared to Theorem 7, here we have simply omitted the term $A_\sigma^{(n)}$ in the upper bound on $A_\rho^{(n)}$. Note that ε_P need not have any finite moments. For instance, ε_P may follow a Cauchy distribution.

2.5 Numerical experiments

We demonstrate that confidence intervals derived from the cross-fitted, doubly robust average partial effect estimator (2.3) and associated variance estimator (2.4) constructed using the approaches of Sections 2.3 and 2.4 is able to maintain good coverage across a range of settings. As competing methods, we consider a version of (2.3) using a simple numerical difference for the derivative estimate (as suggested in Chernozhukov et al. (2022d, §S5.2)) and a quadratic basis approach for score estimation (similar to Rothenhäusler and Yu (2020)); the method of Rothenhäusler and Yu (2020); and the doubly-robust partially linear regression (PLR) of Chernozhukov et al. (2018, §4.1). Theorem 27 suggests that the basis approaches of Chernozhukov et al. (2022d, §2) and Rothenhäusler and Yu (2020) are similar, and since the latter is easier to implement we use this as a reasonable proxy for the approach of Chernozhukov et al. (2022d). While our estimator may be used with any plug-in machine learning regression, here we make use of gradient boosting for its good predictive power, perform scale estimation via decision tree so that our estimates are bounded away from zero, and perform univariate score estimation via a penalised smoothing spline (Cox, 1985; Ng, 1994, 2003), which has the attractive property of smoothing towards a Gaussian in the sense of Cox (1985, Thm. 4). The precise implementation details are given in Section 2.11.2. For a sanity check we also include the ordinary least squares (OLS), which is expected to do very poorly in general. Code to reproduce our experiments is contained in the R package `drape` available from https://github.com/harveyklyne/drape.

2.5.1 Settings

In all cases we generate $Y = f_P(X, Z) + N(0, 1)$ using a known regression function f_P and predictor distribution $(X, Z) \sim P$, so that we may compute the target parameter $\theta_P = \mathbb{E}_P[f'_P(X, Z)]$ to any degree of accuracy using Monte Carlo. The predictors $(X, Z) \in \mathbb{R} \times \mathbb{R}^p$ are either generated synthetically from a location–scale family or taken from a real data set.

Location–scale families

For these fully simulated settings, we fix $n = 1000$ and

$$\begin{aligned} Z &\sim N(0, \Sigma) \in \mathbb{R}^9 \text{ where } \Sigma_{jj} = 1,\ \Sigma_{jk} = 0.5 \text{ for } j \neq k; \\ X &= m_P(Z) + \sigma_P(Z)\varepsilon_P, \end{aligned}$$

for the following choices of m_P, σ_P, ε_P. We use two step functions for $m_P, \sigma_P : \mathcal{Z} \to \mathbb{R}$.

$$\begin{aligned} m_P(z) &= \mathbb{1}_{z_1 > 0} \\ \sigma_P(z) &= \begin{cases} \sqrt{\frac{3}{2}} & \text{if } z_3 < 0; \\ \frac{1}{\sqrt{2}} & \text{if } z_3 \geq 0. \end{cases} \end{aligned}$$

Note that $\mathbb{E}_P\left[\sigma_P^2(Z)\right] = 1$. We use the following options for the noise ε_P.

$$\varepsilon_{\text{norm}} = N(0, 1) \tag{2.12}$$

$$\varepsilon_{\text{mix2}} = N\left(\pm\frac{1}{\sqrt{2}}, \frac{1}{2}\right) \text{ equiprobably;} \tag{2.13}$$

$$\varepsilon_{\text{mix3}} = N\left(\pm\frac{\sqrt{3}}{\sqrt{2}}, \frac{1}{3}\right) \text{ equiprobably;} \tag{2.14}$$

$$\varepsilon_{\log} = \text{Logistic}\left(0, \frac{\sqrt{3}}{\pi}\right); \tag{2.15}$$

$$\varepsilon_{\text{t4}} = \frac{1}{\sqrt{2}} t_4. \tag{2.16}$$

In all cases ε_P is independent of (X, Z), and has zero mean and unit variance. Since σ_P is not constant, the heavy-tailed settings $\varepsilon_{\log}, \varepsilon_{\text{t4}}$ are not covered by the results in Section 2.4. The score functions for these random variables are plotted in Figure 2.2.

401k dataset

To examine misspecification of the location–scale model for (X, Z), we import the 401k data set from the `DoubleML` R package (Bach et al., 2021). We take X to be the income feature, and Z to be age, education, family size, marriage, two-earner household, defined benefit pension, individual retirement account, home ownership, and 401k availability, giving $p = 10$. We make use of all the observations ($n = 9915$), and centre and scale the predictors before generating the simulated response variables.

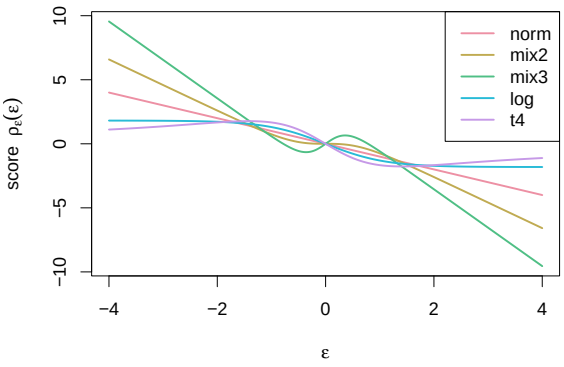


Figure 2.2 Score functions for the choices of distribution for ε_P in our numerical experiments. All these distributions have mean zero and variance one. The pink line corresponds to a standard Gaussian (2.12), the gold and green lines to Gaussian mixtures (2.13) and (2.14), the blue line to the logistic distribution (2.15), and the purple line to the Student's t distribution with 4 degrees of freedom (2.16).

Simulated responses

For the choices of regression function f_P, first define the following sinusoidal and sigmoidal families of functions:

$$f_{\text{sino}}(u; a) = \exp(-u^2/2)\sin(au);$$
$$f_{\text{sigm}}(u; s) = (1 + \exp(-su))^{-1};$$

for $u \in \mathbb{R}$, $a, s > 0$. We use the following choices for $f_P : \mathbb{R} \times \mathcal{Z} \to \mathbb{R}$, giving partially linear, additive, and interaction settings:

$$f_{\text{plm}}(x, z) = x + f_{\text{sigm}}(z_2; s = 1) + f_{\text{sino}}(z_2; a = 1); \tag{2.17}$$
$$f_{\text{add}}(x, z) = f_{\text{sigm}}(x; s = 1) + f_{\text{sino}}(x; a = 1) + f_{\text{sino}}(z_2; a = 3); \tag{2.18}$$
$$f_{\text{int}}(x, z) = f_{\text{sigm}}(x; s = 3) + f_{\text{sino}}(x; a = 3) + f_{\text{sino}}(z_2; a = 3) + x \times z_2. \tag{2.19}$$

2.5.2 Results

We examine the coverage and median width of estimated confidence intervals for each of the 5 methods in each of the 18 settings described. Figures 2.3, 2.4, and 2.5 show nominal 95% confidence intervals from each of 1000 repeated experiments. We find that our method achieves at least 85% coverage in each of the 18 settings trialed. The numerical difference and quadratic basis approach performs reasonably well, but tends to under cover and in the worst case has coverage below 80%. As one would expect, the doubly-robust partially

linear regression does very well when the partially linear model is correctly specified (2.17) — achieving full coverage with narrow confidence intervals — but risks completely losing coverage when the response is non-linear in X. Interestingly, the quadratic basis approach of Rothenhäusler and Yu (2020) displayed a similar tendency. The ordinary least squares approach did not achieve close to the specified coverage in any setting.

The coverage level measures the amount of bias in the estimator, and the median confidence interval width is a measure of the variability. In additional results which we do not include here, we find that our multivariate score estimation procedure reduces the bias as compared to the high-dimensional basis approach, and our resmoothing reduces the variance compared to numerical differencing. Taken together, our proposed estimator performs well in all settings considered.

2.6 Discussion

The average partial effect is of interest in nonparametric regression settings, giving a parametric summary of the effect of a predictor. In this work we have suggested a framework to enable the use of arbitrary machine learning regression procedures when doing inference on average partial effects. We propose kernel resmoothing of a first-stage machine learning method to yield a new, differentiable regression estimate. Theorem 4 demonstrates the attractive properties of this approach for a range of kernel bandwidths. We further advocate location-scale modelling for multivariate (conditional) score estimation, which we prove reduces this challenging problem to the better studied, univariate case in settings of interest (Theorems 6, 7, and 8). Our proofs rely on a novel result of independent interest: that Lipschitz score functions yield sub-Gaussian random variables (Theorem 5).

We confirm that our methods work well in practice, including when our location-scale modelling assumption is misspecified, with a numerical study in Section 2.5. We find that our proposals for conditional mean and score estimation successfully reduce the bias and variance of the resulting doubly-robust estimator, when compared to existing procedures. Our method achieves the best coverage of all the estimators considered. We hope that our method will see use in practical data applications, and we share an implementation in the R package `drape` (Doubly-Robust Average Partial Effects) available from https://github.com/harveyklyne/drape.

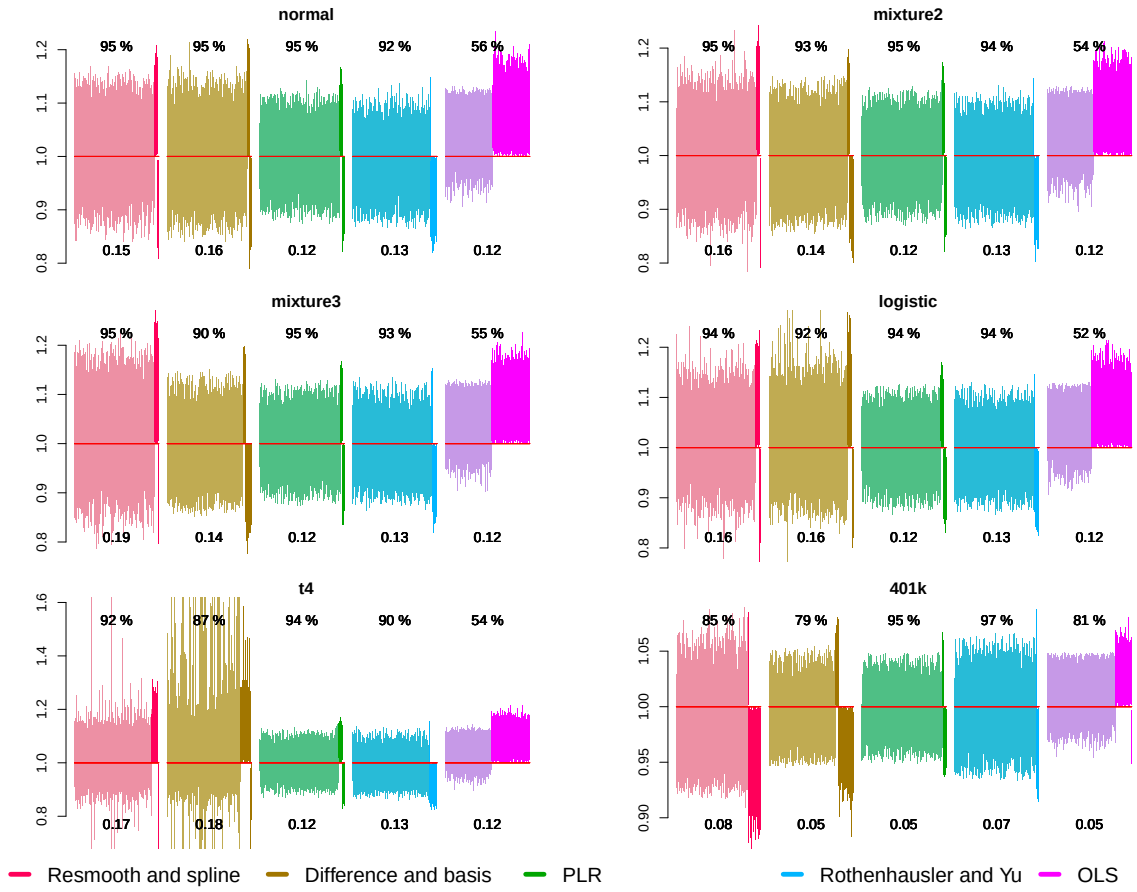


Figure 2.3 Estimated confidence intervals from the partially linear model experiment (2.17). The subplots correspond to different settings for predictor (X, Z) generation. The different colours refer to the different methods. The red horizontal lines correspond to the population level target parameter. The percentages above each subplot refer to the achieved coverage (specified level 95%), and the decimals below give the median confidence interval width. In the bottom left subplot — setting (2.16) — some of the confidence intervals for the numerical difference and basis score approach extend substantially beyond the plotting limits.

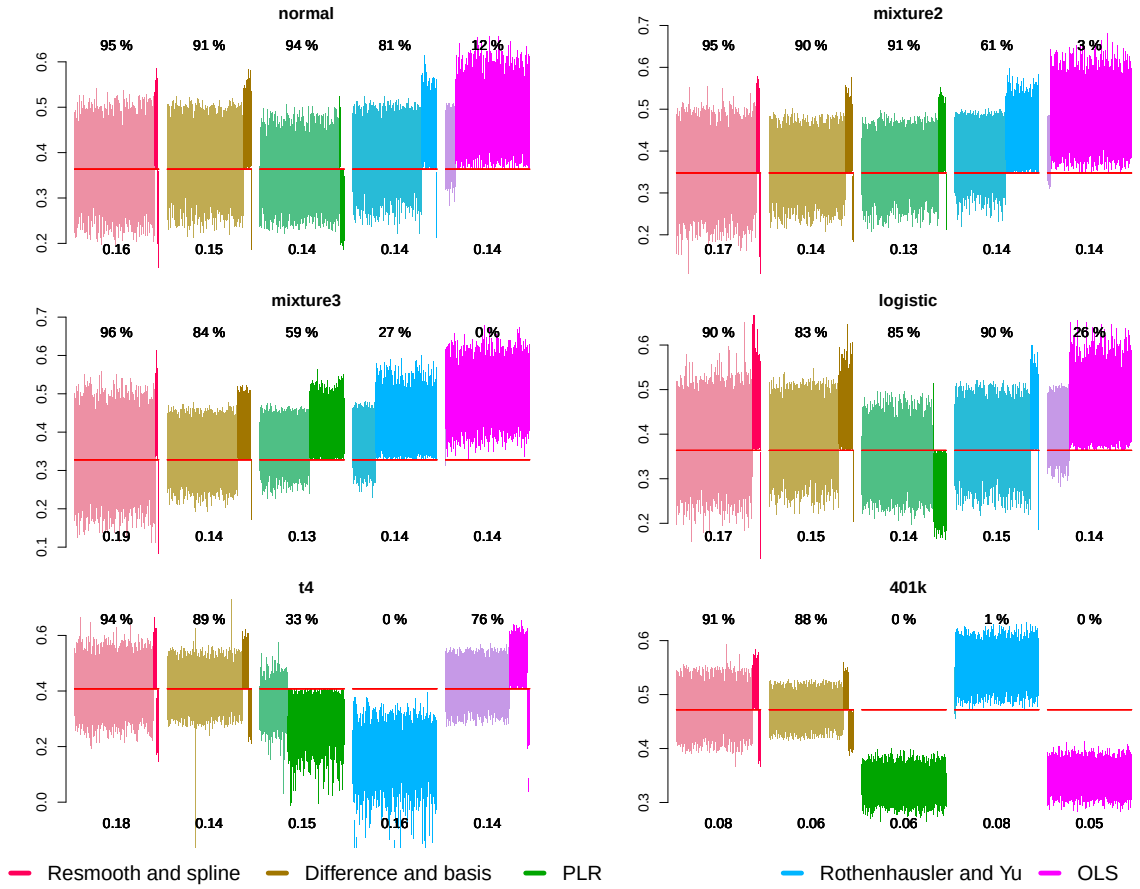


Figure 2.4 Estimated confidence intervals from the additive model experiment (2.18). The subplots correspond to different settings for predictor (X, Z) generation. The different colours refer to the different methods. The red horizontal lines correspond to the population level target parameter. The percentages above each subplot refer to the achieved coverage (specified level 95%), and the decimals below give the median confidence interval width.

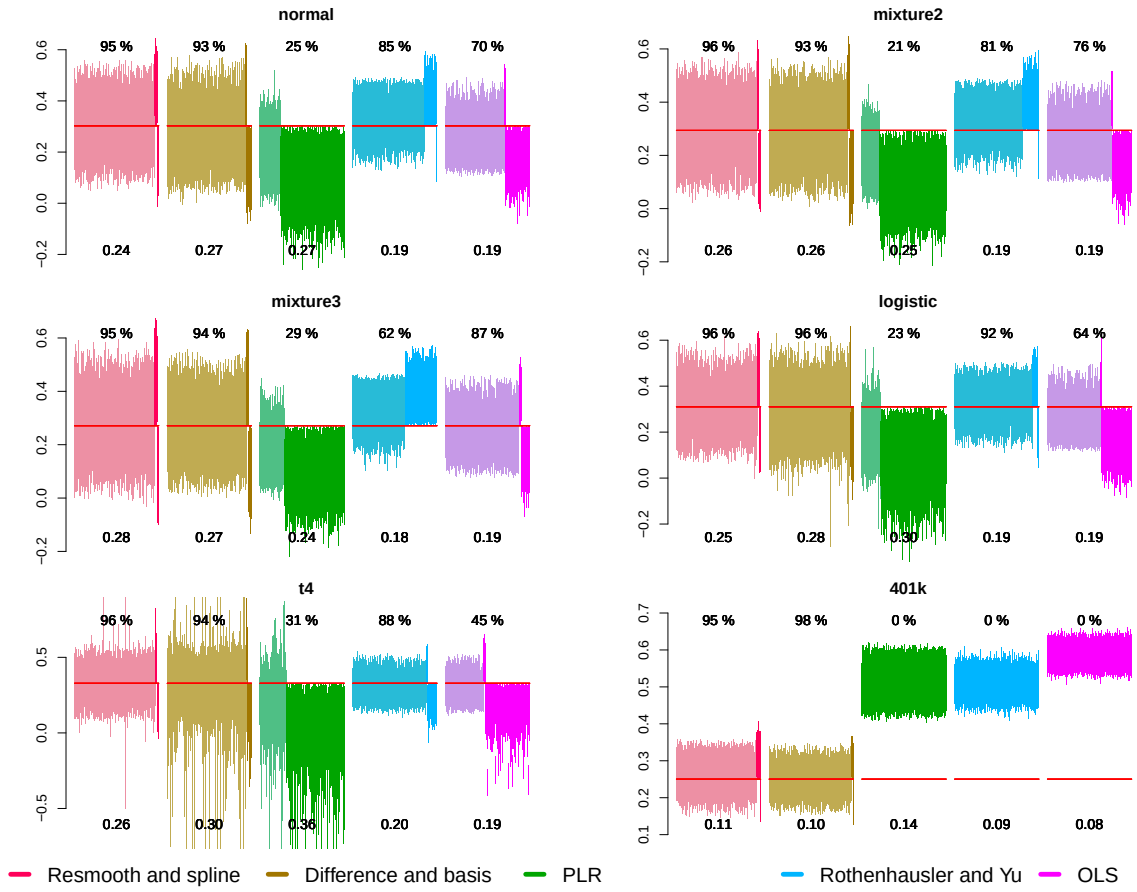


Figure 2.5 Estimated confidence intervals from the interaction model experiment (2.19). The subplots correspond to different settings for predictor (X, Z) generation. The different colours refer to the different methods. The red horizontal lines correspond to the population level target parameter. The percentages above each subplot refer to the achieved coverage (specified level 95%), and the decimals below give the median confidence interval width. In the bottom left subplot — setting (2.16) — some of the confidence intervals for the numerical difference and basis score approach extend substantially beyond the plotting limits.

2.7 Proofs in Section 2.2

2.7.1 Proof of Proposition 2

Proof. Fix some $(x_{-j}, z) \in \mathbb{R}^{d-1} \times \mathcal{Z}$ where z has positive marginal density. By the product rule,

$$\begin{aligned}\nabla_j\Big(g(x,z)p_P(x \mid z)\Big) &= \nabla_j g(x,z)p_P(x \mid z) + g(x,z)\nabla_j p_P(x \mid z) \\ &= \nabla_j g(x,z)p_P(x \mid z) + \rho_{P,j}(x,z)g(x,z)p_P(x \mid z).\end{aligned}$$

Therefore writing

$$q_{a,b}(x,z) := \{\nabla_j g(x,z)p_P(x \mid z) + \rho_{P,j}(x,z)g(x,z)p_P(x \mid z)\}\mathbb{1}_{[a,b]}(x_j),$$

for any $-\infty < a < b < \infty$ we have

$$\begin{aligned}\int_{-\infty}^{\infty} q_{a,b}(x,z)\,dx_j &= \int_a^b \nabla_j\Big(g(x,z)p_P(x \mid z)\Big)\,dx_j \\ &= g(b,x_{-j},z)p_P(b,x_{-j} \mid z) - g(a,x_{-j},z)p_P(a,x_{-j} \mid z).\end{aligned}$$

Note that

$$q_{a,b}(x,z) \le |\nabla_j g(x,z)p_P(x \mid z) + \rho_{P,j}(x,z)g(x,z)p_P(x \mid z)|$$

and

$$\int_{-\infty}^{\infty} |\nabla_j g(x,z)p_P(x \mid z) + \rho_{P,j}(x,z)g(x,z)p_P(x \mid z)|\,dx_j < \infty,$$

the final inequality coming from Fubini's theorem and condition (2.1). Let sequences (a_n) and (b_n) be as in the statement of the theorem. By dominated convergence theorem

$$\int_{-\infty}^{\infty} \{\nabla_j g(x,z)p_P(x \mid z) + \rho_{P,j}(x,z)g(x,z)p_P(x \mid z)\}\,dx_j = \lim_{n\to\infty}\int_{-\infty}^{\infty} q_{a_n,b_n}(x,z)\,dx_j = 0.$$

As this hold for every (x_{-j}, z) for which z has positive marginal density, integrating over x_{-j} and then taking a further expectation over Z proves the claim. □

2.7.2 Proof of Theorem 3

Proof. In an abuse of notation, we refer to the quantities

$$A_f^{(n,k)} := \mathbb{E}_P\Big[\{f_P(X,Z) - \hat{f}^{(n,1)}(X,Z)\}^2 \;\Big|\; D^{(n,1)}\Big],$$

for each fold $k = 1, \ldots, K$. Each $A_f^{(n,k)}$ satisfies the same probabilistic assumptions as $A_f^{(n)} = A_f^{(n,1)}$ due to the equal partitioning and i.i.d. data. Likewise we define $A_\rho^{(n,k)}, E_f^{(n,k)}, E_\rho^{(n,k)}$.

To show the first conclusion we first highlight the term which converges to a standard normal distribution, and then deal with the remainder. Note that the lower bound on the minimum eigenvalue of Σ_P corresponds to an upper bound on the maximal eigenvalue of $(\Sigma_P)^{-1/2}$. Denote the random noise in Y as

$$\begin{aligned}\xi_P &= Y - f_P(X, Z);\\ \xi_{P,i} &= y_i - f_P(x_i, z_i),\end{aligned}$$

so that $\mathbb{E}_P(\xi_P \mid X, Z) = 0$.

With these preliminaries, we have

$$\sqrt{n}(\Sigma_P)^{-1/2}\Big(\hat{\theta}^{(n)} - \theta_P\Big) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\Sigma_P)^{-1/2}\psi_P(y_i, x_i, z_i) + (\Sigma_P)^{-1/2}\sum_{k=1}^{K} R_P^{(n,k)},$$

where the uniform central limit theorem (Lemma 9) applies to the first term and

$$\begin{aligned}R_P^{(n,k)} := \frac{1}{\sqrt{n}}\sum_{i\in I^{(n,k)}} \Big\{&\nabla \hat{f}^{(n,k)}(x_i, z_i) - \hat{\rho}^{(n,k)}(x_i, z_i)\{y_i - \hat{f}^{(n,k)}(x_i, z_i)\}\\ &- \nabla f_P(x_i, z_i) + \rho_P(x_i, z_i)\xi_{P,i}\Big\}.\end{aligned}$$

Note that, conditionally on $D^{(n,k)}$, each summand of $R_P^{(n,k)}$ is i.i.d. To show that $R_P^{(n,k)} = o_{\mathcal{P}}(1)$, we fix some element $j \in \{1, \ldots, d\}$ and decompose

$$R_{P,j}^{(n,k)} = a^{(n,k)} - b_f^{(n,k)} + b_\rho^{(n,k)}, \tag{2.20}$$

where

$$\begin{aligned}a^{(n,k)} &:= \frac{1}{\sqrt{n}}\sum_{i\in I^{(n,k)}}\{\rho_{P,j}(x_i, z_i) - \hat{\rho}_j^{(n,k)}(x_i, z_i)\}\{f_P(x_i, z_i) - \hat{f}^{(n,k)}(x_i, z_i)\};\\ b_f^{(n,k)} &:= \frac{1}{\sqrt{n}}\sum_{i\in I^{(n,k)}}[\nabla_j f_P(x_i, z_i) - \nabla_j \hat{f}^{(n,k)}(x_i, z_i)\\ &\qquad + \rho_{P,j}(x_i, z_i)\{f_P(x_i, z_i) - \hat{f}^{(n,k)}(x_i, z_i)\}];\\ b_\rho^{(n,k)} &:= \frac{1}{\sqrt{n}}\sum_{i\in I^{(n,k)}}\{\rho_{P,j}(x_i, z_i) - \hat{\rho}_j^{(n,k)}(x_i, z_i)\}\xi_{P,i}.\end{aligned}$$

We now show that each term is $o_{\mathcal{P}}(1)$, so Lemma 11 yields the first conclusion.

By the Cauchy–Schwarz inequality, we have

$$\begin{aligned}\mathbb{E}_P[|a^{(n,k)}| \mid D^{(n,k)}] &\leq \sqrt{n}\mathbb{E}_P[|\rho_{P,j}(X,Z) - \hat{\rho}_j^{(n,k)}(X,Z)||f_P(X,Z) - \hat{f}^{(n,k)}(X,Z)| \mid D^{(n,k)}] \\ &\leq \sqrt{nA_f^{(n,k)}A_\rho^{(n,k)}} = o_{\mathcal{P}}(1),\end{aligned}$$

so $a^{(n,k)}$ is $o_{\mathcal{P}}(1)$ by Lemma 12. Note that each summand of $b_\rho^{(n,k)}$ is mean-zero conditionally on X and Z. This means that

$$\begin{aligned}\mathbb{E}_P[(b_\rho^{(n,k)})^2 \mid D^{(n,k)}] &= \mathbb{E}_P[\{\rho_{P,j}(X,Z) - \hat{\rho}_j^{(n,k)}(X,Z)\}^2\mathbb{E}_P(\xi_P^2 \mid X,Z) \mid D^{(n,k)}] \\ &\leq E_\rho^{(n,k)} = o_{\mathcal{P}}(1).\end{aligned}$$

Again using Lemma 12 we have that $b_\rho^{(n,k)} = o_{\mathcal{P}}(1)$.

We now apply a similar argument to $b_f^{(n,k)}$, using Proposition 2 to show that each summand is mean zero. Given $\epsilon > 0$, noting that both $A_f^{(n,k)}$ and $E_f^{(n,k)}$ are $O_{\mathcal{P}}(1)$, we have there exists M and $N \in \mathbb{N}$ such that for sequences of $D^{(n,k)}$-measurable events $\Omega_{P,n}$ with $\mathbb{P}_P(\Omega_{P,n}) \geq 1 - \epsilon$, for all $n \geq N$,

$$\mathbb{E}_P\Big[\Big|f_P(X,Z) - \hat{f}^{(n,k)}(X,Z)\Big| \;\Big|\; D^{(n,k)}\Big]\mathbb{1}_{\Omega_{P,n}} < M; \tag{2.21}$$

$$\mathbb{E}_P\Big[\Big|\nabla_j f_P(X,Z) - \nabla_j \hat{f}^{(n,k)}(X,Z) + \rho_{P,j}\Big\{f_P(X,Z) - \hat{f}^{(n,k)}(X,Z)\Big\}\Big| \;\Big|\; D^{(n,k)}\Big]\mathbb{1}_{\Omega_{P,n}} < M. \tag{2.22}$$

Now fixing z, x_{-j} (where $x_{-j} \in \mathbb{R}^{d-1}$), we have that the function

$$\Big(f_P\Big((\cdot, x_{-j}), z\Big) - \hat{f}^{(n,k)}\Big((\cdot, x_{-j}), z\Big)\Big)p\Big((\,\cdot\,, x_{-j}) \;\Big|\; z\Big)$$

is continuous, where we understand $(u, x_{-j}) = (x_1, \ldots, x_{j-1}, u, x_{j+1}, \ldots, x_d)$ for $u \in \mathbb{R}$. We may therefore apply Lemma 24 to both

$$\begin{aligned}t &\mapsto \Big(f_P\Big((t, x_{-j}), z\Big) - \hat{f}^{(n,k)}\Big((t, x_{-j}), z\Big)\Big)p\Big((t, x_{-j}) \;\Big|\; z\Big) \\ t &\mapsto \Big(f_P\Big((-t, x_{-j}), z\Big) - \hat{f}^{(n,k)}\Big((-t, x_{-j}), z\Big)\Big)p\Big((-t, x_{-j}) \;\Big|\; z\Big).\end{aligned}$$

This, in combination with (2.21) implies that on $\Omega_{P,n}$ and for each $n \geq N$, there exist $D^{(n,k)}$-measurable sequences $a_{P,m} \to -\infty$, $b_{P,m} \to \infty$ such that

$$\lim_{m\to\infty} \Big\{ \Big(f_P\big((b_{P,m}, x_{-j}), z\big) - \hat{f}^{(n,k)}\big((b_{P,m}, x_{-j}), z\big)\Big) p\big((b_{P,m}, x_{-j}) \,\big|\, z\big) \\ - \Big(f_P\big((a_{P,m}, x_{-j}), z\big) - \hat{f}^{(n,k)}\big((a_{P,m}, x_{-j}), z\big)\Big) p\big((a_{P,m}, x_{-j}) \,\big|\, z\big) \Big\} = 0. \tag{2.23}$$

Equations (2.22 and 2.23) verify that we may apply Proposition 2 conditionally on $D^{(n,k)}$. Therefore, for all n sufficiently large,

$$\mathbb{E}_P\Big[\nabla_j f_P(X,Z) - \nabla_j \hat{f}^{(n,k)}(X,Z) + \rho_{P,j}\Big\{f_P(X,Z) - \hat{f}^{(n,k)}(X,Z)\Big\} \,\Big|\, D^{(n,k)}\Big] \mathbb{1}_{\Omega_{P,n}} \; = 0$$

and hence

$$\begin{aligned} \mathbb{E}_P[(b_f^{(n,k)})^2 \mid D^{(n,k)}] \mathbb{1}_{\Omega_{P,n}} &= \mathbb{E}_P([\nabla_j f_P(X,Z) - \nabla_j \hat{f}^{(n,k)}(X,Z) \\ &\qquad + \rho_{P,j}(X,Z)\{f_P(X,Z) - \hat{f}^{(n,k)}(X,Z)\}]^2 \mid D^{(n,k)}) \\ &\leq E_f^{(n,k)}. \end{aligned}$$

Now

$$\mathbb{P}_P(b_f^{(n,k)} > \epsilon) \leq \mathbb{P}_P(b_f^{(n,k)} \mathbb{1}_{\Omega_{P,n}} > \epsilon) + \mathbb{P}_P(\Omega_{P,n}^c) \leq \mathbb{P}_P(b_f^{(n,k)} \mathbb{1}_{\Omega_{P,n}} > \epsilon) + \epsilon.$$

Lemma 12 shows that the first term above converges to 0, uniformly in P, and so $b_f^{(n,k)}$ is $o_{\mathcal{P}}(1)$.

Turning now to the second conclusion, we aim to show that $\hat{\Sigma}^{(n)} - \Sigma_P = o_{\mathcal{P}}(1)$. We introduce notation for the following random functions:

$$\hat{\psi}^{(n,k)}(y,x,z) := \nabla \hat{f}^{(n,k)}(x,z) - \hat{\rho}^{(n,k)}(x,z)\{y - \hat{f}^{(n,k)}(x,z)\} - \hat{\theta}^{(n)}.$$

We will focus on an individual element $(\hat{\Sigma}^{(n)} - \Sigma_P)_{l,m}$, $1 \leq l, m \leq d$, and make use of Lemma 10. We first check that

$$\sup_{P\in\mathcal{P}} \mathbb{E}_P\Big[\Big|\psi_{P,l}(Y,X,Z)\psi_{P,m}(Y,X,Z) - \mathbb{E}_P[\psi_{P,l}(Y,X,Z)\psi_{P,m}(Y,X,Z)]\Big|^{1+\tilde{\eta}}\Big] \leq \tilde{c},$$

for some $\tilde{\eta}, \tilde{c} > 0$. Indeed, due to the convexity of $x \mapsto |x|^{1+\tilde{\eta}}$,

$$\begin{aligned}
&\mathbb{E}_P\Big[\Big|\psi_{P,l}(Y,X,Z)\psi_{P,m}(Y,X,Z) - \mathbb{E}_P[\psi_{P,l}(Y,X,Z)\psi_{P,m}(Y,X,Z)]\Big|^{1+\tilde{\eta}}\Big] \\
&\qquad \le 2^{\tilde{\eta}}\Big\{\mathbb{E}_P\Big[\Big|\psi_{P,l}(Y,X,Z)\psi_{P,m}(Y,X,Z)\Big|^{1+\tilde{\eta}}\Big] \\
&\qquad\qquad + \Big|\mathbb{E}_P[\psi_{P,l}(Y,X,Z)\psi_{P,m}(Y,X,Z)]\Big|^{1+\tilde{\eta}}\Big\} \\
&\qquad \le 2^{1+\tilde{\eta}}\mathbb{E}_P\Big[\Big|\psi_{P,l}(Y,X,Z)\psi_{P,m}(Y,X,Z)\Big|^{1+\tilde{\eta}}\Big] \\
&\qquad \le 2^{1+\tilde{\eta}}\mathbb{E}_P\Big[\|\psi_P(Y,X,Z)\|_2^{2+2\tilde{\eta}}\Big].
\end{aligned}$$

The first inequality is $|(a+b)/2|^{1+\tilde{\eta}} \le (|a|^{1+\tilde{\eta}} + |b|^{1+\tilde{\eta}})/2$, the second is Jensen's inequality, and the final inequality is $|ab| \le (a^2+b^2)/2$. Therefore the condition is satisfied for $\tilde{\eta} = \eta/2$, $\tilde{c} = 2^{1+\eta/2}c_2$.

We are now ready to decompose the covariance estimation error.

$$\begin{aligned}
(\hat{\Sigma}^{(n)} - \Sigma_P)_{l,m} &= \frac{1}{n}\sum_{k=1}^{K}\sum_{i\in I^{(n,k)}} \hat{\psi}_l^{(n,k)}(y_i,x_i,z_i)\hat{\psi}_m^{(n,k)}(y_i,x_i,z_i) \\
&\quad - \mathbb{E}_P[\psi_{P,l}(Y,X,Z)\psi_{P,m}(Y,X,Z)] \\
&= \frac{1}{n}\sum_{i=1}^{n}\Big[\psi_{P,l}(y_i,x_i,z_i)\psi_{P,m}(y_i,x_i,z_i) - \mathbb{E}_P[\psi_{P,l}(Y,X,Z)\psi_{P,m}(Y,X,Z)]\Big] \\
&\quad + \frac{1}{K}\sum_{k=1}^{K} S_P^{(n,k)},
\end{aligned}$$

where the first term is $o_{\mathcal{P}}(1)$ by Lemma 10 and

$$S_P^{(n,k)} := \frac{K}{n}\sum_{i\in I^{(n,k)}}\Big[\hat{\psi}_l^{(n,k)}(y_i,x_i,z_i)\hat{\psi}_m^{(n,k)}(y_i,x_i,z_i) - \psi_{P,l}(y_i,x_i,z_i)\psi_{P,m}(y_i,x_i,z_i)\Big].$$

We show that $S_P^{(n,k)} = o_{\mathcal{P}}(1)$ using the following identity for $a_1, a_2, b_1, b_2 \in \mathbb{R}$,

$$a_1b_1 - a_2b_2 = (a_1-a_2)(b_1-b_2) + a_2(b_1-b_2) + b_2(a_1-a_2),$$

and then applying the Cauchy–Schwarz inequality to each term.

$$
\begin{aligned}
\left|S_P^{(n,k)}\right| &= \left|\frac{K}{n}\sum_{i\in I^{(n,k)}} \hat{\psi}_l^{(n,k)}(y_i,x_i,z_i)\hat{\psi}_m^{(n,k)}(y_i,x_i,z_i) - \psi_{P,l}(y_i,x_i,z_i)\psi_{P,m}(y_i,x_i,z_i)\right| \\
&\le \left|\frac{K}{n}\sum_{i\in I^{(n,k)}} \left\{\hat{\psi}_l^{(n,k)}(y_i,x_i,z_i) - \psi_{P,l}(y_i,x_i,z_i)\right\}\right. \\
&\qquad\qquad \left.\left\{\hat{\psi}_m^{(n,k)}(y_i,x_i,z_i) - \psi_{P,m}(y_i,x_i,z_i)\right\}\right| \\
&\quad + \left|\frac{K}{n}\sum_{i\in I^{(n,k)}} \left\{\hat{\psi}_l^{(n,k)}(y_i,x_i,z_i) - \psi_{P,l}(y_i,x_i,z_i)\right\}\psi_{P,m}(y_i,x_i,z_i)\right| \\
&\quad + \left|\frac{K}{n}\sum_{i\in I^{(n,k)}} \left\{\hat{\psi}_m^{(n,k)}(y_i,x_i,z_i) - \psi_{P,m}(y_i,x_i,z_i)\right\}\psi_{P,l}(y_i,x_i,z_i)\right| \\
&\le \left[\frac{K}{n}\sum_{i\in I^{(n,k)}} \left\{\hat{\psi}_l^{(n,k)}(y_i,x_i,z_i) - \psi_{P,l}(y_i,x_i,z_i)\right\}^2\right]^{1/2} \\
&\qquad\qquad \left[\frac{K}{n}\sum_{i\in I^{(n,k)}} \left\{\hat{\psi}_m^{(n,k)}(y_i,x_i,z_i) - \psi_{P,m}(y_i,x_i,z_i)\right\}^2\right]^{1/2} \\
&\quad + \left[\frac{K}{n}\sum_{i\in I^{(n,k)}} \left\{\hat{\psi}_l^{(n,k)}(y_i,x_i,z_i) - \psi_{P,l}(y_i,x_i,z_i)\right\}^2\right]^{1/2} \\
&\qquad\qquad \left[\frac{K}{n}\sum_{i\in I^{(n,k)}} \psi_{P,m}(y_i,x_i,z_i)^2\right]^{1/2} \\
&\quad + \left[\frac{K}{n}\sum_{i\in I^{(n,k)}} \left\{\hat{\psi}_m^{(n,k)}(y_i,x_i,z_i) - \psi_{P,m}(y_i,x_i,z_i)\right\}^2\right]^{1/2} \\
&\qquad\qquad \left[\frac{K}{n}\sum_{i\in I^{(n,k)}} \psi_{P,l}(y_i,x_i,z_i)^2\right]^{1/2}.
\end{aligned}
$$

Therefore it suffices to show that, for each $l = 1, \ldots, d$,

$$
\begin{aligned}
T_{P,1}^{(n,k)} &:= \frac{K}{n}\sum_{i\in I^{(n,k)}} \psi_{P,l}(y_i,x_i,z_i)^2 = O_{\mathcal{P}}(1); \\
T_{P,2}^{(n,k)} &:= \frac{K}{n}\sum_{i\in I^{(n,k)}} \left\{\hat{\psi}_l^{(n,k)}(y_i,x_i,z_i) - \psi_{P,l}(y_i,x_i,z_i)\right\}^2 = o_{\mathcal{P}}(1).
\end{aligned}
$$

To this end, Lemma 10 gives $T_{P,1}^{(n,k)} = (\Sigma_P)_{(l,l)} + o_{\mathcal{P}}(1)$. Moreover, similarly to equation (2.20) and using the inequality $(a+b+c+d)^2 \leq 4(a^2+b^2+c^2+d^2)$,

$$\begin{aligned}
T_{P,2}^{(n,k)} &= \frac{K}{n} \sum_{i \in I^{(n,k)}} \left\{ \hat{\psi}_l^{(n,k)}(y_i, x_i, z_i) - \psi_{P,l}(y_i, x_i, z_i) \right\}^2 \\
&= \frac{K}{n} \sum_{i \in I^{(n,k)}} \Big[\left\{ \rho_{P,l}(x_i, z_i) - \hat{\rho}_l^{(n,k)}(x_i, z_i) \right\} \left\{ f_P(x_i, z_i) - \hat{f}^{(n,k)}(x_i, z_i) \right\} \\
&\quad - \nabla_l f_P(x_i, z_i) + \nabla_l \hat{f}^{(n,k)}(x_i, z_i) - \rho_{P,l}(x_i, z_i) \left\{ f_P(x_i, z_i) - \hat{f}^{(n,k)}(x_i, z_i) \right\} \\
&\quad + \left\{ \rho_{P,l}(x_i, z_i) - \hat{\rho}_l^{(n,k)}(x_i, z_i) \right\} \xi_{P,i} - \hat{\theta}_l^{(n)} + \theta_{P,l} \Big]^2 \\
&\leq 4\{ \tilde{a}^{(n,k)} + \tilde{b}_f^{(n,k)} + \tilde{b}_\rho^{(n,k)} + (\hat{\theta}_l^{(n)} - \theta_{P,l})^2 \},
\end{aligned}$$

where

$$\begin{aligned}
\tilde{a}^{(n,k)} &:= \frac{K}{n} \sum_{i \in I^{(n,k)}} \left\{ \rho_{P,l}(x_i, z_i) - \hat{\rho}_l^{(n,k)}(x_i, z_i) \right\}^2 \left\{ f_P(x_i, z_i) - \hat{f}^{(n,k)}(x_i, z_i) \right\}^2; \\
\tilde{b}_f^{(n,k)} &:= \frac{K}{n} \sum_{i \in I^{(n,k)}} \Big[\nabla_l f_P(x_i, z_i) - \nabla_l \hat{f}^{(n,k)}(x_i, z_i) \\
&\qquad\qquad + \rho_{P,l}(x_i, z_i) \left\{ f_P(x_i, z_i) - \hat{f}^{(n,k)}(x_i, z_i) \right\} \Big]^2; \\
\tilde{b}_\rho^{(n,k)} &:= \frac{K}{n} \sum_{i \in I^{(n,k)}} \left\{ \rho_{P,l}(x_i, z_i) - \hat{\rho}_l^{(n,k)}(x_i, z_i) \right\}^2 \xi_{P,i}^2.
\end{aligned}$$

Since $n^{-1/2}(\hat{\theta}^{(n)} - \theta_P)$ is uniformly asympotically Gaussian, we have that $(\hat{\theta}_l^{(n)} - \theta_{P,l})^2 = O_{\mathcal{P}}(n^{-1})$. For $\tilde{a}^{(n,k)}$, $\tilde{b}_f^{(n,k)}$ and $\tilde{b}_\rho^{(n,k)}$ we use Lemma 12, noting that conditionally on $D^{(n,k)}$ each summand is i.i.d.

Using the identity $\sum_i a_i b_i \leq (\sum_i a_i)(\sum_i b_i)$ for positive sequences (a_i) and (b_i), we have

$$\left| \tilde{a}^{(n,k)} \right| \leq \frac{n}{K} \tilde{a}_\rho^{(n,k)} \tilde{a}_f^{(n,k)},$$

for

$$\begin{aligned}
\tilde{a}_\rho^{(n,k)} &:= \frac{K}{n} \sum_{i \in I^{(n,k)}} \left\{ \rho_{P,l}(x_i, z_i) - \hat{\rho}_l^{(n,k)}(x_i, z_i) \right\}^2; \\
\tilde{a}_f^{(n,k)} &:= \frac{K}{n} \sum_{i \in I^{(n,k)}} \left\{ f_P(x_i, z_i) - \hat{f}^{(n,k)}(x_i, z_i) \right\}^2.
\end{aligned}$$

Finally,

$$\begin{aligned}
\mathbb{E}_P\Big(\big|\tilde{a}_f^{(n,k)}\big| \;\Big|\; D^{(n,k)}\Big) &= \mathbb{E}_P\Big[\Big\{f_P(X,Z) - \hat{f}^{(n,k)}(X,Z)\Big\}^2 \;\Big|\; D^{(n,k)}\Big] \\
&\leq A_f^{(n,k)}; \\
\mathbb{E}_P\Big(\big|\tilde{a}_\rho^{(n,k)}\big| \;\Big|\; D^{(n,k)}\Big) &= \mathbb{E}_P\Big[\Big\{\rho_{P,l}(X,Z) - \hat{\rho}_l^{(n,k)}(X,Z)\Big\}^2 \;\Big|\; D^{(n,k)}\Big] \\
&\leq A_\rho^{(n,k)}; \\
\mathbb{E}_P\Big(\big|\tilde{b}_f^{(n,k)}\big| \;\Big|\; D^{(n,k)}\Big) &= \mathbb{E}_P\Big(\Big[\nabla_l f_P(X,Z) - \nabla_l \hat{f}^{(n,k)}(X,Z) \\
&\qquad + \rho_{P,l}(X,Z)\Big\{f_P(X,Z) - \hat{f}^{(n,k)}(X,Z)\Big\}\Big]^2 \;\Big|\; D^{(n,k)}\Big) \\
&\leq E_f^{(n,k)}; \\
\mathbb{E}_P\Big(\big|\tilde{b}_\rho^{(n,k)}\big| \;\Big|\; D^{(n,k)}\Big) &= \mathbb{E}_P\Big[\Big\{\rho_{P,l}(X,Z) - \hat{\rho}_l^{(n,k)}(X,Z)\Big\}^2 \xi_P^2 \;\Big|\; D^{(n,k)}\Big] \\
&\leq E_\rho^{(n,k)}.
\end{aligned}$$

This suffices to show that $T_{P,2}^{(n,k)} = o_{\mathcal{P}}(1)$, so $\hat{\Sigma}^{(n)} - \Sigma_P = o_{\mathcal{P}}(1)$.

