An Inferential Perspective on Data Depth

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Subhabrata Majumdar

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Snigdhansu Chatterjee, Adviser

May 2017

Truth is ever to be found in simplicity, and not in the multiplicity and confusion of things.

- Issac Newton

Simplicity is the final achievement. After one has played a vast quantity of notes and more notes, it is simplicity that emerges as the crowning reward of art.

- Frédéric Chopin

Acknowledgements

I would like to thank my advisor, Prof. Snigdhansu Chatterjee for everything he has done for me. This dissertation would not have been possible without the very fundamental ideas behind it that came out of our discussions during the first two years of my PhD. Anshu da gave me the freedom to pursue those ideas and shape them as I wanted. I am thankful for the professional support he provided, as well as for always being available to listen to my ramblings. I am going to miss those weekly meetings in which we talked about anything and everything statistics.

I want to thank Prof. Saonli Basu of the Division of Biostatistics, with whom I collaborated during my last year. She has been an invaluable mentor, and discussions with her helped me learn about many applied aspects of the problems we had worked on. Thanks to her for being a reviewer of this dissertation as well. Special thanks to Profs. Lan Wang and Xiaotong Shen for reviewing the thesis and being on the final exam committee, and Prof. Gongjun Xu for being in my oral committee.

I would like to thank other professors in the department, who have always been helpful in answering questions, and the staff in the department office for their support regarding official matters. Thanks to other students in the department, for being there with feedback and discussions: especially Abhirup, Adam, Aaron, Daniel, Dootika and Sakshi.

I would like to acknowledge the National Science Foundation Climate Expeditions grant IIS-1029711, which provided me funding for 3 semesters, and the University of Minnesota Graduate School's Interdisciplinary Doctoral Fellowship (IDF), which supported me during my final year.

I consider myself very fortunate to have worked with a number of collaborators

while still being a graduate student. Thanks to Profs. Matt McGue and Mike Miller of the Department of Psychology for their help during the IDF collaboration. Outside the campus, I am grateful to Prof. Subhash C. Basak of University of Minnesota Duluth for helping me write my first paper back in 2012 and all subsequent collaborations. Rayid Ghani of University of Chicago and Kush Varshney of IBM Research have been inspirations in developing my approach to collaborative research.

Coming to personal life, I would like to thank my friends for keeping me grounded and connected to the world outside. Special thanks to the three musketeers- Abhirup Mallik, Somnath Kundu and Suvankar Biswas, for being part of many happy memories during the past five years. I am grateful to Abhishek Nandy, for all his support during my first two years. Thanks to Amit da, Arja, Deepashree, Shriya, Tallin di, Taraswi di, Tushar and many more people associated with the Bengali Student Society of Minnesota, life in Minneapolis has been so much enjoyable.

I would like to thank my teachers from school, Prabhat Kusum Sarkar for introducing me to statistics, and Tapas Kumar Dhar for helping me build the necessary mathematical foundation. The past five years would not have been possible without the all-encompassing influence of Indian Statistical Institute (ISI), where I finished my bachelors and masters degree. This is too short a space to list all the avenues the ISI connection has helped me through in terms of personal and professional networking. I am thankful for being part of this ISI family.

Finally, I want to thank the three people closest to me for always being there no matter what. My girlfriend, Rajeshwaree Chatterjee, has been a constant source of motivation and fresh perspectives on life, even from half a world away in Kolkata. To the other two persons, my parents, I owe more than anyone else in this world. In spite of coming from humble beginnings, they have the highest respect for the intellectual pursuit, and there is no substitute for the sacrifices they made in bringing me up. Ma and Baba, I dedicate this thesis to you.

Dedication

This dissertation is dedicated to my parents:

Samita Mazumder, my mother- for being patient while teaching a little boy the alphabets of mathematics, and instilling in me the value of hard work; and

Satyabrota Mazumder, my father- for the many life lessons, and wholeheartedly supporting me in every major decision I have made.