It remains to show the final conclusion. By Lemma 11, it is enough to show that

$$(\Sigma_P)^{-1/2}(\hat{\Sigma}^{(n)})^{1/2} = I + o_{\mathcal{P}}(1).$$

We have that the maximal eigenvalue of $(\Sigma_P)^{-1/2}$ is uniformly bounded above, and further that

$$\begin{aligned}
(\Sigma_P)^{-1/2}(\hat{\Sigma}^{(n)})^{1/2} &= (\Sigma_P)^{-1/2}\Big\{(\hat{\Sigma}^{(n)})^{1/2} - (\Sigma_P)^{1/2} + (\Sigma_P)^{1/2}\Big\} \\
&= I + (\Sigma_P)^{-1/2}\Big\{(\hat{\Sigma}^{(n)})^{1/2} - (\Sigma_P)^{1/2}\Big\}.
\end{aligned}$$

Therefore it remains to check that

$$(\hat{\Sigma}^{(n)})^{1/2} = (\Sigma_P)^{1/2} + o_{\mathcal{P}}(1).$$

By Horn and Johnson (1985, Eqn. (7.2.13)),

$$\begin{aligned}
\Big\|(\hat{\Sigma}^{(n)})^{1/2} - (\Sigma_P)^{1/2}\Big\|_2 &\leq \Big\|(\Sigma_P)^{-1/2}\Big\|_2 \|\hat{\Sigma}^{(n)} - \Sigma_P\|_2 \\
&\leq c_1^{-1/2}\|\hat{\Sigma}^{(n)} - \Sigma_P\|_2.
\end{aligned}$$

Hence $(\hat{\Sigma}^{(n)})^{1/2} - (\Sigma_P)^{1/2} = o_{\mathcal{P}}(1)$. This completes the proof. □

2.7.3 Auxiliary lemmas

Lemma 9 (Shah and Peters (2020, Supp. Lem. 18), vectorised)**.** *Let $\mathcal{P}$ be a family of distributions for $\zeta \in \mathbb{R}^d$ and suppose $\zeta_1, \zeta_2, \ldots$ are i.i.d. copies. For each $n \in \mathbb{N}$, let $S_n = n^{-1/2} \sum_{i=1}^n \zeta_i$. Suppose that for all $P \in \mathcal{P}$, we have $E_P(\zeta) = 0$, $\mathrm{Var}_P(\zeta) = I$, and $\mathbb{E}_P(\|\zeta\|_2^{2+\eta}) \leq c$ for some $c, \eta > 0$. Then we have that*

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}} \sup_{t\in\mathbb{R}^d} |\mathbb{P}_P(S_n \leq t) - \Phi(t)| = 0.$$

Proof. For each n, let $P_n \in \mathcal{P}$ satisfy

$$\sup_{P\in\mathcal{P}} \sup_{t\in\mathbb{R}^d} |\mathbb{P}_P(S_n \leq t) - \Phi(t)| \leq \sup_{t\in\mathbb{R}^d} |\mathbb{P}_{P_n}(S_n \leq t) - \Phi(t)| + n^{-1}.$$

Let $Y_{n,i}$ be equal in distribution to $n^{-1/2}\zeta_i$ under P_n. We check the conditions to apply van der Vaart (1998, Prop. 2.27). Indeed, $Y_{n,1}, \ldots, Y_{n,n}$ are i.i.d. for each n, and $\sum_{i=1}^n \mathrm{Var}(Y_{n,i}) = \sum_{i=1}^n n^{-1}\mathrm{Var}_{P_n}(\zeta) = I$. Finally, for any $\epsilon > 0$ we have

$$\begin{aligned}
\sum_{i=1}^n \mathbb{E}\Big(\|Y_{n,1}\|_2^2 \mathbb{1}_{\{\|Y_{n,i}\|_2 > \epsilon\}}\Big) &= \mathbb{E}_{P_n}\Big(\|\zeta\|_2^2 \mathbb{1}_{\{\|\zeta\|_2 > \sqrt{n}\epsilon\}}\Big) \\
&\leq \Big[\mathbb{E}_{P_n}(\|\zeta\|_2^{2+\eta})\Big]^{2/(2+\eta)} \Big[\mathbb{E}_{P_n}\Big(\mathbb{1}_{\{\|\zeta\|_2 > \sqrt{n}\epsilon\}}^{(2+\eta)/\eta}\Big)\Big]^{\eta/(2+\eta)} \\
&\leq c^{2/(2+\eta)} [\mathbb{P}_{P_n}(\|\zeta\|_2 > \sqrt{n}\epsilon)]^{\eta/(2+\eta)} \\
&\leq c^{2/(2+\eta)} [\mathbb{E}_{P_n}(\|\zeta\|_2)/(\sqrt{n}\epsilon)]^{\eta/(2+\eta)} \\
&\leq c\epsilon^{-\eta/(2+\eta)} n^{-\eta/(4+2\eta)} \to 0.
\end{aligned}$$

Here the first inequality is due to Hölder, the third due to Markov and the second and fourth are applying the assumption $\mathbb{E}_P(\|\zeta\|_2^{2+\eta}) \leq c$. □

Lemma 10 (Shah and Peters (2020, Supp. Lem. 19))**.** *Let $\mathcal{P}$ be a family of distributions for $\zeta \in \mathbb{R}$ and suppose $\zeta_1, \zeta_2, \ldots$ are i.i.d. copies. For each $n \in \mathbb{N}$, let $S_n = n^{-1} \sum_{i=1}^n \zeta_i$. Suppose that for all $P \in \mathcal{P}$, we have $E_P(\zeta) = 0$, and $\mathbb{E}_P(|\zeta|^{1+\eta}) \leq c$ for some $c, \eta > 0$. Then we have that for all $\epsilon > 0$,*

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}} \mathbb{P}_P(|S_n| > \epsilon) = 0.$$

Lemma 11. *Let $\mathcal{P}$ be a family of distributions that determines the law of sequences $(V_n)_{n\in\mathbb{N}}$ and $(W_n)_{n\in\mathbb{N}}$ of random vectors in $\mathbb{R}^d$ and $(M_n)_{n\in\mathbb{N}}$ random matrices in $\mathbb{R}^{d\times d}$. Suppose*

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}} \sup_{t\in\mathbb{R}^d} |\mathbb{P}_P(V_n \leq t) - \Phi(t)| = 0.$$

Then we have the following.

(a) If $W_n = o_{\mathcal{P}}(1)$ we have

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}} \sup_{t\in\mathbb{R}^d} |\mathbb{P}_P(V_n + W_n \le t) - \Phi(t)| = 0.$$

(b) If $M_n = I + o_{\mathcal{P}}(1)$ we have

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}} \sup_{t\in\mathbb{R}^d} |\mathbb{P}_P(M_n^{-1}V_n \le t) - \Phi(t)| = 0.$$

Proof. We first show that for any $\delta > 0$, $\sup_{t\in\mathbb{R}^d} |\Phi(t+\delta) - \Phi(t)| \le d\delta$. Indeed, letting $Z \sim N(0, I)$ in $\mathbb{R}^d$,

$$\begin{aligned}
|\Phi(t+\delta) - \Phi(t)| &= \mathbb{P}(\cap_j\{Z_j \le t_j + \delta\}) - \mathbb{P}(\cap_j\{Z_j \le t_j\}) \\
&= \mathbb{P}(\cup_j\{Z_j \in (t_j, t_j + \delta]\}) \\
&\le \sum_j \mathbb{P}(Z_j \in (t_j, t_j + \delta]) \\
&\le d\delta.
\end{aligned}$$

The final line follows because the univariate standard normal c.d.f. has Lipschitz constant $1/\sqrt{2\pi} < 1$.

Now consider the setup of (a). Given $\epsilon > 0$ let N be such that for all $n \ge N$ and for all $P \in \mathcal{P}$,

$$\sup_{t\in\mathbb{R}^d} |\mathbb{P}_P(V_n \le t) - \Phi(t)| < \epsilon/3 \quad \text{and} \quad \mathbb{P}_P[\|W_n\|_\infty > \epsilon/(3d)] < \epsilon/3.$$

Then

$$\begin{aligned}
\mathbb{P}_P(V_n + W_n \le t) - \Phi(t) &= \mathbb{P}_P(\cap_j\{V_{nj} + W_{nj} \le t_j\}) - \Phi(t) \\
&\le \mathbb{P}_P[(\cap_j\{V_{nj} \le t_j + \epsilon/(3d)\}) \cup (\cup_j\{W_{nj} < -\epsilon/(3d)\})] - \Phi(t) \\
&\le \mathbb{P}[V_n \le t + \epsilon/(3d)] + \mathbb{P}[\|W_n\|_\infty > \epsilon/(3d)] - \Phi(t) \\
&< \epsilon/3 + \Phi[t + \epsilon/(3d)] - \Phi(t) + \epsilon/3 < \epsilon,
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{P}_P(V_n + W_n \le t) - \Phi(t) \\
&\qquad = 1 - \Phi(t) - \mathbb{P}_P(\cup_j \{V_{nj} + W_{nj} > t_j\}) \\
&\qquad \ge 1 - \Phi(t) - \mathbb{P}_P[\cup_j (\{V_{nj} > t_j - \epsilon/(3d)\} \cup \{W_{nj} > \epsilon/(3d)\})] \\
&\qquad = 1 - \Phi(t) - \mathbb{P}_P[(\cup_j \{V_{nj} > t_j - \epsilon/(3d)\}) \cup \{\|W_n\|_\infty > \epsilon/(3d)\}] \\
&\qquad \ge 1 - \Phi(t) - \mathbb{P}_P(\cup_j \{V_{nj} > t_j - \epsilon/(3d)\}) - \mathbb{P}_P[\|W_n\|_\infty > \epsilon/(3d)] \\
&\qquad > \mathbb{P}_P[V_n \le t - \epsilon/(3d)] - \Phi(t) - \epsilon/3 \\
&\qquad > -\epsilon/3 + \Phi[t - \epsilon/(3d)] - \Phi(t) - \epsilon/3 > -\epsilon.
\end{aligned}$$

Thus for all $n \ge N$ and $P \in \mathcal{P}$,

$$\sup_{t \in \mathbb{R}^d} |\mathbb{P}_P(V_n + W_n \le t) - \Phi(t)| < \epsilon.$$

To prove (b), it suffices to show that $(M_n^{-1} - I)V_n = o_{\mathcal{P}}(1)$ and then apply (a). We have that $M_n - I$ is $o_{\mathcal{P}}(1)$, and so the sequence

$$\|M_n - I\|_\infty := \sup_{x : \|x\|_\infty = 1} \|(M_n - I)x\|_\infty = o_{\mathcal{P}}(1).$$

By Golub and Van Loan (2013, Thm. 2.3.4), when $\|M_n - I\|_\infty < 1$, then M_n is nonsingular and

$$\|M_n^{-1} - I\|_\infty \le \frac{\|M_n - I\|_\infty}{1 - \|M_n - I\|_\infty}.$$

Now

$$\begin{aligned}
\sup_{P \in \mathcal{P}} \mathbb{P}_P(\|M_n^{-1} - I\|_\infty > \epsilon) &\le \sup_{P \in \mathcal{P}} \mathbb{P}_P\left(\frac{\|M_n - I\|_\infty}{1 - \|M_n - I\|_\infty} > \epsilon\right) \\
&= \sup_{P \in \mathcal{P}} \mathbb{P}_P(\|M_n - I\|_\infty > \epsilon/(1+\epsilon)) \to 0,
\end{aligned}$$

so $\|M_n^{-1} - I\|_\infty$ is also $o_{\mathcal{P}}(1)$.

Now we can show that the sequence $\|(M_n^{-1} - I)V_n\|_\infty$ is $o_{\mathcal{P}}(1)$. Indeed given $\epsilon > 0$, let $\delta > 0$ be such that $\Phi(\epsilon/\delta) > 1 - \epsilon/3$, and let N be such that for all $n \ge N$ and for all $P \in \mathcal{P}$,

$$\sup_{t \in \mathbb{R}^d} |\mathbb{P}_P(V_n \le t) - \Phi(t)| < \epsilon/3 \quad \text{and} \quad \mathbb{P}_P(\|M_n^{-1} - I\|_\infty > \delta) < \epsilon/3.$$

Then

$$\begin{aligned}\mathbb{P}_P(\|(M_n^{-1}-I)V_n\|_\infty > \epsilon) &\leq \mathbb{P}_P(\|M_n^{-1}-I\|_\infty \|V_n\|_\infty > \epsilon) \\ &\leq \mathbb{P}_P(\{\|M_n^{-1}-I\|_\infty > \delta\} \cup \{\|V_n\|_\infty > \epsilon/\delta\}) \\ &\leq \mathbb{P}_P(\|M_n^{-1}-I\|_\infty > \delta) + 1 - \mathbb{P}_P(V_n \leq \epsilon/\delta) \\ &< \epsilon/3 + 1 - \Phi(\epsilon/\delta) + \epsilon/3 < \epsilon.\end{aligned}$$

This suffices to show that the sequence of random vectors $(M_n^{-1}-I)V_n$ is $o_{\mathcal{P}}(1)$, so we are done by (a). □

Lemma 12. *Let X_m and Y_m be sequences of random vectors governed by laws in some set $\mathcal{P}$, let $\|\cdot\|$ be any norm and $q \geq 1$.*

(a) If $\mathbb{E}_P(\|X_m\|^q \mid Y_m) = o_{\mathcal{P}}(1)$, then $\|X_m\| = o_{\mathcal{P}}(1)$.

(b) If $\mathbb{E}_P(\|X_m\|^q \mid Y_m) = O_{\mathcal{P}}(1)$, then $\|X_m\| = O_{\mathcal{P}}(1)$.

Proof. In both cases we work with a bounded version of $\|X_m\|$, and apply Markov's inequality.

Let $\mathbb{E}_P(\|X_m\|^q \mid Y_m) = o_{\mathcal{P}}(1)$. Given $\epsilon > 0$,

$$\begin{aligned}\mathbb{P}_P[\|X_m\| > \epsilon] &= \mathbb{P}_P[\|X_m\|^q > \epsilon^q] \\ &= \mathbb{P}_P[(\|X_m\|^q \wedge 2\epsilon^q) > \epsilon^q] \\ &\leq \epsilon^{-q}\mathbb{E}_P[\|X_m\|^q \wedge 2\epsilon^q] \\ &= \epsilon^{-q}\mathbb{E}_P[\mathbb{E}_P(\|X_m\|^q \mid Y_m) \wedge 2\epsilon^q].\end{aligned}$$

Writing $W_m = \mathbb{E}_P(\|X_m\|^q \mid Y_m) \wedge 2\epsilon^q$, we have that $W_m = o_{\mathcal{P}}(1)$ and $|W_m| \leq 2\epsilon^q$ almost surely. Taking supremum over $\mathcal{P}$ and applying Shah and Peters (2020, Supp. Lem. 25) (uniform bounded convergence), we have

$$\sup_{P\in\mathcal{P}} \mathbb{P}_P(\|X_m\| > \epsilon) \leq \epsilon^{-q} \sup_{P\in\mathcal{P}} \mathbb{E}_P(W_m) \to 0.$$

The second conclusion is similar. Let $\mathbb{E}_P(\|X_m\|^q \mid Y_m) = O_{\mathcal{P}}(1)$. Given $\epsilon > 0$ and for M to be fixed later, we have

$$\begin{aligned}\mathbb{P}_P[\|X_m\| > M] &= \mathbb{P}_P[\|X_m\|^q > M^q] \\ &= \mathbb{P}_P[(\|X_m\|^q \wedge 2M^q) > M^q] \\ &\leq M^{-q}\mathbb{E}_P[\|X_m\|^q \wedge 2M^q] \\ &= M^{-q}\mathbb{E}_P[\mathbb{E}_P(\|X_m\|^q \mid Y_m) \wedge 2M^q].\end{aligned}$$

Now let $W_m := \mathbb{E}_P(\|X_m\|^q \mid Y_m) \wedge 2M^q$. Note that for any $\tilde{M} > 0$,

$$\begin{aligned} W_m &= W_m \mathbb{1}_{\{\mathbb{E}_P(\|X_m\|^q|Y_m) \leq \tilde{M}\}} + W_m \mathbb{1}_{\{\mathbb{E}_P(\|X_m\|^q|Y_m) > \tilde{M}\}} \\ &\leq \tilde{M} + 2M^q \mathbb{1}_{\{\mathbb{E}_P(\|X_m\|^q|Y_m) > \tilde{M}\}} \end{aligned}$$

almost surely. Since $\mathbb{E}_P(\|X_m\|^q \mid Y_m) = O_{\mathcal{P}}(1)$, we may choose $\tilde{M}$ so that

$$\sup_{m \in \mathbb{N}} \sup_{P \in \mathcal{P}} \mathbb{P}_P[\mathbb{E}_P(\|X_m\|^q \mid Y_m) > \tilde{M}] < \epsilon/3,$$

and then choose $M > (3\tilde{M}/\epsilon)^{1/q}$. Again applying Shah and Peters (2020, Supp. Lem. 25), we have

$$\begin{aligned} \sup_{m \in \mathbb{N}} \sup_{P \in \mathcal{P}} \mathbb{P}_P(\|X_m\| > M) &\leq M^{-q} \sup_{m \in \mathbb{N}} \sup_{P \in \mathcal{P}} \mathbb{E}_P(W_m) \\ &\leq M^{-q}(\tilde{M} + 2M^q\epsilon/3) \\ &< \epsilon/3 + 2\epsilon/3 = \epsilon. \end{aligned}$$

□

2.8 Proof of Theorem 4

Proof. Let

$$\sup_{P \in \mathcal{P}} \sup_{x,z} \left|\rho'_P(x,z)\right| =: C < \infty.$$

Using the inequality $(a+b)^2 \leq 2\Big(a^2 + b^2\Big)$, we decompose the quantities of interest (2.6).

$$\begin{aligned} A_f^{(n)} &= \mathbb{E}_P\Big(\Big[f_P(X,Z) - \{f_P(\cdot,Z) * K_h\}(X) \\ &\qquad + \{f_P(\cdot,Z) * K_h\}(X) - \hat{f}^{(n,1)}(X,Z) \Big]^2 \;\Big|\; D^{(n,1)} \Big) \qquad (2.24) \\ &\leq 2\mathbb{E}_P\Big([f_P(X,Z) - \{f_P(\cdot,Z) * K_h\}(X)]^2 \Big) \\ &\quad + 2\mathbb{E}_P\Big(\Big[\{f_P(\cdot,Z) * K_h\}(X) - \Big\{ \tilde{f}^{(n,1)}(\cdot,Z) * K_h \Big\}(X) \Big]^2 \;\Big|\; D^{(n,1)} \Big) \\ &= 2\mathbb{E}_P([f_P(X,Z) - \{f_P(\cdot,Z) * K_h\}(X)]^2) \\ &\quad + 2\mathbb{E}_P\Big(\Big[\Big\{ f_P(\cdot,Z) - \tilde{f}^{(n,1)}(\cdot,Z) \Big\} * K_h \Big]^2(X) \;\Big|\; D^{(n,1)} \Big) \\ &\leq 2\mathbb{E}_P\Big(\sup_x [f_P(x,Z) - \{f_P(\cdot,Z) * K_h\}(x)]^2 \Big) \\ &\quad + 2\mathbb{E}_P\Big(\Big[\Big\{ f_P(\cdot,Z) - \tilde{f}^{(n,1)}(\cdot,Z) \Big\} * K_h \Big]^2(X) \;\Big|\; D^{(n,1)} \Big). \qquad (2.25) \end{aligned}$$

Similarly,

$$\begin{aligned}
E_f^{(n)} &\le 2\mathbb{E}_P\Big[\Big(f_P'(X,Z) - \{f_P(\cdot,Z) * K_h\}'(X) \\
&\qquad + \rho_P(X,Z)[f_P(X,Z) - \{f_P(\cdot,Z) * K_h\}(X)]\Big)^2\Big] \\
&\quad + 2\mathbb{E}_P\Bigg[\Bigg(\Big[\Big\{f_P(\cdot,Z) - \tilde{f}^{(n,1)}(\cdot,Z)\Big\} * K_h\Big]'(X) \\
&\qquad + \rho_P(X,Z)\Big[\Big\{f_P(\cdot,Z) - \tilde{f}^{(n,1)}(\cdot,Z)\Big\} * K_h\Big](X)\Bigg)^2 \;\Bigg|\; D^{(n,1)}\Bigg] \\
&\le 4\mathbb{E}_P\Big(\sup_x [f_P'(x,Z) - \{f_P(\cdot,Z) * K_h\}'(x)]^2\Big) \\
&\quad + 4\mathbb{E}_P\Big(\mathbb{E}_P\Big[\rho_P^2(X,Z) \;\Big|\; Z\Big] \sup_x [f_P(x,Z) - \{f_P(\cdot,Z) * K_h\}(x)]^2\Big) \\
&\quad + 4\mathbb{E}_P\Bigg[\Big[\Big\{f_P(\cdot,Z) - \tilde{f}^{(n,1)}(\cdot,Z)\Big\} * K_h\Big]'^2(X) \;\Bigg|\; D^{(n,1)}\Bigg] \\
&\quad + 4\mathbb{E}_P\Big(\rho_P^2(X,Z)\Big[\Big\{f_P(\cdot,Z) - \tilde{f}^{(n,1)}(\cdot,Z)\Big\} * K_h\Big]^2(X) \;\Big|\; D^{(n,1)}\Big). \qquad (2.26)
\end{aligned}$$

By Theorem 5, $\mathbb{E}_P\Big[\rho_P^2(X,Z) \;\Big|\; Z = z\Big]$ is bounded by C for almost every $z \in \mathcal{Z}$. We wish to apply Lemma 13 to the quantities

$$f_P(x,z) - \{f_P(\cdot,z) * K_h\}(x); \quad f_P'(x,z) - \{f_P(\cdot,z) * K_h\}'(x).$$

To this end, note that by a Taylor expansion

$$|f_P(x+hw,z)| \le |f_P(x,z)| + h|w||f_P'(x,z)| + \frac{C_P(z)}{2} w^2 h^2. \qquad (2.27)$$

Since f_P, f_P' are real-valued, both $|f_P(x,z)|$ and $|f_P'(x,z)|$ are finite for any fixed x, z. The conditions for Lemma 13 follow. Now we have that

$$\begin{aligned}
\Big|\{f_P(\cdot,z) * K_h\}(x) - f_P(x,z)\Big| &= \Big|\mathbb{E}[f_P(x+hW,z) - f_P(x,z)]\Big| \\
&\le \left|\mathbb{E}\left[hW f_P'(x,z) + \frac{h^2W^2}{2} \sup_{t\in\mathbb{R}} \Big|f_P''(t,z)\Big|\right]\right| \\
&\le \frac{C_P(z)}{2} h^2.
\end{aligned}$$

In the second line we have applied equation (2.27) and the third line $\mathbb{E}(W) = 0, \mathbb{E}\Big(W^2\Big) = 1$. Similarly,

$$\begin{aligned}
&\Big|\{f_P(\cdot, z) * K_h\}'(x) - f'_P(x, z)\Big| \\
&\qquad = \left|\frac{1}{h}\mathbb{E}[W f_P(x + hW, z) - h f'_P(x, z)]\right| \\
&\qquad \leq \left|\frac{1}{h}\mathbb{E}\left[W f_P(x, z) + h\Big(W^2 - 1\Big) f'_P(x, z) + \frac{h^2|W^3|}{2} \sup_{t\in\mathbb{R}} \Big|f''_P(t, z)\Big|\right]\right| \\
&\qquad \leq \frac{\sqrt{2}C_P(z)}{\sqrt{\pi}} h,
\end{aligned}$$

noting that $\mathbb{E}\Big(|W|^3\Big) = 2\sqrt{2/\pi}$. The choice of $h = cn^{-\gamma}$ for any

$$\gamma \geq \alpha/4 \tag{2.28}$$

yields the desired rates on the respective terms in equations (2.25, 2.26).

Write

$$\Delta_{P,n}(x, z) := f_P(x, z) - \tilde{f}^{(n,1)}(x, z).$$

It remains to demonstrate the following rates

$$\mathbb{E}_P\Big[\{\Delta_{P,n}(\cdot, Z) * K_h\}^2(X) \;\Big|\; D^{(n,1)}\Big] = O_{\mathcal{P}}\Big(\tilde{A}_f^{(n)}\Big); \tag{2.29}$$

$$\mathbb{E}_P\Big[\{\Delta_{P,n}(\cdot, Z) * K_h\}'^2(X) \;\Big|\; D^{(n,1)}\Big] = o_{\mathcal{P}}(1); \tag{2.30}$$

$$\mathbb{E}_P\Big[\rho_P^2(X, Z)\{\Delta_{P,n}(\cdot, Z) * K_h\}^2(X) \;\Big|\; D^{(n,1)}\Big] = o_{\mathcal{P}}(1); \tag{2.31}$$

which we do by proving bounds in terms of

$$\begin{aligned}
\mathbb{E}_P\Big[\Delta_{P,n}^2(X, Z) \;\Big|\; D^{(n,1)}\Big] &= \tilde{A}_f^{(n)}; \\
\Big(\mathbb{E}_P\Big[|\Delta_{P,n}(X, Z)|^{2+\eta} \;\Big|\; D^{(n,1)}\Big]\Big)^{\frac{2}{2+\eta}} &= \tilde{B}_f^{(n)}.
\end{aligned} \tag{2.32}$$

We work on the event that $\Delta_{P,n}$ is bounded, which happens with high probability by assumption. This will enable us to use dominated convergence to exchange various limits below. Recall that we are not assuming any smoothness of $\tilde{f}^{(n,1)}$ or $\Delta_{P,n}$. Since we are working under the event that $\Delta_{P,n}$ is bounded over (x, z), we have that $\Delta_{P,n} * K_n$ is bounded by the same bound as $\Delta_{P,n}$, and also due to Lemma 13 we have that $(\Delta_{P,n} * K_h)'$ exists and is bounded. By assumption and Theorem 5 we have that $\rho'_P(x, z)$ and all moments of $\rho_P(X, Z)$ are bounded.

We first show that (2.31) follows from (2.29, 2.30). Due to the aforementioned bounds and Lemma 23, we can apply Proposition 2 as follows.

$$\begin{aligned}
&\mathbb{E}_P\Big[\rho_P^2(X,Z)\{\Delta_{P,n}(\cdot,Z) * K_h\}^2(X) \;\Big|\; D^{(n,1)}\Big] \\
&\quad = -\mathbb{E}_P\Big[\rho_P'(X,Z)\{\Delta_{P,n}(\cdot,Z) * K_h\}^2(X) \\
&\qquad\qquad + 2\rho_P(X,Z)\{\Delta_{P,n}(\cdot,Z) * K_h\}'(X)\{\Delta_{P,n}(\cdot,Z) * K_h\}(X) \;\Big|\; D^{(n,1)}\Big] \\
&\quad \leq \sup_{x,z} |\rho_P'(x,z)| \mathbb{E}_P\Big[\{\Delta_{P,n}(\cdot,Z) * K_h\}^2(X) \;\Big|\; D^{(n,1)}\Big] \\
&\qquad + \sup_{x,z} |\Delta_{P,n}(x,z)| \Big(\mathbb{E}_P[\rho_P^2(X,Z)]\Big)^{1/2} \Big(\mathbb{E}_P\Big[\{\Delta_{P,n}(\cdot,Z) * K_h\}'^2(X) \;\Big|\; D^{(n,1)}\Big]\Big)^{1/2}
\end{aligned}$$

The second line is due to the Hölder and Cauchy–Schwarz inequalities. All the random quantities above are integrable due to the stated bounds. It remains to show (2.29, 2.30).

We start with (2.29). By Lemma 13, conditional Jensen's inequality, and Fubini's theorem,

$$\begin{aligned}
\mathbb{E}_P\Big[\{\Delta_{P,n}(\cdot,Z) * K_h\}^2(X) \;\Big|\; D^{(n,1)}\Big] &= \mathbb{E}_P\Big[\mathbb{E}\{\Delta_{P,n}(X+hW,Z) \mid X, Z, D^{(n,1)}\}^2 \;\Big|\; D^{(n,1)}\Big] \\
&\leq \mathbb{E}_P\Big[\mathbb{E}\Big\{\Delta_{P,n}^2(X+hW,Z) \mid X, Z, D^{(n,1)}\Big\} \;\Big|\; D^{(n,1)}\Big] \\
&= \mathbb{E}\Big[\mathbb{E}_P\Big\{\Delta_{P,n}^2(X+hW,Z) \;\Big|\; W, D^{(n,1)}\Big\} \;\Big|\; D^{(n,1)}\Big].
\end{aligned}$$

Define a new function $\phi_{P,n} : \mathbb{R} \to \mathbb{R}$ by $\phi_{P,n}(t) = \mathbb{E}_P\Big[\Delta_{P,n}^2(X+t,Z) \;\Big|\; D^{(n,1)}\Big]$, so $\phi_{P,n}(0) = \tilde{A}_f^{(n)}$. We will show later in the proof that $\phi_{P,n}$ is twice differentiable, which we assume to be true for now. By a Taylor expansion, for each fixed $h > 0$, $w \in \mathbb{R}$ we have

$$\phi_{P,n}(hw) \leq \phi_{P,n}(0) + hw\phi_{P,n}'(0) + \frac{h^2w^2}{2} \sup_{|t| \leq h|w|} \Big|\phi_{P,n}''(t)\Big|.$$

We will also show later that the remainder term is integrable with respect to the Gaussian density. Taking expectations over W yields

$$\begin{aligned}
\mathbb{E}\Big[\phi_{P,n}(hW) \;\Big|\; D^{(n,1)}\Big] &\leq \phi_{P,n}(0) + h\mathbb{E}(W)\phi_{P,n}'(0) + \frac{h^2}{2}\int_{\mathbb{R}} w^2K(w) \sup_{|t| \leq h|w|} \Big|\phi_{P,n}''(t)\Big| \, dw \\
&= \phi_{P,n}(0) + \frac{h^2}{2}\int_{\mathbb{R}} w^2K(w) \sup_{|t| \leq h|w|} \Big|\phi_{P,n}''(t)\Big| \, dw. \qquad (2.33)
\end{aligned}$$

In the final line we have used $\mathbb{E}(W) = 0$.

Now considering the quantity (2.30), Lemma 13 implies that

$$\{\Delta_{P,n}(\cdot,z) * K_h\}'(x) = \frac{1}{h}\mathbb{E}\Big[W\Delta_{P,n}(x+hW,z) \;\Big|\; D^{(n,1)}\Big].$$

Similarly to above,

$$\begin{aligned}\mathbb{E}_P\Big([\{\Delta_{P,n}(\cdot,Z)*K_h\}'(X)]^2 \;\Big|\; D^{(n,1)}\Big) \\ &= h^{-2}\mathbb{E}_P\Big[\mathbb{E}\Big\{W\Delta_{P,n}(X+hW,Z) \;\Big|\; X,Z,D^{(n,1)}\Big\}^2 \;\Big|\; D^{(n,1)}\Big] \\ &\le h^{-2}\mathbb{E}_P\Big[\mathbb{E}\Big\{W^2\Delta_{P,n}^2(X+hW,Z) \;\Big|\; X,Z,D^{(n,1)}\Big\} \;\Big|\; D^{(n,1)}\Big] \\ &= h^{-2}\mathbb{E}\Big[W^2\mathbb{E}_P\Big\{\Delta_{P,n}^2(X+hW,Z) \;\Big|\; W,D^{(n,1)}\Big\} \;\Big|\; D^{(n,1)}\Big].\end{aligned}$$

Moreover,

$$\begin{aligned}h^{-2}\mathbb{E}\Big[W^2\phi_{P,n}(hW) \;\Big|\; D^{(n,1)}\Big] &= h^{-2}\mathbb{E}\Big(W^2\Big)\phi_{P,n}(0) + h^{-1}\mathbb{E}\Big(W^3\Big)\phi'_{P,n}(0) \\ &\quad + \frac{1}{2}\int_{\mathbb{R}} w^4K(w)\sup_{|t|\le h|w|}\Big|\phi''_{P,n}(t)\Big|\,dw \\ &= h^{-2}\phi_{P,n}(0) + \frac{1}{2}\int_{\mathbb{R}} w^4K(w)\sup_{|t|\le h|w|}\Big|\phi''_{P,n}(t)\Big|\,dw. \qquad (2.34)\end{aligned}$$

In the final line we have used $\mathbb{E}\Big(W^2\Big) = 1$, $\mathbb{E}\Big(W^3\Big) = 0$.

It remains to check that $\phi_{P,n}$ is twice differentiable and compute its derivatives. By a change of variables $u = x + t$,

$$\begin{aligned}\phi_{P,n}(t) &= \mathbb{E}_P\Big[\mathbb{E}_P\Big\{\Delta_{P,n}^2(X+t,Z) \;\Big|\; Z,D^{(n,1)}\Big\} \;\Big|\; D^{(n,1)}\Big] \\ &= \mathbb{E}_P\left[\int_{\mathbb{R}}\Delta_{P,n}^2(x+t,Z)p_P(x \mid Z)\,dx \;\middle|\; D^{(n,1)}\right] \\ &= \mathbb{E}_P\left[\int_{\mathbb{R}}\Delta_{P,n}^2(u,Z)p_P(u-t \mid Z)\,du \;\middle|\; D^{(n,1)}\right].\end{aligned}$$

The conditional density p_P is assumed twice differentiable, so the integrand is twice differentiable with respect to t. The bound on $\Delta_{P,n}$ and conclusion of Lemma 14 allow us to interchange the differentiation and expectation operators using Aliprantis and Burkinshaw (1990, Thm. 20.4). Differentiating $\phi_{P,n}$ twice gives

$$\begin{aligned}\phi''_{P,n}(t) &= \partial_t^2\mathbb{E}_P\left[\int_{\mathbb{R}}\Delta_{P,n}^2(u,Z)p_P(u-t \mid Z)\,du \;\middle|\; D^{(n,1)}\right] \\ &= \mathbb{E}_P\left[\int_{\mathbb{R}}\Delta_{P,n}^2(u,Z)p''_P(u-t \mid Z)\,du \;\middle|\; D^{(n,1)}\right]. \qquad (2.35)\end{aligned}$$

Note that

$$\begin{aligned}\rho'_P(x,z) &= \left(\frac{p'_P(x \mid z)}{p_P(x \mid z)}\right)' \\ &= \frac{p''_P(x \mid z)}{p_P(x \mid z)} - \left(\frac{p'_P(x \mid z)}{p_P(x \mid z)}\right)^2 \\ &= \frac{p''_P(x \mid z)}{p_P(x \mid z)} - \rho_P^2(x,z). \end{aligned} \tag{2.36}$$

Applying equation (2.36), the Lipschitz property of ρ_P, and Lemma 26 to the interior of (2.35) yields

$$\begin{aligned}&\left| \int_{\mathbb{R}} \Delta_{P,n}^2(u,z) p''_P(u-t \mid z)\, du \right| \\ &\quad = \left| \int_{\mathbb{R}} \Delta_{P,n}^2(u,z) \big\{ \rho'_P(u-t,z) + \rho_P^2(u-t,z) \big\}\, p_P(u-t \mid z)\, du \right| \\ &\quad = \left| \int_{\mathbb{R}} \Delta_{P,n}^2(u,z) \Big[\rho'_P(u-t,z) + \{\rho_P(u-t,z) - \rho_P(u,z) + \rho_P(u,z)\}^2 \Big] \right. \\ &\qquad \left. \frac{p_P(u-t \mid z)}{p_P(u \mid z)} p_P(u \mid z)\, du \right| \\ &\quad \le \int_{\mathbb{R}} \Delta_{P,n}^2(u,z) \Big[C + \{C|t| + \rho_P(u,z)\}^2 \Big]\, \exp\left(-t\rho_P(u,z) + \frac{C}{2}t^2 \right) p_P(u \mid z)\, du \\ &\quad \le \int_{\mathbb{R}} \Delta_{P,n}^2(u,z) \big\{ C + 2C^2t^2 + 2\rho_P^2(u,z) \big\}\, \exp\left(-t\rho_P(u,z) + \frac{C}{2}t^2 \right) p_P(u \mid z)\, du \\ &\quad = \mathbb{E}_P \left[\Delta_{P,n}^2(X,z) \big\{ C + 2C^2t^2 + 2\rho_P^2(X,z) \big\} \right. \\ &\qquad\qquad \left. \exp\left(-t\rho_P(X,z) + \frac{C}{2}t^2 \right) \,\middle|\, Z = z, D^{(n,1)} \right]. \end{aligned}$$

The penultimate line uses $(a+b)^2 \le 2(a^2+b^2)$. Plugging this in to (2.35) and using Fubini's theorem gives

$$\begin{aligned}\left|\phi''_{P,n}(t)\right| &\le \mathbb{E}_P \left[\Delta_{P,n}^2(X,Z) \big\{ C + 2C^2t^2 + 2\rho_P^2(X,Z) \big\}\, \exp\left(-t\rho_P(X,Z) + \frac{C}{2}t^2 \right) \,\middle|\, D^{(n,1)} \right] \\ &= \big(C + 2C^2t^2 \big) \exp\left(\frac{C}{2}t^2 \right) \mathbb{E}_P \Big[\Delta_{P,n}^2(X,Z)\, \exp\big(-t\rho_P(X,Z) \big) \,\Big|\, D^{(n,1)} \Big] \\ &\quad + 2\exp\left(\frac{C}{2}t^2 \right) \mathbb{E}_P \Big[\Delta_{P,n}^2(X,Z) \rho_P^2(X,Z)\, \exp\big(-t\rho_P(X,Z) \big) \,\Big|\, D^{(n,1)} \Big]. \end{aligned} \tag{2.37}$$

We will use Hölder's inequality to bound this in terms of (2.32). Pick q_1 to be any integer strictly larger than $(2+\eta)/\eta$, and set q_2 so that $1/q_1 + 1/q_2 = \eta/(2+\eta)$. Applying Hölder's inequality to (2.37) twice,

$$\begin{aligned}
\left|\phi''_{P,n}(t)\right| &\leq \left(C + 2C^2t^2\right)\exp\left(\frac{C}{2}t^2\right)\left(\mathbb{E}_P\left[\exp\left(-\frac{2+\eta}{\eta}t\rho_P(X,Z)\right)\right]\right)^{\frac{\eta}{2+\eta}}\tilde{B}_f^{(n)} \\
&\quad + 2\exp\left(\frac{C}{2}t^2\right)\left(\mathbb{E}_P\left[|\rho_P(X,Z)|^{\frac{2(2+\eta)}{\eta}}\exp\left(-\frac{2+\eta}{\eta}t\rho_P(X,Z)\right)\right]\right)^{\frac{\eta}{2+\eta}}\tilde{B}_f^{(n)} \\
&\leq \exp\left(\frac{C}{2}t^2\right)\left\{C + 2C^2t^2 + 2\Big(\mathbb{E}_P\Big[\rho_P^{2q_1}(X,Z)\Big]\Big)^{\frac{1}{q_1}}\right\} \\
&\qquad \left(\mathbb{E}_P\left[\exp\left(-q_2t\rho_P(X,Z)\right)\right]\right)^{\frac{1}{q_2}}\tilde{B}_f^{(n)}.
\end{aligned}$$

We are now in a position to apply Theorem 5. Recalling the moment generating function bound for sub-Gaussian random variables, we have

$$\begin{aligned}
\phi''_{P,n}(t) &\leq \exp\left(\frac{C}{2}t^2\right)\left\{C + 2C^2t^2 + 2\Big(C^{q_1}(2q_1-1)!!\Big)^{\frac{1}{q_1}}\right\}\Big(\exp\left(q_2^2C^2t^2\right)\Big)^{\frac{1}{q_2}}\tilde{B}_f^{(n)} \\
&\leq c_1\Big(1+t^2\Big)\exp\Big(c_2t^2\Big)\tilde{B}_f^{(n)},
\end{aligned}$$

for some constants $c_1, c_2 > 0$ depending on C and η but not on P or n.

Returning to equations (2.33, 2.34), we have

$$\begin{aligned}
\mathbb{E}\Big[\phi_{P,n}(hW)\;\Big|\;D^{(n,1)}\Big] &\leq \tilde{A}_f^{(n)} + c_1h^2\tilde{B}_f^{(n)}\int_{\mathbb{R}} w^2(1+h^2w^2)\exp(c_2h^2w^2)K(w)\,dw. \\
h^{-2}\mathbb{E}\Big[W^2\phi_{P,n}(hW)\;\Big|\;D^{(n,1)}\Big] &\leq h^{-2}\tilde{A}_f^{(n)} + c_1h^2\tilde{B}_f^{(n)}\int_{\mathbb{R}} w^4(1+h^2w^2)\exp(c_2h^2w^2)K(w)\,dw.
\end{aligned}$$

For all $0 < h < \frac{1}{2\sqrt{c_2}}$, the final integrals are bounded by a constant.

Hence the choice of $h = cn^{-\gamma}$ for any

$$(\alpha-\beta)/2 \leq \gamma < \alpha/2$$

yields the desired rates on (2.29, 2.30). Combining this with (2.28) gives the final range

$$\max\{\alpha/4, (\alpha-\beta)/2\} \leq \gamma < \alpha/2$$

to achieve the desired rates in (2.25, 2.26). □

2.8.1 Auxiliary lemmas

Lemma 13. *Let $W \sim N(0,1)$ be a standard Gaussian random variable independent of (X, Z), and fix $h > 0$. Let $g : \mathbb{R} \times \mathcal{Z} \to \mathbb{R}$ be such that $\mathbb{E}|g(x + hW, z)| < \infty$ for all (x, z). Then we have that for each z*

$$\{g(\cdot, z) * K_h\}(x) = \mathbb{E}[g(x + hW, z)]$$

is absolutely continuous in x, and for almost every x its derivative exists. If $\mathbb{E}\Big[|g(x + hW, z)|^{1+\eta}\Big] < \infty$ for some $\eta > 0$ then the derivative is given by

$$\{g(\cdot, z) * K_h\}'(x) = \frac{1}{h}\mathbb{E}[W g(x + hW, z)].$$

Proof. Recall that the convolution operator is

$$\{g(\cdot, z) * K_h\}(x) = \int_R g(u, z) K_h(x - u)\, du.$$

We check the conditions for interchanging differentiation and integration operators (Aliprantis and Burkinshaw, 1990, Thm. 20.4). The integrand $g(u, z)K_h(x - u)$ is integrable in u with respect to the Lebesgue measure for each (x, z), since

$$\begin{aligned}
\int_{\mathbb{R}} |g(u, z)| K_h(x - u)\, du &= \int_{\mathbb{R}} |g(u, z)|\, \frac{1}{h}\, K\left(\frac{x - u}{h}\right) du \\
&= \int_{\mathbb{R}} |g(x + hw, z)| K(-w)\, dw \\
&= \int_{\mathbb{R}} |g(x + hw, z)| K(w)\, dw \\
&= \mathbb{E}[|g(x + hW, z)|] < \infty.
\end{aligned}$$

Due to the smoothness of the Gaussian kernel, $g(u, z)K_h(x - u)$ is absolutely continuous in x for each (u, z). Furthermore it has x-derivative

$$g(u, z)K_h'(x - u) = -g(u, z)\left(\frac{x - u}{h^2}\right) K_h(x - u).$$

Fix x_0 and $V = [x_0 - h, x_0 + h]$. It remains to find a Lebesgue integrable function $G : \mathbb{R} \to \mathbb{R}$ such that

$$\left| g(u, z)\left(\frac{x - u}{h^2}\right) K_h(x - u) \right| \leq G(u)$$

for all $x \in V$ and $u \in \mathbb{R}$. Now for any $x \in V$, $u \in \mathbb{R}$,

$$\begin{aligned}
&\left|g(u,z)\left(\frac{x-u}{h^2}\right)K_h(x-u)\right| \\
&\quad= \left|g(u,z)\left(\frac{x-x_0+x_0-u}{h^2}\right)\frac{K_h(x-u)}{K_h(x_0-u)}K_h(x_0-u)\right| \\
&\quad= |g(u,z)|\left|\frac{x-x_0+x_0-u}{h^2}\right|\exp\left(-\frac{(x-x_0)^2}{2}-(x-x_0)(x_0-u)\right)K_h(x_0-u) \\
&\quad\leq |g(u,z)|\frac{h+|x_0-u|}{h^2}\exp\left(-\frac{h^2}{2}+h|x_0-u|\right)K_h(x_0-u) \\
&\quad=: G(u).
\end{aligned}$$

Moreover, recalling the symmetry of K and using a change of variables $w = (u-x_0)/h$,

$$\int_{\mathbb{R}} G(u)\,du = \frac{1}{h}\exp\left(-\frac{h^2}{2}\right)\int_{\mathbb{R}}|g(x_0+hw,z)|(1+|w|)\exp\left(|w|h^2\right)K(w)\,dw$$

We now apply Hölder's inequality twice. Pick $q_1, q_2 > 1$ be such that $1/q_1+1/q_2 = \eta/(1+\eta)$. Now,

$$\begin{aligned}
\int_{\mathbb{R}} G(u)\,du &\leq \frac{1}{h}\exp\left(-\frac{h^2}{2}\right)\left(\mathbb{E}\left[|g(x_0+hW,z)|^{1+\eta}\right]\right)^{\frac{1}{1+\eta}} \\
&\qquad\left(\int_{\mathbb{R}}(1+|w|)^{\frac{1+\eta}{\eta}}\exp\left(\frac{1+\eta}{\eta}|w|h^2\right)K(w)\,dw\right)^{\frac{\eta}{1+\eta}} \\
&\leq \frac{1}{h}\exp\left(-\frac{h^2}{2}\right)\left(\mathbb{E}\left[|g(x_0+hW,z)|^{1+\eta}\right]\right)^{\frac{1}{1+\eta}} \\
&\qquad\left(\mathbb{E}\left[(1+|W|)^{q_1}\right]\right)^{\frac{1}{q_1}}\left(\mathbb{E}\left[\exp\left(q_2|w|h^2\right)\right]\right)^{\frac{1}{q_2}}.
\end{aligned}$$

We have that $\mathbb{E}\left[|g(x_0+hW,z)|^{1+\eta}\right]$ is finite by assumption, $\mathbb{E}\left[(1+|W|)^{q_1}\right]$ is a Gaussian moment so is finite, and $(\mathbb{E}\left[\exp\left(q_2|w|h^2\right)\right]$ is bounded in terms of the Gaussian moment generating function. Hence G is Lebesgue integrable.

Finally we check the claimed identities. Using a change of variables $u = x+hw$, and recalling the symmetry of K, we have that

$$\begin{aligned}
\{g(\cdot,z)*K_h\}(x) &= \int_{\mathbb{R}} g(u,z)\,\frac{1}{h}\,K\left(\frac{x-u}{h}\right)du \\
&= \int_{\mathbb{R}} g(x+hw,z)K(w)\,dw \\
&= \mathbb{E}[g(x+hW,z)],
\end{aligned}$$

and

$$\begin{aligned}\{g(\cdot,z) * K_h\}'(x) &= \int_{\mathbb{R}} g(u,z) K_h'(x-u)\, du \\ &= -\frac{1}{h^2}\int_{\mathbb{R}} g(u,z)\left(\frac{x-u}{h}\right) K\left(\frac{x-u}{h}\right) du \\ &= \frac{1}{h}\int_{\mathbb{R}} g(x+hw,z) w K(w)\, dw \\ &= \frac{1}{h}\mathbb{E}[W g(x+hW,z)].\end{aligned}$$

□

Lemma 14. *Let p be a twice differentiable density on $\mathbb{R}$, with $\sup_{x\in\mathbb{R}} |\partial_x^2 \log p(x)| = \sup_{x\in\mathbb{R}} |\rho'(x)| \le C < \infty$. Then for every $t_0 \in \mathbb{R}$ there exists a neighbourhood V of t_0 and Lebesgue integrable function g such that*

$$|p'(x-t)|,\ |p''(x-t)| \le g(x)$$

for all $x \in \mathbb{R}$ and $t \in V$.

Proof. Fix t_0 and $V = [t_0 - 1, t_0 + 1]$. We will make use of Lemma 26. Indeed for any $t \in V$,

$$\begin{aligned}|p'(x-t)| &= |\rho(x-t)|p(x-t) \\ &= |\rho(x-t) - \rho(x-t_0) + \rho(x-t_0)|\frac{p(x-t)}{p(x-t_0)}p(x-t_0) \\ &\le \Big\{C|t-t_0| + |\rho(x-t_0)|\Big\}\frac{p(x-t)}{p(x-t_0)}p(x-t_0) \\ &\le \Big\{C|t-t_0| + |\rho(x-t_0)|\Big\}\exp\left(|t-t_0|\ |\rho(x-t_0)| + \frac{(t-t_0)^2 C}{2}\right)p(x-t_0) \\ &\le \Big\{C + |\rho(x-t_0)|\Big\}\exp\left(|\rho(x-t_0)| + \frac{C}{2}\right)p(x-t_0).\end{aligned}$$

Similarly,

$$\begin{aligned}
|p''(x-t)| &= |\rho'(x-t) + \rho^2(x-t)|p(x-t) \\
&\leq \Big[C + \{\rho(x-t) - \rho(x-t_0) + \rho(x-t_0)\}^2\Big] \frac{p(x-t)}{p(x-t_0)} p(x-t_0) \\
&\leq \Big[C + 2C^2(t-t_0)^2 + 2\rho^2(x-t_0)\Big] \frac{p(x-t)}{p(x-t_0)} p(x-t_0) \\
&\leq \Big[C + 2C^2(t-t_0)^2 + 2\rho^2(x-t_0)\Big] \\
&\qquad \exp\left(|t-t_0|\,|\rho(x-t_0)| + \frac{(t-t_0)^2 C}{2}\right) p(x-t_0) \\
&\leq \Big[C + 2C^2 + 2\rho^2(x-t_0)\Big] \exp\left(|\rho(x-t_0)| + \frac{C}{2}\right) p(x-t_0).
\end{aligned}$$

In the third line we have used the inequality $(a+b)^2 \leq 2(a^2+b^2)$.

Taking g to be the maximum of the two bounds, it suffices to check that the function $|\rho(x-t_0)|^k \exp(|\rho(x-t_0)|)p(x-t_0)$ is Lebesgue integrable with respect to x for $k = 0, 1, 2$. Using the change of variables $y = x - t_0$ and the Cauchy–Schwarz inequality,

$$\begin{aligned}
\int_{\mathbb{R}} |\rho(x-t_0)|^k \exp(|\rho(x-t_0)|)p(x-t_0)\, dx &= \int_{\mathbb{R}} |\rho(y)|^k \exp(|\rho(y)|)p(y)\, dx \\
&= \mathbb{E}\Big[|\rho(X)|^k \exp(|\rho(X)|)\Big] \\
&\leq \Big(\mathbb{E}\Big[\rho^{2k}(X)\Big]\Big)^{\frac{1}{2}} \Big(\mathbb{E}\Big[\exp(2|\rho(X)|)\Big]\Big)^{\frac{1}{2}},
\end{aligned}$$

where $X \sim p$. By Theorem 5,

$$\mathbb{E}\Big[\rho^{2k}(X)\Big] \leq C^k(2k-1)!!$$

for $k = 1, 2$ and moreover $\rho(X)$ is sub-Gaussian with parameter $\sqrt{2C}$, so

$$\begin{aligned}
\mathbb{E}\Big[\exp(2|\rho(X)|)\Big] &\leq \mathbb{E}\Big[\exp(2\rho(X))\Big] + \mathbb{E}\Big[\exp(-2\rho(X))\Big] \\
&\leq 2\exp(4C).
\end{aligned}$$

This completes the proof. □

2.9 Proofs relating to Section 2.4

Our proofs make use of the following representations of $p_{\hat{\varepsilon}}(\epsilon)$ and $\rho_{\hat{\varepsilon}}(\epsilon)$. We first note that

$$\varepsilon_P = \hat{\varepsilon}^{(n)} + u_\sigma^{(n)}(Z)\,\hat{\varepsilon}^{(n)} + u_m^{(n)}(Z),$$

where we recall

$$u_\sigma^{(n)}(z) := \frac{\hat{\sigma}^{(n)}(z) - \sigma_P(z)}{\sigma_P(z)}; \quad u_m^{(n)}(z) := \frac{\hat{m}^{(n)}(z) - m_P(z)}{\sigma_P(z)}.$$

Recall that since we do not have access to samples of ε_P, only $\hat{\varepsilon}^{(n)}$, our goal is to show that the score functions of these two variables are similar. Conditionally on $D^{(n)}$, and for each fixed $\epsilon \in \mathbb{R}$ and $z \in \mathcal{Z}$, the estimated residual $\hat{\varepsilon}^{(n)}$ and covariates Z have joint density

$$\begin{aligned} p_{\hat{\varepsilon},Z}(\epsilon, z) &= p_{\varepsilon,Z}\Big(\epsilon + u_\sigma^{(n)}(z)\epsilon + u_m^{(n)}(z), z\Big) \\ &= p_\varepsilon\Big(\epsilon + u_\sigma^{(n)}(z)\epsilon + u_m^{(n)}(z)\Big)\, p_Z(z), \end{aligned}$$

where the first equality is via a change-of-variables and the second is using the independence of ε and Z. Integrating over z, we have that the marginal density of $\hat{\varepsilon}^{(n)}$, conditionally on $D^{(n)}$, is

$$p_{\hat{\varepsilon}}(\epsilon) = \mathbb{E}_P\Big[p_\varepsilon\Big(\epsilon + u_\sigma^{(n)}(Z)\epsilon + u_m^{(n)}(Z)\Big) \;\Big|\; D^{(n)}\Big].$$

If p'_ϵ is bounded and $\mathbb{E}[|u_\sigma(Z)| \mid D^{(n)}] < \infty$ then the estimated residual score function is

$$\begin{aligned} \rho_{\hat{\varepsilon}}(\epsilon) &= \frac{p'_{\hat{\varepsilon}}(\epsilon)}{p_{\hat{\varepsilon}}(\epsilon)} \\ &= \frac{\mathbb{E}_P\Big[\Big\{1 + u_\sigma^{(n)}(Z)\Big\}p'_\varepsilon\Big(\epsilon + u_\sigma^{(n)}(Z)\epsilon + u_m^{(n)}(Z)\Big) \;\Big|\; D^{(n)}\Big]}{\mathbb{E}_P\Big[p_\varepsilon\Big(\epsilon + u_\sigma^{(n)}(Z)\epsilon + u_m^{(n)}(Z)\Big) \;\Big|\; D^{(n)}\Big]}, \end{aligned}$$

by differentiating under the integral sign (see, for example, Aliprantis and Burkinshaw (1990, Thm. 20.4)).

2.9.1 Proof of Theorem 5

Proof. By Wainwright (2019, Thm. 2.6), the moment bound is sufficient to show sub-Gaussianity. Note that when X is symmetrically distributed, its density $p(\cdot)$ is anti-symmetric. Thus its score function $\rho(\cdot)$ is anti-symmetric, and so the random variable $\rho(X)$ is symmetrically distributed.

We prove the moment bound by induction. Suppose it is true for all $1 \le j < k$ for some $k \ge 1$. By the product rule,

$$\begin{aligned} \Big(\rho^{2k-1}(x)p(x)\Big)' &= \rho^{2k-1}(x)p'(x) + (2k-1)\rho'(x)\rho^{2k-2}(x)p(x) \\ &= \rho^{2k}(x)p(x) + (2k-1)\rho'(x)\rho^{2k-2}(x)p(x). \end{aligned}$$

Therefore for any $-\infty < a < b < \infty$ we have

$$\int_a^b \rho^{2k}(x)p(x)\,dx = \rho^{2k-1}(b)p(b) - \rho^{2k-1}(a)p(a) - (2k-1)\int_a^b \rho'(x)\rho^{2k-2}(x)p(x)\,dx. \quad (2.38)$$

We have that $\mathbb{E}[\rho^{2k-2}(X)] < \infty$ by the induction hypothesis if $k \geq 2$ and trivially if $k = 1$. By Lemma 25 we can choose sequences $a_n \to -\infty$, $b_n \to \infty$ such that

$$\lim_{n\to\infty} \left\{\rho^{2k-1}(b_n)p(b_n) - \rho^{2k-1}(a_n)p(a_n)\right\} = 0.$$

By Hölder's inequality, we have that

$$\begin{aligned}\int_{\mathbb{R}} \left|\rho'(x)\rho^{2k-2}(x)p(x)\right| dx &\leq C\int_{\mathbb{R}} \rho^{2k-2}(x)p(x)\,dx \\ &\leq \begin{cases} C^k(2k-3)!! \text{ if } k \geq 2 \text{ by the induction hypothesis;} \\ C \text{ if } k = 1. \end{cases}\end{aligned}$$

Therefore dominated convergence gives

$$\begin{aligned}\lim_{n\to\infty}\left|(2k-1)\int_{a_n}^{b_n} \rho'(x)\rho^{2k-2}(x)p(x)\,dx\right| &= \left|(2k-1)\int_{\mathbb{R}} \rho'(x)\rho^{2k-2}(x)p(x)\,dx\right| \\ &\leq C^k(2k-1)!!.\end{aligned}$$

Finally, we can assume without loss of generality that the sequences (a_n) and (b_n) are both monotone, for example by relabelling their monotone sub-sequences. Now, for each $x \in \mathbb{R}$ the sequence $\mathbb{1}_{[a_n,b_n]}(x)\rho^{2k}(x)p(x)$ is increasing in n. The monotone convergence theorem thus gives

$$\lim_{n\to\infty}\int_{a_n}^{b_n} \rho^{2k}(x)p(x)\,dx = \mathbb{E}\left[\rho^{2k}(X)\right].$$

Taking the limit in equation (2.38) yields

$$\begin{aligned}\mathbb{E}\left[\rho^{2k}(X)\right] &= \lim_{n\to\infty}\int_{a_n}^{b_n} \rho^{2k}(x)p(x)\,dx \\ &= \lim_{n\to\infty}\Bigg\{\rho^{2k-1}(b_n)p(b_n) - \rho^{2k-1}(a_n)p(a_n) \\ &\qquad - (2k-1)\int_{a_n}^{b_n} \rho'(x)\rho^{2k-2}(x)p(x)\,dx\Bigg\} \\ &\leq C^k(2k-1)!!,\end{aligned}$$

as claimed. □

2.9.2 Proof of Theorem 6

Proof. Define

$$\bar{\rho}_P^{(n)}(x,z) = \frac{1}{\sigma_P(z)}\,\rho_\varepsilon\!\left(\frac{x-\hat{m}^{(n)}(z)}{\sigma_P(z)}\right).$$

Using the inequality $(a+b)^2 \le 2(a^2+b^2)$, we have

$$\begin{aligned}
A_\rho^{(n)} &= \mathbb{E}_P\Big[\Big\{\rho_P(X,Z) - \bar{\rho}_P^{(n)}(X,Z) + \bar{\rho}_P^{(n)}(X,Z) - \hat{\rho}^{(n)}(X,Z)\Big\}^2 \;\Big|\; D^{(n)}\Big] \\
&\le 2\mathbb{E}_P\Big[\Big\{\rho_P(X,Z) - \bar{\rho}_P^{(n)}(X,Z)\Big\}^2 \;\Big|\; D^{(n)}\Big] + 2\mathbb{E}_P\Big[\Big\{\bar{\rho}_P^{(n)} - \hat{\rho}^{(n)}(X,Z)\Big\}^2 \;\Big|\; D^{(n)}\Big].
\end{aligned}$$

The first term readily simplifies using Hölder's inequality and the Lipschitz property of ρ_ε.

$$\begin{aligned}
&\mathbb{E}_P\Big[\Big\{\rho_P(X,Z) - \bar{\rho}_P^{(n)}(X,Z)\Big\}^2 \;\Big|\; D^{(n)}\Big] \\
&\qquad = \mathbb{E}_P\left[\frac{1}{\sigma_P^2(Z)}\left\{\rho_\varepsilon\!\left(\frac{x-m_P(z)}{\sigma_P(z)}\right) - \rho_\varepsilon\!\left(\frac{x-\hat{m}^{(n)}(z)}{\sigma_P(z)}\right)\right\}^2 \;\middle|\; D^{(n)}\right] \\
&\qquad \le \left(\frac{\|\rho_\varepsilon\|_{Lip}}{\inf_z \sigma_P(z)}\right)^2 \mathbb{E}_P\Big[u_m^{(n)2}(Z) \;\Big|\; D^{(n)}\Big] \\
&\qquad = \left(\frac{\|\rho_\varepsilon\|_{Lip}}{\inf_z \sigma_P(z)}\right)^2 A_m^{(n)}.
\end{aligned}$$

We now expand the second term, working on the arbitrarily high-probability event that $D^{(n)}$ is such that both $\frac{\sigma_P(z)}{\hat{\sigma}^{(n)}(z)}$ and $|u_m^{(n)}(z)|$ are bounded, for all n sufficiently large.

$$\begin{aligned}
&\mathbb{E}_P\Big[\Big\{\bar{\rho}_P^{(n)} - \hat{\rho}^{(n)}(X,Z)\Big\}^2 \;\Big|\; D^{(n)}\Big] \\
&\qquad = \mathbb{E}_P\left[\left\{\frac{1}{\sigma_P(Z)} - \frac{1}{\hat{\sigma}^{(n)}(Z)}\right\}^2 \rho_\varepsilon^2\!\left(\frac{x-\hat{m}^{(n)}(z)}{\hat{\sigma}^{(n)}(z)}\right) \;\middle|\; D^{(n)}\right] \\
&\qquad = \mathbb{E}_P\left[\left\{\frac{1}{\sigma_P(Z)} - \frac{1}{\hat{\sigma}^{(n)}(Z)}\right\}^2 \rho_\varepsilon^2\Big(\varepsilon + u_m^{(n)}(Z)\Big) \;\middle|\; D^{(n)}\right] \\
&\qquad = \mathbb{E}_P\left[\frac{1}{\sigma_P^2(Z)}\left\{\frac{\hat{\sigma}^{(n)}(Z)}{\sigma_P(Z)}\right\}^2 u_\sigma^{(n)2}(Z)\rho_\varepsilon^2\Big(\varepsilon + u_m^{(n)}(Z)\Big) \;\middle|\; D^{(n)}\right].
\end{aligned}$$

Applying the Lipschitz property of ρ_ε and using the inequality $(a+b)^2 \le 2(a^2+b^2)$,

$$\begin{aligned}
\rho_\varepsilon^2\Big(\varepsilon + u_m^{(n)}(Z)\Big) &\le \Big(|\rho_\varepsilon(\varepsilon)| + \|\rho_\varepsilon\|_{Lip}\,|u_m^{(n)}(Z)|\Big)^2 \\
&\le 2\rho_\varepsilon^2(\varepsilon) + 2\|\rho_\varepsilon\|_{Lip}^2\, u_m^{(n)2}(Z).
\end{aligned}$$

Recalling that ε is independent of Z we deduce

$$\begin{aligned}
&\mathbb{E}_P\Big[\Big\{\bar{\rho}_P^{(n)}-\hat{\rho}^{(n)}(X,Z)\Big\}^2 \;\Big|\; D^{(n)}\Big] \\
&\qquad \le \mathbb{E}_P\left[\frac{1}{\sigma_P^2(Z)}\left\{\frac{\hat{\sigma}^{(n)}(Z)}{\sigma_P(Z)}\right\}^2 u_\sigma^{(n)2}(Z)\Big\{2\rho_\varepsilon^2(\varepsilon)+2\|\rho_\varepsilon\|_{Lip}^2\, u_m^{(n)2}(Z)\Big\} \;\middle|\; D^{(n)}\right] \\
&\qquad \le \frac{1}{\Big(\inf_z \sigma_P^2(z)\Big)^2}\left(\sup_z \frac{\hat{\sigma}^{(n)}(z)}{\sigma_P(z)}\right)^2\left(2\mathbb{E}_P\Big[\rho_\varepsilon^2(\varepsilon)\Big]+2\|\rho_\varepsilon\|_{Lip}^2\left(\sup_z|u_m^{(n)}|\right)^2\right) \\
&\qquad\qquad \times \mathbb{E}_P\Big[u_\sigma^{(n)2}(Z)\;\Big|\; D^{(n)}\Big] \\
&\qquad = \frac{1}{\Big(\inf_z \sigma_P^2(z)\Big)^2}\left(\sup_z \frac{\hat{\sigma}^{(n)}(z)}{\sigma_P(z)}\right)^2\left(2\mathbb{E}_P\Big[\rho_\varepsilon^2(\varepsilon)\Big]+2\|\rho_\varepsilon\|_{Lip}^2\left(\sup_z|u_m^{(n)}|\right)^2\right) A_\sigma^{(n)}.
\end{aligned}$$

This suffices to prove the claim. □

2.9.3 Proof of Theorem 7

Proof. The assumptions on $u_m^{(n)}$ and $u_\sigma^{(n)}$ mean that for any $\epsilon>0$ we can find N, C_m, C_σ such that for any $n \ge N$, with uniform probability at least $1-\epsilon$, the data $D^{(n)}$ is such that

$$\sup_z\Big|u_m^{(n)}(z)\Big| \le C_m; \quad \sup_z\Big|u_\sigma^{(n)}(z)\Big| \le C_\sigma. \tag{2.39}$$

It suffices to show that under this event, we can find a uniform constant C (not depending on P or n) such that

$$A_\rho^{(n)} \le C\Big(A_m^{(n)}+A_\sigma^{(n)}+A_{\hat{\varepsilon}}^{(n)}\Big).$$

Fix $P\in\mathcal{P}$ and $D^{(n)}$ such that (2.39) holds. We decompose $A_\rho^{(n)}$ so as to consider the various sources of error separately.

$$\begin{aligned}
A_\rho^{(n)} &= \mathbb{E}_P\Big[\Big\{\rho_P(X,Z)-\hat{\rho}^{(n)}(X,Z)\Big\}^2 \;\Big|\; D^{(n)}\Big] \\
&= \mathbb{E}_P\left[\left\{\frac{1}{\sigma_P(Z)}\rho_\varepsilon(\varepsilon_P)-\frac{1}{\hat{\sigma}^{(n)}(Z)}\hat{\rho}_{\hat{\varepsilon}}^{(n)}\Big(\hat{\varepsilon}^{(n)}\Big)\right\}^2 \;\middle|\; D^{(n)}\right] \\
&= \mathbb{E}_P\left[\frac{1}{\hat{\sigma}^{(n)2}(Z)}\left\{\left(\frac{\hat{\sigma}^{(n)}(Z)}{\sigma_P(Z)}-1\right)\rho_\varepsilon(\varepsilon_P)+\rho_\varepsilon(\varepsilon_P)-\hat{\rho}_{\hat{\varepsilon}}^{(n)}\Big(\hat{\varepsilon}^{(n)}\Big)\right\}^2 \;\middle|\; D^{(n)}\right] \\
&= \mathbb{E}_P\left[\frac{1}{\hat{\sigma}^{(n)2}(Z)}\Big\{-u_\sigma^{(n)}(Z)\rho_\varepsilon(\varepsilon_P)+\rho_\varepsilon(\varepsilon_P)-\rho_{\hat{\varepsilon}}(\varepsilon_P)\right. \\
&\qquad\qquad \left.+\rho_{\hat{\varepsilon}}(\varepsilon_P)-\rho_{\hat{\varepsilon}}(\hat{\varepsilon}^{(n)})+\rho_{\hat{\varepsilon}}(\hat{\varepsilon}^{(n)})-\hat{\rho}_{\hat{\varepsilon}}^{(n)}\Big(\hat{\varepsilon}^{(n)}\Big)\Big\}^2 \;\middle|\; D^{(n)}\right].
\end{aligned}$$

Note that

$$\frac{1}{\hat{\sigma}^{(n)}(z)} = \frac{1}{\left\{1 - u_\sigma^{(n)}(z)\right\}\sigma_P(z)} \leq \frac{1}{1 - C_\sigma} \frac{1}{\inf_{P\in\mathcal{P}} \inf_{z\in\mathcal{Z}} \sigma_P(z)} < \infty.$$

Applying Hölder's inequality and $(a+b+c+d)^2 \leq 4(a^2+b^2+c^2+d^2)$, we deduce

$$\begin{aligned} A_\rho^{(n)} \leq 4 \frac{1}{(1-C_\sigma)^2} \frac{1}{\left(\inf_{P\in\mathcal{P}} \inf_{z\in\mathcal{Z}} \sigma_P(z)\right)^2} \Bigg\{ & \mathbb{E}_P\left[u_\sigma^{(n)2}(Z)\rho_\varepsilon^2(\varepsilon_P) \;\middle|\; D^{(n)}\right] \\ & + \mathbb{E}_P\left[\left\{\rho_\varepsilon(\varepsilon_P) - \rho_{\hat{\varepsilon}}(\varepsilon_P)\right\}^2 \;\middle|\; D^{(n)}\right] \\ & + \mathbb{E}_P\left[\left\{\rho_{\hat{\varepsilon}}(\varepsilon_P) - \rho_{\hat{\varepsilon}}(\hat{\varepsilon}^{(n)})\right\}^2 \;\middle|\; D^{(n)}\right] \\ & + \mathbb{E}_P\left[\left\{\rho_{\hat{\varepsilon}}(\hat{\varepsilon}^{(n)}) - \hat{\rho}_{\hat{\varepsilon}}^{(n)}\left(\hat{\varepsilon}^{(n)}\right)\right\}^2 \;\middle|\; D^{(n)}\right] \Bigg\}. \end{aligned} \tag{2.40}$$

We consider the expectations in (2.40) separately. For the first term, the independence of ε_P and Z and Theorem 5 imply

$$\mathbb{E}_P\left[u_\sigma^{(n)2}(Z)\rho_\varepsilon^2(\varepsilon_P) \;\middle|\; D^{(n)}\right] = \mathbb{E}_P\left[\rho_\varepsilon^2(\varepsilon_P)\right] A_\sigma^{(n)} \leq C_\rho A_\sigma^{(n)}.$$

Lemma 15 applies to the second term. To apply Lemma 18 to the third term, we note that

$$\left(\mathbb{E}_P(\varepsilon_P^8)\right)^{\frac{1}{8}} \leq (768 C_\varepsilon^8)^{\frac{1}{8}} < 3C_\varepsilon$$

by Lemma 16. The fourth term is equal to $A_{\hat{\varepsilon}}^{(n)}$ by definition. This completes the proof. □

2.9.4 Auxiliary lemmas

Lemma 15. *Let P be such that p_ε is twice differentiable on $\mathbb{R}$, with*

$$\sup_{\epsilon\in\mathbb{R}} |\partial_\epsilon^2 \log p_\varepsilon(\epsilon)| = \sup_{\epsilon\in\mathbb{R}} |\rho_\varepsilon'(\epsilon)| \leq C_\rho,$$

p_ε' is bounded, and ε_P is sub-Gaussian with parameter C_ε. Further assume that $D^{(n)}$ is such that $\sup_{z\in\mathcal{Z}} |u_m^{(n)}(z)| \leq C_m$ and $\sup_{z\in\mathcal{Z}} |u_\sigma^{(n)}(z)| \leq C_\sigma$. If

$$\sqrt{C_\rho} C_\sigma C_\varepsilon \leq \frac{1}{18}$$

then there exists a constant C, depending only on $C_\rho, C_m, C_\sigma, C_\varepsilon$, such that

$$\mathbb{E}_P\Big[\Big\{\rho_\varepsilon(\varepsilon_P) - \rho_{\hat{\varepsilon}}(\varepsilon_P)\Big\}^2 \;\Big|\; D^{(n)}\Big] \le C\Big(A_m^{(n)} + A_\sigma^{(n)}\Big).$$

Proof. For ease of notation, write $Q^{(n)}$ for the distribution of $\Big(u_m^{(n)}(Z), u_\sigma^{(n)}(Z)\Big)$ conditionally on $D^{(n)}$, and let $(U_m, U_\sigma) \sim Q^{(n)}$. Therefore

$$A_m^{(n)} = \mathbb{E}_{Q^{(n)}}\Big(U_m^2\Big); \quad A_\sigma^{(n)} = \mathbb{E}_{Q^{(n)}}\Big(U_\sigma^2\Big).$$

The conditions on p_ε and U_σ are sufficient to interchange differentiation and expectation operators as follows (Aliprantis and Burkinshaw, 1990, Thm. 20.4).

$$\begin{aligned}
\rho_{\hat{\varepsilon}}(\epsilon) &= \frac{\frac{\partial}{\partial\epsilon}\mathbb{E}_{Q^{(n)}}\Big[p_\varepsilon(\epsilon + U_\sigma\epsilon + U_m)\Big]}{\mathbb{E}_{Q^{(n)}}\Big[p_\varepsilon(\epsilon + U_\sigma\epsilon + U_m)\Big]} \\
&= \frac{\mathbb{E}_{Q^{(n)}}\Big[(1+U_\sigma)p'_\varepsilon(\epsilon + U_\sigma\epsilon + U_m)\Big]}{\mathbb{E}_{Q^{(n)}}\Big[p_\varepsilon(\epsilon + U_\sigma\epsilon + U_m)\Big]}.
\end{aligned}$$

We may decompose the approximation error as follows.

$$\begin{aligned}
|\rho_{\hat{\varepsilon}}(\epsilon) - \rho_\varepsilon(\epsilon)| &= \left|\frac{\mathbb{E}_{Q^{(n)}}[(1+U_\sigma)p'_\varepsilon(\epsilon + U_\sigma\epsilon + U_m)]}{\mathbb{E}_{Q^{(n)}}[p_\varepsilon(\epsilon + U_\sigma\epsilon + U_m)]} - \rho_\varepsilon(\epsilon)\right| \\
&= \left|\frac{\mathbb{E}_{Q^{(n)}}[(1+U_\sigma)\{\rho_\varepsilon(\epsilon + U_\sigma\epsilon + U_m) - \rho_\varepsilon(\epsilon)\}\; p_\varepsilon(\epsilon + U_\sigma\epsilon + U_m)]}{\mathbb{E}_{Q^{(n)}}[p_\varepsilon(\epsilon + U_\sigma\epsilon + U_m)]}\right. \\
&\quad \left. + \frac{\mathbb{E}_{Q^{(n)}}[U_\sigma\; p_\varepsilon(\epsilon + U_\sigma\epsilon + U_m)]}{\mathbb{E}_{Q^{(n)}}[p_\varepsilon(\epsilon + U_\sigma\epsilon + U_m)]}\rho_\varepsilon(\epsilon)\right| \\
&\le C_\rho \frac{\mathbb{E}_{Q^{(n)}}[|(1+U_\sigma)(U_\sigma\epsilon + U_m)|\; p_\varepsilon(\epsilon + U_\sigma\epsilon + U_m)]}{\mathbb{E}_{Q^{(n)}}[p_\varepsilon(\epsilon + U_\sigma\epsilon + U_m)]} \\
&\quad + \frac{\mathbb{E}_{Q^{(n)}}[|U_\sigma|\; p_\varepsilon(\epsilon + U_\sigma\epsilon + U_m)]}{\mathbb{E}_{Q^{(n)}}[p_\varepsilon(\epsilon + U_\sigma\epsilon + U_m)]}|\rho_\varepsilon(\epsilon)| \\
&\le \left\{C_\rho\Big(\mathbb{E}_{Q^{(n)}}\Big[(1+U_\sigma)^2(U_\sigma\epsilon + U_m)^2\Big]\Big)^{1/2} + |\rho_\varepsilon(\epsilon)|\Big(\mathbb{E}_{Q^{(n)}}\Big(U_\sigma^2\Big)\Big)^{1/2}\right\} \\
&\quad \cdot \frac{\Big(\mathbb{E}_{Q^{(n)}}\Big[p_\varepsilon^2(\epsilon + U_\sigma\epsilon + U_m)\Big]\Big)^{1/2}}{\mathbb{E}_{Q^{(n)}}[p_\varepsilon(\epsilon + U_\sigma\epsilon + U_m)]} \\
&=: R_1(\epsilon)R_2(\epsilon).
\end{aligned}$$

The first inequality uses the Lipschitz property of ρ_ε. The second applies the Cauchy–Schwarz inequality.

We will show that the first term in the product is dominated by $\mathbb{E}_{Q^{(n)}}\left(U_\sigma^2\right)+\mathbb{E}_{Q^{(n)}}\left(U_m^2\right)$, and that the second term is bounded. Indeed,

$$\begin{aligned}
R_1^2(\epsilon) &\le 2C_\rho^2\mathbb{E}_{Q^{(n)}}\left[(1+U_\sigma)^2(U_\sigma\epsilon+U_m)^2\right]+2\rho_\varepsilon^2(\epsilon)\mathbb{E}_{Q^{(n)}}\left(U_\sigma^2\right)\\
&\le 2C_\rho^2(1+C_\sigma)^2\mathbb{E}_{Q^{(n)}}\left[(U_\sigma\epsilon+U_m)^2\right]+2\rho_\varepsilon^2(\epsilon)\mathbb{E}_{Q^{(n)}}\left(U_\sigma^2\right)\\
&\le 4C_\rho^2(1+C_\sigma)^2\mathbb{E}_{Q^{(n)}}\left(U_\sigma^2\right)\epsilon^2+4C_\rho^2(1+C_\sigma)^2\mathbb{E}_{Q^{(n)}}\left(U_m^2\right)+2\rho_\varepsilon^2(\epsilon)\mathbb{E}_{Q^{(n)}}\left(U_\sigma^2\right).
\end{aligned}$$

The first and third inequalities are $(a+b)^2\le 2(a^2+b^2)$, and the second is the almost sure bound $|U_\sigma|\le C_\sigma$.

For any $\epsilon\in\mathbb{R}$ such that $p_\varepsilon(\epsilon)>0$, and for any constant $c_1>0$ (to be chosen later),

$$\begin{aligned}
R_2^2(\epsilon) &\le \left(\frac{\sup_{|u_m|\le C_m\,,\,|u_\sigma|\le C_\sigma} p_\varepsilon(\epsilon+u_\sigma\epsilon+u_m)/p_\varepsilon(\epsilon)}{\inf_{|u_m|\le C_m\,,\,|u_\sigma|\le C_\sigma} p_\varepsilon(\epsilon+u_\sigma\epsilon+u_m)/p_\varepsilon(\epsilon)}\right)^2\\
&\le \exp\left\{4|C_m+C_\sigma\epsilon|\;|\rho_\varepsilon(\epsilon)|+2C_\rho(C_m+C_\sigma\epsilon)^2\right\}\\
&\le \exp\left\{\frac{\rho_\varepsilon^2(\epsilon)}{C_\rho c_1}+(4c_1+2)C_\rho(C_m+C_\sigma\epsilon)^2\right\}.
\end{aligned}$$

The first line is a supremum bound for the ratio of expectations, the second is the application of Lemma 26, and the third uses that for all $c>0$,

$$0\le\left(\frac{a}{\sqrt{c}}-2\sqrt{c}b\right)^2 \implies 4ab=2\left(\frac{a}{\sqrt{c}}\right)(2\sqrt{c}b)\le\frac{a^2}{c}+4cb^2.$$

Using the above and Hölder's inequality, we have that for any $c_2>0$ (to be chosen later),

$$\begin{aligned}
&\mathbb{E}_P\left[\left\{\rho_\varepsilon(\varepsilon_P)-\rho_{\hat\varepsilon}(\varepsilon_P)\right\}^2\;\middle|\;D^{(n)}\right]\\
&\qquad\le \mathbb{E}_P\left[R_1^2(\varepsilon_P)R_2^2(\varepsilon_P)\;\middle|\;D^{(n)}\right]\\
&\qquad\le \left(\mathbb{E}_P\left[R_1^{\frac{2(1+c_2)}{c_2}}(\varepsilon_P)\;\middle|\;D^{(n)}\right]\right)^{\frac{c_2}{1+c_2}}\left(\mathbb{E}_P\left[R_2^{2(1+c_2)}(\varepsilon_P)\;\middle|\;D^{(n)}\right]\right)^{\frac{1}{1+c_2}}.
\end{aligned}$$

By the triangle inequality (for the $L_{(1+c_2)/c_2}(P)$ norm),

$$\begin{aligned}
\left(\mathbb{E}_P\left[R_1^{\frac{2(1+c_2)}{c_2}}(\varepsilon_P)\;\middle|\;D^{(n)}\right]\right)^{\frac{c_2}{1+c_2}} &\le 4C_\rho^2(1+C_\sigma)^2\mathbb{E}_{Q^{(n)}}\left(U_\sigma^2\right)\left(\mathbb{E}_P\left[\varepsilon_P^{\frac{2(1+c_2)}{c_2}}\right]\right)^{\frac{c_2}{1+c_2}}\\
&\quad+4C_\rho^2(1+C_\sigma)^2\mathbb{E}_{Q^{(n)}}\left(U_m^2\right)\\
&\quad+2\mathbb{E}_{Q^{(n)}}\left(U_\sigma^2\right)\left(\mathbb{E}_P\left[\rho_\varepsilon^{\frac{2(1+c_2)}{c_2}}(\varepsilon_P)\right]\right)^{\frac{c_2}{1+c_2}}.
\end{aligned}$$

By Hölder's inequality, for any $c_3 > 0$ (to be chosen later),

$$\begin{aligned}
&\mathbb{E}_P\Big[R_2^{2(1+c_2)}(\varepsilon_P) \;\Big|\; D^{(n)}\Big] \\
&\qquad \le \mathbb{E}_P\left[\exp\left\{\frac{(1+c_2)\rho_\varepsilon^2(\varepsilon_P)}{C_\rho c_1} + (1+c_2)(4c_1+2)C_\rho(C_m + C_\sigma\varepsilon_P)^2\right\}\right] \\
&\qquad \le \left(\mathbb{E}_P\left[\exp\left\{\frac{(1+c_3)(1+c_2)\rho_\varepsilon^2(\varepsilon_P)}{C_\rho c_1}\right\}\right]\right)^{\frac{1}{1+c_3}} \\
&\qquad\quad \cdot \left(\mathbb{E}_P\left[\exp\left\{\frac{(1+c_3)(1+c_2)(4c_1+2)C_\rho(C_m + C_\sigma\varepsilon_P)^2}{c_3}\right\}\right]\right)^{\frac{c_3}{1+c_3}} \\
&\qquad \le \left(\mathbb{E}_P\left[\exp\left\{\frac{(1+c_3)(1+c_2)\rho_\varepsilon^2(\varepsilon_P)}{C_\rho c_1}\right\}\right]\right)^{\frac{1}{1+c_3}} \\
&\qquad\quad \cdot \left(\mathbb{E}_P\left[\exp\left\{\frac{2(1+c_3)(1+c_2)(4c_1+2)C_\rho(C_m^2 + C_\sigma^2\varepsilon_P^2)}{c_3}\right\}\right]\right)^{\frac{c_3}{1+c_3}} \\
&\qquad =: \left(\mathbb{E}_P\left[\exp\left\{\frac{\lambda_\rho\rho_\varepsilon^2(\varepsilon_P)}{2\left(\sqrt{2C_\rho}\right)^2}\right\}\right]\right)^{\frac{1}{1+c_3}} \exp\Big\{2(1+c_2)(4c_1+2)C_\rho C_m^2)\Big\} \\
&\qquad\quad \cdot \left(\mathbb{E}_P\left[\exp\left\{\frac{\lambda_\varepsilon\varepsilon_P^2}{2C_\varepsilon^2}\right\}\right]\right)^{\frac{c_3}{1+c_3}};
\end{aligned}$$

the final inequality uses $(a+b)^2 \le 2(a^2+b^2)$ and the monotonicity of the exponential function; and in the final equality the newly defined quantities are

$$\begin{aligned}
\lambda_\rho &:= \frac{4(1+c_3)(1+c_2)}{c_1}, \\
\lambda_\varepsilon &:= \frac{4C_\varepsilon^2(1+c_3)(1+c_2)(4c_1+2)C_\rho C_\sigma^2}{c_3}.
\end{aligned}$$

To apply Lemma 17, we must choose $c_1, c_2, c_3 > 0$ such that both $\lambda_\rho, \lambda_\varepsilon \in [0,1)$. The choice

$$(c_1, c_2, c_3) = \left(9, \frac{1}{16}, 1\right)$$

suffices for

$$\lambda_\rho = \frac{17}{18}, \quad \lambda_\varepsilon \le 1 - \frac{1}{18^2}.$$

Hence

$$\mathbb{E}_P\Big[\Big\{\rho_\varepsilon(\varepsilon_P)-\rho_{\hat\varepsilon}(\varepsilon_P)\Big\}^2\;\Big|\;D^{(n)}\Big]\le\Bigg\{4C_\rho^2(1+C_\sigma)^2\mathbb{E}_{Q^{(n)}}\Big(U_\sigma^2\Big)\Big(\mathbb{E}_P\Big(\varepsilon_P^{34}\Big)\Big)^{\frac{1}{17}}$$
$$+4C_\rho^2(1+C_\sigma)^2\mathbb{E}_{Q^{(n)}}\Big(U_m^2\Big)$$
$$+2\mathbb{E}_{Q^{(n)}}\Big(U_\sigma^2\Big)\Big(\mathbb{E}_P\Big[\rho_\varepsilon^{34}(\varepsilon_P)\Big]\Big)^{\frac{1}{17}}\Bigg\}$$
$$\cdot\left(\mathbb{E}_P\left[\exp\left\{\frac{\lambda_\rho\rho_\varepsilon^2(\varepsilon_P)}{2\Big(\sqrt{2C_\rho}\Big)^2}\right\}\right]\right)^{\frac{8}{17}}\exp\Big(76C_\rho C_m^2\Big)$$
$$\cdot\left(\mathbb{E}_P\left[\exp\left\{\frac{\lambda_\varepsilon\varepsilon_P^2}{2C_\varepsilon^2}\right\}\right]\right)^{\frac{8}{17}}.$$

Finally, by Theorem 5 and Lemmas 16 and 17 we have the bounds

$$\Big(\mathbb{E}_P\Big[\rho_\varepsilon^{34}(\varepsilon_P)\Big]\Big)^{\frac{1}{17}}\le C_\rho(33!!)^{\frac{1}{17}}<13C_\rho;$$
$$\Big(\mathbb{E}_P\Big(\varepsilon_P^{34}\Big)\Big)^{\frac{1}{17}}\le\Big(34\cdot2^{17}C_\varepsilon^{34}\Gamma(17)\Big)^{\frac{1}{17}}=2C_\varepsilon^2(34\cdot16!)^{\frac{1}{17}}<15C_\varepsilon^2;$$
$$\left(\mathbb{E}_P\left[\exp\left\{\frac{\lambda_\rho\rho_\varepsilon^2(\varepsilon_P)}{2\Big(\sqrt{2C_\rho}\Big)^2}\right\}\right]\right)^{\frac{8}{17}}\le\left(\frac{1}{\sqrt{1-\lambda_\rho}}\right)^{\frac{8}{17}}<4;$$
$$\left(\mathbb{E}_P\left[\exp\left\{\frac{\lambda_\varepsilon\varepsilon_P^2}{2C_\varepsilon^2}\right\}\right]\right)^{\frac{8}{17}}\le\left(\frac{1}{\sqrt{1-\lambda_\varepsilon}}\right)^{\frac{8}{17}}<2.$$

This gives the final bound

$$\begin{aligned}\mathbb{E}_P\Big[\Big\{\rho_\varepsilon(\varepsilon_P)-\rho_{\hat\varepsilon}(\varepsilon_P)\Big\}^2\;\Big|\;D^{(n)}\Big]&\\ \le\Big\{480C_\rho^2(1+C_\sigma)^2C_\varepsilon^2+208C_\rho\Big\}\exp\Big(76C_\rho C_m^2\Big)\;\mathbb{E}_{Q^{(n)}}\Big(U_\sigma^2\Big)&\\ +32C_\rho^2(1+C_\sigma)^2\exp\Big(76C_\rho C_m^2\Big)\;\mathbb{E}_{Q^{(n)}}\Big(U_m^2\Big)&\\ =\Big\{480C_\rho^2(1+C_\sigma)^2C_\varepsilon^2+208C_\rho\Big\}\exp\Big(76C_\rho C_m^2\Big)\;A_\sigma^{(n)}&\\ +32C_\rho^2(1+C_\sigma)^2\exp\Big(76C_\rho C_m^2\Big)\;A_m^{(n)}.&\end{aligned}$$

□

Lemma 16. *Let X be mean-zero and sub-Gaussian with parameter $\sigma>0$. Then for any $p>0$,*

$$\mathbb{E}\Big(|X|^p\Big)\le p2^{\frac{p}{2}}\sigma^p\Gamma\Big(\frac{p}{2}\Big),$$

where $\Gamma(x) = \int_0^\infty u^{x-1}\exp(-u)\,du$ *is the gamma function.*

Proof. By the Chernoff bound we have that

$$\mathbb{P}(|X| > t) \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right).$$

We are now able to make use of the tail probability formula for expectation.

$$\begin{aligned}
\mathbb{E}\Big(|X|^p\Big) &= \int_0^\infty \mathbb{P}\Big(|X|^p > s\Big)\,ds \\
&= \int_0^\infty \mathbb{P}\Big(|X| > s^{-p}\Big)\,ds \\
&= \int_0^\infty pt^{p-1}\mathbb{P}\Big(|X| > t\Big)\,dt \\
&\leq \int_0^\infty pt^{p-1}2\exp\left(-\frac{t^2}{2\sigma^2}\right)dt \\
&= \int_0^\infty \sigma^2 p(2\sigma^2 u)^{\frac{p}{2}-1}2\exp(-u)\,du \\
&= p2^{\frac{p}{2}}\sigma^p\int_0^\infty u^{\frac{p}{2}-1}\exp(-u)\,du.
\end{aligned}$$

The third line makes the substitution $t = s^{-p}$, the fifth $u = t^2/2\sigma^2$. Recalling the definition of the Gamma function, we are done. □

Lemma 17 (Wainwright (2019) Thm. 2.6)**.** *Let* X *be mean-zero and sub-Gaussian with parameter* $\sigma > 0$*. Then*

$$\mathbb{E}\left[\exp\left(\frac{\lambda X^2}{2\sigma^2}\right)\right] \leq \frac{1}{\sqrt{1-\lambda}} \text{ for all } \lambda \in [0,1).$$

Lemma 18. *Let* P *be such that* p_ε *is twice differentiable on* $\mathbb{R}$*, with*

$$\sup_{\epsilon\in\mathbb{R}}|\partial_\epsilon^2 \log p_\varepsilon(\epsilon)| = \sup_{\epsilon\in\mathbb{R}}|\rho_\varepsilon'(\epsilon)| \leq C_\rho,$$

p_ε' *and* p_ε'' *are both bounded, and* $\Big(\mathbb{E}_P(\varepsilon_P^8)\Big)^{\frac{1}{8}} = C_\varepsilon < \infty$*. Further assume that* $D^{(n)}$ *is such that* $\sup_{z\in\mathcal{Z}}|u_m^{(n)}(z)| \leq C_m$ *and* $\sup_{z\in\mathcal{Z}}|u_\sigma^{(n)}(z)| \leq C_\sigma$ *for almost every* $z \in \mathcal{Z}$*. Then there exists a constant* C*, depending only on* $C_\rho, C_m, C_\sigma, C_\varepsilon$*, such that*

$$\mathbb{E}_P\left[\Big\{\rho_{\hat{\varepsilon}}(\varepsilon_P) - \rho_{\hat{\varepsilon}}\big(\hat{\varepsilon}^{(n)}\big)\Big\}^2 \,\Big|\, D^{(n)}\right] \leq C\Big(A_m^{(n)} + A_\sigma^{(n)}\Big).$$

Proof. For ease of notation, write $Q^{(n)}$ for the distribution of $\Big(u_m^{(n)}(Z), u_\sigma^{(n)}(Z)\Big)$ conditionally on $D^{(n)}$, and let $(U_m, U_\sigma) \sim Q^{(n)}$. Therefore

$$A_m^{(n)} = \mathbb{E}_{Q^{(n)}}\Big(U_m^2\Big); \quad A_\sigma^{(n)} = \mathbb{E}_{Q^{(n)}}\Big(U_\sigma^2\Big).$$

The proof proceeds by first bounding the derivative of $\rho_{\hat{\varepsilon}}$. The conditions on p_ε and U_σ are sufficient to interchange differentiation and expectation operators as follows (Aliprantis and Burkinshaw, 1990, Thm. 20.4).

$$\begin{aligned}\rho_{\hat{\varepsilon}}(\epsilon) &= \frac{\frac{\partial}{\partial\epsilon}\mathbb{E}_{Q^{(n)}}\Big[p_\varepsilon(\epsilon+U_\sigma\epsilon+U_m)\Big]}{\mathbb{E}_{Q^{(n)}}\Big[p_\varepsilon(\epsilon+U_\sigma\epsilon+U_m)\Big]}\\ &= \frac{\mathbb{E}_{Q^{(n)}}\Big[(1+U_\sigma)p'_\varepsilon(\epsilon+U_\sigma\epsilon+U_m)\Big]}{\mathbb{E}_{Q^{(n)}}\Big[p_\varepsilon(\epsilon+U_\sigma\epsilon+U_m)\Big]},\end{aligned}$$

and further,

$$\begin{aligned}\rho'_{\hat{\varepsilon}}(\epsilon) &= \frac{\partial}{\partial\epsilon}\frac{\mathbb{E}_{Q^{(n)}}\Big[(1+U_\sigma)p'_\varepsilon(\epsilon+U_\sigma\epsilon+U_m)\Big]}{\mathbb{E}_{Q^{(n)}}\Big[p_\varepsilon(\epsilon+U_\sigma\epsilon+U_m)\Big]}\\ &= \frac{\frac{\partial}{\partial\epsilon}\mathbb{E}_{Q^{(n)}}\Big[(1+U_\sigma)p'_\varepsilon(\epsilon+U_\sigma\epsilon+U_m)\Big]}{\mathbb{E}_{Q^{(n)}}\Big[p_\varepsilon(\epsilon+U_\sigma\epsilon+U_m)\Big]} - \rho^2_{\hat{\varepsilon}}(\epsilon)\\ &= \frac{\mathbb{E}_{Q^{(n)}}\Big[(1+U_\sigma)^2p''_\varepsilon(\epsilon+U_\sigma\epsilon+U_m)\Big]}{\mathbb{E}_{Q^{(n)}}\Big[p_\varepsilon(\epsilon+U_\sigma\epsilon+U_m)\Big]}\\ &\quad - \frac{\Big(\mathbb{E}_{Q^{(n)}}\Big[(1+U_\sigma)p'_\varepsilon(\epsilon+U_\sigma\epsilon+U_m)\Big]\Big)^2}{\Big(\mathbb{E}_{Q^{(n)}}\Big[p_\varepsilon(\epsilon+U_\sigma\epsilon+U_m)\Big]\Big)^2}\\ &= \frac{\mathbb{E}_{Q^{(n)}}\Big[(1+U_\sigma)^2\Big\{\rho'_\varepsilon(\epsilon+U_\sigma\epsilon+U_m)+\rho^2_\varepsilon(\epsilon+U_\sigma\epsilon+U_m)\Big\}p_\varepsilon(\epsilon+U_\sigma\epsilon+U_m)\Big]}{\mathbb{E}_{Q^{(n)}}\Big[p_\varepsilon(\epsilon+U_\sigma\epsilon+U_m)\Big]}\\ &\quad - \frac{\Big(\mathbb{E}_{Q^{(n)}}\Big[(1+U_\sigma)\rho_\varepsilon(\epsilon+U_\sigma\epsilon+U_m)p_\varepsilon(\epsilon+U_\sigma\epsilon+U_m)\Big]\Big)^2}{\Big(\mathbb{E}_{Q^{(n)}}\Big[p_\varepsilon(\epsilon+U_\sigma\epsilon+U_m)\Big]\Big)^2}\end{aligned}$$