Abstract

Data depth provides a plausible extension of robust univariate quantities like ranks, order statistics and quantiles in multivariate setup. Although depth has gained visibility and has seen many applications in recent years, especially in classification problems for multivariate and functional data, its generalizability and utility in achieving traditional parametric inferential goals is largely unexplored. In this thesis we develop several approaches to address this. In particular, firstly we define an evaluation map function that is more general than data depth, and establish several results in a parametric modelling context using a broad definition of a statistical model. A fast algorithm for covariate selection using data depths as evaluation functions arises as a special case of this. We demonstrate applications of this framework on data from diverse fields: namely climate science, medical imaging and behavioral genetics. Secondly we propose a multivariate rank transformation using data depth and use them for robust inference in location and scale problems in elliptical distributions. Thirdly, we lay out a depth-based regularization framework in multi-response regression, and derive a new method of nonconvex penalized sparse regression in the multitask situation. Across the thesis, several simulation studies and real data examples demonstrate the effectiveness of the methods developed here.

Contents

List of Tables

List of Figures

List of notations

Notation	Meaning
$\mathbb{E}$	Expectation of a random variable (scalar/ vector/ matrix valued)
$\mathbb{V}$	Variance of a scalar-valued random variable or covariance matrix of a vector-valued random variable
$\mathbb{P}$	Probability of some event
$\rightsquigarrow$	Convergence in distribution
$a_n \asymp b_n$	$a_n = O(b_n)$ and $b_n = O(a_n)$
$\|.\|$	Euclidean norm for vectors, Frobenius norm for matrices, unless otherwise stated
$a_n = O_{P_n}(b_n)$	a_n/b_n converges to some $c \in \mathbb{R}$ in probability conditional on the data as $n \to \infty$
$a_n = o_{P_n}(b_n)$	a_n/b_n converges to some 0 in probability conditional on the data as $n \to \infty$
$a_n \xrightarrow{P_n} b_n$	a_n converges to b_n in probability conditional on the data as $n \to \infty$

Chapter 1

Introduction

1.1 Background

The nonparametric concept of data depth had first been proposed by Tukey (1975) when he introduced the halfspace depth. The motivation behind this was to formulate a unified framework for nonparametric inference in multivariate concept: in particular, the multivariate equivalent of methods based on signs and ranks, order statistics, quantiles and outlyingness functions.

Given a dataset, the depth of a given point in the sample space measures how far inside the data cloud the point exists, i.e. it is a measure of centrality of the point with respect to the data. An overview of statistical depth functions can be found in Zuo and Serfling (2000). Depth-based methods have gained popularity in the past two decades, for robust nonparametric classification (Jornsten, 2004; Ghosh and Chaudhuri, 2005; Dutta and Ghosh, 2012; Sguera et al., 2014). In parametric estimation, depth-weighted means (Zuo et al., 2004) and covariance matrices (Zuo and Cui, 2005) provide high-breakdown point as well as efficient estimators, although they do involve choice of a suitable weight function and tuning parameters. As Liu and Singh (1997) have shown, it is also possible to use statistical depth functions in hypothesis testing and an alternate notion of p-values. Approaching data depth

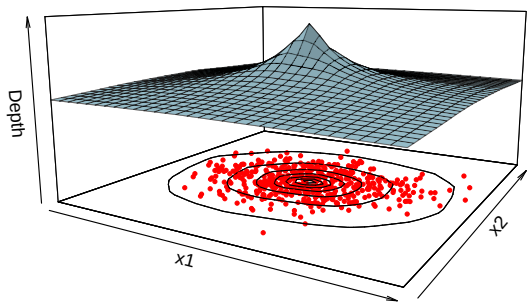


Figure 1.1: Depth is a scalar measure of how much inside a point is with respect to a data cloud: 500 points from $\mathcal{N}_2((0,0)^T, \text{diag}(2,1))$

as from the perspective of breakdown points, Rousseeuw and Hubert (1999) also introduced the concept of regression depth, which was later generalized by Mizera (2002).

1.2 Definition and examples

For any multivariate distribution F taking values $\tilde{\mathbb{R}}^p$ (or a subset of it), the depth of a point $\mathbf{x} \in \mathbb{R}^p$, say $D(\mathbf{x}, F_{\mathbf{X}})$ is any real-valued function that provides a 'center outward ordering' of $\mathbf{x}$ with respect to F (Zuo and Serfling, 2000). Figure 1.1 gives an intuition of data depth for samples from a bivariate normal distribution. As demonstrated by the contours and plot of values, a point close to the center, which coincides with the mean for elliptical distributions, has high depth. In other words, the point is situated deep inside the data/ underlying distribution. In comparison, a point closer to the periphery shall have less depth.