In the third line we have made use of the identities

$$\begin{aligned}p'_\varepsilon(\epsilon) &= \rho_\varepsilon(\epsilon)p_\varepsilon(\epsilon);\\ p''_\varepsilon(\epsilon) &= \Big\{\rho'_\varepsilon(\epsilon)+\rho^2_\varepsilon(\epsilon)\Big\}p_\varepsilon(\epsilon).\end{aligned}$$

We now apply both the triangle and Hölder inequalities to deduce

$$\begin{aligned}\Big|\rho'_{\hat{\varepsilon}}(\epsilon)\Big| &\leq \sup_{|u_m|\leq C_m,|u_\sigma|\leq C_\sigma}(1+u_\sigma)^2\Big\{\Big|\rho'_\varepsilon(\epsilon+u_\sigma\epsilon+u_m)\Big|+2\rho^2_\varepsilon(\epsilon+u_\sigma\epsilon+u_m)\Big\}\\ &\leq (1+C_\sigma)^2\left\{C_\rho+2\sup_{|u_m|\leq C_m,|u_\sigma|\leq C_\sigma}\rho^2_\varepsilon(\epsilon+u_\sigma\epsilon+u_m)\right\}.\end{aligned}$$

Now we apply a Taylor expansion as follows.

$$
\begin{aligned}
&\mathbb{E}_P\Big[\Big\{\rho_{\hat{\varepsilon}}(\varepsilon_P)-\rho_{\hat{\varepsilon}}(\hat{\varepsilon})\Big\}^2 \;\Big|\; D^{(n)}\Big] \\
&\quad= \mathbb{E}_P\Big(\mathbb{E}_{Q^{(n)}}\Big[\Big\{\rho_{\hat{\varepsilon}}(\varepsilon_P)-\rho_{\hat{\varepsilon}}(\varepsilon_P+U_\sigma\varepsilon_P+U_m)\Big\}^2 \;\Big|\; \varepsilon_P,\ D^{(n)}\Big] \;\Big|\; D^{(n)}\Big) \\
&\quad\le \mathbb{E}_P\Bigg(\mathbb{E}_{Q^{(n)}}\Big[(U_\sigma\varepsilon_P+U_m)^2 \;\Big|\; \varepsilon_P\Big] \\
&\qquad\qquad \cdot\Big\{\sup_{|u_m|\le C_m,|u_\sigma|\le C_\sigma}\rho'_{\hat{\varepsilon}}(\varepsilon_P+u_\sigma\varepsilon_P+u_m)\Big\}^2 \;\Big|\; D^{(n)}\Bigg) \\
&\quad\le (1+C_\sigma)^2\mathbb{E}_P\Bigg(\mathbb{E}_{Q^{(n)}}\Big[(U_\sigma\varepsilon_P+U_m)^2 \;\Big|\; \varepsilon_P\Big] \\
&\qquad\qquad \cdot\Big\{C_\rho+2\sup_{|\eta_m|\le 2C_m+C_\sigma C_m,|\eta_\sigma|\le 2C_\sigma+C_\sigma^2}\rho_\varepsilon^2(\varepsilon_P+\eta_\sigma\varepsilon_P+\eta_m)\Big\}^2 \;\Big|\; D^{(n)}\Bigg) \\
&\quad\le 2(1+C_\sigma)^2\mathbb{E}_P\Bigg[\Big\{\mathbb{E}_{Q^{(n)}}\Big(U_\sigma^2\Big)\varepsilon_P^2+\mathbb{E}_{Q^{(n)}}\Big(U_m^2\Big)\Big\} \\
&\qquad\qquad \cdot\Big\{C_\rho+2\sup_{|\eta_m|\le 2C_m+C_\sigma C_m,|\eta_\sigma|\le 2C_\sigma+C_\sigma^2}\rho_\varepsilon^2(\varepsilon_P+\eta_\sigma\varepsilon_P+\eta_m)\Big\}^2 \;\Big|\; D^{(n)}\Bigg] \\
&\quad\le 2(1+C_\sigma)^2\Bigg(\mathbb{E}_P\Big[\Big\{\mathbb{E}_{Q^{(n)}}\Big(U_\sigma^2\Big)\varepsilon_P^2+\mathbb{E}_{Q^{(n)}}\Big(U_m^2\Big)\Big\}^2 \;\Big|\; D^{(n)}\Big]\Bigg)^{\frac{1}{2}} \\
&\qquad\cdot\Bigg(\mathbb{E}_P\Bigg[\Big\{C_\rho+2\sup_{|\eta_m|\le 2C_m+C_\sigma C_m,|\eta_\sigma|\le 2C_\sigma+C_\sigma^2}\rho_\varepsilon^2(\varepsilon_P+\eta_\sigma\varepsilon_P+\eta_m)\Big\}^4\Bigg]\Bigg)^{\frac{1}{2}} \\
&\quad\le 2(1+C_\sigma)^2\Big[\mathbb{E}_{Q^{(n)}}\Big(U_\sigma^2\Big)\Big\{\mathbb{E}_P\Big(\varepsilon_P^4\Big)\Big\}^{\frac{1}{2}}+\mathbb{E}_{Q^{(n)}}\Big(U_m^2\Big)\Big] \\
&\qquad\cdot\Bigg(\mathbb{E}_P\Bigg[\Big\{C_\rho+2\sup_{|\eta_m|\le 2C_m+C_\sigma C_m,|\eta_\sigma|\le 2C_\sigma+C_\sigma^2}\rho_\varepsilon^2(\varepsilon_P+\eta_\sigma\varepsilon_P+\eta_m)\Big\}^4\Bigg]\Bigg)^{\frac{1}{2}} \\
&\quad\le 2(1+C_\sigma)^2\Big[\mathbb{E}_{Q^{(n)}}\Big(U_\sigma^2\Big)\Big\{\mathbb{E}_P\Big(\varepsilon_P^4\Big)\Big\}^{\frac{1}{2}}+\mathbb{E}_{Q^{(n)}}\Big(U_m^2\Big)\Big] \\
&\qquad\cdot\Bigg(\mathbb{E}_P\Bigg[\Big\{C_\rho+2\Big(|\rho_\varepsilon(\varepsilon_P)|+C_\rho(2C_m+C_\sigma C_m)+C_\rho(2C_\sigma+C_\sigma^2)\varepsilon_P\Big)^2\Big\}^4\Bigg]\Bigg)^{\frac{1}{2}} \\
&\quad\le 2(1+C_\sigma)^2\Big[\mathbb{E}_{Q^{(n)}}\Big(U_\sigma^2\Big)\Big\{\mathbb{E}_P\Big(\varepsilon_P^4\Big)\Big\}^{\frac{1}{2}}+\mathbb{E}_{Q^{(n)}}\Big(U_m^2\Big)\Big] \\
&\qquad\cdot\Bigg(\mathbb{E}_P\Big[\Big\{C_\rho+6\rho_\varepsilon^2(\varepsilon_P)+6C_\rho^2(2C_m+C_\sigma C_m)^2+6C_\rho^2(2C_\sigma+C_\sigma^2)^2\varepsilon_P^2\Big\}^4\Big]\Bigg)^{\frac{1}{2}} \\
&\quad\le 2(1+C_\sigma)^2\Big[\mathbb{E}_{Q^{(n)}}\Big(U_\sigma^2\Big)\Big\{\mathbb{E}_P\Big(\varepsilon_P^4\Big)\Big\}^{\frac{1}{2}}+\mathbb{E}_{Q^{(n)}}\Big(U_m^2\Big)\Big] \\
&\qquad\cdot\Bigg(C_\rho+6\Big(\mathbb{E}_P\Big[\rho_\varepsilon^8(\varepsilon_P)\Big]\Big)^{\frac{1}{4}}+6C_\rho^2(2C_m+C_\sigma C_m)^2+6C_\rho^2(2C_\sigma+C_\sigma^2)^2\Big(\mathbb{E}_P\Big[\varepsilon_P^8\Big]\Big)^{\frac{1}{4}}\Bigg)^2
\end{aligned}
$$

where we have made use of the triangle inequalities for $L_2(P)$ and $L_4(P)$, and also the inequalities, $\{(a+b)/2\}^2 \leq (a^2+b^2)/2$ and $\{(a+b+c)/3\}^2 \leq (a^2+b^2+c^2)/3$.

Finally, by Theorem 5 and the assumed eighth moment of ε_P, we have that

$$6\Big(\mathbb{E}_P\Big[\rho_\varepsilon^8(\varepsilon_P)\Big]\Big)^{\frac{1}{4}} \leq 6 \cdot 105^{\frac{1}{4}} C_\rho < 20 C_\rho;$$
$$\Big\{\mathbb{E}_P\Big(\varepsilon_P^4\Big)\Big\}^{\frac{1}{2}} \leq \Big\{\mathbb{E}_P\Big(\varepsilon_P^8\Big)\Big\}^{\frac{1}{4}} = C_\varepsilon^2.$$

Hence

$$\mathbb{E}_P\Big[\Big\{\rho_{\hat{\varepsilon}}(\varepsilon_P) - \rho_{\hat{\varepsilon}}(\hat{\varepsilon})\Big\}^2 \;\Big|\; D^{(n)}\Big] \leq 2(1+C_\sigma)^2 C_\rho^2 \Big\{21 + C_\rho(2+C_\sigma)^2(C_m^2 + C_\sigma^2 C_\varepsilon^2)\Big\}^2 \cdot \Big\{4C_\varepsilon^2 \mathbb{E}_{Q^{(n)}}\Big(U_\sigma^2\Big) + \mathbb{E}_{Q^{(n)}}\Big(U_m^2\Big)\Big\}.$$

□

Proof of Theorem 8

Proof. The assumption on $u_m^{(n)}$ means that for any $\epsilon > 0$ we can find N, C_m such that for any $n \geq N$, with uniform probability at least $1-\epsilon$, the data $D^{(n)}$ is such that

$$\sup_z \Big|u_m^{(n)}(z)\Big| \leq C_m. \tag{2.41}$$

It suffices to show that under this event, we can find a uniform constant C (not depending on P or n) such that

$$A_\rho^{(n)} \leq C\Big(A_m^{(n)} + A_{\hat{\varepsilon}}^{(n)}\Big).$$

Fix $P \in \mathcal{P}$ and $D^{(n)}$ such that (2.41) holds. We decompose $A_\rho^{(n)}$ so as to consider the various sources of error separately.

$$\begin{aligned}
A_\rho^{(n)} &= \mathbb{E}_P\Big[\Big\{\rho_\varepsilon(\varepsilon_P) - \hat{\rho}_{\hat{\varepsilon}}^{(n)}\Big(\hat{\varepsilon}^{(n)}\Big)\Big\}^2 \;\Big|\; D^{(n)}\Big] \\
&= \mathbb{E}_P\Big[\Big\{\rho_\varepsilon(\varepsilon_P) - \rho_{\hat{\varepsilon}}(\varepsilon_P) + \rho_{\hat{\varepsilon}}(\varepsilon_P) - \rho_{\hat{\varepsilon}}(\hat{\varepsilon}^{(n)}) + \rho_{\hat{\varepsilon}}(\hat{\varepsilon}^{(n)}) - \hat{\rho}_{\hat{\varepsilon}}^{(n)}\Big(\hat{\varepsilon}^{(n)}\Big)\Big\}^2 \;\Big|\; D^{(n)}\Big] \\
&= 3\mathbb{E}_P\Big[\Big\{\rho_\varepsilon(\varepsilon_P) - \rho_{\hat{\varepsilon}}(\varepsilon_P)\Big\}^2 \;\Big|\; D^{(n)}\Big] \\
&\quad + 3\mathbb{E}_P\Big[\Big\{\rho_{\hat{\varepsilon}}(\varepsilon_P) - \rho_{\hat{\varepsilon}}(\hat{\varepsilon}^{(n)})\Big\}^2 \;\Big|\; D^{(n)}\Big] \\
&\quad + 3\mathbb{E}_P\Big[\Big\{\rho_{\hat{\varepsilon}}(\hat{\varepsilon}^{(n)}) - \hat{\rho}_{\hat{\varepsilon}}^{(n)}\Big(\hat{\varepsilon}^{(n)}\Big)\Big\}^2 \;\Big|\; D^{(n)}\Big],
\end{aligned} \tag{2.42}$$

where the final inequality is $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$.

We consider the expectations in (2.42) separately. Lemma 19 applies to the first term. Lemma 20 applies to the second term. The third term is equal to $A_{\hat{\varepsilon}}^{(n)}$ by definition. This completes the proof. □

Lemma 19. *Let P be such that p_ε is twice differentiable on $\mathbb{R}$, with*

$$\sup_{\epsilon\in\mathbb{R}} |\partial_\epsilon^2 \log p_\varepsilon(\epsilon)| = \sup_{\epsilon\in\mathbb{R}} |\rho'_\varepsilon(\epsilon)| \leq C_\rho$$

and p'_ε bounded. Further assume that $D^{(n)}$ is such that $\sup_{z\in\mathcal{Z}} |u_m^{(n)}(z)| \leq C_m$. Then there exists a constant C, depending only on C_ρ, C_m, such that

$$\mathbb{E}_P\Big[\Big\{\rho_\varepsilon(\varepsilon_P) - \rho_{\hat{\varepsilon}}(\varepsilon_P)\Big\}^2 \;\Big|\; D^{(n)}\Big] \leq C A_m^{(n)}.$$

Proof. For ease of notation, write $Q^{(n)}$ for the distribution of $u_m^{(n)}(Z)$ conditionally on $D^{(n)}$, and let $U \sim Q^{(n)}$. Therefore

$$A_m^{(n)} = \mathbb{E}_{Q^{(n)}}\Big(U^2\Big).$$

The condition on p_ε is sufficient to interchange differentiation and expectation operators as follows (Aliprantis and Burkinshaw, 1990, Thm. 20.4).

$$\begin{aligned}
\rho_{\hat{\varepsilon}}(\epsilon) &= \frac{\frac{\partial}{\partial\epsilon}\mathbb{E}_{Q^{(n)}}[p_\varepsilon(\epsilon+U)]}{\mathbb{E}_{Q^{(n)}}[p_\varepsilon(\epsilon+U)]} \\
&= \frac{\mathbb{E}_{Q^{(n)}}\Big[p'_\varepsilon(\epsilon+U)\Big]}{\mathbb{E}_{Q^{(n)}}[p_\varepsilon(\epsilon+U)]}.
\end{aligned}$$

We may decompose the approximation error as follows.

$$\begin{aligned}
|\rho_{\hat{\varepsilon}}(\epsilon) - \rho_\varepsilon(\epsilon)| &= \left|\frac{\mathbb{E}_{Q^{(n)}}\Big[p'_\varepsilon(\epsilon+U)\Big]}{\mathbb{E}_{Q^{(n)}}[p_\varepsilon(\epsilon+U)]} - \rho_\varepsilon(\epsilon)\right| \\
&= \left|\frac{\mathbb{E}_{Q^{(n)}}[\{\rho_\varepsilon(\epsilon+U) - \rho_\varepsilon(\epsilon)\}\, p_\varepsilon(\epsilon+U)]}{\mathbb{E}_{Q^{(n)}}[p_\varepsilon(\epsilon+U)]}\right| \\
&\leq C_\rho \frac{\mathbb{E}_{Q^{(n)}}[|U| p_\varepsilon(\epsilon+U)]}{\mathbb{E}_{Q^{(n)}}[p_\varepsilon(\epsilon+U)]} \\
&\leq C_\rho \Big(\mathbb{E}_{Q^{(n)}}\Big(U^2\Big)\Big)^{1/2} \frac{\Big(\mathbb{E}_{Q^{(n)}}\Big[p_\varepsilon^2(\epsilon+U)\Big]\Big)^{1/2}}{\mathbb{E}_{Q^{(n)}}[p_\varepsilon(\epsilon+U)]} \\
&=: C_\rho \Big(A_m^{(n)}\Big)^{1/2} R(\epsilon).
\end{aligned}$$

The first inequality uses the Lipschitz property of ρ_ε. The second applies the Cauchy–Schwarz inequality.

Now, for every $\epsilon \in \mathbb{R}$ with $p_\varepsilon(\epsilon) > 0$,

$$\begin{aligned} R^2(\epsilon) &\leq \left(\frac{\sup_{|u| \leq C_m} p_\varepsilon(\epsilon + u)/p_\varepsilon(\epsilon)}{\inf_{|u| \leq C_m} p_\varepsilon(\epsilon + u)/p_\varepsilon(\epsilon)} \right)^2 \\ &\leq \exp\Big\{ 4C_m |\rho_\varepsilon(\epsilon)| + 2C_\rho C_m^2 \Big\}. \end{aligned}$$

The first line is a supremum bound for the ratio of expectations, the second is the application of Lemma 26. Since $\exp(|x|) \leq \exp(x) + \exp(-x)$, this yields the bound

$$\begin{aligned} &\mathbb{E}_P\Big[\Big\{ \rho_\varepsilon(\varepsilon_P) - \rho_{\hat{\varepsilon}}(\varepsilon_P) \Big\}^2 \;\Big|\; D^{(n)} \Big] \\ &\qquad \leq C_\rho^2 \exp\Big(2C_\rho C_m^2 \Big) \Big(\mathbb{E}_P[\exp\{4C_m \rho_\varepsilon(\varepsilon_P)\}] + \mathbb{E}_P[\exp\{-4C_m \rho_\varepsilon(\varepsilon_P)\}] \Big)\; A_m^{(n)}. \end{aligned}$$

By Theorem 5, $\rho_\varepsilon(\varepsilon_P)$ is sub-Gaussian with parameter $\sqrt{2C_\rho}$, so for all $\lambda \in \mathbb{R}$ we have

$$\mathbb{E}_P[\exp\{\lambda \rho_\varepsilon(\varepsilon_P)\}] \leq \exp(\lambda^2 C_\rho).$$

Thus

$$\mathbb{E}_P\Big[\Big\{ \rho_\varepsilon(\varepsilon_P) - \rho_{\hat{\varepsilon}}(\varepsilon_P) \Big\}^2 \;\Big|\; D^{(n)} \Big] \leq 2C_\rho^2 \exp(18 C_\rho C_m^2)\; A_m^{(n)}.$$

□

Lemma 20. *Let P be such that p_ε is twice differentiable on $\mathbb{R}$, with*

$$\sup_{\epsilon \in \mathbb{R}} |\partial_\epsilon^2 \log p_\varepsilon(\epsilon)| = \sup_{\epsilon \in \mathbb{R}} |\rho'_\varepsilon(\epsilon)| \leq C_\rho,$$

and p'_ε and p''_ε both bounded. Further assume that $D^{(n)}$ is such that $\sup_{z \in \mathcal{Z}} |u_m^{(n)}(z)| \leq C_m$ for almost every $z \in \mathcal{Z}$. Then there exists a constant C, depending only on C_ρ, C_m, such that

$$\mathbb{E}_P\Big[\Big\{ \rho_{\hat{\varepsilon}}(\varepsilon_P) - \rho_{\hat{\varepsilon}}(\hat{\varepsilon}^{(n)}) \Big\}^2 \;\Big|\; D^{(n)} \Big] \leq C\Big(A_m^{(n)} + A_\sigma^{(n)} \Big).$$

Proof. For ease of notation, write $Q^{(n)}$ for the distribution of $u_m^{(n)}(Z)$ conditionally on $D^{(n)}$, and let $U \sim Q^{(n)}$. Therefore

$$A_m^{(n)} = \mathbb{E}_{Q^{(n)}}\Big(U^2 \Big).$$

The proof proceeds by first bounding the derivative of $\rho_{\hat{\varepsilon}}$. The conditions on p_ε are sufficient to interchange differentiation and expectation operators as follows (Aliprantis

and Burkinshaw, 1990, Thm. 20.4).

$$\begin{aligned}\rho_{\hat{\varepsilon}}(\epsilon) &= \frac{\frac{\partial}{\partial\epsilon}\mathbb{E}_{Q^{(n)}}[p_\varepsilon(\epsilon+U)]}{\mathbb{E}_{Q^{(n)}}[p_\varepsilon(\epsilon+U)]} \\ &= \frac{\mathbb{E}_{Q^{(n)}}\left[p'_\varepsilon(\epsilon+U)\right]}{\mathbb{E}_{Q^{(n)}}\left[p_\varepsilon(\epsilon+U)\right]},\end{aligned}$$

and further,

$$\begin{aligned}\rho'_{\hat{\varepsilon}}(\epsilon) &= \frac{\partial}{\partial\epsilon}\frac{\mathbb{E}_{Q^{(n)}}\left[p'_\varepsilon(\epsilon+U)\right]}{\mathbb{E}_{Q^{(n)}}\left[p_\varepsilon(\epsilon+U)\right]} \\ &= \frac{\frac{\partial}{\partial\epsilon}\mathbb{E}_{Q^{(n)}}\left[p'_\varepsilon(\epsilon+U)\right]}{\mathbb{E}_{Q^{(n)}}[p_\varepsilon(\epsilon+U)]} - \rho^2_{\hat{\varepsilon}}(\epsilon) \\ &= \frac{\mathbb{E}_{Q^{(n)}}\left[p''_\varepsilon(\epsilon+U)\right]}{\mathbb{E}_{Q^{(n)}}[p_\varepsilon(\epsilon+U)]} - \frac{\left(\mathbb{E}_{Q^{(n)}}\left[p'_\varepsilon(\epsilon+U)\right]\right)^2}{\left(\mathbb{E}_{Q^{(n)}}\left[p_\varepsilon(\epsilon+U_\sigma\epsilon+U_m)\right]\right)^2} \\ &= \frac{\mathbb{E}_{Q^{(n)}}\left[\left\{\rho'_\varepsilon(\epsilon+U)+\rho^2_\varepsilon(\epsilon+U)\right\}p_\varepsilon(\epsilon+U)\right]}{\mathbb{E}_{Q^{(n)}}\left[p_\varepsilon(\epsilon+U)\right]} - \frac{\left(\mathbb{E}_{Q^{(n)}}[\rho_\varepsilon(\epsilon+U)p_\varepsilon(\epsilon+U)]\right)^2}{\left(\mathbb{E}_{Q^{(n)}}[p_\varepsilon(\epsilon+U)]\right)^2}\end{aligned}$$

In the third line we have made use of the identities

$$\begin{aligned}p'_\varepsilon(\epsilon) &= \rho_\varepsilon(\epsilon)p_\varepsilon(\epsilon); \\ p''_\varepsilon(\epsilon) &= \left\{\rho'_\varepsilon(\epsilon)+\rho^2_\varepsilon(\epsilon)\right\}p_\varepsilon(\epsilon).\end{aligned}$$

We now apply both the triangle and Hölder inequalities to deduce

$$\begin{aligned}\left|\rho'_{\hat{\varepsilon}}(\epsilon)\right| &\leq \sup_{|u|\leq C_m}\left\{\left|\rho'_\varepsilon(\epsilon+u)\right|+2\rho^2_\varepsilon(\epsilon+u)\right\} \\ &\leq C_\rho + 2\sup_{|u|\leq C_m}\rho^2_\varepsilon(\epsilon+u).\end{aligned}$$

Now we apply a Taylor expansion as follows, noting that ε_P is independent of U conditionally on $D^{(n)}$.

$$\begin{aligned}
\mathbb{E}_P\Big[\Big\{\rho_{\hat{\varepsilon}}(\varepsilon_P) - \rho_{\hat{\varepsilon}}(\hat{\varepsilon}^{(n)})\Big\}^2 \;\Big|\; D^{(n)}\Big] &= \mathbb{E}_P\Big(\mathbb{E}_{Q^{(n)}}\Big[\{\rho_{\hat{\varepsilon}}(\varepsilon_P) - \rho_{\hat{\varepsilon}}(\varepsilon_P + U)\}^2 \;\Big|\; \varepsilon_P,\ D^{(n)}\Big] \;\Big|\; D^{(n)}\Big) \\
&\leq \mathbb{E}_P\left(\mathbb{E}_{Q^{(n)}}(U^2)\Big\{\sup_{|u|\leq C_m} \rho'_{\hat{\varepsilon}}(\varepsilon_P + u)\Big\}^2 \;\middle|\; D^{(n)}\right) \\
&= A_m^{(n)}\mathbb{E}_P\left(\Big\{\sup_{|u|\leq C_m} \rho'_{\hat{\varepsilon}}(\varepsilon_P + u)\Big\}^2 \;\middle|\; D^{(n)}\right) \\
&\leq A_m^{(n)}\mathbb{E}_P\left(\Big\{C_\rho + 2\sup_{|\eta|\leq 2C_m} \rho^2_{\hat{\varepsilon}}(\varepsilon_P + \eta)\Big\}^2 \;\middle|\; D^{(n)}\right) \\
&\leq A_m^{(n)}\mathbb{E}_P\Big[\Big\{C_\rho + 2\Big(|\rho_\varepsilon(\varepsilon_P)| + 2C_\rho C_m\Big)^2\Big\}^2\Big] \\
&\leq A_m^{(n)}\ \mathbb{E}_P\Big[\Big\{C_\rho + 4\rho^2_\varepsilon(\varepsilon_P) + 16C_\rho^2 C_m^2\Big\}^2\Big] \\
&\leq A_m^{(n)}\ \Big(3C_\rho^2 + 48\mathbb{E}_P\Big[\rho^4_\varepsilon(\varepsilon_P)\Big] + 768C_\rho^4 C_m^4\Big),
\end{aligned}$$

where we have made use of the inequalities $(a+b)^2 \leq 2(a^2+b^2)$ and $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$. Finally, by Theorem 5, $\mathbb{E}_P\Big[\rho^4_\varepsilon(\varepsilon_P)\Big] \leq 4C_\rho^2$. Hence

$$\mathbb{E}_P\Big[\Big\{\rho_{\hat{\varepsilon}}(\varepsilon_P) - \rho_{\hat{\varepsilon}}(\hat{\varepsilon}^{(n)})\Big\}^2 \;\Big|\; D^{(n)}\Big] \leq A_m^{(n)}\ \Big(147C_\rho^2 + 768C_\rho^4 C_m^4\Big).$$

□

2.10 Auxiliary lemmas

Lemma 21. *If p is a twice differentiable density function on $\mathbb{R}$ with score ρ defined everywhere and $\sup_{x\in\mathbb{R}} |\rho'(x)| \leq C$, then $\sup_{x\in\mathbb{R}} p(x) \leq 2\sqrt{2C}$.*

Proof. Suppose, for a contradiction, that $\sup_{x\in\mathbb{R}} p(x) > 2\sqrt{2C}$. Pick $x_0 < x_1'$ such that $0 < p(x_0) < \sqrt{2C}$ and $p(x_1') > 2\sqrt{2C}$, and further that x_1' is not the maximiser of p on the interval $[x_0, x_1']$. We set x_1 to be the maximiser of p in (x_0, x_1'), and observe that $p(x_1) =: M > 2\sqrt{2C}$, $p'(x_1) = 0$, and $p''(x_1) < 0$.

Now let $x_- = \sup\{x < x_1 : p(x) \leq M/2\} > x_0$, the final inequality following from the intermediate value theorem. Note that as

$$1 \geq \int_{x_-}^{x_1} p(x)\,dx \geq \frac{M}{2}(x_1 - x_-),$$

$x_1 - x_- \leq 2/M$. We also have that $p(x_1) - p(x_-) \geq M/2$, so there must be a point $\tilde{x} \in [x_-, x_1]$ where $p'(\tilde{x}) \geq M^2/4$. Now because $p'(x_1) = 0$ there must also be a point $x_* \in [\tilde{x}, x_1]$ with $p''(x_*) \leq -M^3/8$.

Finally we may employ the assumption on $|\rho'|$ to bound $p''(x_*)$ from below. Noting that $p(x_*) \leq M$ as $x_* \in [x_0, x_1]$, we have

$$\begin{aligned} -M^3/8 \geq p''(x_*) &= \rho'(x_*)p(x_*) + \rho^2(x_*)p(x_*) \\ &\geq \rho'(x_*)p(x_*) \geq -CM. \end{aligned}$$

□

Corollary 22. *If p is a twice differentiable density function on $\mathbb{R}$ with score ρ defined everywhere and $\sup_{x\in\mathbb{R}} |\rho'(x)| \leq C$ then $\inf_{x\in\mathbb{R}} p''(x) \geq -2\sqrt{2}C^{3/2}$.*

Proof. This follows from Lemma 21 and $p''(x) = \rho'(x)p(x) + \rho^2(x)p(x) \geq \rho'(x)p(x)$. □

Lemma 23. *If p is a twice differentiable density function on $\mathbb{R}$ and $\sup_{x\in\mathbb{R}} |\rho'(x)| \leq C$, then $p(x) \to 0$ as $|x| \to \infty$.*

Proof. Note first that by Lemma 21 we know that $p(x)$ is uniformly bounded. Suppose then, for contradiction, that $\limsup_{|x|\to\infty} p(x) =: 2\epsilon > 0$. Then for any $M \geq 0$ we can find x_0 with $|x_0| > M$ and $p(x_0) \geq \epsilon$. We will show that the integral of $p(x)$ over a finite interval containing x_0 is bounded below. This means that we can choose non-overlapping intervals $I_1, \ldots, I_N$ such that

$$\int_{\mathbb{R}} p(x)\, dx \geq \sum_{n=1}^{N} \int_{I_n} p(x)\, dx > 1,$$

a contradiction.

Since p' is continuous, we have that $|p'(x_0)| < \infty$. By Corollary 22, $\inf_{x\in\mathbb{R}} p''(x) \geq -2\sqrt{2}C^{3/2}$. Using a Taylor expansion, we can fit a negative quadratic beneath the curve p at x_0. Integrating this quadratic over the region where it is positive gives the bound. Indeed,

$$\begin{aligned} p(x) &\geq p(x_0) + (x - x_0)p'(x_0) - \sqrt{2}C^{3/2}(x - x_0)^2 \\ &= p(x_0) + \frac{\big(p'(x_0)\big)^2}{4\sqrt{2}C^{3/2}} - \sqrt{2}C^{3/2}\left(x - x_0 - \frac{p'(x_0)}{2\sqrt{2}C^{3/2}}\right)^2 \\ &\geq \epsilon - \sqrt{2}C^{3/2}\left(x - x_0 - \frac{p'(x_0)}{2\sqrt{2}C^{3/2}}\right)^2 =: f(x). \end{aligned}$$

The quadratic $f(x)$ has roots

$$a := x_0 + \frac{p'(x_0)}{2\sqrt{2}C^{3/2}} - \frac{\sqrt{\epsilon}}{2^{1/4}C^{3/4}};$$
$$b := x_0 + \frac{p'(x_0)}{2\sqrt{2}C^{3/2}} + \frac{\sqrt{\epsilon}}{2^{1/4}C^{3/4}}.$$

Thus (a, b) is a finite interval containing x_0 and

$$\int_a^b p(x)\,dx \geq \int_a^b f(x)\,dx = \frac{2^{7/4}\epsilon^{3/2}}{3C^{3/4}}. \qquad \square$$

Lemma 24. *Let $f : \mathbb{R} \to \mathbb{R}$ be a continuous function. Then at least one of the following holds.*

(a) There exists a sequence $a_n \to \infty$ such that $f(a_n) \to 0$.

(b) There exists $A \in \mathbb{R}$ and $\epsilon > 0$ such that $f(x) > \epsilon$ for all $x \geq A$, and in particular $\int_A^\infty f(x)\,dx = \infty$.

(c) There exists $A \in \mathbb{R}$ and $\epsilon > 0$ such that $f(x) < -\epsilon$ for all $x \geq A$, and in particular $\int_A^\infty (-f(x))\,dx = \infty$.

Proof. If $\liminf_{x\to\infty} f(x) > 0$ then clearly (b) occurs while if $\limsup_{x\to\infty} f(x) < 0$ then (c) occurs. Thus we may assume that $\liminf_{x\to\infty} f(x) \leq 0 \leq \limsup_{x\to\infty} f(x)$. If either of these inequalities are equalities, then (a) occurs, so we may assume they are both strict. However in this case, as $\{f(x) : x \geq A\}$ has infinitely many positive points and negative points for all $A \geq 0$, by the intermediate value theorem, we must have that (a) occurs. $\square$

Lemma 25. *Let p be a twice differentiable density function on $\mathbb{R}$ with score ρ defined everywhere, and let k a non-negative integer. If $\sup_{x\in\mathbb{R}} |\rho'(x)| \leq C$ and $\mathbb{E}[\rho^{2k}(X)] < \infty$, then there exist sequences $a_n \to -\infty$ and $b_n \to \infty$ such that $\rho^{2k+1}(a_n)p(a_n) \to 0$ and $\rho^{2k+1}(b_n)p(b_n) \to 0$.*

Proof. Write $f(x) = \rho^{2k+1}(x)p(x)$. Since f is continuous, we may apply Lemma 24 to both f and $x \mapsto f(-x)$ to conclude that either the statement of the lemma holds, or one of the following hold for some $B \in \mathbb{R}$ and $\epsilon > 0$:

(a) $f(x) > \epsilon$ for all $x \geq B$,

(b) $f(x) < -\epsilon$ for all $x \geq B$

or one of the above with $x \geq B$ replaced with $x \leq B$. Let us suppose for a contradiction that (a) occurs (the other cases are similar), so in particular

$$\int_B^\infty f(x)\,dx = \infty. \tag{2.43}$$

If $k = 0$, then

$$\int_B^\infty f(x)\,dx = \int_B^\infty p'(x)\,dx = \lim_{b\to\infty}\int_B^b p'(x)\,dx = \lim_{b\to\infty} p(b) - p(B);$$

here the penultimate equality follows from monotone convergence and the final equality follows from the fundamental theorem of calculus. By Lemma 23 however, this is finite, a contradiction. If instead $k \geq 1$, then for any $b \geq B$ we have that

$$\begin{aligned}\rho^{2k}(b)p(b) - \rho^{2k}(B)p(B) &= \int_B^b \rho^{2k}(x)p'(x)\,dx + 2k\int_B^b \rho'(x)\rho^{2k-1}(x)p(x)\,dx. \\ &= \int_B^b \rho^{2k+1}(x)p(x)\,dx + 2k\int_B^b \rho'(x)\rho^{2k-1}(x)p(x)\,dx. \qquad (2.44)\end{aligned}$$

We will take the limit as $b \to \infty$. Since $\rho^{2k}(x)p(x)$ is non-negative and we have that $\mathbb{E}[\rho^{2k}(X)] < \infty$, we can choose an increasing sequence $b_n \to \infty$ satisfying $\rho^{2k}(b_n)p(b_n) \leq 1$ for every n.

Note that for each n and for every $x \in \mathbb{R}$, Hölder's inequality gives

$$\left|\mathbb{1}_{[B,b_n]}(x)\rho'(x)\rho^{2k-1}(x)p(x)\right| \leq C|\rho(x)|^{2k-1}p(x).$$

By Jensen's inequality, $\mathbb{E}\left[|\rho(X)|^{2k-1}\right] < \infty$. Thus, by dominated convergence theorem,

$$\begin{aligned}\lim_{n\to\infty} 2k\int_B^{b_n} \rho'(x)\rho^{2k-1}(x)p(x)\,dx &= 2k\int_B^\infty \rho'(x)\rho^{2k-1}(x)p(x)\,dx \\ &\leq 2kC\int_B^\infty |\rho(x)|^{2k-1}p(x)\,dx \\ &\leq 2kC\,\mathbb{E}\left[|\rho(X)|^{2k-1}\right] < \infty.\end{aligned}$$

Now (2.44) implies that

$$\lim_{n\to\infty}\int_B^{b_n} f(x)\,dx < \infty.$$

But we assumed that $f(x) \geq \epsilon > 0$ for all $x \geq B$, so for each fixed $x \in \mathbb{R}$ the integrand $\mathbb{1}_{[B,b_n]}(x)f(x)$ is increasing as a function of n. Therefore monotone convergence implies that

$$\int_B^\infty f(x)\,dx < \infty,$$

contradicting (2.43). □

Lemma 26. *If p is a twice differentiable density on $\mathbb{R}$ with score ρ defined everywhere such that $\sup_{x\in\mathbb{R}}|\rho'(x)| \leq C$, then for any $x, u \in \mathbb{R}$ such that $p(x) > 0$,*

$$\exp\left\{u\rho(x) - \frac{u^2}{2}C\right\} \leq \frac{p(x+u)}{p(x)} \leq \exp\left\{u\rho(x) + \frac{u^2}{2}C\right\}.$$

Proof. The inequality is proved via a Taylor expansion on $\log p(x+u)$ around $u = 0$. Indeed,

$$\log p(x+u) = \log p(x) + u\rho(x) + \frac{u^2}{2}\rho'(\eta)$$

for some $\eta \in [x - |u|, x + |u|]$. Rearranging and taking absolute values gives the bound

$$\left| \log\left(\frac{p(x+u)}{p(x)}\right) - u\rho(x) \right| \leq \frac{u^2}{2} C.$$

Since the exponential function is increasing, this suffices to prove the claim. □

2.11 Additional points

2.11.1 Linear score functions

Some works have made the simplifying assumption that

$$\rho_P(x, z) = \beta_P^T b(x, z) \tag{2.45}$$

for some known basis $b(x, z)$. This has some theoretical appeal, since any ρ_P can be represented in this way for some bases, and the score estimation problem is made parametric. Practically, however, even with domain knowledge it can be hard to choose a good basis. When (X, Z) are of moderate to large dimension, there are limited interactions that one can practically allow — for instance a quadratic basis may be feasible, but a multivariate kernel basis not. If the chosen basis contains the vector x, then it transpires that the linearity assumption (2.45) is equivalent to assuming a certain conditional Gaussian linear model for x given the other basis elements (see Theorem 27 below). This provides additional insight into the method of Rothenhäusler and Yu (2020), which is based on the debiased lasso (van de Geer et al., 2014; Zhang and Zhang, 2014).

Theorem 27. *Let $b(x, z) = \Big(x, g^T(x, z)\Big)^T \in \mathbb{R}^m$ for some $g : \mathbb{R} \times \mathcal{Z} \to \mathbb{R}^{m-1}$ be such that $\mathbb{E}\Big[b(X, Z)b(X, Z)^T\Big]$ is positive definite, $\mathbb{E}\Big[\|b(X, Z)\|_2^2\Big] < \infty$, $\mathbb{E}|\partial_x b(X, Z)| < \infty$, and for almost every $z \in \mathcal{Z}$ we have that $b(\cdot, z)$ and $\partial_x b(\cdot, z)$ are absolutely continuous and $\lim_{|x|\to\infty} b(x, z)p(x \mid z) = 0$. Define the linearly transformed basis functions*

$$\bar{g}(x, z) := g(x, z) - x\Big(\mathbb{E}\Big[\partial_x g(X, Z)\Big]\Big) \in \mathbb{R}^{m-1}.$$

We have that $\rho(x, z) = \beta^T b(x, z)$ for some $\beta \in \mathbb{R}^m$ if and only if $\rho(x, z) = \tilde{\rho}(x, \bar{g}(x, z))$, where $\tilde{\rho}$ is the score function corresponding to the related multivariate Gaussian linear model:

$$(X, g) \overset{d}{=} \Big(X, \bar{g}(X, Z)\Big); \quad X \mid g \sim N\Big(\gamma^T g, S\Big).$$

Here $\gamma \in \mathbb{R}^{(m-1)}$ and $S > 0$ do not depend on (X, g).

Proof. First assume that $X \mid g$ has the stated conditional distribution. Then

$$\begin{aligned}\tilde{\rho}(x, g) &= \partial_x \log \tilde{p}(x \mid g) \\ &= -S^{-1}(x - \gamma^T g) \\ &= \begin{pmatrix} -S^{-1} & S^{-1}\gamma \end{pmatrix}^T \begin{pmatrix} x \\ g \end{pmatrix},\end{aligned}$$

so indeed $\tilde{\rho}(x, \bar{g}(x, z))$ is in the linear span of $\{x, \bar{g}(x, z)\}$, and hence that of $b(x, z)$.

Now let $\rho(x, z) = \beta^T b(x, z)$, and denote by β_x and β_g the first and last $m-1$ components of β respectively. Define the transformed variables

$$\begin{aligned}\bar{b}(x, z) &:= \begin{pmatrix} x \\ \bar{g}(x, z) \end{pmatrix} = \begin{pmatrix} 1 & 0_{1\times(m-1)} \\ -\mathbb{E}\big[\partial_x g(X, Z)\big] & I_{(m-1)\times(m-1)} \end{pmatrix} b(x, z); \qquad (2.46) \\ \bar{\beta}_x &:= \beta_x + \beta_g^T \mathbb{E}\big[\partial_x g(X, Z)\big]; \\ \bar{\beta} &:= \begin{pmatrix} \bar{\beta}_x \\ \beta_g \end{pmatrix}.\end{aligned}$$

By the decomposition (2.46) we see that $\mathbb{E}\big[\bar{b}(X, Z)\bar{b}^T(X, Z)\big]$ inherits the positive definiteness of $\mathbb{E}\big[b(X, Z)b^T(X, Z)\big]$. Then we have that

$$\rho(x, z) = \bar{\beta}^T \bar{b}(x, z); \quad \mathbb{E}\big[\partial_x \bar{b}^T(X, Z)\big] = \begin{pmatrix} 1 & 0_{(m-1)\times 1} \end{pmatrix}.$$

The conditions on b mean that ρ satisfies the conditions of Cox (1985, Prop. 1) conditionally on Z, so $\bar{\beta}$ minimises

$$\mathbb{E}\big[(\bar{\beta}^T \bar{b}(X, Z))^2 + 2\partial_x \bar{\beta}^T \bar{b}(X, Z)\big] = \bar{\beta}^T \mathbb{E}\big[\bar{b}(X, Z)\bar{b}^T(X, Z)\big]\bar{\beta} + 2\bar{\beta}^T \Big(\mathbb{E}\big[\partial_x \bar{b}^T(X, Z)\big]\Big)^T.$$

Hence

$$\mathbb{E}\big[\bar{b}(X, Z)\bar{b}^T(X, Z)\big]\bar{\beta} + \begin{pmatrix} 1 \\ 0_{(m-1)\times 1} \end{pmatrix} = 0.$$

Using the Schur complement identity for the inverse, we find that $\bar{\beta}$ takes the following form:

$$\begin{aligned}
\bar{\beta} &= -\begin{pmatrix} 1 \\ \gamma \end{pmatrix} S^{-1}, \\
\gamma &= \Big(\mathbb{E}\big[\bar{g}(X,Z)\bar{g}^T(X,Z)\big]\Big)^{-1}\mathbb{E}\big[\bar{g}(X,Z)X^T\big], \\
S &= \mathbb{E}\Big[\big\{X - \gamma^T\bar{g}(X,Z)\big\}X^T\Big].
\end{aligned}$$

Therefore we have that

$$\rho(x,z) = -S^{-1}\{x - \gamma^T\bar{g}(x,z)\}.$$

Finally, note that $\gamma \in \mathbb{R}^{m-1}$ satisfies

$$\mathbb{E}\Big[\big\{X - \gamma^T\bar{g}(X,Z)\big\}\bar{g}^T(X,Z)\Big] = 0.$$

This implies both that γ minimises $\mathbb{E}\Big[\{X - \gamma^T\bar{g}(X,Z)\}^2\Big] = S$. This suffices to prove that

$$\rho(x,z) = -S^{-1}\{x - \gamma^T\bar{g}(x,z)\} = \tilde{\rho}(x, \bar{g}(x,z)).$$

□

2.11.2 Explicit estimators for numerical experiments

In order reduce the computational burden, we pre-tune all hyperparameters on 1000 datasets, each of which we split into training and testing. This includes all gradient boosting regression parameters, the various spline degrees of freedom and the Lasso tuning parameters of the basis approaches.

Resmooth and spline

Let $\tilde{f}^{(n,k)}$ and $\hat{m}^{(n,k)}$ be gradient boosting regressions (`xgboost` package (Chen and Guestrin, 2016)) of Y on (X,Z) and X on Z respectively, using the out-of-fold data $D^{(n,k)}$. Further let $\hat{\sigma}^{(n,k)}$ be the a decision tree (`partykit` package (Hothorn and Zeileis, 2015)) regression of the squared in-sample residuals of X on Z, and $\hat{\rho}_{\hat{\varepsilon}}^{(n,k)}$ be a univariate spline score estimate (our implementation) using the scaled in-sample residuals.

Let $\hat{\theta}^{(n)}, \hat{\Sigma}^{(n)}$ be as in (2.3) where

$$\begin{aligned}
\hat{f}^{(n,k)}(x,z) &= \sum_{j=1}^{J} \tilde{f}^{(n,k)}(x + hw_j, z) q_j; \\
\nabla \hat{f}^{(n,k)}(x,z) &= \frac{1}{h} \sum_{j=1}^{J} w_j \tilde{f}^{(n,k)}(x + hw_j, z) q_j; \\
\hat{\rho}^{(n,k)}(x,z) &= \frac{1}{\hat{\sigma}^{(n,k)}(z)} \, \hat{\rho}_{\hat{\varepsilon}}^{(n,k)} \left(\frac{x - \hat{m}^{(n,k)}(z)}{\hat{\sigma}^{(n,k)}(z)} \right).
\end{aligned}$$

Here we approximate Gaussian expectations via numerical integration, using a deterministic set of pairs (w_j, q_j) such that, for functions g,

$$\mathbb{E}[g(W)] \approx \sum_{j=1}^{J} g(w_j) q_j.$$

We have used $J = 101$, $\{w_j\}$ to be an evenly spaced grid on $[-5, 5]$, and q_j to be proportional to the standard normal density at w_j, scaled so that $\sum_{j=1}^{J} q_j = 1$.

We took the set of bandwidths $\mathcal{H}$ in Algorithm 1 to be

$$\frac{\exp(-5)}{2\sqrt{3}} \hat{\sigma}_X, \, \frac{\exp(-4.8)}{2\sqrt{3}} \hat{\sigma}_X, \ldots, \frac{\exp(2)}{2\sqrt{3}} \hat{\sigma}_X,$$

where $\hat{\sigma}_X$ denotes the empirical standard deviation of the X-variable.

Difference and basis

We form an estimator as in (2.3). Let $\tilde{f}^{(n,k)}$ be a gradient boosting regression (`xgboost` package (Chen and Guestrin, 2016)) of Y on (X, Z) using the out-of-fold data $D^{(n,k)}$. Set basis b to the quadratic basis for $(X, Z) \in \mathbb{R}^p$, omitting the X term:

$$b(x,z) = (1, x^2, xz_1, \ldots, xz_{p-1}, z_1, z_1^2, z_1 z_2, \ldots, z_1 z_{p-1}, z_2, z_2^2, z_2 z_3, \ldots, \ldots, z_{p-1}, z_{p-1}^2).$$

Let $\hat{\beta}^{(n,k)}$ be the lasso coefficient (`glmnet` package) when regressing X on $b(X, Z)$ using $D^{(n,k)}$, and $\hat{\sigma}^{(n,k)}$ be the in-sample variance estimate, computed using the product of X and the X on Z residuals.

Let $\hat{\theta}^{(n)}, \hat{\Sigma}^{(n)}$ be as in (2.3) where

$$\begin{aligned}\hat{f}^{(n,k)}(x,z) &= \tilde{f}^{(n,k)}(x,z);\\ \nabla\hat{f}^{(n,k)}(x,z) &= \frac{\tilde{f}^{(n,k)}\left(x+\frac{D}{2},z\right)-\tilde{f}^{(n,k)}\left(x-\frac{D}{2},z\right)}{D};\\ \hat{\rho}^{(n,k)}(x,z) &= -\frac{1}{\left(\hat{\sigma}^{(n,k)}\right)^2}\left(x_i - b(x_i,z_i)^T\hat{\beta}^{(n,k)}\right).\end{aligned}$$

Here D is set to one quarter of the (population) marginal standard deviation of X.

Partially linear regression

We consider a doubly-robust partially linear regression as in Chernozhukov et al. (2018, §4.1), implemented in the `DoubleML` R package (Bach et al., 2021). The partially linear regression makes the simplifying assumption that $\mathbb{E}_P(Y \mid X, Z) = \theta_P X + g_P(Z)$. When this relationship is misspecified, procedures which minimise the sum of squares target the quantity

$$\theta_P^* = \frac{\mathbb{E}_P[\mathrm{Cov}_P\{X, \mathbb{E}_P(Y \mid X, Z) \mid Z\}]}{\mathbb{E}_P[\mathrm{Var}_P(X \mid Z)]}$$

(Vansteelandt and Dukes, 2022); this does not equal the average partial effect $\theta_P = \mathbb{E}_P[f'_P(X,Z)]$ in general.

The nuisance functions g_P and $\mathbb{E}_P(X \mid Z)$ may be modelled via plug-in machine learning, so again we use gradient boosting (`xgboost` package (Chen and Guestrin, 2016)). Hyperparameter pre-tuning for g_P estimation is done by regressing $Y - \theta_P X$ on Z. Here we have used θ_P instead of the unknown θ_P^* for convenience, but we do not expect this to be critical.

Rothenhäusler and Yu (2020)

The estimator of Rothenhäusler and Yu (2020) is based on the debiased lasso (van de Geer et al., 2014; Zhang and Zhang, 2014). As they recommend, we use a quadratic basis for $Z \in \mathbb{R}^{p-1}$,

$$b(z) = (1, z_1, z_1^2, z_1z_2, \ldots, z_1z_{p-1}, z_2, z_2^2, z_2z_3, \ldots, \ldots, z_{p-1}, z_{p-1}^2).$$

We perform the lasso regressions using `glmnet` (Friedman et al., 2010).

2.11.3 Spline score estimation

We use the univariate estimator of Cox (1985), which we implemented according to Ng (1994, 2003). Let $X_1, \ldots, X_n$ be univariate random variables, which we are treating as an

i.i.d. sample. For our application to multivariate score estimation these are the in-sample residuals.

Theorem 28 (Ng (1994, 2003))**.** *Assume that* $X_1 < X_2 < \ldots < X_n$*. Define*

$$
\begin{aligned}
h &:= (X_2 - X_1, \ldots, X_n - X_{n-1}) \in \mathbb{R}^{n-1} \\
wih &:= \left(\frac{w_1}{h_1}, \ldots, \frac{w_{n-2}}{h_{n-2}}, \frac{w_{n-1} + w_n}{h_{n-1}} \right) \in \mathbb{R}^{n-1} \\
wh &:= \left(w_1 h_1, \ldots, w_{n-2} h_{n-2}, \left(w_{n-1} - \frac{w_n}{2} \right) h_{n-1} \right) \in \mathbb{R}^{n-1} \\
a &:= \Big((wih, 0) - (0, wih) \Big) \in \mathbb{R}^n \\
c &:= \left(\frac{wh[-(n-1)] + 2wh[-1]}{3} \right) \in \mathbb{R}^{n-2} \\
R &:= \begin{bmatrix}
\frac{2}{3}(h_1 + h_2) & \frac{h_2}{3} & 0 & \cdots & \cdots & 0 \\
\frac{h_2}{3} & \frac{2}{3}(h_2 + h_3) & \frac{h_3}{3} & \ddots & & \vdots \\
0 & \frac{h_3}{3} & \ddots & \ddots & \ddots & \vdots \\
\ldots & \ddots & \ddots & \ddots & \frac{h_{n-3}}{3} & 0 \\
\ldots & & \ddots & \frac{h_{n-3}}{3} & \frac{2}{3}(h_{n-3} + h_{n-2}) & \frac{h_{n-2}}{3} \\
0 & \cdots & \cdots & 0 & \frac{h_{n-2}}{3} & \frac{2}{3}(h_{n-2} + h_{n-1})
\end{bmatrix} \\
&\in \mathbb{R}^{(n-2)\times(n-2)} \\
Q &:= \begin{bmatrix}
\frac{1}{h_1} & 0 & \cdots & \cdots & \cdots & 0 \\
-\left(\frac{1}{h_1} + \frac{1}{h_2} \right) & \frac{1}{h_2} & \ddots & & & \vdots \\
\frac{1}{h_2} & -\left(\frac{1}{h_2} + \frac{1}{h_3} \right) & \ddots & \ddots & & \vdots \\
0 & \frac{1}{h_3} & \ddots & \ddots & \ddots & \vdots \\
\vdots & \ddots & \ddots & \ddots & \frac{1}{h_{n-3}} & 0 \\
\vdots & & \ddots & \ddots & -\left(\frac{1}{h_{n-3}} + \frac{1}{h_{n-2}} \right) & \frac{1}{h_{n-2}} \\
\vdots & & & \ddots & \frac{1}{h_{n-2}} & -\left(\frac{1}{h_{n-2}} + \frac{1}{h_{n-2}} \right) \\
0 & \cdots & \cdots & \cdots & 0 & \frac{1}{h_{n-1}}
\end{bmatrix} \\
&\in \mathbb{R}^{n\times(n-2)} \\
Y &:= \left\{ diag\left(\frac{1}{w_1}, \ldots, \frac{1}{w_n} \right) \right\} \Big(a + QR^{-1}c \Big) \in \mathbb{R}^n.
\end{aligned}
$$

Then the minimiser of the spline score objective function (Cox, 1985),

$$
J(\rho) := \frac{1}{\sum_{i=1}^n w_i} \sum_{i=1}^n w_i \Big\{ \rho(X_i)^2 + 2\rho'(X_i) \Big\} + \lambda \int_{\mathbb{R}} \rho''(x)^2 \, dx,
$$

is also the minimiser of the classical smoothing spline objective function

$$\tilde{J}(f) := \frac{1}{\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} w_i \{Y_i - f(X_i)\}^2 + \lambda \int_{\mathbb{R}} f''(x)^2 \, dx,$$

for which implementations exist.

Note that in scaling, $Y \sim 1/h$. This creates numerical problems when n is large, and hence h is small. Our solution is to group nearby data points and use the weightings introduced here.

Chapter 3

Conditional independence testing with structured categorical data

3.1 Introduction

Categorical data are common across quantitative fields, arising naturally from clinical diagnoses and treatment actions, opinion surveys, and census style personal information. Conditional independence testing is important for modern data analysis, with applications to significance testing and causal structure learning. Conditional independence testing between categorical variables can be challenging when there are a large number of categories, as some combinations may be infrequent in the data. Having said this, categorical variables of interest often possess additional structure, such as occupations being implicitly arranged by sector, and ordinal responses to Likert-style opinion surveys. This domain knowledge can be used to borrow strength between related categories, such as by combining some categories into larger groups. In this work we propose a novel test for conditional independence testing between categorical variables X and Y given a third variable Z of arbitrary type. Our procedure makes use of the structure of X and Y to adaptively increase power whilst maintaining valid size.

Conditional independence testing is a hard problem. Shah and Peters (2020) prove that when Z has a continuous component, any test for conditional independence between X and Y given Z which holds size against all null hypotheses cannot have power against any alternative. This means that it is necessary to restrict the class of null distributions under consideration, and this may be done conveniently in terms of the convergence rates of machine learning estimation procedures. The starting point for our test is a categorical version of the Generalised Covariance Measure (GCM) of Shah and Peters (2020). Ankan and Textor (2022) motivate the use of the GCM for categorical conditional independence testing due to its of simplicity, symmetry (we reject $X \perp\!\!\!\perp Y \mid Z$ if and only if we reject $Y \perp\!\!\!\perp X \mid Z$), computational efficiency, calibration guarantees, and power over diverse

alternatives. Let X and Y take values in some finite sets $\{1, \ldots, d_X\}$ and $\{1, \ldots, d_Y\}$ respectively, and allow Z to range over some set $\mathcal{Z} \subseteq \mathbb{R}^p$. We reduce the problem of testing conditional independence $X \perp\!\!\!\perp Y \mid Z$ to testing whether the vector

$$\Big(\mathbb{E}_P\Big[\mathrm{Cov}_P(\mathbb{1}\{X = j\}, \mathbb{1}\{Y = k\} \mid Z)\Big] \ : \ j \in \{1, \ldots, d_X\}, \ k \in \{1, \ldots, d_Y\}\Big) \in \mathbb{R}^{d_X d_Y} \tag{3.1}$$

is zero. Tests based on estimating (3.1) have power against a wide range of alternatives, but do not make use of the structure of the categorical variables X and Y.

For ordinal X and Y, Li and Shepherd (2010) consider several tests of conditional independence based on parametric estimates of the conditional distributions of $X \mid Z$ and $Y \mid Z$. Their tests compare various summary statistics of the independence null $\mathbb{P}(X = j, Y = k \mid Z) = \mathbb{P}(X = j \mid Z)\mathbb{P}(Y = k \mid Z)$ to the observed data distribution, in such a way that they utilise the ordinal structure. Ankan and Textor (2022) extend this work to more general propensity estimators, and advocate testing for conditional covariance between a certain transformation of (X, Z) and (Y, Z) (Li and Shepherd, 2010, 2012). Their method does not have power against all alternatives with (3.1) non-zero, for instance when the distribution of $Y \mid (X, Z)$ is symmetric. Liu et al. (2021) view ordinal X and Y as discretisations of some unobserved continuous latent variables. If the function linking these follows a parametric model, they are able to estimate and sample from the latent distributions for each of X and Y. This reduces the problem to univariate continuous conditional independence testing, but it may be unclear how to choose this parametric model practically.

We instead consider testing for conditional independence using various partitions of the labels of X and Y. As a motivating example, suppose we collect 1000 responses to a survey data where the variable X represents which of 30 common jobs — divided into 5 sectors — is closest to that of the respondent, Y represents their agreement with a certain statement on a scale of 1–10, and Z contains some other factors. Suppose that there is a conditional dependence between employment sectors and agreement strength, given Z, but it is weak enough that tests based on estimating (3.1), which is 300-dimensional, do not have sufficient power to reject the null. It may be that most of the conditional dependence information is captured at a coarser level of sectors and disagreement / neutral / agreement with the statement. In this case a test using an estimate of the following lower dimensional summary statistic could have more power:

$$\Big(\mathbb{E}_P\Big[\mathrm{Cov}_P(\mathbb{1}\{X \in A_l\}, \mathbb{1}\{Y \in B_m\} \mid Z)\Big] \ : \ l \in \{1, \ldots, |A|\}, m \in \{1, \ldots, |B|\}\Big) \in \mathbb{R}^{|A||B|}, \tag{3.2}$$

where $A := \Big(\{\text{jobs in } l\text{th sector}\} \ : \ l = 1, \ldots, 5\Big)$ is a partition of $\{1, \ldots, 30\}$ and $B := \Big(\{1, 2, 3\}, \{4, 5, 6, 7\}, \{8, 9, 10\}\Big)$ is a partition of $\{1, \ldots, 10\}$, with $|A||B| = 15$. Our

primary contribution is a test which chooses partitions A and B automatically based on the data so as to adaptively increase its power.

We draw attention to two specific challenges that we face. First, the dimensionality of the problem may cause difficulties when inverting estimated (co-)variances, which is required by some existing methods (Ankan and Textor, 2022; Shah and Peters, 2020). Second, adaptive label merging may result in double-dipping so standard asymptotics do not apply. We perform calibration via a bootstrap procedure. An alternative would be to use one portion of the sample to determine a promising partition, and then apply this to the rest of the data and appeal to standard results (notwithstanding the concerns about (co-)variance inversion).

3.1.1 Our contributions and organisation of the chapter

In Section 3.2.4 we introduce an adaptive test for conditional independence between structured categorical variables X and Y, in the presence of an arbitrary variable Z (Algorithm 4). Our test greedily searches for a partition of the labels which enhances the test's power, and is able to leverage structural knowledge of X and Y to restrict the search space. Our test is calibrated using a novel bootstrap algorithm, and we prove that our method controls size asymptotically under weak conditions (Theorem 29). In Section 3.3 we demonstrate empirically that our method controls size in finite-sample settings where existing methods fail. We further demonstrate that our adaptive search procedure does indeed improve power against a non-adaptive version of our method. We additionally show that when the structural information is misspecified, we do not have substantially worse power than the non-adaptive methods. We describe a fast implementation of our search procedure in Section 6 (Algorithm 6). Code is shared in the R package `catci` (CATegorical Conditional Independence) available from https://github.com/harveyklyne/catci.

3.1.2 Other related work

There has been substantial recent progress on conditional independence testing when all of X, Y, and Z are continuous. Shah and Peters (2020) and Scheidegger et al. (2022) estimate the conditional means, and test for correlation of the residuals. Zhou et al. (2020), Petersen and Hansen (2021), Cai et al. (2022) estimate the conditional cumulative distribution functions, and test for independence of the partial copulas. This may be extended to ordinal X and Y using a continuous version of the cumulative distribution function (Brockwell, 2007).

As noted in Shah and Peters (2020), when Z takes values in a finite set one may test for conditional independence by dividing the data up depending on the value of Z, and performing marginal independence tests between X and Y on each subset. Li and

Shepherd (2010) discuss some methods when X and Y are ordinal. Alternative strategies exist, with Marx and Vreeken (2019) instead estimating a Kullback-Leibler divergence between the conditional independence null and alternative distributions (although they do not formally calibrate this). When Z is continuous one option is to divide the state space into strata, reducing to the finite case. These strata must be small enough so that (X, Y) within each strata have similar distributions conditional on Z, but such tests lose power if there are too many strata. Calibration can also be problematic for the marginal independence tests if some of the strata contain relatively few data points.

Bootstrap methods (Efron, 1979) have been used to calibrate a wide range of statistical estimators and tests (Efron and Tibshirani, 1994), with early theoretical work including Beran (1986); Bickel and Freedman (1981); Hall (1992); Romano (1988); Singh (1981). Nonparametric bootstraps are known to be inconsistent in some nonregular problems, e.g. Beran (1997); Samworth (2003); Shao (1994, 1996). Krinsky and Robb (1986) motivate the use of parametric bootstaps to do inference on non-linear functionals of asymptotically Gaussian random variables, for which the delta method (based on a linear approximation) performs poorly, although they do not prove theory. Similarly our test statistic asymptotically follows a piecewise-continuous transformation of a multivariate Gaussian distribution, which allows us to perform valid inference via a parametric bootstrap. There has been recent interest in doing bootstrap inference in high-dimensional regimes, notably Chernozhukov et al. (2013). See Chernozhukov et al. (2023a) for a review of high-dimensional bootstraps.

3.1.3 Notation

Let $\mathcal{M}^{(d_X, d_Y)} \subset \mathbb{R}^{d_X d_Y \times d_X d_Y}$ be the set of symmetric positive semidefinite matrices on $\mathbb{R}^{d_X d_Y \times d_X d_Y}$. We understand inequalities between vectors to apply elementwise. For notational convenience, we write $\tilde{X}$ and $\tilde{Y}$ for the one-hot encodings

$$\tilde{X} := \Big(\mathbb{1}\{X = 1\}, \ldots, \mathbb{1}\{X = d_X\}\Big) \in \{0, 1\}^{d_X};$$
$$\tilde{Y} := \Big(\mathbb{1}\{Y = 1\}, \ldots, \mathbb{1}\{Y = d_Y\}\Big) \in \{0, 1\}^{d_Y}.$$

Note that $\sum_{j=1}^{d_X} \tilde{X}_j = \sum_{k=1}^{d_Y} \tilde{Y}_k = 1$ almost surely.

As in Lundborg et al. (2022), given a family of sequences of real-valued random variables $(W_{P,n})_{P \in \mathcal{P}, n \in \mathbb{N}}$ taking values in a finite-dimensional vector space and whose distributions are determined by $P \in \mathcal{P}$, we write $W_{P,n} = o_{\mathcal{P}}(1)$ if $\sup_{P \in \mathcal{P}} \mathbb{P}_P(|W_{P,n}| > \epsilon) \to 0$ for every $\epsilon > 0$. Similarly, we write $W_{P,n} = O_{\mathcal{P}}(1)$ if, for any $\epsilon > 0$, there exist $M_\epsilon, N_\epsilon > 0$ such that $\sup_{n \geq N_\epsilon} \sup_{P \in \mathcal{P}} \mathbb{P}_P(|W_{P,n}| > M_\epsilon) < \epsilon$. Given a second family of sequences of random variables $(V_{P,n})_{P \in \mathcal{P}, n \in \mathbb{N}}$, we write $W_{P,n} = o_{\mathcal{P}}(V_{P,n})$ if there exists $R_{P,n}$ with $W_{P,n} = V_{P,n} R_{P,n}$ and $R_{P,n} = o_{\mathcal{P}}(1)$; likewise, we write $W_{P,n} = O_{\mathcal{P}}(V_{P,n})$

if $W_{P,n} = V_{P,n}R_{P,n}$ and $R_{P,n} = O_{\mathcal{P}}(1)$. If $W_{P,n}$ is vector or matrix-valued, we write $W_{P,n} = o_{\mathcal{P}}(1)$ if $\|W_{P,n}\| = o_{\mathcal{P}}(1)$ for some norm, and similarly $O_{\mathcal{P}}(1)$. By the equivalence of norms for finite-dimensional vector spaces, if this holds for some norm then it holds for all norms.

3.2 Structured categorical conditional independence testing via greedy label merging

Our adaptive conditional independence test works as follows. We use the data (X, Y, Z) to form an asymptotically Gaussian estimate $\hat{T}$ of (3.1), and corresponding covariance estimate $\hat{\Sigma}$. This means that $\hat{T}$ is mean-zero under the null $X \perp\!\!\!\perp Y \mid Z$. We pick a function $\phi : \bigcup_{d=1}^{\infty} \mathbb{R}^d \times \mathbb{R}^{d\times d} \to \mathbb{R}$ which will be used to quantify deviation from centrality, with larger outputs representing larger deviations from centrality of the input. A straightforward approach would be to use some norm of $\hat{T}$, such as the Euclidean norm $q : (t, \sigma) \mapsto \|t\|_2$. These do not make use of the covariance estimate, so we expect the cumulative distribution of some norm of $\hat{T}$ to have better properties. Consideration must also be given to the computational efficiency, in light of the stepwise procedure we describe next. We later give our recommended choice in (3.9).

Taking $\hat{T}$, $\hat{\Sigma}$, and ϕ we next define a sequence of partitions

$$\left(A^{(1)}, B^{(1)}\right), \left(A^{(2)}, B^{(2)}\right), \ldots, \left(A^{(L)}, B^{(L)}\right)$$

of $\{1, \ldots, d_X\}$ and $\{1, \ldots, d_Y\}$ respectively for some $L \in \mathbb{N}$. We do this in a greedy stepwise fashion, which we describe below. When there are no restrictions on the permissible merges we keep merging sequentially until both sets of labels are partitioned in two, leading to $L = d_X + d_Y - 3$. Note that the partitioned generalised covariance (3.2) is a linear transformation of the raw generalised covariance (3.1), and we write $\Pi^{(A,B)}$ for the corresponding matrix. We take $A^{(1)}$ and $B^{(1)}$ to be the singleton partitions. At each step $s = 2, \ldots, L$, if $|A^{(s)}|, |B^{(s)}| > 2$ we take as candidates the partitions

$$\left(A^{(s)}, \left\{B^{(s)}_{m_1} \cup B^{(s)}_{m_2}\right\} \cup \left\{B^{(s)}_{m} \ : \ m \notin \{m_1, m_2\}\right\}\right)$$

for $1 \leq m_1 < m_2 \leq |B^{(s)}|$ and also

$$\left(\left\{A^{(s)}_{l_1} \cup A^{(s)}_{l_2}\right\} \cup \left\{A^{(s)}_{l} \ : \ l \notin \{l_1, l_2\}\right\}, B^{(s)}\right)$$

for $1 \leq l_1 < l_2 \leq |B^{(s)}|$. If $|A^{(s)}|$ or $|B^{(s)}|$ equals 2 we fix that partition (so only consider merging the other), and if both partitions are of size 2 then we stop (L is assumed to be

such that we stop after the Lth step, if at all). We set $\big(A^{(s+1)}, B^{(s+1)}\big)$ to be the candidate which maximises the criterion

$$\phi\Big(\Pi^{(A,B)}\hat{T}, \Pi^{(A,B)}\hat{\Sigma}\Pi^{(A,B)T}\Big).$$

In this way we have defined a mapping

$$q : (t, \sigma) \mapsto \left(\phi\Big(\Pi^{(A^{(s)},B^{(s)})}t, \Pi^{(A^{(s)},B^{(s)})}\sigma\Pi^{(A^{(s)},B^{(s)})T}\Big) \ : \ s = 1, \dots, L\right),$$

where we understand that each $\big(A^{(s)}, B^{(s)}\big)$ is itself a function of (t, σ). Given structural information about X and Y we may choose to restrict the permissible choices for $\big(A^{(s)}, B^{(s)}\big)$, which we discuss in Section 3.2.2. Write $\hat{M} := q\big(\hat{T}, \hat{\Sigma}\big) \in \mathbb{R}^L$. We wish to test each element of $\hat{M}$ for significance against its own population distribution.

Treating $\hat{\Sigma}$ as fixed, we introduce a centred Gaussian $T \overset{d}{:=} N\big(0, \hat{\Sigma}\big)$, and draw independent copies of T to form a parametric bootstrap sample $T^{(1)}, \dots, T^{(B)}$. Suppose the null $X \perp\!\!\!\perp Y \mid Z$ is true. Under relatively weak conditions, the bootstrap samples should have the same asymptotic distribution as the observed vector $\hat{T}$. Let $M^{(b)} := q\big(T^{(b)}, \hat{\Sigma}\big)$, recalling that the sequence of partitions selected depends on the value of $T^{(b)}$, so needs not be the same. Under weak assumptions the transformed sample $\{M_l^{(b)} \ : \ b = 1, \dots, B\}$ should have the same asymptotic distribution as $\hat{M}_l$, for each $l = 1, \dots, L$. We set $\hat{F}_l$ to be the bootstrap empirical cumulative distribution function for $\hat{M}_l$, and take $\hat{G} := \max_{l=1,\dots,L} \hat{F}_l$. At this point we could use a Bonferroni correction to calibrate the $\hat{G}$, considering $1 - \hat{F}_1, \dots, 1 - \hat{F}_L$ to be p-values from L separate tests. This can be overly conservative, so instead we re-use the bootstrap sample to compute bootstrap versions $F_l^{(b)}$ and $G^{(b)}$, and compute the bootstrap p-value for $\hat{G}$.

In the rest of this section we make precise our test (Algorithm 4), and prove that our p-value has the appropriate distribution uniformly within a class of null distributions $P \in \mathcal{P}$ (Theorem 29). We include additional results for our calibration procedure (Algorithm 3, Theorem 30) and our centrality test (Algorithm 5, Theorem 31) in Section 3.5.

3.2.1 Reduction to location testing

Following the hardness result of Shah and Peters (2020), we restrict attention to distributions P for which we can estimate the conditional probabilities

$$f_{P,j}(z) := \mathbb{P}_P(X = j \mid Z = z) = \mathbb{P}_P(\tilde{X}_j = 1 \mid Z = z), \quad j = 1, \dots, d_X;$$
$$g_{P,k}(z) := \mathbb{P}_P(Y = k \mid Z = z) = \mathbb{P}_P(\tilde{Y}_k = 1 \mid Z = z), \quad k = 1, \dots, d_Y,$$

at some reasonable rates (Assumption 1). Since X and Y are discrete, P satisfies the null hypothesis $X \perp\!\!\!\perp Y \mid Z$ if and only if

$$\mathbb{P}_P(X = j, Y = k \mid Z = z) = f_{P,j}(z) g_{P,k}(z)$$

for almost every (j, k, z), i.e. $\mathrm{Cov}_P\Big(\tilde{X}_j, \tilde{Y}_k \mid Z\Big)$ is almost surely zero. Let $\mu_P \in \mathbb{R}^{d_X d_Y}$ have entries

$$\mu_{P,j,k} := \mathbb{E}_P\Big[\mathrm{Cov}_P\Big(\tilde{X}_j, \tilde{Y}_k \mid Z\Big)\Big], \tag{3.3}$$

where we have indexed $d_X d_Y$ dimensional space as

$$(1,1), \ldots, (1, d_Y), (2,1), \ldots, (2, d_Y), \ldots, (d_X, 1), \ldots, (d_X, d_Y).$$

If P satisfies the null hypothesis $X \perp\!\!\!\perp Y \mid Z$ we must have $\mu_P = 0$. One may also consider the weighted version $\mathbb{E}_P\Big[\mathrm{Cov}_P\Big(\tilde{X}_j, \tilde{Y}_k \mid Z\Big) w(Z)\Big]$, for some $w : \mathcal{Z} \to \mathbb{R}$. This equals zero under the null for any choice of w, but can have power in settings where our method does not (Scheidegger et al., 2022).

In our estimation of μ_P we make use of cross-fitting (Chernozhukov et al., 2018), however sample splitting is not always necessary. Given a sequence of i.i.d. data sets

$$D^{(n)} := \{(X_i, Y_i, Z_i) \ : \ i = 1, \ldots, n\},$$

define an N-fold partition $\Big(I^{(n,r)}\Big)_{r=1,\ldots,N}$ of $\{1, \ldots, n\}$ for some N fixed (practically we take $N = 5$). For simplicity, we assume that n is a multiple of N and each subset is of equal size n/N. Let the propensity score estimates $\Big\{\hat{f}_j^{(n,r)}, \hat{g}_k^{(n,r)} \ : \ j = 1, \ldots, d_X, k = 1, \ldots, d_Y\Big\}$ be estimated using data

$$D^{(n,r)} := \Big\{(y_i, x_i, z_i) \ : \ i \in \{1, \ldots, n\} \setminus I^{(n,r)}\Big\}.$$

We have found gradient boosting multinomial regressions (`xgboost` package (Chen and Guestrin, 2016)) to do well in practice. For each data point define one-hot encodings

$$\begin{aligned} \tilde{X}_{ij} &:= \mathbb{1}\{X_i = j\}, \quad j = 1, \ldots, d_X; \\ \tilde{Y}_{ik} &:= \mathbb{1}\{Y_i = k\}, \quad k = 1, \ldots, d_Y. \end{aligned}$$

The cross-fitted, doubly-robust generalised covariance estimate is $\hat{T}^{(n)} = \Big(\hat{T}^{(n)}_{j,k}\Big) \in \mathbb{R}^{d_X d_Y}$, where

$$\hat{T}^{(n)}_{j,k} := \frac{1}{n} \sum_{r=1}^{N} \sum_{i \in I^{(n,r)}} \Big\{\tilde{X}_{ij} - \hat{f}_j^{(n,r)}(Z_i)\Big\}\Big\{\tilde{Y}_{ik} - \hat{g}_k^{(n,r)}(Z_i)\Big\}, \tag{3.4}$$

with corresponding variance estimator $\hat{\Sigma}^{(n)} = \big(\hat{\Sigma}^{(n)}_{(j,k)(j',k')}\big) \in \mathbb{R}^{d_X d_Y \times d_X d_Y}$,

$$\hat{\Sigma}^{(n)}_{(j,k)(j',k')} := \frac{1}{n}\sum_{r=1}^{N}\sum_{i \in I^{(n,r)}} \Big[\Big\{\tilde{X}_{ij} - \hat{f}_j^{(n,r)}(Z_i)\Big\}\Big\{\tilde{Y}_{ik} - \hat{g}_k^{(n,r)}(Z_i)\Big\} - \hat{T}^{(n)}_{j,k}\Big] \\ \Big[\Big\{\tilde{X}_{ij'} - \hat{f}_{j'}^{(n,r)}(Z_i)\Big\}\Big\{\tilde{Y}_{ik'} - \hat{g}_{k'}^{(n,r)}(Z_i)\Big\} - \hat{T}^{(n)}_{j',k'}\Big]. \quad (3.5)$$

Additionally, define $\Sigma_P \in \mathbb{R}^{d_X d_Y \times d_X d_Y}$ to be the population covariance matrix, which has elements

$$\Sigma_{P,(j,k)(j',k')} = \mathrm{Cov}_P\Big[\Big\{\tilde{X}_j - f_{P,j}(Z)\Big\}\Big\{\tilde{Y}_k - g_{P,k}(Z)\Big\}, \Big\{\tilde{X}_{j'} - f_{P,j'}(Z)\Big\}\Big\{\tilde{Y}_{k'} - g_{P,k'}(Z)\Big\}\Big]. \quad (3.6)$$

If P satisfies certain weak assumptions, then $\hat{T}^{(n)}$ converges in distribution to a Gaussian vector with mean μ_P and covariance matrix Σ_P, and additionally $\hat{\Sigma}^{(n)}$ converges in probability to Σ_P.

Remark 3.2.1. *Any other test statistic which is central under the null and has a consistent covariance estimator may be used, such as the Generalised Covariance Measure (Shah and Peters, 2020). We have found the additional normalisation step therein to sometimes yield worse finite-sample properties (see Figure 3.1).*

3.2.2 Greedy label merging

Let $A = (A_1, \ldots, A_{d'_X})$ be a partition of $\{1, \ldots, d_X\}$, with $|A| = d'_X \leq d_X$, and let $(B_1, \ldots, B_{d'_Y})$ be a similar partition of $\{1, \ldots, d_Y\}$. Consider the linear operation from $\mathbb{R}^{d_X d_Y} \to \mathbb{R}^{d'_X d'_Y}$ defined by

$$\Big(x_{j,k} \,:\, j = 1, \ldots, d_X, k = 1, \ldots, d_Y\Big) \\ \mapsto \left(\sum_{j \in A_l}\sum_{k \in B_m} x_{j,k} \,:\, l = 1, \ldots, d'_X,\ m = 1, \ldots, d'_Y\right),$$

which we have already seen in the discussion around (3.2). This transformation may be represented using a matrix $\Pi^{(A,B)} \in \{0,1\}^{d'_X d'_Y \times d_X d_Y}$ defined by

$$\Pi^{(A,B)}_{(l,m)(j,k)} = \mathbb{1}\{j \in A_l,\ k \in B_m\}. \quad (3.7)$$

Write

$$\mathcal{C}_{(d_X,d_Y)(d'_X,d'_Y)} := \Big\{\Pi^{(A,B)} \,:\, A \text{ a } d'_X\text{-partition of } \{1, \ldots, d_X\}, \\ B \text{ a } d'_Y\text{-partition of } \{1, \ldots, d_Y\}\Big\}$$

for the set of such transformation matrices, for $d'_X, d'_Y \geq 2$.

Recall that we use a sequence of nested partitions to test if a random vector has mean zero at various scales. We select these partitions in a greedy fashion, maximising the observed deviation from centrality at each stage. We incorporate structural knowledge of X and Y by restricting the search space for possible merges, which improves both power and run-time. If X is ordinal then we would only consider merging subsets of labels $j_1 < j_2$ where $j_2 = j_1 + 1$. If the labels of X have some hierarchical structure, then we consider the labels to be leaves on some tree. We only consider merges between siblings at each stage. One may also wish to restrict the number of labels being merged, either my fixing a maximum group size or a total number of merges. Denote the combined structural assumptions on the labels of X and Y by an element $\mathbb{S}$ in a space $\mathcal{S}$, which may be anything. Our theoretical results do not depend on $\mathbb{S}$, so are robust to misspecification. Indeed, in Section 3.3 we consider misspecification of $\mathbb{S}$, and find that our procedure still performs well.

Remark 3.2.2. *One may think of $\mathcal{S}$ as the set of all permissible partitions (A, B), which takes values in a finite set.*

We seek a measure of deviation from centrality amongst Gaussian vectors (we will discuss calibrating this into a formal test later). One such is the chi-square cumulative distribution function

$$(t, \sigma) \mapsto \mathbb{P}\Big(\chi^2_{\text{rank}(\sigma)} \leq t^T \sigma^+ t\Big), \tag{3.8}$$

where σ is a (non-zero) covariance matrix and σ^+ denotes its generalised inverse. When $T \sim N(0, \Sigma)$ for Σ positive semidefinite with rank $r > 0$, the random variable $T^T \Sigma^+ T$ is equal in distribution to a chi-square distribution with r degrees of freedom, denoted χ^2_r. Equation (3.8) maps (T, Σ) to the probability that an independently sampled χ^2_r variable is less than or equal to the observed $T^T \Sigma^+ T$, with larger probabilities being evidence that the true mean of T is non-zero. Inverting estimated covariance matrices is numerically both expensive and unstable, and the naive test based on (3.8) has very poor finite sample properties (see Figure 3.1). We instead work with an approximation to (3.8) due to Box (1954),

$$\phi : (t, \sigma) \mapsto \mathbb{P}\Big(g(\sigma)\chi^2_{h(\sigma)} \leq \|t\|_2^2\Big), \tag{3.9}$$

where $g(\sigma) = \text{tr}(\sigma^2)/\text{tr}(\sigma)$ and $h(\sigma) = \text{tr}(\sigma)^2/\text{tr}(\sigma^2)$. This has additional attractive properties from the perspective of our adaptive test, which enables a fast implementation (see Section 3.6.1).

We understand ϕ as a function acting on vectors and non-zero covariance matrices of arbitrary dimension,

$$\phi : \bigcup_{d_X, d_Y \geq 2} \mathbb{R}^{d_X d_Y} \times \left(\mathcal{M}^{(d_X, d_Y)} \setminus \{0\} \right) \to [0, 1].$$

In Algorithm 2 we detail our greedy merging query function, which we use to map $\left(\hat{T}^{(n)}, \hat{\Sigma}^{(n)} \right)$ to a vector of criteria $\hat{M}^{(n)} \in \mathbb{R}^L$ for some $L \in \mathbb{N}$. When there are no restrictions on the permissible merges we keep merging until $d_X = d_Y = 2$, leading to $L = d_X + d_Y - 3$. The version we give here is not intended to be efficient, but rather to make clear our desired output. See Algorithm 6 for our actual implementation.

Input: Statistic vector $T \in \mathbb{R}^{d_X d_Y}$, covariance matrix $\Sigma \in \mathcal{M}^{(d_X, d_Y)}$, X-dimension d_X, Y-dimension d_Y, structural information $\mathbb{S} \in \mathcal{S}$.

Output: Vector of criteria $M \in [0, 1]^L$, with larger values corresponding to larger deviations from centrality.

Set $M_1 = \phi(T, \Sigma)$.

while *continuing merging is consistent with the structure* $\mathbb{S}$ **do**

  for *potential pairs* $1 \leq j_1 < j_2 \leq d_X$ *of* X*-labels to merge* **do**

    if *merging* (j_1, j_2) *is consistent with the structure* $\mathbb{S}$ **then**

      Set $A = \left\{ \{1\}, \ldots, \{j_1 - 1\}, \{j_1, j_2\}, \ldots, \{j_2 - 1\}, \{j_2 + 1\}, \ldots, \{d_X\} \right\}$ and $B = \left\{ \{1\}, \ldots, \{d_Y\} \right\}$;

      Compute $T_X^{(j_1, j_2)} := \Pi^{(A,B)} T$ and $\Sigma_X^{(j_1, j_2)} := \Pi^{(A,B)} \Sigma \Pi^{(A,B)T}$.

    end

  end

  for *potential pairs* $1 \leq k_1 < k_2 \leq d_Y$ *of* Y*-labels to merge* **do**

    if *merging* (k_1, k_2) *is consistent with the structure* $\mathbb{S}$ **then**

      Set $B = \left\{ \{1\}, \ldots, \{k_1 - 1\}, \{k_1, k_2\}, \ldots, \{k_2 - 1\}, \{k_2 + 1\}, \ldots, \{d_Y\} \right\}$ and $A = \left\{ \{1\}, \ldots, \{d_X\} \right\}$;

      Compute $T_Y^{(k_1, k_2)} := \Pi^{(A,B)} T$ and $\Sigma_Y^{(k_1, k_2)} := \Pi^{(A,B)} \Sigma \Pi^{(A,B)T}$.

    end

  end

  Set $M_l = \max \left\{ \phi\left(T_X^{(j_1, j_2)}, \Sigma_X^{(j_1, j_2)} \right), \phi\left(T_Y^{(k_1, k_2)}, \Sigma_Y^{(k_1, k_2)} \right) : \right.$ $\left.$ permitted (j_1, j_2) and $(k_1, k_2) \right\}$;

  Update (T, Σ) to be the maximising arguments;

  Update d_X, d_Y, and $\mathbb{S}$ to be consistent with the new partitions.

end

Algorithm 2: Greedy merging query function.

3.2.3 Bootstrap calibration

In Section 3.2.2 we discuss transforming a vector with unknown mean into a sequence of criteria, with larger elements representing deviations from centrality. It remains to calibrate this into a formal test, which we do in a manner which simultaneously tests each of the resulting criteria individually. We make use of a parametric bootstrap, which we use twice to control for the multiple testing. In this way we do not lose as much power as, say, a Bonferonni correction.

Given an observed random vector of criteria $\hat{M}^{(n)}$ and bootstrap versions $M_n^{(1)}, \ldots, M_n^{(B)}$ all in $\mathbb{R}^L$, we define a p-value in $[0, 1]$ as follows. We first use the empirical bootstrap sample to estimate the marginal cumulative distribution function $\hat{F}_l$, $l = 1, \ldots, L$, of each of the elements of $\hat{M}^{(n)}$. This tells us how extreme each of the observed criteria are (note that the criteria need not have the same scaling). We then take the maximum of these as our test statistic for the second stage, which we calibrate by reusing the original bootstrap sample. Details are given in Algorithm 3. In practice, Algorithm 3 can have problems when L is not small compared to B, in which case several of the $G_n^{(b)}$ take the same value in $\{1/B, 2/B, \ldots, 1\}$. There are various ways one could deal with ties. We suggest using a continuous version of the empirical cumulative distribution function, see Algorithm 7.

Input: Test vector $\hat{M}^{(n)} \in \mathbb{R}^L$, number of bootstrap samples $B \in \mathbb{N}$, bootstrap sample $M_n^{(1)}, \ldots, M_n^{(B)} \in \mathbb{R}^L$.

Output: P-value $p \in [0, 1]$.

for $l = 1, \ldots, L$ **do**
  Set $\hat{F}_{n,l} = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}\left\{M_{n,l}^{(b)} \leq \hat{M}_l^{(n)}\right\}$;
end
Set $\hat{G}_n = \max_{l=1,\ldots,L} \hat{F}_{n,l}$.
for $b' = 1, \ldots, B$ **do**
  for $l = 1, \ldots, L$ **do**
    Set $F_{n,l}^{(b')} = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}\left\{M_{n,l}^{(b)} \leq M_{n,l}^{(b')}\right\}$;
  end
  Set $G_n^{(b')} = \max_{l=1,\ldots,L} F_{n,l}^{(b')}$.
end
Set $p = 1 - \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}\left\{G_n^{(b)} \leq \hat{G}_n\right\}$.

Algorithm 3: Bootstrap calibration. The first stage quantities $\hat{F}_{n,1}, \ldots, \hat{F}_{n,L}$ quantify how extreme each of the observed criteria are. We take the maximum of these — $\hat{G}_n$ — as our test statistic for the second stage, which we calibrate by reusing the empirical bootstrap samples.

3.2.4 Conditional independence test and asymptotic properties

We are now familiar with all the elements of our conditional independence test for structured categorical X and Y given arbitrary Z. Our full procedure is Algorithm 4. Theorem 29 provides a guarantee on the asymptotic validity of our test.

Input: Independent, identically distributed data $D^{(n)} = \{(X_i, Y_i, Z_i) \,:\, i = 1, \dots, n\}$, number of bootstrap samples $B \in \mathbb{N}$, structural information $\mathbb{S} \in \mathcal{S}$.
Output: P-value $p \in [0, 1]$.
Set $\hat{T}^{(n)}$ as the cross-fitted, doubly-robust generalised covariance measure (3.4) and $\hat{\Sigma}^{(n)}$ as the corresponding covariance estimate (3.5).
Compute criterion $\hat{M}^{(n)} = q\Big(\hat{T}^{(n)}, \hat{\Sigma}^{(n)}, d_X, d_Y, \mathbb{S}\Big)$, where q is the output of Algorithm 2.
for $b = 1, \dots, B$ **do**
  Draw parametric bootstrap $T_n^{(b)} \sim N\Big(0, \hat{\Sigma}_n\Big)$, for example using Algorithm 8.
  Compute bootstrap criterion $M_n^{(b)} = q\Big(T_n^{(b)}, \hat{\Sigma}^{(n)}, d_X, d_Y, \mathbb{S}\Big)$.
end
Set $p = p\Big(\hat{M}^{(n)}, B, M_n^{(1)}, \dots, M_n^{(B)}\Big)$, where p is the output of Algorithm 3.
Algorithm 4: Our conditional independence test for structured categorical data.

We now introduce the assumptions we require for Theorem 29.

Assumption 1. *Define the following sequences of random variables:*

$$E_f^{(n)} := \max_{j=1,\dots,d_X} \mathbb{E}_P\Big[\{f_{P,j}(Z) - \hat{f}_j^{(n,1)}(Z)\}^2 \;\Big|\; D^{(n,1)}\Big],$$
$$E_g^{(n)} := \max_{k=1,\dots,d_Y} \mathbb{E}_P\Big[\{g_{P,k}(Z) - \hat{g}_k^{(n,1)}(Z)\}^2 \;\Big|\; D^{(n,1)}\Big],$$

note that we have suppressed P-dependence in the quantities defined above. Let $\mathcal{P}$ be such that all of the following hold. The covariance matrix Σ_P defined in (3.6) *exists for every $P \in \mathcal{P}$. The remainder terms defined above satisfy:*

$$E_f^{(n)} = o_{\mathcal{P}}(1); \quad E_g^{(n)} = o_{\mathcal{P}}(1); \quad E_f^{(n)} E_g^{(n)} = o_{\mathcal{P}}(n^{-1}). \tag{3.10}$$

The assumptions on $E_f^{(n)}$ and $E_g^{(n)}$ are relatively weak and standard; for example they are satisfied if each of $E_f^{(n)}$, $E_g^{(n)}$ converge at the nonparametric rate $o_{\mathcal{P}}(n^{-1/2})$. For example, consider the case where $\mathcal{Z} = \mathbb{R}^p$ and each $f_{P,j}$ is $s > 0$ Hölder smooth, i.e., writing $m := \lceil s \rceil - 1$, for every $\alpha := (\alpha_1, \dots, \alpha_p)$ with $\alpha_1 + \cdots + \alpha_p = m$ and $\alpha_j \in \mathbb{Z}_{\geq 0}$, the partial derivatives (assumed to exist) satisfy

$$\left| \frac{\partial^\alpha f_{P,j}}{\partial^{\alpha_1} z_1 \cdots \partial^{\alpha_p} z_p}(z) - \frac{\partial^\alpha f_{P,j}}{\partial^{\alpha_1} z_1 \cdots \partial^{\alpha_p} z_p}(z') \right| \leq C \|z - z'\|_2^{s-m}$$

for all $P \in \mathcal{P}$ and $z, z' \in \mathbb{R}^p$. Then we can expect that $E_f^{(n)} = O_{\mathcal{P}}(n^{-2s/(2s+p)})$ for appropriately chosen regression procedures; see for example Györfi et al. (2002). Then when $s > p/2$, this is $o_{\mathcal{P}}(n^{-1/2})$. Moreover, a faster rate for $E_f^{(n)}$ permits a slower rate for $E_g^{(n)}$ and vice versa.

Recall that $\sum_{j=1}^{d_X} \tilde{X}_j = \sum_{k=1}^{d_Y} \tilde{Y}_k = 1$ almost surely. This implies that $\ker(\Sigma) \subset \mathbb{R}^{d_X d_Y}$ contains at least the $(d_X + d_Y - 1)$-dimensional space spanned by

$$u^{(j)}_{j',k'} := \begin{cases} 1 & \text{if } j' = j; \\ 0 & \text{otherwise;} \end{cases} \tag{3.11}$$

$$v^{(k)}_{j',k'} := \begin{cases} 1 & \text{if } k' = k; \\ 0 & \text{otherwise.} \end{cases} \tag{3.12}$$

Write $\mathcal{K}^{(d_X,d_Y)} := \operatorname{span}\{u^{(j)}, v^{(k)} \ : \ j = 1, \ldots, d_X, \ k = 1, \ldots, d_Y\} \subset \mathbb{R}^{d_X d_Y}$.

Assumption 2. *For each $P \in \mathcal{P}$, the population covariance matrix Σ_P defined in* (3.6) *is equal to some fixed Σ which satisfies* $\ker(\Sigma) = \mathcal{K}^{(d_X,d_Y)}$*, i.e. the kernel of Σ is as small as possible given the categorical nature of X and Y.*

The kernel condition in Assumption 2 is equivalent to asking for the upper $(d_X - 1)(d_Y - 1)$ block of the covariance matrix Σ to be positive definite, which is relatively weak and standard. One may relax the assumption that Σ is fixed under stronger assumptions on the other parts of our procedure, see Remark 3.5.1.