In order to standardizing this notion Liu (1990) outlined the desirable properties of a statistical depth function:

(P1) *Affine invariance*: $D(A\mathbf{x} + \mathbf{b}, \mathbf{A}F + \mathbf{b}) = D(\mathbf{x}, F)$ for any $\mathbf{A} \in \mathbb{R}^{p\times p}, \mathbf{b} \in \mathbb{R}^p$. Here $\mathbf{A}F + \mathbf{b}$ is a slight abuse of notation, and denotes the distribution of $\mathbf{AX} + \mathbf{b}$

where $\mathbf{X} \sim F$;

(P2) *Maximality at center*: $D(\boldsymbol{\theta}, F) = \sup_{\mathbf{x}\in\mathbb{R}^p} D(\mathbf{x}, F)$ for F having center of symmetry $\boldsymbol{\theta}$. This point is called the *deepest point* of the distribution.;

(P3) *Decreasing from deepest point along any ray*: $D(\mathbf{x}; F) \leqslant D(\boldsymbol{\theta} + a(\mathbf{x} - \boldsymbol{\theta}), F)$;

(P4) *Vanishing at infinity*: $D(\mathbf{x}; F) \to 0$ as $\|\mathbf{x}\| \to \infty$.

In (P2) the types of symmetry considered can be central symmetry, angular symmetry and halfspace symmetry. Also for multimodal probability distributions, i.e. distributions with multiple local maxima in their probability density functions, properties (P2) and (P3) are actually restrictive towards the formulation of a reasonable depth function that captures the shape of the data cloud. Finally we think affine invariance is an artifact of depth functions being formulated keeping robustness with respect to elliptical distributions in mind, and in most practical cases a location and scale invariance suffices. Furthermore, because of their formulations, technical properties like quasi-concavity, Lipschitz continuity, uniform convergence rise naturally in different definitions of data depth (Liu, 1990; Zuo and Serfling, 2000; Mosler, 2013).

It should be noted here that likelihood is not same as depth. Although in the univariate case many of these are essentially functions of the cumulative distribution function, and indeed for elliptical multivariate distributions depth contours coincide with density contours, unlike depths, likelihood is a local property. It is sensitive to multimodality, does not measure 'inlyingness' to a distribution in general and the maximum likelihood point may not be a central point according to any definition of symmetry (Serfling, 2006).

Some popular measures of data depth available in the literature and extensively used in nonparametric and semiparametric inference are as follows:

- **Halfspace depth** (HD: (Tukey, 1975)) is defined as the minimum probability

of all halfspaces containing a point. In our notations,

$$HD(\mathbf{x}, F) = \inf_{\mathbf{u}\in\mathbb{R}^p;\mathbf{u}\neq\mathbf{0}} P(\mathbf{u}^T\mathbf{X} \geqslant \mathbf{u}^T\mathbf{x})$$

- **Mahalanobis depth** (MhD: Liu et al. (1999)) is based on the Mahalanobis distance of $\mathbf{x}$ to $\boldsymbol{\mu}$ with respect to Σ: $d_{\boldsymbol{\Sigma}}(\mathbf{x}, \boldsymbol{\mu}) = \sqrt{(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$. It is defined as

$$MhD(\mathbf{X}, F) = \frac{1}{1 + d^2_{\boldsymbol{\Sigma}}(\mathbf{x}, \boldsymbol{\mu})}$$

- **Projection depth** (PD: Zuo (2003)) is another depth function based on an outlyingness function. Here that function is

$$O(\mathbf{x}, F) = \sup_{\|\mathbf{u}\|=1} \frac{|\mathbf{u}^T\mathbf{x} - m(\mathbf{u}^T\mathbf{X})|}{s(\mathbf{u}^T\mathbf{X})}$$

  where m and s are some univariate measures location and scale, respectively. Given this the depth at $\mathbf{x}$ is defined as $PD(\mathbf{x}, F) = 1/(1 + O(\mathbf{x}, F))$.