Theorem 29. *Let $\mathcal{P} \subset \{P \ : \ X \perp\!\!\!\perp Y \mid Z\}$ be the class of null distributions satisfying Assumptions 1 and 2. Let $p : (\{1, \ldots, d_X\} \times \{1, \ldots, d_Y\} \times \mathcal{Z})^n \times \mathbb{N} \times \mathcal{S} \to [0, 1]$ be the output of Algorithm 4. Then the random variable $p\Big(D^{(n)}, B, \mathbb{S}\Big)$ is asymptotically uniformly distributed, uniformly in the class $\mathcal{P}$:*

$$\lim_{n,B\to\infty} \sup_{P\in\mathcal{P}} \sup_{u\in[0,1]} \left| \mathbb{P}_P\left(p\Big(D^{(n)}, B, \mathbb{S}\Big) \leq u \right) - u \right| = 0.$$

The result of Theorem 29 is independent of the structural information $\mathbb{S}$. Practitioners may choose any structure that they believe will improve the power of the test over their data (see discussion in Section 3.2.2), without worrying if this is "correctly specified" in any sense. Proving asymptotic power guarantees is beyond the scope of this chapter. In order to derive the asymptotic properties of our test in Theorem 29 we make use of the fact that the bootstrap samples are asymptotically equivalent to the test statistic. In this way we avoid having to derive the probabilities of each possible sequence of label merges, which would be tedious and depend on the structure $\mathbb{S}$. This trick doesn't apply

when considering objects with different limiting distributions, for instance when the null hypothesis is violated. We instead demonstrate the power of our test through a numerical study.

3.3 Numerical experiments

We demonstrate that our adaptive conditional independence test for structured categorical data is able to control size and improve power in a wide range of settings, under both hierarchical and ordinal assumptions. We additionally show that when the structural assumptions are misspecified, our greedy approach does no worse than corresponding non-adaptive approaches. We are primarily interested in our adaptive test Algorithm 4, which uses the greedy merging query function Algorithm 2, and where we assume either a hierarchical or ordinal structure on X and Y. We compare to two similar but non-adaptive approaches, which fix $L = 1$ and set $q : (t, \sigma) \mapsto \|t\|_2$ (Euclidean norm) and $\|t\|_\infty$ (max norm) respectively. Note that the Euclidean criterion is equivalent to the approximate chi-square value $q : (t, \sigma) \mapsto \phi(t, \sigma)$ as in (3.9) (i.e. our approach with search depth $L = 1$), since the latter is a strictly increasing function of the former (for fixed σ). We additionally compare to the multivariate Generalised Covariance Measure (mGCM) (Shah and Peters, 2020), and in the ordinal case also the conditional independence test of Ankan and Textor (2022). While our test may be used with any plug-in machine learning multinomial regressions, here we make use of gradient boosting for its good predictive power (`xgboost` package (Chen and Guestrin, 2016)). The precise implementation details are given in Section 3.6. For a sanity check we also include the chi-square test (3.8) — which is expected to do poorly when the dimension is not very small with respect to the sample size — and a linear multinomial regression of Y on X and Z, which is expected to do poorly in general. Code to reproduce our experiments is made available in the R package `catci` (CATegorical Conditional Independence) available from https://github.com/harveyklyne/catci.

In all cases we fix $n = 1000$ and $d_X = d_Y = 8$. We draw $Z \sim N(0, S) \in \mathbb{R}^5$, where $S_{jj} = 1$, $S_{jk} = 0.5$ for $j \neq k$, and then draw (X, Y) conditionally on Z from a known joint probability mass function. Define the following propensity functions $p : \mathbb{R} \to [0, 1]^8$,

satisfying $\sum_{j=1}^{8} p_j(z) = 1$ for every $z \in \mathbb{R}$:

$$p_j^{(lin)}(z) := \frac{2j + 5 + (-4j + 18)w_1(z)}{112}; \tag{3.13}$$

$$p_j^{(vee)}(z) := \frac{-|2j - 9| + 10 + \big(2|2j - 9| - 8\big)w_1(z)}{48}; \tag{3.14}$$

$$p_j^{(hat)}(z) := \frac{2\mathbb{1}\{3 \le j \le 6\} + 1 + \big(-4\mathbb{1}\{3 \le j \le 6\} + 2\big)w_1(z)}{16}; \tag{3.15}$$

$$p_j^{(sig)}(z) := \Phi\left(\frac{c_{j+1}^{(sig)} - w_1(z)}{2}\right) - \Phi\left(\frac{c_j^{(sig)} - w_1(z)}{2}\right); \tag{3.16}$$

$$p_j^{(sin)}(z) := \Phi\left(\frac{c_{j+1}^{(sin)} - w_2(z)}{2}\right) - \Phi\left(\frac{c_j^{(sin)} - w_2(z)}{2}\right). \tag{3.17}$$

Here, $w_1(z) := (1 + \exp(-3z))^{-1}$ and $w_2(z) := \exp\big(-z^2/2\big)\sin(z)$ are weighting functions, Φ is the standard Gaussian cumulative distribution function, and $c^{(sig)}$ and $c^{(sin)}$ are the following vectors in $(\mathbb{R} \cup \{\pm\infty\})^9$:

$$c^{(sig)} := \big(-\infty \quad -2.0 \quad -1.0 \quad 0.0 \quad 0.5 \quad 1.0 \quad 2.0 \quad 3.0 \quad +\infty\big);$$
$$c^{(sin)} := \big(-\infty \quad -2.3 \quad -1.2 \quad -0.6 \quad 0.0 \quad 0.6 \quad 1.2 \quad 2.3 \quad +\infty\big).$$

These propensity functions have been chosen so that all label probabilities are bounded away from zero, and so as to vary in complexity in both z and j.

For the size control experiments, we draw X and Y given Z from the product of the probability mass functions,

$$\mathbb{P}(X = j, Y = k \mid Z = z) = p_j^{(X)}(z_1)p_k^{(Y)}(z_2), \tag{3.18}$$

for choices of $p^{(X)}$ and $p^{(Y)}$ in (3.13–3.17). Note that the elements Z_1 and Z_2 — and hence X and Y — are correlated, but that X and Y are conditionally independent given Z.

In Figure 3.1 we plot the empirical quantiles of the various tests' p-values (through 1000 repeats) against those of the target uniform distribution. If the methods are well-calibrated, we expect the corresponding plots to be linear. We find that our adaptive test Algorithm 4 is correctly calibrated in all settings, as are the non-adaptive versions with Euclidean and max norms query functions. We further find that the method of Ankan and Textor (2022) is well-calibrated. As expected, the naive chi-square and multinormial tests do not adequately control size. We were surprised that the mGCM also failed here, particularly in light of Shah and Peters (2020, Thm. 9). The mGCM is equivalent to a

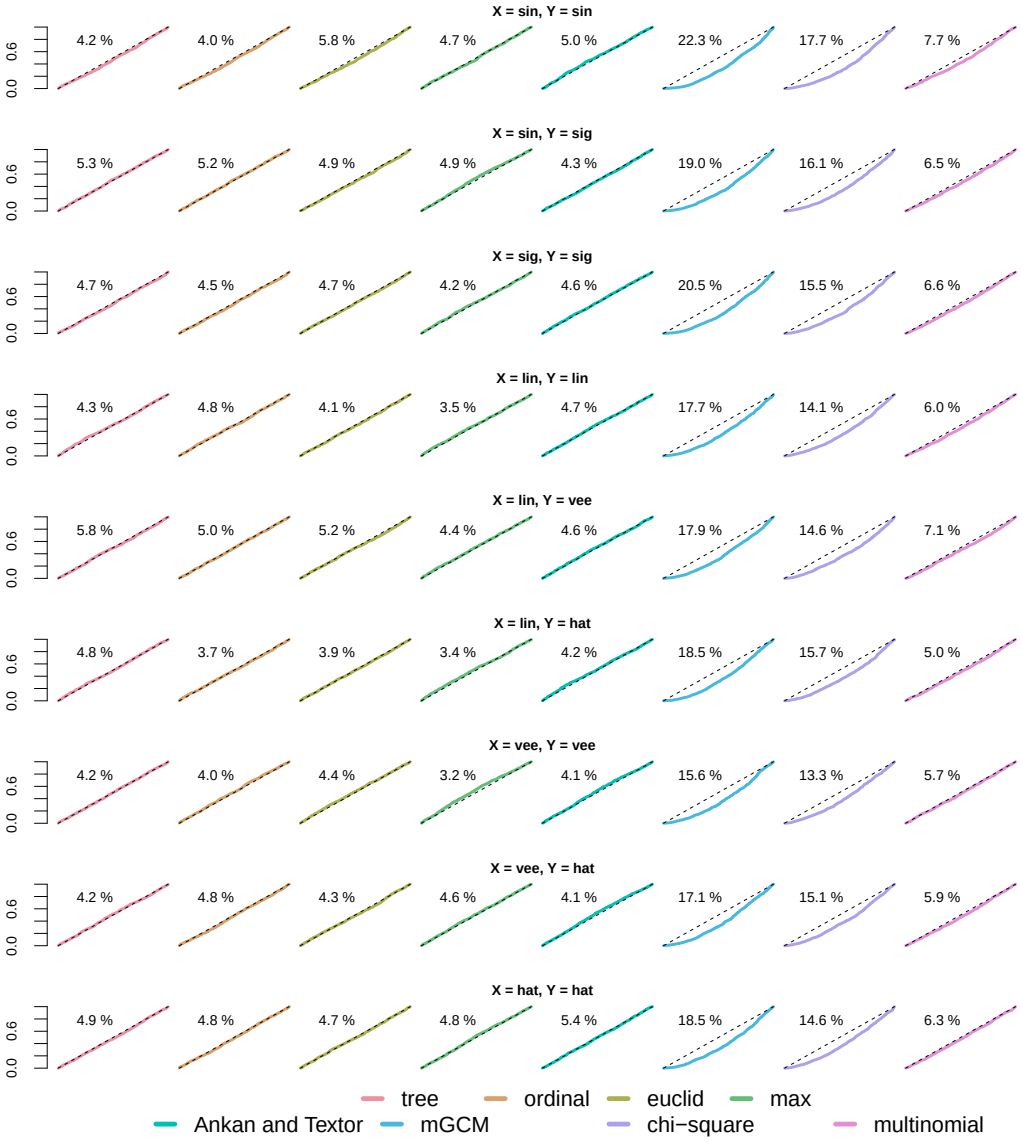


Figure 3.1 Quantile–quantile plots of p-values produced by various methods under various null settings, each with 1000 repeats. The p-values should be uniformly distributed (dashed black lines), with 5% having values less than 0.05 (text percentages). The plots labelled "tree" (red) and "ordinal" (orange) refer to our procedure (Algorithm 4) with the structural information $\mathbb{S}$ on both X and Y corresponding to either a binary tree or an ordinal structure. The plots labelled "euclid" (yellow) and "max" (green) refer to our calibration procedure (Algorithm 3) applied to the Euclidean norm and max norm criteria respectively, "Ankan and Textor" (turquoise) refers to the ordinal method of Ankan and Textor (2022), "mGCM" (blue) refers to the method of Shah and Peters (2020), "chi-square" (lilac) refers to the cumulative χ^2 distribution function (3.8), "multinomial" (pink) refers to one minus the observed significance of a linear multinomial regression of Y on (X, Z).

modified version of the "max" procedure, using the normalised statistics

$$\hat{T}^{(n)}_{j,k} \mapsto \frac{\hat{T}^{(n)}_{j,k}}{\sqrt{\hat{\Sigma}^{(n)}_{(j,k)(j,k)}}}; \quad \hat{\Sigma}^{(n)}_{(j,k)(j',k')} \mapsto \frac{\hat{\Sigma}^{(n)}_{(j,k)(j',k')}}{\sqrt{\hat{\Sigma}^{(n)}_{(j,k)(j,k)}\hat{\Sigma}^{(n)}_{(j',k')(j',k')}}}.$$

Whilst this does not involve estimating a full precision matrix, it is still inverting estimated variances.

We next examine the power of the well-calibrated methods. We consider two cases where $X \not\perp\!\!\!\perp Y \mid Z$, motivated by the hierarchical and ordinal data structures we have discussed. We set

$$\mathbb{P}(X = j, Y = k \mid Z = z) = p^{(X)}_j(z_1)p^{(Y)}_k(z_2) + \lambda d_{j,k}, \tag{3.19}$$

where $\lambda > 0$ controls the conditional dependence strength and $d \in \mathbb{R}^{8\times 8}$ is one of the following options:

$$d^{(tree)} := \begin{pmatrix} -1.1 & -1.1 & -0.9 & -0.9 & 0.9 & 0.9 & 1.1 & 1.1 \\ -1.1 & -1.1 & -0.9 & -0.9 & 0.9 & 0.9 & 1.1 & 1.1 \\ -0.9 & -0.9 & -1.1 & -1.1 & 1.1 & 1.1 & 0.9 & 0.9 \\ -0.9 & -0.9 & -1.1 & -1.1 & 1.1 & 1.1 & 0.9 & 0.9 \\ 0.9 & 0.9 & 1.1 & 1.1 & -1.1 & -1.1 & -0.9 & -0.9 \\ 0.9 & 0.9 & 1.1 & 1.1 & -1.1 & -1.1 & -0.9 & -0.9 \\ 1.1 & 1.1 & 0.9 & 0.9 & -0.9 & -0.9 & -1.1 & -1.1 \\ 1.1 & 1.1 & 0.9 & 0.9 & -0.9 & -0.9 & -1.1 & -1.1 \end{pmatrix}; \tag{3.20}$$

$$d^{(step)} := \begin{pmatrix} -1 & -1 & 1 & 1 & 1 & 1 & -1 & -1 \\ -1 & -1 & 1 & 1 & 1 & 1 & -1 & -1 \\ -1 & -1 & 1 & 1 & 1 & 1 & -1 & -1 \\ -1 & -1 & 1 & 1 & 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \end{pmatrix}. \tag{3.21}$$

Note that the rows and columns of each d sum to one, and we only consider values of the strength parameter λ for which all of the propensity scores are strictly positive.

When examining the conditional dependence setting (3.20), we apply our greedy query function (Algorithm 2) under a binary tree structural assumption on X and Y. Figure 3.2 demonstrates that our hierarchical-based adaptive search procedure significantly improves power compared to the non-adaptive versions in all X and Y settings under consideration.

Turning to setting (3.21), we make an ordinal structural assumption on X and Y in Algorithm 5. Figure 3.3 demonstrates that our ordinal-based adaptive search procedure improves power compared to the non-adaptive versions in all X and Y settings under consideration. We find that the method of Ankan and Textor (2022) does not have power in the ordinal setting (3.21).

Finally, we consider our procedure under complete misspecification of the structural information $\mathbb{S}$. We repeat the settings (3.20–3.21), but randomly permute the labels of X and Y once the data is generated. We find that our adaptive procedure does not perform substantially worse than any of the non-adaptive procedures, see Figures 3.4 and 3.5.

3.4 Discussion

Conditional independence testing is both interesting and challenging with categorical data. Any conditional independence test must restrict the null space, which is convenient to do through the convergence rates of machine learning methods. We suggest a test which has power against a diverse range of alternatives and is able to make use of structural information about the categorical variables (Algorithm 4). Our test efficiently searches for a partition of the categorical labels which maximises its power, and is calibrated using a single parametric bootstrap sample. We prove that our test asymptotically controls size uniformly within a broad class of null distributions (Theorem 29). In Section 3.3 we show that our test controls size in finite-sample settings where other methods fail, and is more powerful than non-adaptive versions and existing methods. We hope that our method will see use in practical data applications, and we share an implementation in the R package `catci` (CATegorical Conditional Independence) available from https://github.com/harveyklyne/catci.

3.5 Additional asymptotic results

3.5.1 Calibration procedure

In Section 3.2.4 we introduce Algorithm 3 for testing whether a random vector has any significantly large entries, given a bootstrap sample. Such tests are of independent interest, so we provide a more general result in Theorem 30. In this section, we allow $\hat{M}^{(n)}$ to be any random vector in $\mathbb{R}^L$.

Write $\mathbb{P}_n^B$ for the empirical distribution of the bootstrap samples $M_n^{(b)}$, so for any function $w : \mathbb{R}^L \to \mathbb{R}$ we define

$$\mathbb{P}_n^B(w(M_n^{(b)}) \leq x) := \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}\Big\{w\Big(M_n^{(b)}\Big) \leq t\Big\},$$

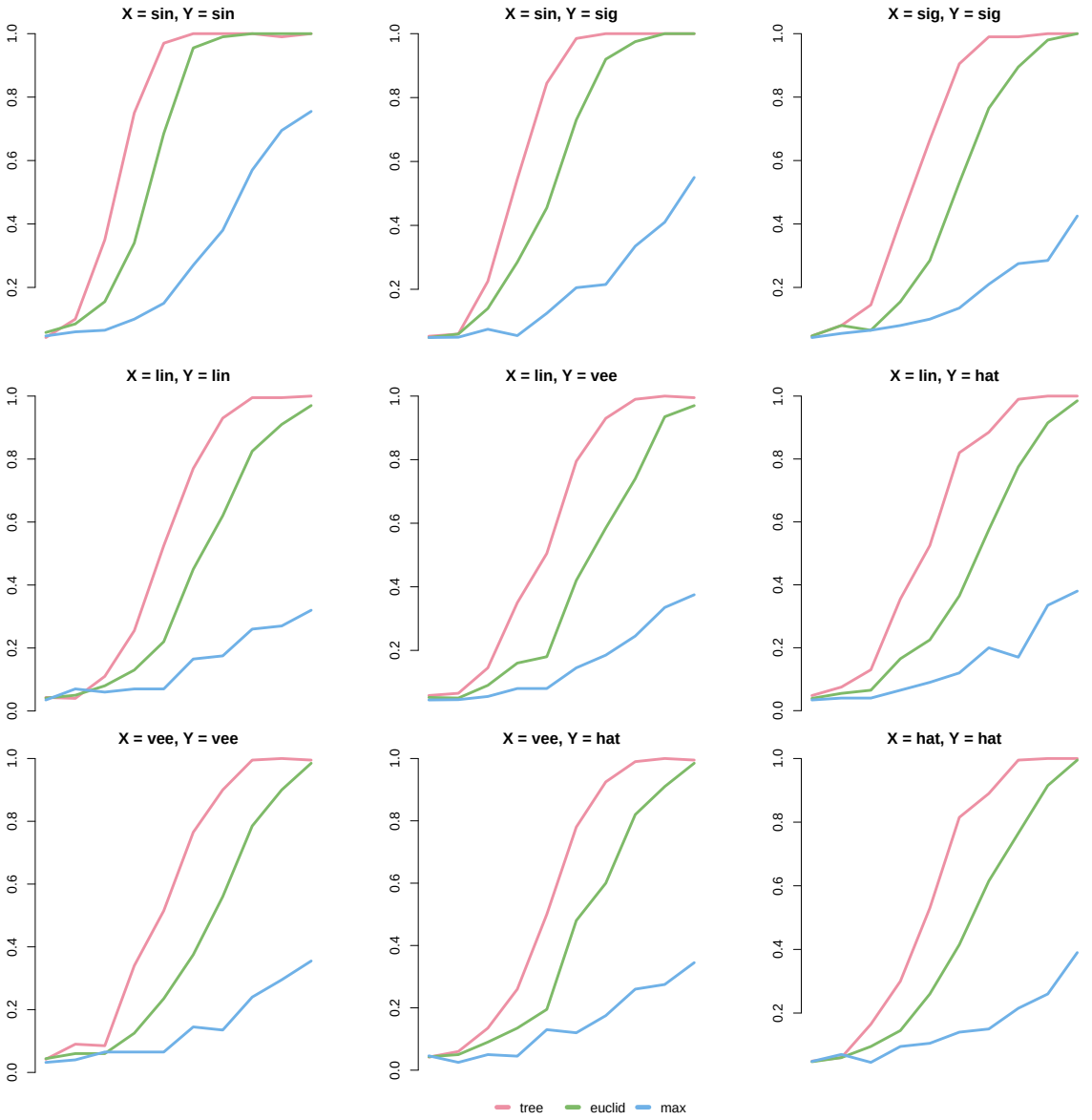


Figure 3.2 Plots of rejection rates at the 5% size level versus conditional dependence strength λ, from 200 repeats. X and Y have the binary tree conditional dependence structure (3.20), given Z. The plot labelled "tree" (red) refers to our procedure (Algorithm 4) with the structural information $\mathbb{S}$ on both X and Y corresponding to a binary tree structure. The plots labelled "euclid" (green) and "max" (blue) refer to our calibration procedure (Algorithm 3) applied to the Euclidean norm and max norm criteria respectively.

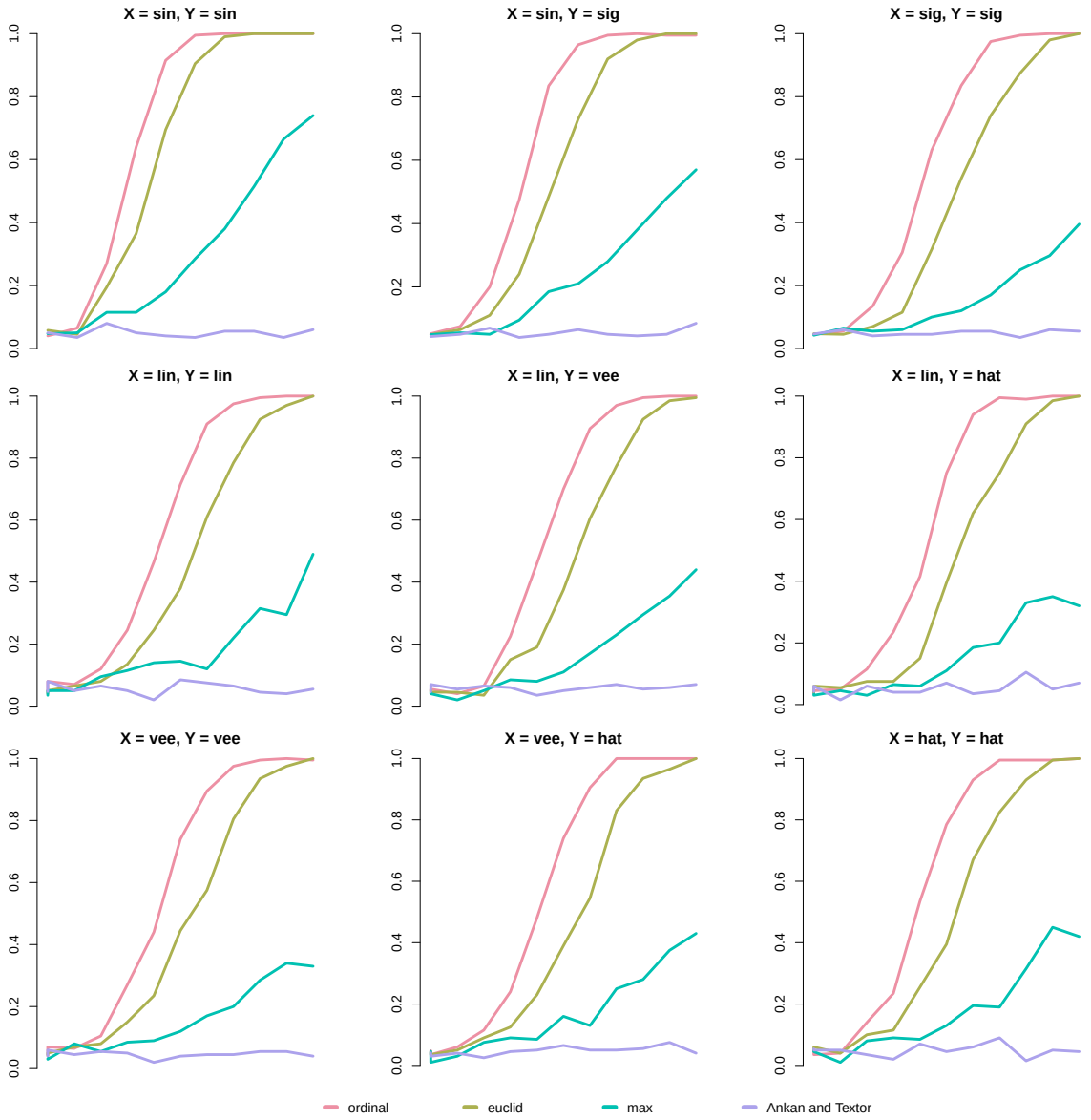


Figure 3.3 Plots of rejection rates at the 5% size level versus conditional dependence strength λ, from 200 repeats. X and Y have the step function conditional dependence (3.21), given Z. The plot labelled "ordinal" (red) refers to our procedure (Algorithm 4) with the structural information $\mathbb{S}$ on both X and Y corresponding to an ordinal structure. The plots labelled "euclid" (gold) and "max" (blue) refer to our calibration procedure (Algorithm 3) applied to the Euclidean norm and max norm criteria respectively, "Ankan and Textor" (lilac) refers to the ordinal method of Ankan and Textor (2022).

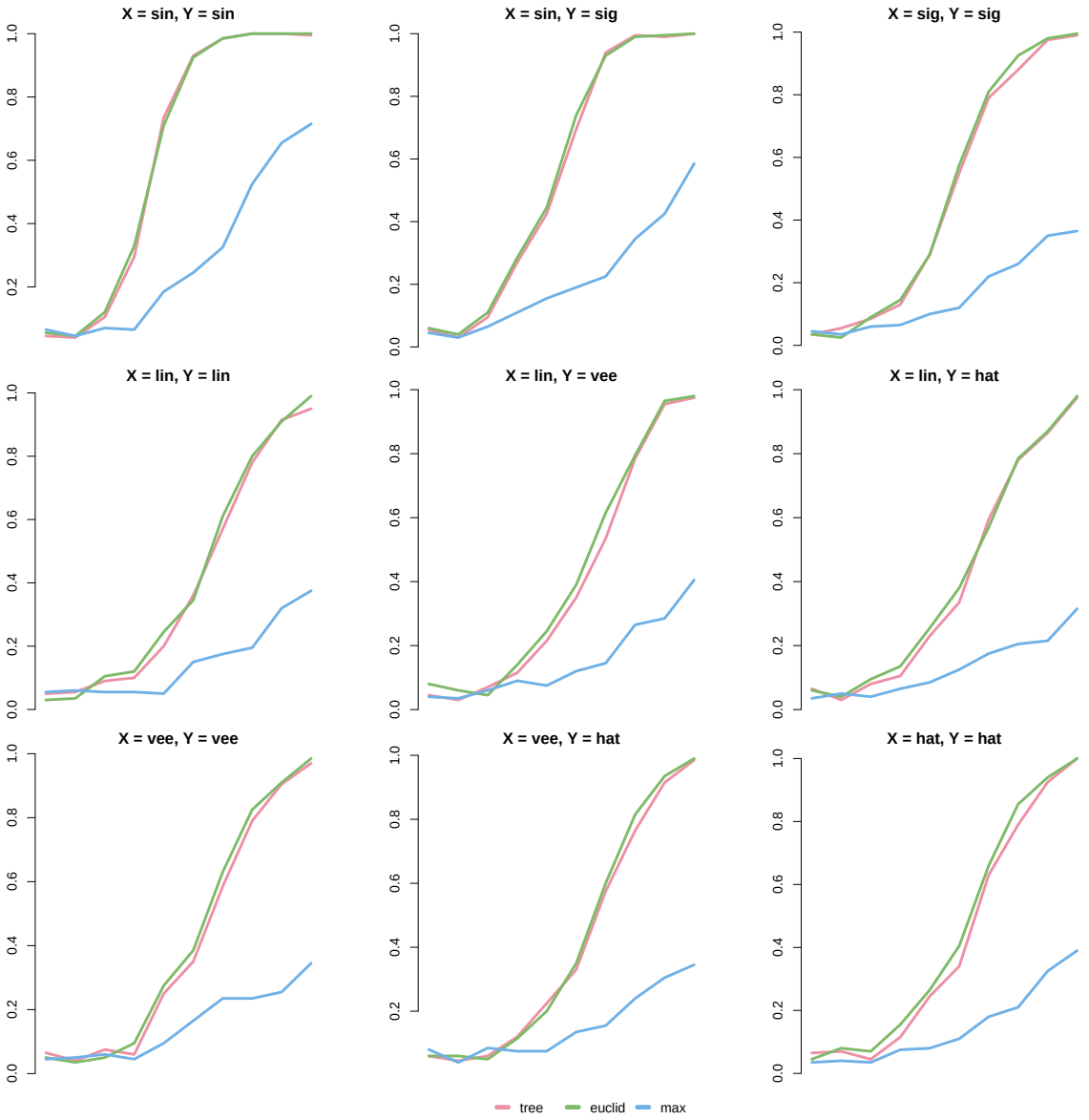


Figure 3.4 Plots of rejection rates at the 5% size level versus conditional dependence strength λ, from 200 repeats. These settings are identical to those of Figure 3.2, but with X and Y labels randomly permuted. The plot labelled "tree" (red) refers to our procedure (Algorithm 4) with the structural information $\mathbb{S}$ on both X and Y corresponding to a binary tree structure. The plots labelled "euclid" (green) and "max" (blue) refer to our calibration procedure (Algorithm 3) applied to the Euclidean norm and max norm criteria respectively.

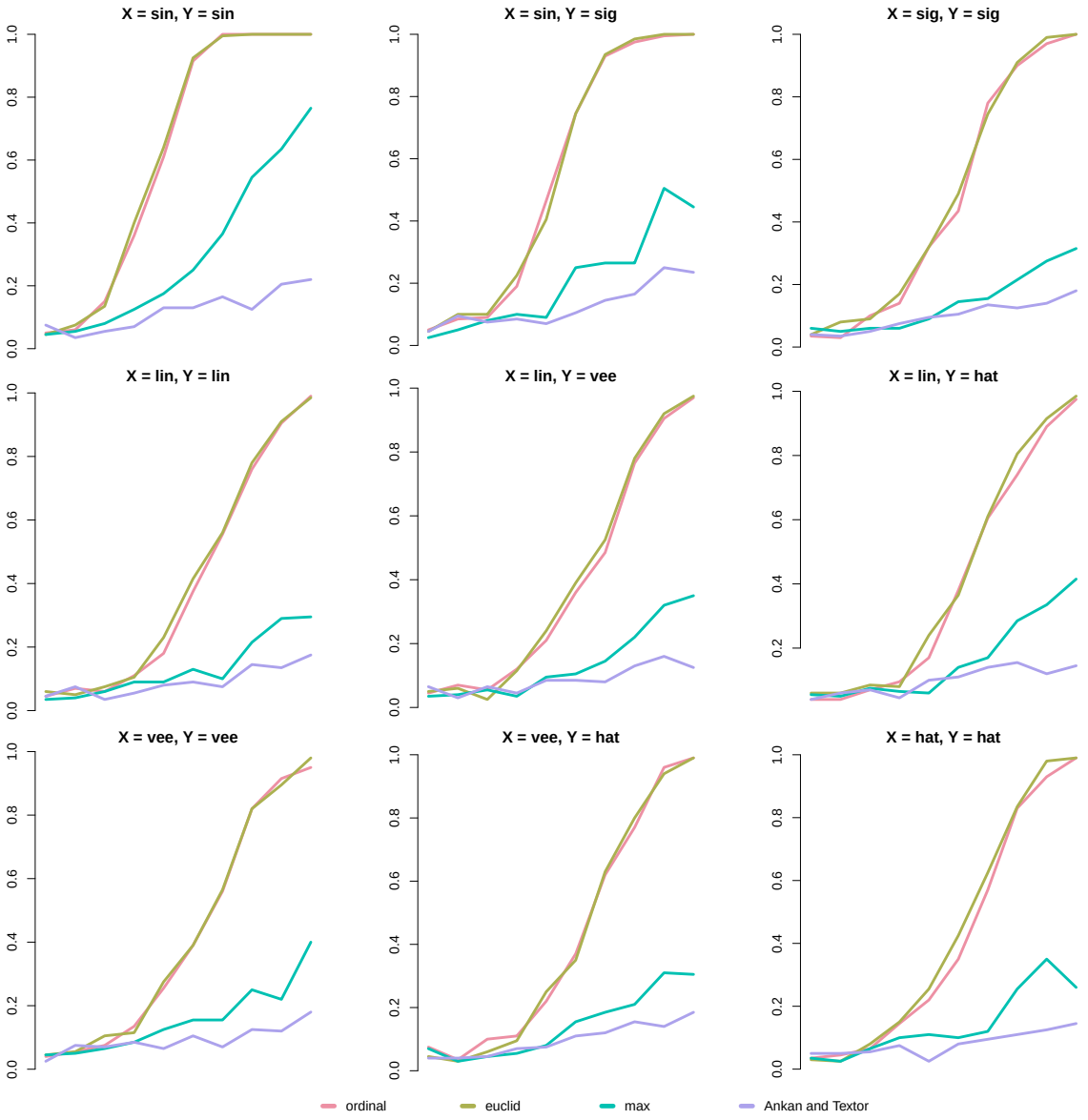


Figure 3.5 Plots of rejection rates at the 5% size level versus conditional dependence strength λ, from 200 repeats. These settings are identical to those of Figure 3.3, but with X and Y labels randomly permuted. The plot labelled "ordinal" (red) refers to our procedure (Algorithm 4) with the structural information $\mathbb{S}$ on both X and Y corresponding to an ordinal structure. The plots labelled "euclid" (gold) and "max" (blue) refer to our calibration procedure (Algorithm 3) applied to the Euclidean norm and max norm criteria respectively, "Ankan and Textor" (lilac) refers to the ordinal method of Ankan and Textor (2022).

for $x \in \mathbb{R}$.

Assumption 3. *For some class of laws $\mathcal{P}$ governing the data $D^{(n)}$, the vector of criteria $\hat{M}^{(n)}$ converges uniformly in distribution to some random variable M, whose distribution does not depend on P. That is, as $n \to \infty$,*

$$\sup_{P \in \mathcal{P}} \sup_{m \in \mathbb{R}^L} \left| \mathbb{P}_P\Big(\hat{M}^{(n)} \leq m\Big) - \mathbb{P}(M \leq m) \right| \to 0.$$

Furthermore, for each $l = 1, \ldots, L$ the limiting marginal distribution function $F_l(x) := \mathbb{P}(M_l \leq x)$ is continuous for $x \in \mathbb{R}$.

Remark 3.5.1. *One may prove a version of Theorem 30 which covers classes of distributions where the limiting random variable M_P depends on the distribution P. Our proof makes use of the continuous mapping theorem, and sufficient conditions exist under stronger continuity assumptions (e.g. Kasy (2018)).*

Assumption 4. *The distribution function of the bootstrap vectors $F_{n,l}(x) := \mathbb{P}_P\Big(M_{n,l}^{(b)} \leq x \;\Big|\; D^{(n)}\Big)$, which is a random function determined by the data $D^{(n)}$, converges uniformly in probability to F_l. That is, for each $l = 1, \ldots, L$ and $\epsilon > 0$, as $n \to \infty$ it holds that*

$$\sup_{P \in \mathcal{P}} \mathbb{P}_P\left(\sup_{x \in \mathbb{R}} \left| F_{n,l}(x) - F_l(x) \right| > \varepsilon \right) \to 0.$$

Additionally define $G(m) := \max_{l=1,\ldots,L} F_l(m_l)$ for $m \in \mathbb{R}^d$ and $\psi(u) = \mathbb{P}(G(M) \leq u)$ for $u \in [0,1]$. As $n \to \infty$,

$$\sup_{P \in \mathcal{P}} \mathbb{P}_P\left(\sup_{u \in [0,1]} \left| \mathbb{P}_P\left(G\Big(M_n^{(b)}\Big) \leq u \;\middle|\; D^{(n)} \right) - \psi(u) \right| > \epsilon \right) \to 0.$$

Theorem 30. *Under Assumptions 3 and 4, the output of Algorithm 3 converges uniformly in probability to a uniform random variable on $[0,1]$. Indeed, as $n, B \to \infty$,*

$$\sup_{P \in \mathcal{P}} \sup_{u \in [0,1]} \left| \mathbb{P}_P\Big(p\Big(\hat{M}^{(n)}, M_n^{(1)}, \ldots, M_n^{(B)}\Big) \leq u \Big) - u \right| \to 0.$$

3.5.2 Gaussian location testing

In Section 3.2.4 we reduce the problem of conditional independence testing to testing whether an asymptotically multivariate Gaussian vector has mean zero. Such tests are of independent interest, so we provide a more general statement in Algorithm 5. Recall that we are writing $\mathcal{M}^{(d_X, d_Y)} \subset \mathbb{R}^{d_X d_Y \times d_X d_Y}$ for the set of symmetric positive semidefinite

matrices on $\mathbb{R}^{d_X d_Y \times d_X d_Y}$. In this section, we allow q to be any user-specified query function

$$q : \bigcup_{d_X, d_Y \geq 2} \mathbb{R}^{d_X d_Y} \times \Big(\mathcal{M}^{(d_X, d_Y)} \setminus \{0\} \Big) \to [0,1]^L.$$

Input: $\hat{T}^{(n)}$, $\hat{\Sigma}^{(n)}$, q, B
Output: P-value $p \in [0,1]$.
Compute criterion $\hat{M}^{(n)} = q\Big(\hat{T}^{(n)}, \hat{\Sigma}^{(n)}\Big)$.
for $b = 1, \ldots, B$ **do**
  Draw parametric bootstrap $T_n^{(b)} \sim N\Big(0, \hat{\Sigma}_n\Big)$, for example using Algorithm 8.
  Compute bootstrap criterion $M_n^{(b)} = q\Big(T_n^{(b)}, \hat{\Sigma}^{(n)}\Big)$.
end
Set $p = p\Big(\hat{M}^{(n)}, B, M_n^{(1)}, \ldots, M_n^{(B)}\Big)$, where p is the output of Algorithm 3.

Algorithm 5: Testing whether an asymptotically Gaussian vector has mean zero using query function $q : \mathbb{R}^d \times \mathbb{R}^{d \times d} \to \mathbb{R}^L$ and our calibration procedure (Algorithm 3).

Our basic assumption is that $\hat{T}^{(n)}$ is asymptotically Gaussian and $\hat{\Sigma}^{(n)}$ is consistent. We seek conditions on the query function $q : (\hat{T}^{(n)}, \hat{\Sigma}^{(n)}) \mapsto \hat{M}^{(n)}$ such that Assumptions 3 and 4 hold.

Assumption 5. *For some class of laws $\mathcal{P}$ governing the data, the vector $\hat{T}^{(n)}$ converges uniformly in distribution to a centred Gaussian with non-zero covariance matrix $\Sigma \in \mathcal{M}^{(d_X, d_Y)}$ which does not depend on P, and the estimated covariance matrix $\hat{\Sigma}^{(n)}$ converges uniformly in probability to Σ. That is, for any $\epsilon > 0$ and as $n \to \infty$,*

$$\sup_{P \in \mathcal{P}} \sup_{t \in \mathbb{R}^d} \Big| \mathbb{P}_P\Big(\hat{T}^{(n)} \leq t\Big) - \mathbb{P}(T \leq t) \Big| \to 0; \quad \sup_{P \in \mathcal{P}} \mathbb{P}_P\Big(\Big\| \hat{\Sigma}^{(n)} - \Sigma \Big\| > \epsilon \Big) \to 0,$$

where $T \overset{d}{=} N(0, \Sigma)$ and $\| \cdot \|$ is any norm on $\mathbb{R}^{d_X d_Y \times d_X d_Y}$.

Assumption 6. *The query function $q : \mathbb{R}^{d_X d_Y} \times \mathbb{R}^{d_X d_Y \times d_X d_Y} \to \mathbb{R}^L$ is continuous at every point of a set $\mathcal{T}$ such that $\mathbb{P}\Big((T, \Sigma) \in \mathcal{T}\Big) = 1$.*

Define the following functions for $x \in \mathbb{R}$, $u \in [0,1]$, and $\sigma \in \mathcal{M}^{(d_X, d_Y)} \setminus 0$:

$$F_l(x; \sigma) = \mathbb{P}\Big(q_l\Big(N(0, \sigma), \sigma\Big) \leq x \Big);$$
$$\psi(u; \sigma) = \mathbb{P}\Big(G\Big(q\Big(N(0, \sigma), \sigma\Big)\Big) \leq u \Big),$$

recalling that $G(m) = \max_{l=1,\ldots,L} F_l(m_l) = \max_{l=1,\ldots,L} F_l(m_l; \Sigma)$. Note that when σ is a random variable, $F_l(\cdot; \sigma)$ and $\psi(\cdot; \sigma)$ are random functions (i.e. we take the cumulative distribution functions conditional on σ).

Assumption 7. *The functions $F_l(\cdot; \Sigma)$, $l = 1, \ldots, L$ are continuous.*

Theorem 31. *Under Assumptions 5, 6, and 7, the output of Algorithm 5 converges uniformly in probability to a uniform random variable on* $[0,1]$. *Indeed, as* $n, B \to \infty$,

$$\sup_{P \in \mathcal{P}} \sup_{u \in [0,1]} \left| \mathbb{P}_P \Big(p\Big(\hat{T}^{(n)}, \hat{\Sigma}^{(n)}, q, B\Big) \leq u \Big) - u \right| \to 0.$$

3.6 Implementation

In order reduce the computational burden, we pre-tune all hyperparameters for the experiments in Section 3.3 on 1000 datasets, each of which we split into training and testing. This includes all gradient boosting regression parameters.

3.6.1 Fast greedy merging

Recall our criterion of interest (3.9), which takes input a vector t and matrix σ. Note that it is sufficient to just know the summary quantities $\|t\|_2^2$, $\mathrm{tr}(\sigma)$, $\mathrm{tr}(\sigma^2)$. We are therefore interested in computing these quantities for the transformed vector Πt and matrix $\Pi \sigma \Pi^T$ in an efficient manner. Note that $\mathrm{tr}(\sigma^2)$ equals the sum of the squared elements of σ.

Recalling the stepwise greedy nature of Algorithm 2, consider merging two X-labels, $j_1 < j_2$ (merging two Y-labels is similar). Then the corresponding transformation $\Pi \in \mathcal{C}_{(d_X, d_Y)(d_X - 1, d_Y)}$ has elements

$$\Pi_{(l,m)(j,k)} = \begin{cases} \mathbb{1}\{j = l, k = m\} & \text{if } l \in \{1, \ldots, j_1 - 1, j_1 + 1, \ldots, j_2 - 1\}; \\ \mathbb{1}\{j \in \{j_1, j_2\}, k = m\} & \text{if } l = j_1; \\ \mathbb{1}\{j = l + 1, k = m\} & \text{if } l \in \{j_2, \ldots, d_X - 1\}. \end{cases}$$

It follows that, for each $m, m' \in \{1, \dots, d_Y\}$,

$$(\Pi t)_{l,m} = \begin{cases} t_{l,m} & \text{if } l \in \{1, \dots, j_1 - 1, j_1 + 1, \dots, j_2 - 1\}; \\ t_{j_1,m} + t_{j_2,m} & \text{if } l = j_1; \\ t_{l+1,m} & \text{if } l \in \{j_2, \dots, d_X - 1\}; \end{cases} \tag{3.22}$$

$$\left(\Pi\sigma\Pi^T\right)_{(l,m)(l',m')} = \begin{cases} \sigma_{(l,m)(l',m')} & \text{if } l, l' \in \{1, \dots, j_1 - 1, \\ & \qquad j_1 + 1, \dots, j_2 - 1\}; \\ \sigma_{(j_1,m)(l',m')} + \sigma_{(j_2,m)(l',m')} & \text{if } l = j_1, \\ & \quad l' \in \{1, \dots, j_1 - 1, \\ & \qquad j_1 + 1, \dots, j_2 - 1\}; \\ \sigma_{(l+1,m)(l',m')} & \text{if } l \in \{j_2, \dots, d_X - 1\}, \\ & \quad l' \in \{1, \dots, j_1 - 1, \\ & \qquad j_1 + 1, \dots, j_2 - 1\}; \\ \sigma_{(l,m)(j_1,m')} + \sigma_{(l,m)(j_2,m')} & \text{if } l \in \{1, \dots, j_1 - 1, \\ & \qquad j_1 + 1, \dots, j_2 - 1\}, \\ & \quad l' = j_1; \\ \sigma_{(j_1,m)(j_1,m')} + \sigma_{(j_2,m)(j_1,m')} & \\ +\sigma_{(j_1,m)(j_2,m')} + \sigma_{(j_2,m)(j_2,m')} & \text{if } l = l' = j_1; \\ \sigma_{(l+1,m)(j_1,m')} + \sigma_{(l+1,m)(j_2,m')} & \text{if } l \in \{j_2, \dots, d_X - 1\}, \\ & \quad l' = j_1; \\ \sigma_{(l,m)(l'+1,m')} & \text{if } l \in \{1, \dots, j_1 - 1, \\ & \qquad j_1 + 1, \dots, j_2 - 1\}, \\ & \quad l' \in \{j_2, \dots, d_X - 1\}; \\ \sigma_{(j_1,m)(l'+1,m')} + \sigma_{(j_2,m)(l'+1,m')} & \text{if } l = j_1, \\ & \quad l' \in \{j_2, \dots, d_X - 1\}; \\ \sigma_{(l+1,m)(l'+1,m')} & \text{if } l, l' \in \{j_2, \dots, d_X - 1\}. \end{cases} \tag{3.23}$$

Therefore the updated summary quantities are as follows.

$$\|\Pi t\|_2^2 = \|t\|_2^2 + 2\sum_{k=1}^{d_Y} t_{j_1,k} t_{j_2,k}; \tag{3.24}$$

$$\mathrm{tr}\Big(\Pi\sigma\Pi^T\Big) = \mathrm{tr}(\sigma) + 2\sum_{k=1}^{d_Y} \sigma_{(j_1,k)(j_2,k)} \tag{3.25}$$

$$\begin{aligned}\mathrm{tr}\Big[\Big(\Pi\sigma\Pi^T\Big)^2\Big] = \mathrm{tr}\Big(\sigma^2\Big) &+ 4\sum_{k,m=1}^{d_Y}\sum_{l=1}^{d_X} \sigma_{(j_1,k)(l,m)}\sigma_{(j_2,k)(l,m)} \\ &+ 2\sum_{k,m=1}^{d_Y} \sigma_{(j_1,k)(j_1,m)}\sigma_{(j_2,k)(j_2,m)} \\ &+ 2\sum_{k,m=1}^{d_Y} \sigma_{(j_1,k)(j_2,m)}\sigma_{(j_2,k)(j_1,m)}.\end{aligned} \tag{3.26, 3.27}$$

We may use these equalities to greatly speed up Algorithm 2. In an abuse of notation, write $\phi\Big(\|t\|_2^2, \mathrm{tr}(\sigma), \mathrm{tr}\Big(\sigma^2\Big)\Big) = \phi(t,\sigma)$.

Input: Statistic vector $T \in \mathbb{R}^{d_X d_Y}$, covariance matrix $\Sigma \in \mathcal{M}^{(d_X, d_Y)}$, X-dimension d_X, Y-dimension d_Y, structural information $\mathbb{S} \in \mathcal{S}$, search depth $L \in \mathbb{N}$.

Output: Vector of criteria $M \in [0,1]^L$, with larger values corresponding to larger deviations from centrality.

Initialise norm-square $\|T\|_2^2$, trace $\mathrm{tr}(\Sigma)$, and trace of the squared covariance $\mathrm{tr}\left(\Sigma^2\right) = \sum_{j,j'=1}^{d_X} \sum_{k,k'=1}^{d_Y} \Sigma^2_{(j,k)(j',k')}$;

Set $M_1 = \phi\Big(\|T\|_2^2, \mathrm{tr}(\Sigma), \mathrm{tr}\left(\Sigma^2\right)\Big)$.

for $l = 2, \ldots, L$ **do**

  for *potential pairs* $1 \le j_1 < j_2 \le d_X$ *of* X*-labels to merge* **do**

    if *merging* (j_1, j_2) *is consistent with the structure* $\mathbb{S}$ **then**

      Compute $\|\Pi^{(A,B)} T\|_2^2$ via (3.25), $\mathrm{tr}\Big(\Pi^{(A,B)} \Sigma \Pi^{(A,B)T}\Big)$ via (3.26), and $\mathrm{tr}\left[\Big(\Pi^{(A,B)} \Sigma \Pi^{(A,B)T}\Big)^2\right]$ via (3.27);

      Set $\phi_X^{(j_1,j_2)} = \phi\left(\|\Pi^{(A,B)} T\|_2^2, \mathrm{tr}\Big(\Pi^{(A,B)} \Sigma \Pi^{(A,B)T}\Big), \mathrm{tr}\left[\Big(\Pi^{(A,B)} \Sigma \Pi^{(A,B)T}\Big)^2\right]\right)$.

    end

  end

  for *potential pairs* $1 \le k_1 < k_2 \le d_Y$ *of* Y*-labels to merge* **do**

    if *merging* (k_1, k_2) *is consistent with the structure* $\mathbb{S}$ **then**

      Compute $\|\Pi^{(A,B)} T\|_2^2$ analogously to (3.25), $\mathrm{tr}\Big(\Pi^{(A,B)} \Sigma \Pi^{(A,B)T}\Big)$ analogously to (3.26), and $\mathrm{tr}\left[\Big(\Pi^{(A,B)} \Sigma \Pi^{(A,B)T}\Big)^2\right]$ analogously to (3.27);

      Set $\phi_Y^{(k_1,k_2)} = \phi\left(\|\Pi^{(A,B)} T\|_2^2, \mathrm{tr}\Big(\Pi^{(A,B)} \Sigma \Pi^{(A,B)T}\Big), \mathrm{tr}\left[\Big(\Pi^{(A,B)} \Sigma \Pi^{(A,B)T}\Big)^2\right]\right)$.

    end

  end

  Set $M_l = \max\left\{\phi_X^{(j_1,j_2)}, \phi_Y^{(k_1,k_2)}\right) \ : \ \text{permitted } (j_1, j_2) \text{ and } (k_1, k_2)\Big\}$;

  For the maximising argument, update T as in (3.22) and Σ as in (3.23);

  Update d_X, d_Y, and $\mathbb{S}$ to be consistent with the new partitions.

end

Algorithm 6: Fast implementation of the greedy merging query function (Algorithm 2).

3.6.2 Continuous version of calibration procedure

Input: Test vector $\hat{M}^{(n)} \in \mathbb{R}^L$, number of bootstrap samples $B \in \mathbb{N}$, bootstrap sample $M_n^{(1)}, \ldots, M_n^{(B)} \in \mathbb{R}^L$.

Output: P-value $p \in [0, 1]$.

for $l = 1, \ldots, L$ **do**

  Draw $U_l \sim \text{Uniform}[0, 1]$;

  Set $\hat{F}_{n,l} = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}\Big\{M_{n,l}^{(b)} < \hat{M}_l^{(n)}\Big\} + \frac{U}{B} \sum_{b=1}^{B} \mathbb{1}\Big\{M_{n,l}^{(b)} = \hat{M}_l^{(n)}\Big\}$;

end

Set $\hat{G}_n = \max_{l=1,\ldots,L} \hat{F}_{n,l}$.

for $b' = 1, \ldots, B$ **do**

  for $l = 1, \ldots, L$ **do**

    Draw $U_l^{(b')} \sim \text{Uniform}[0, 1]$;

    Set $F_{n,l}^{(b')} = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}\Big\{M_{n,l}^{(b)} < M_{n,l}^{(b')}\Big\} + \frac{U^{(b')}}{B} \sum_{b=1}^{B} \mathbb{1}\Big\{M_{n,l}^{(b)} = M_{n,l}^{(b')}\Big\}$;

  end

  Set $G_n^{(b')} = \max_{l=1,\ldots,L} F_{n,l}^{(b')}$.

end

Draw $U \sim \text{Uniform}[0, 1]$;

Set $p = 1 - \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}\Big\{G_n^{(b)} < \hat{G}_n\Big\} - \frac{U}{B} \sum_{b=1}^{B} \mathbb{1}\Big\{G_n^{(b)} = \hat{G}_n\Big\}$.

Algorithm 7: Continuous version of Algorithm 3, which we have found to have better finite sample properties.

3.6.3 Generating multivariate Gaussian bootstraps

Input: Non-zero covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$, number of samples $B \in \mathbb{N}$ to be generated.

Output: Independent samples $\{T^{(b)} : b = 1, \ldots, B\}$ drawn from $N(0, \Sigma)$.

Compute the thin matrix square root $\Sigma^{1/2} \in \mathbb{R}^{d \times r}$ of Σ, where $r = \text{rank}(\Sigma)$ (we use the `chol` function in base R).

Populate a matrix $Z \in \mathbb{R}^{r \times B}$ with independent standard Gaussian draws.

Compute $T_0 = \Sigma^{1/2} Z \in \mathbb{R}^{d \times B}$.

for $b = 1, \ldots, B$ **do**

  Set $T^{(b)} \in \mathbb{R}^d$ to be the bth column of T_0.

end

Algorithm 8: Procedure to generate samples from a multivariate Gaussian distribution with mean zero and a specified covariance matrix.

3.7 Proof of Theorem 29

For this section, we define

$$\begin{aligned}\mathcal{C}_* &:= \bigcup_{2\le d'_X\le d_X}\ \bigcup_{2\le d'_Y\le d_Y} \mathcal{C}_{(d_X,d_Y)(d'_X,d'_Y)} \\ &= \Big\{\Pi^{(A,B)} \ :\ A \text{ a partition of } \{1,\dots,d_X\},\\ &\qquad\qquad B \text{ a partition of } \{1,\dots,d_Y\},\ |A|,|B|\ge 2\Big\}.\end{aligned}$$

For each $\Pi \in \mathcal{C}_*$, define a function $\phi_\Pi : \mathbb{R}_{\ge 0} \to [0,1]$ by

$$\phi_\Pi(x) := \mathbb{P}\Big(g\big(\Pi\Sigma\Pi^T\big)\chi^2_{h(\Pi\Sigma\Pi^T)} \le x\Big),$$

where $g(\sigma) = \mathrm{tr}\big(\sigma^2\big)/\mathrm{tr}(\sigma)$ and $h(\sigma) = \mathrm{tr}(\sigma)^2/\mathrm{tr}\big(\sigma^2\big)$.

Proof. Algorithm 4 is equivalent to Algorithm 5 with statistic $\hat{T}^{(n)}$ as in (3.4), covariance matrix $\hat{\Sigma}^{(n)}$ as in (3.5), and query function $(t,\sigma) \mapsto q(t,\sigma,d_X,d_Y,\mathbb{S})$. Therefore it suffices to prove that Assumptions 1 and 2 imply Assumptions 5, 6, and 7, and then make use of Theorem 31. Indeed, Lemma 32 gives that Assumptions 1 and 2 imply Assumption 5, Lemma 33 gives that Assumption 2 implies Assumption 6, and Lemma 37 gives that Assumption 2 implies Assumption 7. □

Lemma 32. *Let $\mathcal{P}$ be a class of distributions satisfying Assumption 1. Then the vector $\hat{T}^{(n)}$ defined in* (3.4) *converges uniformly in distribution to a Gaussian with mean μ_P defined in* (3.3) *and variance Σ_P defined in* (3.6)*, and the estimated covariance matrix $\hat{\Sigma}^{(n)}$ defined in* (3.5) *converges uniformly in probability to Σ_P.*

Proof. In an abuse of notation, we refer to the quantities

$$E_f^{(n,r)} := \max_{j=1,\dots,d_X} \mathbb{E}_P\Big[\Big\{f_{P,j}(Z) - \hat{f}^{(n,1)}(Z)\Big\}^2 \ \Big|\ D^{(n,1)}\Big],$$

for each fold $r = 1,\dots,N$. Each $E_f^{(n,r)}$ satisfies the same probabilistic assumptions as $E_f^{(n)} = E_f^{(n,1)}$ due to the equal partitioning and i.i.d. data. Likewise we define $E_g^{(n,r)}$.

To show the first conclusion we first highlight the term which converges to a target Gaussian distribution, and then deal with the remainder. Define the residuals $\varepsilon_P \in [-1,1]^{d_X}$, $\xi_P \in [-1,1]^{d_Y}$ as

$$\begin{aligned}\varepsilon_{P,j} &= \tilde{X}_j - f_{P,j}(Z);\\ \xi_{P,k} &= \tilde{Y}_k - g_{P,k}(Z),\end{aligned}$$

And similarly the sample equivalents $\varepsilon_{P,i}$ and $\xi_{P,i}$. For every j, k we have

$$\begin{aligned}\mathbb{E}_P(\varepsilon_{P,j} \mid Z) &= \mathbb{E}_P(\xi_{P,k} \mid Z) = 0;\\ \mathbb{E}_P(\varepsilon_{P,j}\xi_{P,k}) &= \mu_{P,j,k},\end{aligned}$$

and also for every j', k' we have

$$\mathrm{Cov}_P\Big(\varepsilon_{P,j}\xi_{P,k}, \varepsilon_{P,j'}\xi_{P,k'}\Big) = \Sigma_{P(j,k)(j',k')}.$$

Now

$$\sqrt{n}\Big(\hat{T}^{(n)}_{j,k} - \mu_{P,j,k}\Big) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big\{\varepsilon_{P,ij}\xi_{P,ik} - \mu_{j,k}\Big\} + \sum_{r=1}^{N} R^{(n,r)}_{P,j,k},$$

where the uniform central limit theorem (Lemma 9) applies to the first term and $R_P^{(n,r)} \in \mathbb{R}^{d_X d_Y}$ has elements

$$R^{(n,r)}_{P,j,k} := \frac{1}{\sqrt{n}}\sum_{i \in I^{(n,r)}}\Big\{\tilde{X}_{ij} - \hat{f}_j^{(n,r)}(Z_i)\Big\}\Big\{\tilde{Y}_{ik} - \hat{g}_k^{(n,r)}(Z_i)\Big\} - \varepsilon_{P,ij}\xi_{P,ik}.$$

Note that, conditionally on $D^{(n,r)}$, each summand of $R_P^{(n,r)}$ is i.i.d. To show that $R_P^{(n,r)} = o_{\mathcal{P}}(1)$, we fix some elements $j \in \{1, \ldots, d_X\},\ k \in \{1, \ldots, d_Y\}$ and decompose

$$R^{(n,r)}_{P,j,k} = a^{(n,r)} + b_f^{(n,r)} + b_g^{(n,r)}, \tag{3.28}$$

where

$$\begin{aligned}a^{(n,r)} &:= \frac{1}{\sqrt{n}}\sum_{i \in I^{(n,r)}}\{f_{P,j}(Z_i) - \hat{f}_j^{(n,r)}(Z_i)\}\{g_{P,k}(Z_i) - \hat{g}_k^{(n,r)}(Z_i)\};\\ b_f^{(n,r)} &:= \frac{1}{\sqrt{n}}\sum_{i \in I^{(n,r)}}\{f_{P,j}(Z_i) - \hat{f}_j^{(n,r)}(Z_i)\}\xi_{P,ik};\\ b_g^{(n,r)} &:= \frac{1}{\sqrt{n}}\sum_{i \in I^{(n,r)}}\{g_{P,k}(Z_i) - \hat{g}_k^{(n,r)}(Z_i)\}\varepsilon_{P,ij}.\end{aligned}$$

We now show that each term is $o_{\mathcal{P}}(1)$, so Lemma 11 yields the first conclusion.

By the Cauchy–Schwarz inequality, we have

$$\begin{aligned}\mathbb{E}_P[|a^{(n,r)}| \mid D^{(n,r)}] &\le \sqrt{n}\mathbb{E}_P[|f_{P,j}(Z) - \hat{f}_j^{(n,r)}(Z)||g_{P,k}(Z) - \hat{g}_k^{(n,r)}(Z)| \mid D^{(n,r)}]\\ &\le \sqrt{nE_f^{(n,r)}E_g^{(n,r)}} = o_{\mathcal{P}}(1),\end{aligned}$$

so $a^{(n,r)}$ is $o_{\mathcal{P}}(1)$ by Lemma 12. Note that each summand of $b_f^{(n,r)}$ is mean-zero conditionally on Z. This means that

$$\begin{aligned}\mathbb{E}_P[(b_f^{(n,r)})^2 \mid D^{(n,r)}] &= \mathbb{E}_P[\{f_{P,j}(Z) - \hat{f}_j^{(n,r)}(Z)\}^2 \xi_{P,k}^2 \mid D^{(n,r)}] \\ &\leq E_f^{(n,r)} = o_{\mathcal{P}}(1),\end{aligned}$$

where we have used a supremum bound for $|\xi_{P,k}| \leq 1$. Again using Lemma 12 we have that $b_f^{(n,r)} = o_{\mathcal{P}}(1)$. By an identical argument, we also have that $b_g^{(n,r)} = o_{\mathcal{P}}(1)$.

Turning now to the second conclusion, we aim to show that $\hat{\Sigma}^{(n)} - \Sigma_P = o_{\mathcal{P}}(1)$. We introduce notation for the following random functions on $\{0,1\}^{d_X} \times \{0,1\}^{d_Y} \times \mathcal{Z}$:

$$\hat{\psi}_{j,k}^{(n,r)}(x,y,z) := \Big\{x_j - \hat{f}_j^{(n,r)}(z)\Big\}\Big\{\tilde{y}_k - \hat{g}_k^{(n,r)}(z)\Big\} - \hat{T}_{j,k}^{(n)},$$

and also the population version

$$\psi_{P,j,k}(x,y,z) := \Big\{x_j - f_{P,j}(z)\Big\}\Big\{\tilde{y}_k - g_{P,k}(z)\Big\} - \mu_{P,j,k}.$$

We will focus on an individual element $(\hat{\Sigma}^{(n)} - \Sigma_P)_{(j,k)(j',k')}$,and make use of Lemma 10. Note that the absolute value of $\psi_{P,j,k}$ is at most 2.

We are now ready to decompose the covariance estimation error.

$$\begin{aligned}(\hat{\Sigma}^{(n)} - \Sigma_P)_{(j,k)(j',k')} &= \frac{1}{n}\sum_{r=1}^{N}\sum_{i \in I^{(n,r)}} \hat{\psi}_{j,k}^{(n,r)}(\tilde{X}_i, \tilde{Y}_i, Z_i)\hat{\psi}_{j',k'}^{(n,r)}(\tilde{X}_i, \tilde{Y}_i, Z_i) \\ &\qquad - \mathbb{E}_P\Big[\psi_{P,j,k}(\tilde{X}, \tilde{Y}, Z)\psi_{P,j',k'}(\tilde{X}, \tilde{Y}, Z)\Big] \\ &= \frac{1}{n}\sum_{i=1}^{n}\Big[\psi_{P,j,k}(\tilde{X}_i, \tilde{Y}_i, Z)\psi_{P,j',k'}(\tilde{X}_i, \tilde{Y}_i, Z) \\ &\qquad - \mathbb{E}_P\Big[\psi_{P,j,k}(\tilde{X}, \tilde{Y}, Z)\psi_{P,j',k'}(\tilde{X}, \tilde{Y}, Z)\Big]\Big] \\ &\quad + \frac{1}{N}\sum_{r=1}^{N} S_P^{(n,r)},\end{aligned}$$

where the first term is $o_{\mathcal{P}}(1)$ by Lemma 10 and

$$S_P^{(n,r)} := \frac{K}{n}\sum_{i \in I^{(n,r)}}\Big[\hat{\psi}_{j,k}^{(n,r)}(\tilde{X}_i, \tilde{Y}_i, Z_i)\hat{\psi}_{j',k'}^{(n,r)}(\tilde{X}_i, \tilde{Y}_i, Z_i) - \psi_{P,j,k}(\tilde{X}_i, \tilde{Y}_i, Z)\psi_{P,j',k'}(\tilde{X}_i, \tilde{Y}_i, Z)\Big].$$

We show that $S_P^{(n,r)} = o_{\mathcal{P}}(1)$ using the following identity for $a_1, a_2, b_1, b_2 \in \mathbb{R}$,

$$a_1 b_1 - a_2 b_2 = (a_1 - a_2)(b_1 - b_2) + a_2(b_1 - b_2) + b_2(a_1 - a_2),$$

and then applying the Cauchy–Schwarz inequality to each term.

$$
\begin{aligned}
\left|S_P^{(n,r)}\right| &= \left|\frac{N}{n}\sum_{i\in I^{(n,r)}} \hat{\psi}_{j,k}^{(n,r)}(\tilde{X}_i,\tilde{Y}_i,Z)\hat{\psi}_{j',k'}^{(n,r)}(\tilde{X}_i,\tilde{Y}_i,Z) - \psi_{P,j,k}(\tilde{X}_i,\tilde{Y}_i,Z)\psi_{P,j',k'}(\tilde{X}_i,\tilde{Y}_i,Z)\right| \\
&\le \left|\frac{N}{n}\sum_{i\in I^{(n,r)}} \left\{\hat{\psi}_{j,k}^{(n,r)}(\tilde{X}_i,\tilde{Y}_i,Z) - \psi_{P,j,k}(\tilde{X}_i,\tilde{Y}_i,Z)\right\}\right. \\
&\qquad\qquad \left.\left\{\hat{\psi}_{j',k'}^{(n,r)}(\tilde{X}_i,\tilde{Y}_i,Z) - \psi_{P,j',k'}(\tilde{X}_i,\tilde{Y}_i,Z)\right\}\right| \\
&\quad + \left|\frac{N}{n}\sum_{i\in I^{(n,r)}} \left\{\hat{\psi}_{j,k}^{(n,r)}(\tilde{X}_i,\tilde{Y}_i,Z) - \psi_{P,j,k}(\tilde{X}_i,\tilde{Y}_i,Z)\right\}\psi_{P,j',k'}(\tilde{X}_i,\tilde{Y}_i,Z)\right| \\
&\quad + \left|\frac{N}{n}\sum_{i\in I^{(n,r)}} \left\{\hat{\psi}_{j',k'}^{(n,r)}(\tilde{X}_i,\tilde{Y}_i,Z) - \psi_{P,j',k'}(\tilde{X}_i,\tilde{Y}_i,Z)\right\}\psi_{P,j,k}(\tilde{X}_i,\tilde{Y}_i,Z)\right| \\
&\le \left[\frac{N}{n}\sum_{i\in I^{(n,r)}} \left\{\hat{\psi}_{j,k}^{(n,r)}(\tilde{X}_i,\tilde{Y}_i,Z) - \psi_{P,j,k}(\tilde{X}_i,\tilde{Y}_i,Z)\right\}^2\right]^{1/2} \\
&\qquad \cdot \left[\frac{N}{n}\sum_{i\in I^{(n,r)}} \left\{\hat{\psi}_{j',k'}^{(n,r)}(\tilde{X}_i,\tilde{Y}_i,Z) - \psi_{P,j',k'}(\tilde{X}_i,\tilde{Y}_i,Z)\right\}^2\right]^{1/2} \\
&\quad + \left[\frac{N}{n}\sum_{i\in I^{(n,r)}} \left\{\hat{\psi}_{j,k}^{(n,r)}(\tilde{X}_i,\tilde{Y}_i,Z) - \psi_{P,j,k}(\tilde{X}_i,\tilde{Y}_i,Z)\right\}^2\right]^{1/2} \\
&\qquad \cdot \left[\frac{N}{n}\sum_{i\in I^{(n,r)}} \psi_{P,j',k'}(\tilde{X}_i,\tilde{Y}_i,Z)^2\right]^{1/2} \\
&\quad + \left[\frac{N}{n}\sum_{i\in I^{(n,r)}} \left\{\hat{\psi}_{j',k'}^{(n,r)}(\tilde{X}_i,\tilde{Y}_i,Z) - \psi_{P,j',k'}(\tilde{X}_i,\tilde{Y}_i,Z)\right\}^2\right]^{1/2} \\
&\qquad \cdot \left[\frac{N}{n}\sum_{i\in I^{(n,r)}} \psi_{P,j,k}(\tilde{X}_i,\tilde{Y}_i,Z)^2\right]^{1/2}.
\end{aligned}
$$

Since $|\psi_{P,j,k}|, |\psi_{P,j',k'}| \le 2$, it suffices to show that for each $j = 1,\dots,d_X$, $k = 1,\dots,d_Y$ we have

$$
\frac{N}{n}\sum_{i\in I^{(n,r)}} \left\{\hat{\psi}_{j,k}^{(n,r)}(\tilde{X}_i,\tilde{Y}_i,Z) - \psi_{P,j,k}(\tilde{X}_i,\tilde{Y}_i,Z)\right\}^2 = o_{\mathcal{P}}(1). \tag{3.29}
$$

Similarly to equation (3.28) and using the inequality $(a+b+c+d)^2 \leq 4(a^2+b^2+c^2+d^2)$,

$$\begin{aligned}
&\frac{N}{n}\sum_{i\in I^{(n,r)}}\left\{\hat{\psi}_{j,k}^{(n,r)}(\tilde{X}_i,\tilde{Y}_i,Z)-\psi_{P,j,k}(\tilde{X}_i,\tilde{Y}_i,Z)\right\}^2 \\
&\qquad=\frac{N}{n}\sum_{i\in I^{(n,r)}}\Big[\{f_{P,j}(Z_i)-\hat{f}_j^{(n,r)}(Z_i)\}\{g_{P,k}(Z_i)-\hat{g}_k^{(n,r)}(Z_i)\} \\
&\qquad\qquad\qquad+\{f_{P,j}(Z_i)-\hat{f}_j^{(n,r)}(Z_i)\}\xi_{P,ik} \\
&\qquad\qquad\qquad+\{g_{P,k}(Z_i)-\hat{g}_k^{(n,r)}(Z_i)\}\varepsilon_{P,ij}-T_{j,k}^{(n)}+\mu_{P,j,k})^2\Big]^2 \\
&\qquad\leq 4\{\tilde{a}^{(n,r)}+\tilde{b}_f^{(n,r)}+\tilde{b}_g^{(n,r)}+(\hat{T}_{j,k}^{(n)}-\mu_{P,j,k})^2\},
\end{aligned}$$

where

$$\begin{aligned}
\tilde{a}^{(n,r)} &:= \frac{N}{n}\sum_{i\in I^{(n,r)}}\{f_{P,j}(Z_i)-\hat{f}_j^{(n,r)}(Z_i)\}\{g_{P,k}(Z_i)-\hat{g}_k^{(n,r)}(Z_i)\}; \\
\tilde{b}_f^{(n,r)} &:= \frac{N}{n}\sum_{i\in I^{(n,r)}}\{f_{P,j}(Z_i)-\hat{f}_j^{(n,r)}(Z_i)\}\xi_{P,ik}; \\
\tilde{b}_g^{(n,r)} &:= \frac{N}{n}\sum_{i\in I^{(n,r)}}\{g_{P,k}(Z_i)-\hat{g}_k^{(n,r)}(Z_i)\}\varepsilon_{P,ij}.
\end{aligned}$$

Since $n^{-1/2}(T_{j,k}^{(n)}-\mu_{P,j,k})$ is uniformly asympotically Gaussian, we have that $(T_{j,k}^{(n)}-\mu_{P,j,k})^2 = O_{\mathcal{P}}(n^{-1})$. For $\tilde{a}^{(n,r)}$, $\tilde{b}_f^{(n,r)}$ and $\tilde{b}_\rho^{(n,r)}$ we use Lemma 12, noting that conditionally on $D^{(n,r)}$ each summand is i.i.d.

Using the identity $\sum_i a_i b_i \leq (\sum_i a_i)(\sum_i b_i)$ for positive sequences (a_i) and (b_i), we have

$$\left|\tilde{a}^{(n,r)}\right| \leq \frac{n}{K}\tilde{a}_f^{(n,r)}\tilde{a}_g^{(n,r)},$$

for

$$\begin{aligned}
\tilde{a}_\rho^{(n,r)} &:= \frac{N}{n}\sum_{i\in I^{(n,r)}}\{f_{P,j}(Z_i)-\hat{f}_j^{(n,r)}(Z_i)\}^2; \\
\tilde{a}_f^{(n,r)} &:= \frac{K}{n}\sum_{i\in I^{(n,r)}}\{g_{P,k}(Z_i)-\hat{g}_k^{(n,r)}(Z_i)\}^2.
\end{aligned}$$

Finally, since each $|\varepsilon_{P,ij}|, |\xi_{P,ik}| \leq 1$,

$$\begin{aligned}
\mathbb{E}_P\Big(\Big|\tilde{b}_f^{(n,r)}\Big| \;\Big|\; D^{(n,r)}\Big) \leq \mathbb{E}_P\Big(\Big|\tilde{a}_f^{(n,r)}\Big| \;\Big|\; D^{(n,r)}\Big) &= \mathbb{E}_P\Big[\Big\{f_{P,j}(Z) - \hat{f}_j^{(n,r)}(Z)\Big\}^2 \;\Big|\; D^{(n,r)}\Big] \\
&\leq E_f^{(n,r)}; \\
\mathbb{E}_P\Big(\Big|\tilde{b}_g^{(n,r)}\Big| \;\Big|\; D^{(n,r)}\Big) \leq \mathbb{E}_P\Big(\Big|\tilde{a}_g^{(n,r)}\Big| \;\Big|\; D^{(n,r)}\Big) &= \mathbb{E}_P\Big[\Big\{g_{P,k}(Z) - \hat{g}_k^{(n,r)}(Z)\Big\}^2 \;\Big|\; D^{(n,r)}\Big] \\
&\leq E_g^{(n,r)}.
\end{aligned}$$

This suffices to show (3.29), so $\hat{\Sigma}^{(n)} - \Sigma_P = o_{\mathcal{P}}(1)$. □

Lemma 33. *Let $q : \mathbb{R}^{d_X d_Y} \times \mathbb{R}^{d_X d_Y \times d_X d_Y}$ be the function computed by Algorithm 2, for any fixed $\mathbb{S} \in \mathcal{S}$. Then Assumption 2 implies Assumption 6, i.e. if $\ker(\Sigma) = \mathcal{K}^{(d_X, d_Y)}$ then the pair (T, Σ) falls in the continuity set of q almost surely, for $T \sim N(0, \Sigma)$.*

Proof. The function ϕ is continuous, since $\sigma \mapsto \mathrm{tr}(\sigma)$ and $\sigma \mapsto \mathrm{tr}(\sigma^2)$ are continuous (and positive), and so is

$$(t, g, h) \mapsto \mathbb{P}\Big(g\chi_h^2 \leq \|t\|_2^2\Big) = \frac{1}{2^{h/2}\Gamma(h/2)} \int_{-\infty}^{\|t\|_2^2/g} x^{h/2-1} e^{-x/2} dx.$$

Any merging of labels results in a linear transformation of the vector T. For any fixed transformation $\Pi \in \mathcal{C}_*$, the function

$$(t, \sigma) \mapsto \phi\Big(\Pi t, \Pi\sigma\Pi^T\Big)$$

is continuous. At each stage, we are choosing a transformation $\Pi = \Pi(T, \Sigma)$ as a function of the data. Discontinuities in the map q can only occur at points where we might choose an alternative transformation. We make this precise as follows. Let $\mathcal{D} \subset \mathbb{R}^d$ be the set of discontinuities of $q(\cdot, \Sigma)$. Then,

$$\begin{aligned}
\mathcal{D} &\subseteq \left\{t \;:\; \arg\max_{\Pi \in \mathcal{C}_*} \left\{\phi_\Pi\Big(\|\Pi t\|_2^2\Big)\right\} \text{ non-unique}\right\} \\
&\subseteq \bigcup_{\Pi \neq \Pi' \in \mathcal{C}_*} \left\{t \;:\; \phi_\Pi\Big(\|\Pi t\|_2^2\Big) = \phi_{\Pi'}\Big(\|\Pi' t\|_2^2\Big)\right\}.
\end{aligned}$$

Hence, a union bound gives

$$\mathbb{P}(T \in \mathcal{D}) \leq \sum_{\Pi \neq \Pi' \in \mathcal{C}_*} \mathbb{P}\Big(\phi_\Pi\Big(\|\Pi T\|_2^2\Big) = \phi_{\Pi'}\Big(\|\Pi' T\|_2^2\Big)\Big).$$

This is a finite sum, so it suffices to show that for arbitrary $\Pi \neq \Pi' \in \mathcal{C}_*$,

$$\mathbb{P}\Big(\phi_\Pi\Big(\|\Pi T\|_2^2\Big) = \phi_{\Pi'}\Big(\|\Pi' T\|_2^2\Big)\Big) = 0.$$

By relabelling $\Pi \leftrightarrow \Pi'$ and $X \leftrightarrow Y$ if necessary, we may assume that $\Pi = \Pi^{(A,B)}$ and $\Pi' = \Pi^{(A',B')}$ with $A \neq A'$ and $|A| \geq |A'|$. The random vectors $(\Pi T, \Pi' T)$ are jointly Guassian with zero mean and known variance, so we may compute the distribution of ΠT conditionally on $\Pi' T$. Indeed,

$$\Pi T \mid \left\{\Pi' T = x\right\} \sim N\Big(\mu_{\Pi|\Pi'}(x), \Sigma_{\Pi|\Pi'}\Big),$$

where $\mu_{\Pi|\Pi'}(x) = \Pi\Sigma\Pi'^T\Big(\Pi'\Sigma\Pi'^T\Big)^+ x$ and $\Sigma_{\Pi|\Pi'} = \Pi\Sigma\Pi^T - \Pi\Sigma\Pi'^T\Big(\Pi'\Sigma\Pi'^T\Big)^+\Pi'\Sigma\Pi^T$ is the generalised Schur complement, which is non-zero by Lemma 35. By Lemma 34 we have that $\Pi\Sigma\Pi^T \neq 0$, and so the function ϕ_Π is a bijection from $\mathbb{R}_{\geq 0}$ to $[0, 1)$. Therefore we have the following.

$$\mathbb{P}\Big(\phi_\Pi\big(\|\Pi T\|_2^2\big) = \phi_{\Pi'}\big(\|\Pi' T\|_2^2\big) \;\Big|\; \Pi' T = x\Big) = \mathbb{P}\Big(\|\Pi T\|_2^2 = \phi_\Pi^{-1}\Big(\phi_{\Pi'}\big(\|x\|_2^2\big)\Big) \;\Big|\; \Pi' T = x\Big). \tag{3.30}$$

Since $\Sigma_{\Pi|\Pi'} \neq 0$ the conditional random variable $\|\Pi T\|_2^2 \mid \left\{\Pi' T = x\right\}$ is continuously distributed (Lemma 36), and so the right hand side of (3.30) is identically zero for any $x \in \mathbb{R}^{|A'||B'|}$. □

Lemma 34. *Let $\Sigma \in \mathbb{R}^{d_X d_Y \times d_X d_Y}$ have $\ker(\Sigma) = \mathcal{K}^{(d_X, d_Y)}$. For all $\Pi \in \mathcal{C}_*$, we have $\Pi\Sigma\Pi^T \neq 0$.*

Proof. We have that $\Pi = \Pi^{(A,B)}$ for some partitions A of $\{1, \ldots, d_X\}$ and B of $\{1, \ldots, d_X\}$. Let $d'_X = |A|$ and $d'_Y = |B|$, so we have $d'_X, d'_Y \geq 2$. Recall that

$$\Pi_{(l,m)(j,k)} = \begin{cases} 1 & \text{if } j \in A_l, k \in B_m; \\ 0 & \text{otherwise.} \end{cases}$$

Define $\tilde{e}^{(l,m)} \in \mathbb{R}^{d'_X d'_Y}$, $l \in \{1, \ldots d'_X\}, m \in \{1, \ldots, d'_Y\}$, by

$$\tilde{e}^{(l,m)}_{l',m'} = \begin{cases} 1 & \text{if } l' = l, m' = m; \\ 0 & \text{otherwise.} \end{cases}$$

Further define $e^{(A_l, B_m)} \in \mathbb{R}^{d_X d_Y}$, $l \in \{1, \ldots d'_X\}, m \in \{1, \ldots, d'_Y\}$ by

$$e^{(A_l,B_m)}_{j,k} = \begin{cases} 1 & \text{if } j \in A_l, k \in B_m; \\ 0 & \text{otherwise.} \end{cases}$$

Then $\Pi^T \tilde{e}^{(l,m)} = e^{(A_l,B_m)}$. We will show that $e^{(A_l,B_m)} \notin \mathcal{K}^{(d_X,d_Y)}$. This gives that

$$\tilde{e}^{(l,m)T}\Pi\Sigma\Pi^T\tilde{e}^{(l,m)} = e^{(A_l,B_m)T}\Sigma e^{(A_l,B_m)} \neq 0,$$

and so $\Pi\Sigma\Pi^T \neq 0$.

Suppose, for a contradiction, that $e^{(A_l,B_m)} \in \mathcal{K}^{(d_X,d_Y)}$. Since $d'_X, d'_Y \geq 2$ we may pick $j_1 \in A_l$, $j_2 \notin A_l$, $k_1 \in B_m$, $k_2 \notin B_m$. Therefore

$$\begin{aligned} e^{(A_l,B_m)}_{j_1,k_1} &= 1; \\ e^{(A_l,B_m)}_{j_2,k_1} &= 0; \\ e^{(A_l,B_m)}_{j_1,k_2} &= 0; \\ e^{(A_l,B_m)}_{j_2,k_2} &= 0. \end{aligned}$$

Recalling the definitions ((3.11)–3.12), consider the set

$$\Big\{u^{(j)} \ : \ j \neq j_1\Big\} \cup \Big\{v^{(k)} \ : \ k = 1, \ldots, d_Y\Big\}.$$

This forms a basis for $\mathcal{K}^{(d_X,d_Y)}$, so we may write

$$e^{(A_l,B_m)} = \sum_{j=1}^{d_X} \alpha_j u^{(j)} + \sum_{k=1}^{d_Y} \beta_k v^{(k)},$$

for some $\alpha \in R^{d_X}$ with $\alpha_{j_1} = 0$, and for some $\beta \in \mathbb{R}^{d_Y}$. Note that for each j, k we have $u^{(j')}_{j,k} = 0$ unless $j' = j$, and similarly $v^{(k')}_{j,k} = 0$ unless $k' = k$. First, $\alpha_{j_1} = 0$ and $e^{(A_l,B_m)}_{j_1,k_1} = 1$ imply that $\beta_{k_1} = 1$. Second, $e^{(A_l,B_m)}_{j_2,k_1} = 0$ implies that $\alpha_{j_2} = -1$. Third, $e^{(A_l,B_m)}_{j_2,k_2} = 0$ implies that $\beta_{k_2} = 1$. But then we must have that $e^{(A_l,B_m)}_{j_1,k_2} = 1$, a contradiction. □

Lemma 35. *Let $\Sigma \in \mathbb{R}^{d_X d_Y \times d_X d_Y}$ have $\ker(\Sigma) = \mathcal{K}^{(d_X,d_Y)}$. Let A, A' be partitions of $\{1, \ldots, d_X\}$ and B, B' partitions of $\{1, \ldots, d_Y\}$ such that $A \neq A'$, $|A| \geq |A'|$, and all of $|A|, |A'|, |B|, |B'| \geq 2$. Define $\Pi = \Pi^{(A,B)}$ and $\Pi' = \Pi^{(A',B')}$. Then*

$$\Sigma_{\Pi|\Pi'} := \Pi\Sigma\Pi^T - \Pi\Sigma\Pi'^T\Big(\Pi'\Sigma\Pi'^T\Big)^+\Pi'\Sigma\Pi^T \neq 0.$$

Proof. Suppose for contradiction that $\Sigma_{\Pi|\Pi'} = 0$. We will show that

$$\Big(\Pi - \Pi\Sigma\Pi'^T\Big(\Pi'\Sigma\Pi'^T\Big)^+\Pi'\Big)\Sigma^{1/2} = 0, \tag{3.31}$$

and then demonstrate a vector $y \in \mathbb{R}^{d_X d_Y}$ in the image of Σ such that $\Pi' y = 0$ and $\Pi y \neq 0$. Thus there exists a vector $x \in \mathbb{R}^{d_X d_Y}$ such that $\Sigma x = y$ and

$$\Big(\Pi - \Pi\Sigma\Pi'^T\Big(\Pi'\Sigma\Pi'^T\Big)^+\Pi'\Big)\Sigma^{1/2}\Sigma^{1/2}x = \Pi y - \Pi\Sigma\Pi'^T\Big(\Pi'\Sigma\Pi'^T\Big)^+\Pi' y = \Pi y \neq 0,$$

a contradiction.

To see (3.31), write $C := \Pi\Sigma^{1/2}$ and $D := \Pi'\Sigma^{1/2}$. Then

$$0 = \Sigma_{\Pi|\Pi'} = C\Big(I - D^T\Big(DD^T\Big)^+ D\Big)C.$$

The matrix $I - D^T\Big(DD^T\Big)^+ D$ is symmetric and idempotent, so

$$C\Big(I - D^T\Big(DD^T\Big)^+ D\Big)\Big(I - D^T\Big(DD^T\Big)^+ D\Big)^T C^T = 0,$$

which in turn implies that

$$C\Big(I - D^T\Big(DD^T\Big)^+ D\Big) = 0.$$

This is precisely (3.31).

To construct y in the image of Σ, we first show that we can relabel the elements of A, A', B, B' so that the sets $A_1 \cap A'_1$, $A_2 \cap A'_1$, $B_1 \cap B'_1$, and $B_2 \cap B'_2$ are all non-empty. Indeed, since A and A' are partitions each $j = 1, \ldots, d_X$ is in some intersection $A_l \cap A'_{l'}$. If it were to hold that $A_l \supseteq A'_{l'}$ in each case, then we would have either $|A| < |A'|$ or $A = A'$. Therefore there must be some l, l' such that $A_l \cap A'_{l'}$ is not empty and A_l does not contain $A'_{l'}$. Pick $j' \in A_l^C \cap A'_{l'}$, which must itself be contained in some A_m for $m \neq l$. Relabel $A_l \leftrightarrow A_1$, $A_m \leftrightarrow A_2$, $A'_{l'} \leftrightarrow A'_1$. Similarly, each $k = 1, \ldots, d_Y$ is in some intersection $B_m \cap B'_{m'}$. If any of the $B_m^C \cap B_{m'}^{\prime C}$ are non-empty pick $k' \in B_m^C \cap B_{m'}^{\prime C}$, which must itself be contained in some $B_l \cap B'_{l'}$. Relabel $B_m \leftrightarrow B_1$, $B_l \leftrightarrow B_2$, $B'_{m'} \leftrightarrow B'_1$, $B'_{l'} \leftrightarrow B'_2$. If instead every $B_m^C \cap B_{m'}^{\prime C}$ is empty, it must be that $|B| = |B'| = 2$ with $B = B'$.

Now pick j_1, j_2, k_1, k_2 from each of $A_1 \cap A'_1$, $A_2 \cap A'_1$, $B_1 \cap B'_1$, and $B_2 \cap B'_2$ respectively. Consider the vector $y \in \mathbb{R}^{d_X d_Y}$ defined by

$$y_{j,k} = \begin{cases} 1 & \text{if } (j,k) \in \{(j_1,k_1),(j_2,k_2)\}; \\ -1 & \text{if } (j,k) \in \{(j_1,k_2),(j_2,k_1)\}; \\ 0 & \text{otherwise.} \end{cases}$$

Since Σ is symmetric, the image space of Σ is the orthogonal complement of the kernel. Recalling the definitions (3.11–3.12), we have that $y^T u^{(j)} = y^T v^{(k)}$ for all j, k, so y is in the image of Σ. Furthermore,

$$\Big(\Pi' y\Big)_{1,1} = y_{j_1,k_1} + y_{j_2,k_1} = 0;$$
$$\Big(\Pi' y\Big)_{1,2} = y_{j_1,k_2} + y_{j_2,k_2} = 0,$$

so $\Pi' y = 0$. Finally, we have that $(\Pi y)_{1,1} = y_{j_1,k_1} = 1$, so $\Pi y \neq 0$. This completes the proof. □

Lemma 36. *Let $T \sim N(0,\Sigma)$ for $\Sigma \in \mathcal{M}^{(d_X,d_Y)}$. If $\Sigma \neq 0$, then $\|T\|_2^2$ has a continuous cumulative distribution function.*

Proof. If $\Sigma \neq 0$ then Σ possesses strictly positive eigenvalues $\lambda_1,\dots,\lambda_r$ for $1 \leq r = \operatorname{rank}(\Sigma) \leq d$. The random variable $\|T\|_2^2$ is a weighted sum of chi-squares,

$$\|T\|_2^2 \stackrel{d}{=} \sum_{j=1}^{r} \lambda_j Z_j^2,$$

where $Z_j \sim N(0,1)$ are independent. This follows a continuous distribution on $\mathbb{R}_{\geq 0}$. □

Lemma 37. *Let $q : \mathbb{R}^{d_X d_Y} \times \mathbb{R}^{d_X d_Y \times d_X d_Y} \to \mathbb{R}^L$ be the output of Algorithm 2 for any fixed $\mathbb{S} \in \mathcal{S}$. Then Assumption 2 implies Assumption 7, i.e. if $\ker(\Sigma) = \mathcal{K}^{(d_X,d_Y)}$ then the functions $F_l(\cdot;\Sigma) := \mathbb{P}\Big(q_l\Big(N(0,\Sigma),\Sigma\Big) \leq x\Big)$ are continuous, for $l = 1,\dots,L$.*

Proof. This follows from Lemma 38. Indeed, let Π_l be the transformation selected at the lth stage of Algorithm 2 on input $(t,\Sigma,\mathbb{S})$. Then $q_l(T,\Sigma)$ equals precisely $\tilde{\phi}(T)$ in equation (3.32) with the choice of map

$$t \mapsto \Pi_l(t)\dots\Pi_1(t).$$

□

Lemma 38. *In an abuse of notation let $\Pi : t \mapsto \Pi(t)$ be any map from $\mathbb{R}_X^d d_Y$ to $\mathcal{C}_*$. Write*

$$\tilde{\phi}(t) := \phi_{\Pi(t)}\Big(\|\Pi(t)t\|_2^2\Big) = \phi\Big(\Pi(t)t, \Pi(t)\Sigma\Pi(t)^T\Big). \tag{3.32}$$

If $\ker(\Sigma) = \mathcal{K}^{(d_X,d_Y)}$ then the random variable $\tilde{\phi}(T)$, where $T \sim N(0,\Sigma)$, has continuous cumulative density function.

Proof. Given $\epsilon > 0$, we will choose $\delta > 0$ such that all $x, x' \in [0,1]$ with $|x - x'| < \delta$ satisfy

$$\Big|\mathbb{P}\Big(\tilde{\phi}(T) \leq x\Big) - \mathbb{P}\Big(\tilde{\phi}(T) \leq x'\Big)\Big| < \epsilon. \tag{3.33}$$

Indeed, let $0 \leq x' < x \leq 1$. Now,

$$\begin{aligned}
\Big|\mathbb{P}\Big(\tilde{\phi}(T) \leq x\Big) - \mathbb{P}\Big(\tilde{\phi}(T) \leq x'\Big)\Big| &= \mathbb{P}\Big(x' < \phi_{\Pi(T)}\Big(\|\Pi(T)T\|_2^2\Big) \leq x\Big) \\
&\leq \mathbb{P}\Big(\bigcup_{\Pi \in \mathcal{C}_*} \Big\{x' < \phi_\Pi\Big(\|\Pi T\|_2^2\Big) \leq x\Big\}\Big) \\
&\leq \sum_{\Pi \in \mathcal{C}_*} \Big|\mathbb{P}\Big(\phi_\Pi\Big(\|\Pi T\|_2^2\Big) \leq x\Big) - \mathbb{P}\Big(\phi_\Pi\Big(\|\Pi T\|_2^2\Big) \leq x'\Big)\Big|.
\end{aligned}$$

In the final line we have used a union bound. Lemma 39 gives that each $\phi_\Pi\Big(\|\Pi T\|_2^2\Big)$ has continuous cumulative distribution function. Since the set $\mathcal{C}_*$ is finite, we can choose $\delta > 0$ such that for all $|x - x'| < \delta$,

$$\max_{\Pi \in \mathcal{C}_*} \left| \mathbb{P}\Big(\phi_\Pi\Big(\|\Pi T\|_2^2\Big) \le x\Big) - \mathbb{P}\Big(\phi_\Pi\Big(\|\Pi T\|_2^2\Big) \le x'\Big) \right| < \frac{\epsilon}{|\mathcal{C}_*|}.$$

This suffices to prove (3.33). □

Lemma 39. *Let* $\ker(\Sigma) = \mathcal{K}^{(d_X, d_Y)}$. *Then for all* $\Pi \in \mathcal{C}_*$, *the random variable* $\phi_\Pi\Big(\|\Pi T\|_2^2\Big)$ *has continuous cumulative density function, where* $T \sim N(0, \Sigma)$.

Proof. By Lemma 34, $\Pi\Sigma\Pi^T$ is non-zero. This means that ϕ_Π is continuous and strictly increasing. Hence ϕ_Π^{-1} exists and is continuous. Now, for $x \in (0, 1)$,

$$\begin{aligned}\mathbb{P}\Big(\phi_\Pi\Big(\|\Pi T\|_2^2\Big) \le x\Big) &= \mathbb{P}\Big(\|\Pi T\|_2^2 \le \phi_\Pi^{-1}(x)\Big) \\ &= G_\Pi\Big(\phi_\Pi^{-1}(x)\Big),\end{aligned}$$

where G_Π is the cumulative density function of the random variable $\|T\Pi\|_2^2$. Since $\Pi\Sigma\Pi^T$ is non-zero and symmetric positive semi-definite, Lemma 36 gives that G_Π is continuous. Therefore the cumulative distribution function

$$x \mapsto \mathbb{P}\Big(\phi_\Pi\Big(\|\Pi T\|_2^2\Big) \le x\Big) = G_\Pi\Big(\phi_\Pi^{-1}(x)\Big)$$

is a composition of continuous functions, and so is continuous. □

3.7.1 Proof of Theorem 31

Proof. We verify the conditions of Theorem 30 for the transformed variables $\hat{M}^{(n)} := q\Big(\hat{T}^{(n)}, \hat{\Sigma}^{(n)}\Big)$ and $M := q(T, \Sigma)$. Indeed, we may apply the continuous mapping theorem along arbitrary sequences $P_n \in \mathcal{P}$ (van der Vaart (1998, Thm. 2.3), Kasy (2018, Thm. 1)) to show that Assumptions 5 and 6 imply

$$\lim_{n \to \infty} \sup_{P \in \mathcal{P}} \sup_{m \in \mathbb{R}^L} \left| \mathbb{P}_P\Big(\hat{M}^{(n)} \le m\Big) - \mathbb{P}(M \le m) \right| = 0.$$

Assumption 7 gives the remaining condition for Assumption 3.

It remains to check Assumption 4, which we will show follows from Assumptions 6 and 7. We wish to apply Lemma 40 to each of the following classes of cumulative distribution

functions:

$$F_l(x;\sigma) := \mathbb{P}\Big(q_l\Big(N(0,\sigma),\sigma\Big) \le x\Big), \quad l = 1,\dots,L;$$
$$\psi(u;\sigma) := \mathbb{P}\Big[G\Big(q_l\Big(N(0,\sigma),\sigma\Big)\Big) \le u\Big].$$

We have that $F_l(\cdot;\Sigma)$ and $\psi(\cdot;\Sigma)$ are continuous by Assumption 7 and Lemma 41, so we need to check that the functions F_l and ψ satisfy (3.34). Since $\hat{\Sigma}^{(n)}$ tends uniformly in probability to Σ it suffices to check that $F_l(x;\cdot)$ and $\psi(u;\cdot)$ are continuous at Σ, again by the continuous mapping theorem along arbitrary sequences $P_n \in \mathcal{P}$ (van der Vaart (1998, Thm. 2.3), Kasy (2018, Thm. 1)).

First, fix $x \in \mathbb{R}$ and consider $F_l(x;\cdot)$ for some $l \in \{1,\dots,L\}$. Let Z be a standard Gaussian random variable in $\mathbb{R}^d$. We have that, for any $\delta_0 > 0$,

$$\begin{aligned}
F_l(x;\sigma) &= \mathbb{P}\Big(q\Big(\sigma^{1/2}Z,\sigma\Big) \le x\Big) \\
&= \mathbb{P}\Big(q\Big(\sigma^{1/2}Z,\sigma\Big) - q\Big(\Sigma^{1/2}Z,\Sigma\Big) + q\Big(\Sigma^{1/2}Z,\Sigma\Big) \le x\Big) \\
&\le \mathbb{P}\Big(\Big|q\Big(\sigma^{1/2}Z,\sigma\Big) - q\Big(\Sigma^{1/2}Z,\Sigma\Big)\Big| > \delta_0\Big) + \mathbb{P}\Big(q\Big(\Sigma^{1/2}Z,\Sigma\Big) \le x+\delta_0\Big) \\
&= \mathbb{P}\Big(\Big|q\Big(\sigma^{1/2}Z,\sigma\Big) - q\Big(\Sigma^{1/2}Z,\Sigma\Big)\Big| > \delta_0\Big) + F_l(x+\delta_0;\Sigma),
\end{aligned}$$

where the third line follows from a union bound. Similarly,

$$\begin{aligned}
F_l(x;\sigma) &= 1 - \mathbb{P}\Big(q\Big(\sigma^{1/2}Z,\sigma\Big) - q\Big(\Sigma^{1/2}Z,\Sigma\Big) + q\Big(\Sigma^{1/2}Z,\Sigma\Big) > x\Big) \\
&\ge 1 - \mathbb{P}\Big(\Big|q\Big(\sigma^{1/2}Z,\sigma\Big) - q\Big(\Sigma^{1/2}Z,\Sigma\Big)\Big| > \delta_0\Big) - \mathbb{P}\Big(q\Big(\Sigma^{1/2}Z,\Sigma\Big) > x-\delta_0\Big) \\
&= -\mathbb{P}\Big(\Big|q\Big(\sigma^{1/2}Z,\sigma\Big) - q\Big(\Sigma^{1/2}Z,\Sigma\Big)\Big| > \delta_0\Big) + F_l(x-\delta_0;\Sigma).
\end{aligned}$$

Thus

$$\begin{aligned}
|F_l(x;\sigma) - F_l(x;\Sigma)| \le \mathbb{P}\Big(\Big|q_l\Big(\sigma^{1/2}Z,\sigma\Big) - q_l\Big(\Sigma^{1/2}Z,\Sigma\Big)\Big| > \delta_0\Big) \\
+ \sup_{x' \in [x-\delta_0, x+\delta_0]} |F_l(x';\Sigma) - F_l(x;\Sigma)|
\end{aligned}$$

Given $\epsilon > 0$, we wish to show that there exists $\delta > 0$ such that $\|\sigma - \Sigma\| < \delta$ implies $|F_l(x;\sigma) - F_l(x;\Sigma)| < \epsilon$. Since $F_l(\cdot;\Sigma)$ is continuous, we may choose δ_0 sufficiently small so that

$$\sup_{x' \in [x-\delta_0, x+\delta_0]} |F_l(x';\Sigma) - F_l(x;\Sigma)| < \frac{\epsilon}{2}.$$

We will now show that for $\delta > 0$ sufficiently small and all $\sigma \in \mathcal{M}^{(d_X, d_Y)}$ such that $\|\sigma - \Sigma\| < \delta$,

$$\mathbb{P}\Big(\Big|q_l\Big(\sigma^{1/2}Z, \sigma\Big) - q_l\Big(\Sigma^{1/2}Z, \Sigma\Big)\Big| > \delta_0\Big) < \frac{\epsilon}{2}.$$

By Assumption 6, the random vector $(\Sigma^{1/2}Z, \Sigma)$ takes values on the continuity set of the function q_l with probability one. Therefore there exist $\delta_1, \delta_2 > 0$ such that

$$\|t_1 - t_2\|_2 < \delta_1,\ \|\sigma_1 - \sigma_2\| < \delta_2 \implies |q_l(t_1, \sigma_1) - q_l(t_2, \sigma_2)| < \delta_0.$$

We will choose $\delta \leq \delta_2$, so it suffices to show that, for $\delta > 0$ sufficiently small,

$$\|\sigma - \Sigma\| < \delta \implies \mathbb{P}\Big(\Big\|\sigma^{1/2}Z - \Sigma^{1/2}Z\Big\|_2 > \delta_1\Big) < \frac{\epsilon}{2}.$$

Pick $\epsilon_1 > 0$ sufficiently small so that $\mathbb{P}\Big(\|Z\Big\|_2 > \delta_1/\epsilon_1\Big) < \frac{\epsilon}{2}$. Since the operation of taking the (principal) square root of a positive semidefinite matrix is continuous, there exists $\delta_3 > 0$ such that

$$\|\sigma - \Sigma\|_F < \delta_3 \implies \Big\|\sigma^{1/2} - \Sigma^{1/2}\Big\|_F < \epsilon_1.$$

Finally, by the equivalence of norms on finite dimensional vector spaces, there exists a $c > 0$ depending only on the choice of matrix norm $\|\cdot\|$ such that $\|\sigma - \Sigma\|_F \leq c\|\sigma - \Sigma\|$. Pick $\delta = \delta_3/c > 0$. We have that

$$\begin{aligned}
\|\sigma - \Sigma\| < \delta &\implies \|\sigma - \Sigma\|_F < \delta_3 \\
&\implies \Big\|\sigma^{1/2} - \Sigma^{1/2}\Big\|_F < \epsilon_1 \\
&\implies \mathbb{P}\Big(\Big\|\sigma^{1/2} - \Sigma^{1/2}\Big\|_F \|Z\|_2 > \delta_1\Big) < \frac{\epsilon}{2} \\
&\implies \mathbb{P}\Big(\Big\|\sigma^{1/2}Z - \Sigma^{1/2}Z\Big\|_2 > \delta_1\Big) < \frac{\epsilon}{2}.
\end{aligned}$$

Therefore F_l satisfies (3.34). It remains to show that ψ satisfies (3.34). Since the functions $F_l(\cdot; \Sigma)$ are continuous, the function $(t, \sigma) \mapsto G(q(t, \sigma))$ satisfies Assumption 6 in place of q. Recall that we have already showed that ψ is continuous. Therefore an identical proof to above — with F_l replaced by ψ and $q_l(\cdot, \cdot)$ replaced by $G(q(\cdot, \cdot))$ — yields the result, and Assumption 4 is verified. □

Lemma 40. *Let $F(\cdot; \sigma) : \mathbb{R} \to [0, 1]$ be a class of cumulative density functions indexed by σ, and let $\hat{\Sigma}^{(n)}$ and Σ be such that $F(\cdot; \Sigma)$ is continuous and for each $x \in \mathbb{R}$ and $\epsilon > 0$,*

$$\lim_{n \to \infty} \sup_{P \in \mathcal{P}} \mathbb{P}_P\Big(\Big|F\Big(x; \hat{\Sigma}^{(n)}\Big) - F(x; \Sigma)\Big| > \epsilon\Big) = 0. \tag{3.34}$$

Then for any $\epsilon > 0$,

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}} \mathbb{P}_P\left(\sup_{x\in\mathbb{R}} \left| F\left(x; \hat{\Sigma}^{(n)}\right) - F(x;\Sigma)\right| > \epsilon \right) = 0.$$

Proof. The proof is similar to that of van der Vaart (1998, Lem. 2.11). Fix $K \in \mathbb{N}$ such that $1/K < \epsilon/2$. By the continuity of $F(\cdot;\Sigma)$, there exist points $-\infty = x_0 < x_1 < \ldots < x_K = \infty$ with $F(x_i;\Sigma) = i/K$. By monotonicity, for $x_{i-1} \le x \le x_i$,

$$F\left(x; \hat{\Sigma}^{(n)}\right) - F(x;\Sigma) \le F\left(x_i; \hat{\Sigma}^{(n)}\right) - F(x_{i-1};\Sigma) = F\left(x_i; \hat{\Sigma}^{(n)}\right) - F(x_i;\Sigma) + 1/K$$
$$F\left(x; \hat{\Sigma}^{(n)}\right) - F(x;\Sigma) \ge F\left(x_{i-1}; \hat{\Sigma}^{(n)}\right) - F(x_i;\Sigma) = F\left(x_{i-1}; \hat{\Sigma}^{(n)}\right) - F(x_{i-1};\Sigma) - 1/K.$$

Therefore

$$\sup_{x\in\mathbb{R}} \left| F\left(x; \hat{\Sigma}^{(n)}\right) - F(x;\Sigma)\right| \le \max_{i=1,\ldots,K} \left| F\left(x_i; \hat{\Sigma}^{(n)}\right) - F(x_i;\Sigma)\right| + \epsilon/2.$$

Now making use of a union bound,

$$\begin{aligned}
&\sup_{P\in\mathcal{P}} \mathbb{P}_P\left(\sup_{x\in\mathbb{R}} \left| F\left(x; \hat{\Sigma}^{(n)}\right) - F(x;\Sigma)\right| > \epsilon \right) \\
&\qquad\qquad \le \sup_{P\in\mathcal{P}} \mathbb{P}_P\left(\max_{i=1,\ldots,K} \left| F\left(x_i; \hat{\Sigma}^{(n)}\right) - F(x_i;\Sigma)\right| > \epsilon/2 \right) \\
&\qquad\qquad = \sup_{P\in\mathcal{P}} \mathbb{P}_P\left(\bigcup_{i=1,\ldots,K} \left\{ \left| F\left(x_i; \hat{\Sigma}^{(n)}\right) - F(x_i;\Sigma)\right| > \epsilon/2 \right\} \right) \\
&\qquad\qquad \le \sum_{i=1}^{K} \sup_{P\in\mathcal{P}} \mathbb{P}_P\left(\left| F\left(x_i; \hat{\Sigma}^{(n)}\right) - F(x_i;\Sigma)\right| > \epsilon/2 \right),
\end{aligned}$$

which tends to zero as $n \to \infty$. □

3.7.2 Proof of Theorem 30

Proof. We first introduce some notation. Recall that we write $\mathbb{P}_n^B$ for the empirical law of the bootstrap samples $M_n^{(b)}$. For $x \in \mathbb{R}$, $m \in \mathbb{R}^L$, and $u \in [0,1]$ let

$$\begin{aligned}
F_l(x) &:= \mathbb{P}(M_l \le x); & G(m) &:= \max_{l=1,\ldots,L} F_l(m_l); & \psi(u) &:= \mathbb{P}(G(M) \le u); \\
\hat{F}_{n,l}^B(x) &:= \mathbb{P}_n^B(M_{n,l}^{(b)} \le x); & \hat{G}_n^B(m) &:= \max_{l=1,\ldots,L} \hat{F}_{n,l}^B(m_l); & \hat{\psi}_n^B(u) &:= \mathbb{P}_n^B(\hat{G}_n^B(M_n^{(b)}) \le u).
\end{aligned}$$

In Algorithm 3, we have that $p = 1 - \hat{\psi}_n^B(\hat{G}_n^B(\hat{M}^{(n)}))$. Therefore it suffices to show that

$$\lim_{n,B\to\infty} \sup_{P\in\mathcal{P}} \sup_{u\in[0,1]} \left| \mathbb{P}_P\left(\hat{\psi}_n^B\left(\hat{G}_n^B\left(\hat{M}^{(n)} \right) \right) \le u \right) - u \right| = 0. \tag{3.35}$$

We will make use of the fact that

$$\sup_{u\in[0,1]}\left|\mathbb{P}\Big(\psi(G(M))\le u\Big)-u\right|=0, \tag{3.36}$$

which is immediate from the definition of ψ.

Given $\epsilon>0$, we fix some $P\in\mathcal{P}$ and $u\in[0,1]$. Decompose (3.35) as follows.

$$\begin{aligned}
&\mathbb{P}_P\Big(\hat{\psi}_n^B\Big(\hat{G}_n^B\Big(\hat{M}^{(n)}\Big)\Big)\le u\Big)-u\\
&\qquad=\mathbb{P}_P\Big(\hat{\psi}_n^B\Big(\hat{G}_n^B\Big(\hat{M}^{(n)}\Big)\Big)-\psi\Big(G\Big(\hat{M}^{(n)}\Big)\Big)+\psi\Big(G\Big(\hat{M}^{(n)}\Big)\Big)\le u\Big)-u\\
&\qquad\le\mathbb{P}_P\Big(\psi\Big(G\Big(\hat{M}^{(n)}\Big)\Big)\le u+\epsilon/3\Big)-u\\
&\qquad\quad+\mathbb{P}_P\Big(\Big|\hat{\psi}_n^B\Big(\hat{G}_n^B\Big(\hat{M}^{(n)}\Big)\Big)-\psi\Big(G\Big(\hat{M}^{(n)}\Big)\Big)\Big|>\epsilon/3\Big)\\
&\qquad=\mathbb{P}\Big(\psi(G(M))\le u+\epsilon/3\Big)-u\\
&\qquad\quad+\mathbb{P}_P\Big(\psi\Big(G\Big(\hat{M}^{(n)}\Big)\Big)\le u+\epsilon/3\Big)-\mathbb{P}\Big(\psi(G(M))\le u+\epsilon/3\Big)\\
&\qquad\quad+\mathbb{P}_P\Big(\Big|\hat{\psi}_n^B\Big(\hat{G}_n^B\Big(\hat{M}^{(n)}\Big)\Big)-\psi\Big(G\Big(\hat{M}^{(n)}\Big)\Big)\Big|>\epsilon/3\Big)\\
&\qquad\le\epsilon/3\\
&\qquad\quad+\sup_{u'\in[0,1]}\left|\mathbb{P}\Big(\psi\Big(G\Big(\hat{M}^{(n)}\Big)\Big)\le u'\Big)-\mathbb{P}\Big(\psi(G(M))\le u'\Big)\right|\\
&\qquad\quad+\mathbb{P}_P\Big(\sup_{m\in\mathbb{R}^L}\Big|\hat{\psi}_n^B\Big(\hat{G}_n^B(m)\Big)-\psi(G(m))\Big|>\epsilon/3\Big).
\end{aligned}$$

The second line makes use of a union bound, and the final line applies (3.36). We may produce a similar lower bound, so taking the supremum over P and u we achieve the following.

$$\begin{aligned}
&\sup_{P\in\mathcal{P}}\sup_{u\in[0,1]}\left|\mathbb{P}_P\Big(\hat{\psi}_n^B\Big(\hat{G}_n^B\Big(\hat{M}^{(n)}\Big)\Big)\le u\Big)-u\right|\le\epsilon/3\\
&\qquad+\sup_{P\in\mathcal{P}}\sup_{u\in[0,1]}\left|\mathbb{P}_P\Big(\psi\Big(G\Big(\hat{M}^{(n)}\Big)\Big)\le u\Big)-\mathbb{P}\Big(\psi(G(M))\le u\Big)\right| &(3.37)\\
&\qquad+\sup_{P\in\mathcal{P}}\mathbb{P}_P\Big(\sup_{m\in\mathbb{R}^L}\Big|\hat{\psi}_n^B\Big(\hat{G}_n^B(m)\Big)-\psi(G(m))\Big|>\epsilon/3\Big). &(3.38)
\end{aligned}$$

Assumption 3 and Lemma 41 give that $\psi\circ G$ is continuous, so $\psi\Big(G\Big(\hat{M}^{(n)}\Big)\Big)$ converges uniformly in distribution to $\psi(G(M))$ by Assumption 3 and the continuous mapping theorem along arbitrary sequences $P_n\in\mathcal{P}$ (van der Vaart (1998, Thm. 2.3), Kasy (2018, Thm. 1)). Therefore the quantity (3.37) is at most $\epsilon/3$ for all n sufficiently large. It remains to show that for all n and B sufficiently large, the quantity (3.38) is at most $\epsilon/3$.

We further decompose the bootstrap approximation error in (3.38) as follows. Again fixing $P \in \mathcal{P}$,

$$\begin{aligned}
&\sup_{m\in\mathbb{R}^L} \left|\hat{\psi}_n^B\Big(\hat{G}_n^B(m)\Big) - \psi(G(m))\right| \\
&\qquad = \sup_{m\in\mathbb{R}^L} \left|\hat{\psi}_n^B\Big(\hat{G}_n^B(m)\Big) - \psi\Big(\hat{G}_n^B(m)\Big) + \psi\Big(\hat{G}_n^B(m)\Big) - \psi(G(m))\right| \\
&\qquad \leq \sup_{u\in[0,1]} \left|\hat{\psi}_n^B(u) - \psi(u)\right| + \sup_{m\in\mathbb{R}^L} \left|\psi\Big(\hat{G}_n^B(m)\Big) - \psi(G(m))\right| \\
&\qquad \leq \sup_{u\in[0,1]} \left|\hat{\psi}_n^B(u) - \psi(u)\right| + L \sup_{m\in\mathbb{R}^L} \left|\hat{G}_n^B(m) - G(m)\right|.
\end{aligned}$$

In the final line we have applied Lemma 41. Let $\eta \in (0,1)$ be chosen later. We utilise a union bound in order to apply Lemma 42. Indeed,

$$\begin{aligned}
&\sup_{P\in\mathcal{P}} \mathbb{P}_P\left(\sup_{m\in\mathbb{R}^L} \left|\hat{\psi}_n^B\Big(\hat{G}_n^B(m)\Big) - \psi(G(m))\right| > \epsilon/3 \right) \\
&\qquad \leq \sup_{P\in\mathcal{P}} \mathbb{P}_P\left(\sup_{u\in[0,1]} \left|\hat{\psi}_n^B(u) - \psi(u)\right| > \frac{\eta\epsilon}{3} \right) \\
&\qquad\quad + \sup_{P\in\mathcal{P}} \mathbb{P}_P\left(\sup_{m\in\mathbb{R}^L} \left|\hat{G}_n^B(m) - G(m)\right| > \frac{(1-\eta)\epsilon}{3L} \right) \\
&\qquad \leq 2(L+1)\exp\left\{ -\frac{B\eta^2\epsilon^2}{18(1+L)^2} \right\} + 2L\exp\left\{ -\frac{B(1-\eta)^2\epsilon^2}{18L^2} \right\}.
\end{aligned}$$

The choice $\eta = (1+L)/(1+2L)$ gives the following bound. For all $B \in \mathbb{N}$ and n sufficiently large,

$$\begin{aligned}
&\sup_{P\in\mathcal{P}} \mathbb{P}_P\left(\sup_{m\in\mathbb{R}^L} \left|\hat{\psi}_n^B\Big(\hat{G}_n^B(m)\Big) - \psi(G(m))\right| > \epsilon/3 \right) \\
&\qquad\qquad \leq 2(2L+1)\exp\left\{ -\frac{B\epsilon^2}{18(1+2L)^2} \right\}. \qquad (3.39)
\end{aligned}$$

For B sufficiently large, the quantity (3.39) is at most $\epsilon/3$.

We deduce that for all n, B sufficiently large,

$$\sup_{P\in\mathcal{P}} \sup_{u\in[0,1]} \left| \mathbb{P}_P\Big(\hat{\psi}_n^B\Big(\hat{G}_n^B\Big(\hat{M}^{(n)}\Big)\Big) \leq u \Big) - u \right| \leq \epsilon.$$

This completes the proof. □

Lemma 41. *Let M be a random vector in $\mathbb{R}^L$ with continuously distributed marginals, i.e. the marginal cumulative density functions $F_l(x) = \mathbb{P}(M_l \leq x)$ are continuous on $\mathbb{R}$. Then the function*

$$\psi(u) := \mathbb{P}\left(\max_{l=1,\ldots,L} F_l(M_l) \leq u \right)$$

is Lipschitz continuous on $[0,1]$ *with* $\|\psi\|_{Lip} \leq L$.

Proof. Let $0 \leq a < b \leq 1$. We apply a union bound to deal with the maximum over $l = 1, \ldots, L$ as follows.

$$\begin{aligned}
|\psi(b) - \psi(a)| &= \mathbb{P}\left(a < \max_{l=1,\ldots,L} F_l(M_l) \leq b\right) \\
&\leq \mathbb{P}\left(\bigcup_{l=1,\ldots,L} \{a < F_l(M_l) \leq b\}\right) \\
&\leq \sum_{l=1}^{L} \mathbb{P}(a < F_l(M_l) \leq b).
\end{aligned}$$

Since M has continuous marginals, the random vector $(F_1(M_1), \ldots, F_L(M_L))$ has uniformly distributed marginals. Therefore, for each $l = 1, \ldots, L$,

$$\mathbb{P}(a < F_l(M_l) \leq b) \leq |b - a|.$$

This completes the proof. □

Lemma 42. *Let* $G, \psi, \hat{G}_n^B, \hat{\psi}_n^B$ *be defined as in the proof of Theorem 30, and let Assumption 4 hold. Given* $\epsilon > 0$ *and* $B \in \mathbb{N}$*, there exists* $N_B \in \mathbb{N}$ *such that for all* $n \geq N_B$*,*

$$\sup_{P \in \mathcal{P}} \mathbb{P}_P\left(\sup_{m \in \mathbb{R}^L} \left|\hat{G}_n^B(m) - G(m)\right| > \epsilon\right) \leq 3L \exp\left\{-\frac{B\epsilon^2}{2}\right\}; \tag{3.40}$$

$$\sup_{P \in \mathcal{P}} \mathbb{P}_P\left(\sup_{u \in [0,1]} \left|\hat{\psi}_n^B(u) - \psi(u)\right| > \epsilon\right) \leq 3(L+1) \exp\left\{-\frac{B\epsilon^2}{2(1+L)^2}\right\}. \tag{3.41}$$

Proof. Let $\mathbb{P}_n$ be the population law of the bootstrap samples $M_n^{(b)}$ conditionally on the data, so $M_n^{(b)}$ are independent draws from $\mathbb{P}_n$ for $b = 1, \ldots, B$. Define the following functions based on $\mathbb{P}_n$. For $x \in \mathbb{R}$ and $m \in \mathbb{R}^L$, let

$$F_{n,l}(x) := \mathbb{P}_n(M_{n,l}^{(b)} \leq x); \quad G_n(m) := \max_{l=1,\ldots,L} F_{n,l}(m_l).$$

We begin with the first claim (3.40), decomposing the error as follows.

$$\begin{aligned}
&\sup_{P\in\mathcal{P}} \mathbb{P}_P\Big(\sup_{m\in\mathbb{R}^L} \big|\hat{G}_n^B(m) - G(m)\big| > \epsilon \Big) \\
&\qquad = \sup_{P\in\mathcal{P}} \mathbb{P}_P\Big(\sup_{m\in\mathbb{R}^L} \big|\hat{G}_n^B(m) - G_n(m) + G_n(m) - G(m)\big| > \epsilon \Big) \\
&\qquad \le \sup_{P\in\mathcal{P}} \mathbb{P}_P\Big(\sup_{m\in\mathbb{R}^L} \big|\hat{G}_n^B(m) - G_n(m)\big| > \epsilon/2 \Big) \\
&\qquad\quad + \sup_{P\in\mathcal{P}} \mathbb{P}_P\Big(\sup_{m\in\mathbb{R}^L} \big|G_n(m) - G(m)\big| > \epsilon/2 \Big) \\
&\qquad \le \sum_{l=1}^{L} \sup_{P\in\mathcal{P}} \mathbb{P}_P\Big(\sup_{x\in\mathbb{R}} \big|\hat{F}_{n,l}^B(x) - F_{n,l}(x)\big| > \epsilon/2 \Big) \qquad (3.42) \\
&\qquad\quad + \sum_{l=1}^{L} \sup_{P\in\mathcal{P}} \mathbb{P}_P\Big(\sup_{x\in\mathbb{R}} \big|F_{n,l}(x) - F_l(x)\big| > \epsilon/2 \Big). \qquad (3.43)
\end{aligned}$$

Here, the second and third lines are both union bounds. The quantity (3.42) is bounded using the Dvoretzky–Kiefer–Wolfowitz inequality (Dvoretzky et al., 1956; Massart, 1990), which gives

$$\sup_{P\in\mathcal{P}} \mathbb{P}_P\Big(\sup_{x\in\mathbb{R}} \big|\hat{F}_{n,l}^B(x) - F_{n,l}(x)\big| > \epsilon/2 \Big) \le 2\exp\Big(-\frac{B\epsilon^2}{2} \Big).$$

The quantity (3.43) is handled using Assumption 4. We may choose N_B such that for all $n \ge N_B$ and $l = 1, \dots, L$,

$$\sup_{P\in\mathcal{P}} \mathbb{P}_P\Big(\sup_{x\in\mathbb{R}} \big|F_{n,l}(x) - F_l(x)\big| > \epsilon/2 \Big) \le \exp\Big(-\frac{B\epsilon^2}{2} \Big).$$

Combining these bounds suffices to prove (3.40).

We now turn to the second claim (3.41). Fix $P \in \mathcal{P}$ and $u \in [0,1]$. Let $\tilde{\epsilon} > 0$ be chosen later. By a union bound and the triangle inequality,

$$\begin{aligned}
\psi_n^B(u) &= \mathbb{P}_n^B\Big(\hat{G}_n^B\big(M_n^{(b)}\big) \le u \Big) \\
&= \mathbb{P}_n^B\Big(\hat{G}_n^B\big(M_n^{(b)}\big) - G\big(M_n^{(b)}\big) + G\big(M_n^{(b)}\big) \le u \Big) \\
&\le \mathbb{P}_n^B\Big(G\big(M_n^{(b)}\big) \le u + \tilde{\epsilon} \Big) + \mathbb{1}\Big\{ \sup_{m\in\mathbb{R}^L} \big|\hat{G}_n^B(m) - G(m)\big| > \tilde{\epsilon} \Big\} \\
&\le \psi(u+\tilde{\epsilon}) + \sup_{u\in[0,1]} \Big| \mathbb{P}_n^B\Big(G\big(M_n^{(b)}\big) \le u \Big) - \psi(u) \Big| + \mathbb{1}\Big\{ \sup_{m\in\mathbb{R}^d} \big|\hat{G}_n^B(m) - G(m)\big| > \tilde{\epsilon} \Big\}.
\end{aligned}$$

We may produce a similar lower bound. Next, we use the Lipschitz continuity of ψ (Lemma 41) and take the supremum over u. Indeed,

$$\sup_{u\in[0,1]}\left|\hat{\psi}_n^B(u)-\psi_P(u)\right| \le L\tilde{\epsilon} + \sup_{u\in[0,1]}\left|\mathbb{P}_n^B\Big(G\Big(M_n^{(b)}\Big)\le u\Big)-\psi(u)\right| + \mathbb{1}\Big\{\sup_{m\in\mathbb{R}^d}\left|\hat{G}_n^B(m)-G(m)\right|>\tilde{\epsilon}\Big\}.$$

Pick $\tilde{\epsilon}=\epsilon/(1+L)$. Now applying a union bound,

$$\begin{aligned}
&\sup_{P\in\mathcal{P}}\mathbb{P}_P\Big(\sup_{u\in[0,1]}\left|\hat{\psi}_n^B(u)-\psi(u)\right|>\epsilon\Big)\\
&\quad\le \sup_{P\in\mathcal{P}}\mathbb{P}_P\Big(\sup_{u\in[0,1]}\left|\mathbb{P}_n^B\Big(G\Big(M_n^{(b)}\Big)\le u\Big)-\psi(u)\right| + \mathbb{1}\Big\{\sup_{m\in\mathbb{R}^L}\left|\hat{G}_n^B(m)-G(m)\right|>\tilde{\epsilon}\Big\}>\tilde{\epsilon}\Big)\\
&\quad\le \sup_{P\in\mathcal{P}}\mathbb{P}_P\Big(\sup_{u\in[0,1]}\left|\mathbb{P}_n^B\Big(G\Big(M_n^{(b)}\Big)\le u\Big)-\psi(u)\right|>\tilde{\epsilon}\Big) \qquad (3.44)\\
&\qquad+\sup_{P\in\mathcal{P}}\mathbb{P}_P\Big(\sup_{m\in\mathbb{R}^L}\left|\hat{G}_n^B(m)-G(m)\right|>\tilde{\epsilon}\Big). \qquad (3.45)
\end{aligned}$$

We have already shown that the quantity (3.45) satisfies (3.40). Turning to (3.44),

$$\begin{aligned}
&\sup_{P\in\mathcal{P}}\mathbb{P}_P\Big(\sup_{u\in[0,1]}\left|\mathbb{P}_n^B\Big(G\Big(M_n^{(b)}\Big)\le u\Big)-\psi(u)\right|>\tilde{\epsilon}\Big)\\
&\quad= \sup_{P\in\mathcal{P}}\mathbb{P}_P\Big(\sup_{u\in[0,1]}\left|\mathbb{P}_n^B\Big(G\Big(M_n^{(b)}\Big)\le u\Big)-\mathbb{P}_n\Big(G\Big(M_n^{(b)}\Big)\le u\Big)\right.\\
&\qquad\qquad\left.+\,\mathbb{P}_n\Big(G\Big(M_n^{(b)}\Big)\le u\Big)-\psi(u)\right|>\tilde{\epsilon}\Big)\\
&\quad\le \sup_{P\in\mathcal{P}}\mathbb{P}_P\Big(\sup_{u\in[0,1]}\left|\mathbb{P}_n^B\Big(G\Big(M_n^{(b)}\Big)\le u\Big)-\mathbb{P}_n\Big(G\Big(M_n^{(b)}\Big)\le u\Big)\right|>\tilde{\epsilon}/2\Big) \qquad (3.46)\\
&\qquad+\sup_{P\in\mathcal{P}}\mathbb{P}_P\Big(\sup_{u\in[0,1]}\left|\mathbb{P}_n\Big(G\Big(M_n^{(b)}\Big)\le u\Big)-\psi(u)\right|>\tilde{\epsilon}/2\Big). \qquad (3.47)
\end{aligned}$$

The quantity (3.46) is bounded using the Dvoretzky–Kiefer–Wolfowitz inequality (Dvoretzky et al., 1956; Massart, 1990), which gives

$$\sup_{P\in\mathcal{P}}\mathbb{P}_P\Big(\sup_{u\in[0,1]}\left|\mathbb{P}_n^B\Big(G\Big(M_n^{(b)}\Big)\le u\Big)-\mathbb{P}_n\Big(G\Big(M_n^{(b)}\Big)\le u\Big)\right|>\tilde{\epsilon}/2\Big)\le 2\exp\Big(-\frac{B\tilde{\epsilon}^2}{2}\Big).$$

The quantity (3.47) is handled using Assumption 4. We may choose N_B such that for all $n \geq N_B$ and $l = 1, \ldots, L$,

$$\sup_{P \in \mathcal{P}} \mathbb{P}_P \left(\sup_{u \in [0,1]} \left| \mathbb{P}_n \Big(G\Big(M_n^{(b)}\Big) \leq u \Big) - \psi(u) \right| > \tilde{\epsilon}/2 \right) \leq \exp\left(- \frac{B\tilde{\epsilon}^2}{2} \right).$$

Combining these bounds suffices to prove

$$\sup_{P \in \mathcal{P}} \mathbb{P}_P \left(\sup_{u \in [0,1]} \left| \mathbb{P}_n^B \Big(G\Big(M_n^{(b)}\Big) \leq u \Big) - \psi(u) \right| > \tilde{\epsilon} \right) \leq 3 \exp\left(- \frac{B\tilde{\epsilon}^2}{2} \right),$$

and so

$$\sup_{P \in \mathcal{P}} \mathbb{P}_P \left(\sup_{u \in [0,1]} \left| \hat{\psi}_n^B(u) - \psi(u) \right| > \epsilon \right) \leq 3(L+1) \exp\left(- \frac{B\tilde{\epsilon}^2}{2} \right).$$

This is precisely the claimed bound (3.41). □

Bibliography

Aliprantis, C. D. and Burkinshaw, O. (1990). *Principles of Real Analysis*. Academic Press, 2 edition.

Ankan, A. and Textor, J. (2022). A simple unified approach to testing high-dimensional conditional independences for categorical and ordinal data. *arXiv*, page 2206.04356v1.

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178.

Bach, P., Chernozhukov, V., Kurz, M. S., and Spindler, M. (2021). DoubleML – An object-oriented implementation of double machine learning in R.

Bera, A. K. and Ng, P. T. (1995). Tests for normality using estimated score function. *Journal of Statistical Computation and Simulation*, 52(3):273–287.

Beran, R. (1986). Simulated Power Functions. *The Annals of Statistics*, 14(1):151–173.

Beran, R. (1997). Diagnosing bootstrap success. *Annals of the Institute of Statistical Mathematics*, 49:1–24.

Berrett, T. B. and Samworth, R. J. (2019). Nonparametric independence testing via mutual information. *Biometrika*, 106(3):547–566.

Berrett, T. B. and Samworth, R. J. (2021). USP: an independence test that improves on Pearson's chi-squared and the G-test. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 477(2256):20210549.

Bickel, P. and Freedman, D. (1981). Some Asymptotic Theory for the Bootstrap. *The Annals of Statistics*, 9(6):1196–1217.

Bickel, P. J., Klaassen, C. A., Bickel, P. J., Ritov, Y., Klaassen, J., Wellner, J. A., and Ritov, Y. (1993). *Efficient and adaptive estimation for semiparametric models*, volume 4. Springer.

Bojer, C. S. and Meldgaard, J. P. (2021). Kaggle forecasting competitions: An overlooked learning opportunity. *International Journal of Forecasting*, 37(2):587–603.

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, i. effect of inequality of variance in the one-way classification. *The Annals of Mathematical Statistics*, 25(2):290–302.

Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.

Brockwell, A. E. (2007). Universal residuals: A multivariate transformation. *Statistics & Probability Letters*, 77(14):1473–1478.

Cai, Z., Li, R., and Zhang, Y. (2022). A distribution free conditional independence test with applications to causal discovery. *Journal of Machine Learning Research*, 23.

Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313 – 2351.

Canny, J. (1986). A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6):679–698.

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *The Annals of Statistics*, 41(6):2786–2819.

Chernozhukov, V., Chetverikov, D., Kato, K., and Koike, Y. (2023a). High-dimensional data bootstrap. *Annual Review of Statistics and Its Application*, 10(1):427–449.

Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K., and Robins, J. M. (2022a). Locally robust semiparametric estimation. *Econometrica*, 90(4):1501–1535.

Chernozhukov, V., Newey, W. K., Quintas-Martinez, V., and Syrgkanis, V. (2021). Automatic debiased machine learning via neural nets for generalized linear regression. *arXiv*, page 2104.14737v1.

Chernozhukov, V., Newey, W. K., Quintas-Martinez, V., and Syrgkanis, V. (2022b). Riesznet and forestriesz: Automatic debiased machine learning with neural nets and random forests. In *Proceedings of the Thirty-ninth International Conference on Machine Learning*.

Chernozhukov, V., Newey, W. K., and Singh, R. (2022c). Automatic debiased machine learning of causal and structural effects. *Econometrica*, 90(3):967–1027.

Chernozhukov, V., Newey, W. K., and Singh, R. (2022d). Debiased machine learning of global and local parameters using regularized riesz representers. *The Econometrics Journal*, 25(3):576–601.

Chernozhukov, V., Newey, W. K., and Singh, R. (2023b). A simple and general debiased machine learning theorem with finite-sample guarantees. *Biometrika*, 110(1):257–264.

Chernozhukov, V., Newey, W. K., Singh, R., and Syrgkanis, V. (2020). Adversarial estimation of riesz representers. *arXiv*, page arXiv:2101.00009v1.

Collomb, G. (1979). Conditions nécessaires et suffisantes de convergence uniforme d'un estimateur de la régression, estimation des dérivées de la régression. *Comptes rendus des séances de l'Académie des sciences, Série A*, 288(2):161–163.

Cox, D. (1985). A penalty method for nonparametric estimation of the logarithmic derivative of a density function. *Annals of the Institute of Statistical Mathematics*, 37:271–288.

Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press.

Da Rosa, J. C., Veiga, A., and Medeiros, M. C. (2008). Tree-structured smooth transition regression models. *Computational Statistics & Data Analysis*, 52(5):2469–2488.

Dai, W., Tong, T., and Genton, M. G. (2016). Optimal estimation of derivatives in nonparametric regression. *The Journal of Machine Learning Research*, 17(1):5700–5724.

Díaz, I. and van der Laan, M. (2012). Population intervention causal effects based on stochastic interventions. *Biometrics*, 68(2):541–549.

Dvoretzky, A., Kiefer, J., and Wolfowitz, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, 27(3):642–669.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1–26.

Efron, B. and Tibshirani, R. (1994). *An Introduction to the Bootstrap*. Chapman and Hall/CRC.

Farrell, M., Liang, T., and Misra, S. (2021). Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London, Series A*, 222(594–604):309–368.

Fisher, R. A. (1925). Theory of statistical estimation. *Mathematical Proceedings of the Cambridge Philosophical Society*, 22(5):700–725.

Fisher, R. A. (1958). Cancer and smoking. *Nature*, 182:596.

Fonseca, Y., Medeiros, M., Vasconcelos, G., and Veiga, A. (2018). Boost: Boosting smooth trees for partial effect estimation in nonlinear regressions. *arXiv preprint arXiv:1808.03698*.

Friedman, J., Tibshirani, R., and Hastie, T. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232.

Golub, G. H. and Van Loan, C. F. (2013). *Matrix Computations*. John Hopkins University Press, 4 edition.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT press.

Györfi, L., Kohler, M., Walk, H., et al. (2002). *A distribution-free theory of nonparametric regression*, volume 1. Springer.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer New York.

Härdle, W. and Stoker, T. M. (1989). Investigating smooth multiple regression by the method of average derivatives. *Journal of the American Statistical Association*, 84(408):986–995.

Hines, O., Diaz-Ordaz, K., and Vansteelandt, S. (2021). Parameterising the effect of a continuous exposure using average derivative effects. *arXiv preprint arXiv:2109.13124*.

Hirshberg, D. A. and Wager, S. (2020). Debiased inference of average partial effects in single-index models: Comment on wooldridge and zhu. *Journal of Business & Economic Statistics*, 38(1):19–24.

Hirshberg, D. A. and Wager, S. (2021). Augmented minimax linear estimation. *The Annals of Statistics*, 49(6):3206–3227.

Hoeffding, W. (1948). A Non-Parametric Test of Independence. *The Annals of Mathematical Statistics*, 19(4):546–557.

Horn, R. A. and Johnson, C. R. (1985). *Matrix Analysis*. Cambridge University Press.

Hothorn, T. and Zeileis, A. (2015). partykit: A modular toolkit for recursive partytioning in R. *Journal of Machine Learning Research*, 16:3905–3909.

Kasy, M. (2018). Uniformity and the delta method. *Journal of Econometric Methods*.

Kennedy, E. H. (2016). Semiparametric theory and empirical processes in causal inference. In He, H., Wu, P., and Chen, D.-G. D., editors, *Statistical Causal Inferences and Their Applications in Public Health Research*, pages 141–167. Springer.

Kennedy, E. H. (2022). Semiparametric doubly robust targeted double machine learning: a review. *arXiv preprint arXiv:2203.06469*.

Kennedy, E. H., Ma, Z., McHugh, M. D., and Small, D. S. (2017). Non-parametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 79(4):1229–1245.

Klyne, H. and Shah, R. (2023). Average partial effect estimation using double machine learning. *arXiv*, page 2308.09207.

Krinsky, I. and Robb, A. (1986). On approximating the statistical properties of elasticities. *The Review of Economics and Statistics*, 68(4):715–719.

Li, C. and Shepherd, B. E. (2010). Test of association between two ordinal variables while adjusting for covariates. *Journal of the American Statistical Association*, 105(490):612–620.

Li, C. and Shepherd, B. E. (2012). A new residual for ordinal outcomes. *Biometrika*, 99(2):473–480.

Li, W. (1996). Asymptotic equivalence of estimators of average derivatives. *Economics Letters*, 52(3):241–245.

Liu, D., Li, S., Yu, Y., and Moustaki, I. (2021). Assessing partial association between ordinal variables: Quantification, visualization, and hypothesis testing. *Journal of the American Statistical Association*, 116(534):955–968.

Lundborg, A. R., Kim, I., Shah, R. D., and Samworth, R. J. (2022). The projected covariance measure for assumption-lean variable significance testing. *arXiv*, page 2211.02039v1.

Marx, A. and Vreeken, J. (2019). Testing conditional independence on discrete data using stochastic complexity. In Chaudhuri, K. and Sugiyama, M., editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 496–505. PMLR.

Massart, P. (1990). The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality. *The Annals of Probability*, 18(3):1269–1283.

Molnar, C. (2022). *Interpretable Machine Learning*. Lulu, 2 edition.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142.

Newey, W. K. and Stoker, T. M. (1993). Efficiency of weighted average derivative estimators and index models. *Econometrica*, 61(5):1199–1223.

Neyman, J. (1923). On the Application of Probability Theory to Agricultural Experimentss: Essay on principles. Excerpts reprinted (1990) in English (D. Dabrowska and T. Speed, trans.). *Statistical Science*, 5(4):465–472.

Ng, P. T. (1994). Smoothing spline score estimation. *SIAM Journal on Scientific Computing*, 15(5):1003–1025.

Ng, P. T. (2003). Computing Cox's smoothing spline score estimator. *Northern Arizona University Working Paper*.

Pearl, J. (2009). *Causality*. Cambridge University Press, 2 edition.

Petersen, L. and Hansen, N. R. (2021). Testing conditional independence via quantile regression based partial copulas. *Journal of Machine Learning Research*, 22(70):1–47.

Powell, J. L., Stock, J. H., and Stoker, T. M. (1989). Semiparametric estimation of index coefficients. *Econometrica*, 57(6):1403–1430.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Richardson, T. and Robins, J. (2013). Single World Intervention Graphs (SWIGs): A Unification of the Counterfactual and Graphical Approaches to Causality. Technical Report 128, Center for Statistics and the Social Sciences, University of Washington.

Ritov, Y., Bickel, P., Gamst, A., and Kleijn, B. (2014). The Bayesian Analysis of Complex, High-Dimensional Models: Can It Be CODA? *Statistical Science*, 29(4):619–639.

Robins, J. M. and Ritov, Y. (1997). Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Statistics in Medicine*, 16(3):285–319.

Robins, J. M. and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129.

Robins, J. M. and Rotnitzky, A. (2001). Comment on Peter J. Bickel and Jaimyoung Kwon article. *Statistica Sinica*, 11(4):920–936.

Robins, J. M., Rotnitzky, A., and van der Laan, M. (2000). Comment on S. A. Murphy and A. W. van der Vaart article. *Journal of the American Statistical Association*, 95(450):477–482.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866.

Romano, J. (1988). A bootstrap revival of some nonparametric distance tests. *Journal of the American Statistical Association*, 83(403):698–708.

Rothenhäusler, D. and Yu, B. (2020). Incremental causal effects. *arXiv*, page 1907.13258v4.

Rotnitzky, A., Smucler, E., and Robins, J. M. (2021). Characterization of parameters with a mixed bias property. *Biometrika*, 108(1):231–238.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688–701.

Samworth, R. (2003). A note on methods of restoring consistency to the bootstrap. *Biometrika*, 90(4):985–990.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120.

Scheidegger, C., Hörrmann, J., and Bühlmann, P. (2022). The weighted generalised covariance measure. *Journal of Machine Learning Research*, 23(273):1–68.

Schick, A. (1986). On Asymptotically Efficient Estimation in Semiparametric Models. *The Annals of Statistics*, 14(3):1139–1151.

Schuster, E. and Yakowitz, S. (1979). Contributions to the Theory of Nonparametric Regression, with Application to System Identification. *The Annals of Statistics*, 7(1):139–149.

Shah, R. D. and Bühlmann, P. (2023). Double-estimation-friendly inference for high-dimensional misspecified models. *Statistical Science*, 38(1):68–91.

Shah, R. D. and Peters, J. (2020). The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514–1538.

Shao, J. (1994). Bootstrap sample size in nonregular cases. *Proceedings of the American Mathematical Society*, 122(4):1251–1262.

Shao, J. (1996). Bootstrap model selection. *Journal of the American Statistical Association*, 91(434):655–665.

Singh, K. (1981). On the Asymptotic Accuracy of Efron's Bootstrap. *The Annals of Statistics*, 9(6):1187–1195.

Spirtes, P. and Glymour, C. (1991). An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, 9(1):62–72.

Spirtes, P., Glymour, C., and Scheines, R. (1993). *Causation, prediction, and search*. Springer.

Sriperumbudur, B., Fukumizu, K., Gretton, A., Hyvärinen, A., and Kumar, R. (2017). Density estimation in infinite dimensional exponential families. *Journal of Machine Learning Research*, 18(57):1–59.

Stoker, T. M. (1986). Consistent estimation of scaled coefficients. *Econometrica*, 54(6):1461–1481.

Stoker, T. M. (1990). Equivalence of direct, indirect and slope estimators of average derivatives. In Barnett, W., Powell, J., and Tauchen, G., editors, *Prodeedings of the Fifth International Symposium in Economic Theory and Econometrics*, pages 99–118. Cambridge University Press.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B*, 58(1):267–288.

Tsiatis, A. A. (2006). *Semiparametric theory and missing data*. Springer.

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

van der Vaart, A. W. (2002). Semiparametric statistics. In Bernard, P., editor, *Lectures on Probability Theory and Statistics*, pages 331–457. Springer.

Vansteelandt, S. and Dukes, O. (2022). Assumption-lean inference for generalised linear model parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(3):657–685.

Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Wasserman, L. (2006). *All of Nonparametric Statistics*. Springer-Verlag.

Watson, G. S. (1964). Smooth regression analysis. *Sankhya Series A*, 26(4):359–372.

Wooldridge, J. M. and Zhu, Y. (2020). Inference in approximately sparse correlated random effects probit models with panel data. *Journal of Business & Economic Statistics*, 38(1):1–18.

Wright, S. (1921). Correlation and causation. *Journal of agricultural research*, 20(7):557–585.

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242.

Zhao, Q. and Hastie, T. (2021). Causal interpretations of black-box models. *Journal of Business & Economic Statistics*, 39(1):272–281.

Zhou, Y., Liu, J., and Zhu, L. (2020). Test for conditional independence with application to conditional screening. *Journal of Multivariate Analysis*, 175:104557.