1.3 Why is depth not a thing yet?

Although some articles on data depth are fairly well-cited (e.g. Liu et al. (1999); Vardi and Zhang (2000)), in general it remains an esoteric, at best intriguing, concept in statistical literature. This is partly due to its nonparametric nature and high computational cost. There have been several approaches for calculating HD. A recent paper (Dyckerhoff and Mozharovskyi, 2016) provides a general algorithm that computes exact HD in $O(n^{p-1}\log n)$ time. PD is generally approximated by taking maximum over a number of random projections: and has high variability for small samples. MhD is easy to calculate since the sample mean and covariance matrix are generally used as estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, respectively. However this makes it less

robust with respect to outliers: defeating the purpose of using data depth in many situations.

A more significant reason though, we believe, is that the concept has not generalized enough since its first inception. There have been attempts at defining depth contours for distributions with nonstandard shapes (multimodal, star-shaped etc.) (Paindaveine and bever, 2013; Chernozhukov et al., 2017) as well as using functional depths (Narisetty and Nair, 2016; Sguera et al., 2016). These certainly broaden the domain of application for data depth. However, the scope of using depth, or depth-like quantities, is much larger in statistical inference. It quantifies the proximity of a point in a multivariate space to a probability distribution on the same space. In this spirit, given some hilbert space $\mathcal{H}$, any such proximity measure $D : \mathcal{H} \times \tilde{\mathcal{H}} \mapsto [0, \infty)$, $\tilde{\mathcal{H}}$ being the space of probability measures on $\mathcal{H}$, can be termed a depth function. Such quantities (or more accurately, decreasing transformations of them, e.g. the outlyingness function of Zuo and Serfling (2000)) provide a bridge between point norms and distributional distance measures like the Kullback-Leibler divergence or the Wasserstein metric in appropriate normed spaces. To the best of our knowledge, this generalized notion of point-to-distribution distance/ proximity is absent in the current literature. This thesis is an attempt in leveraging the extra flexibility provided by the above interpretation of data depth functions in diverse inferential scenarios.

1.4 Summary of work

In Chapter 2, we consider a general modelling framework in which several parameters need to be estimated from the data. Our objective here is to characterize subsets of the model space based on a pre-defined criterion, and to estimate this characterization in the presence of data. A concrete example for this can be variable selection in linear regression, where the user needs to find out the most parsimonious subset of

predictors that do not compromise the quality of model fit. For this purpose we introduce a function called the *statistical evaluation map*, which, while essentially serving the same purpose as depth, are based on much weaker assumptions and take into account the potentially expanding space of parameters. In a transformed space, this evaluates a function of estimated parameters corresponding to a specific subspace with respect to the sampling distribution of parameters from the full parameter space (i.e. the full model). The value of this evaluation function will change based on the specific sample the model estimates are based on, so this has a distribution as well, which depends on the sample size. We name the average evaluation function as the e-value of a model: this acts as a quantification of model evidence. Under a very general definition of 'good' and 'bad' models we demonstrate how these e-values can be used to differentiate between these two types. We use resampling to estimate the random distributions we work with: this is essential towards calculating sample version of model e-values. As a special case, when data depths are considered as evaluation maps, some further refinement can be achieved in the bifurcation of set of candidate models for the traditional statistical model selection problem. This results in an extremely fast, almost trivial, algorithm to separate out essential predictors in a regression-like setup.

Although depth functions did serve as our motivation for the above work and our initial results were assuming depths in elliptical sampling distributions of parameters, we later found out that a majority of the results hold in a much more generalized setup, and it is enough to explicitly invoke the usage of depths for variable selection only. As demonstrated in Chapter 3, the method of e-values leads to valuable insights in several real data situations. In Section 3.3 therein, we expand the scope of e-values by considering tail quantiles of evaluation map distributions as e-values instead of their means: this leads to improved detection of weak single nucleotide polymorphism signals in behavioral trait analysis of genetic data from families.

Chapter 4 onwards we take a more mainstream approach. Here we introduce a composition of the spatial sign function (Locantore et al., 1999) with transformations on functions that are essentially the outlyingness maps of Zuo and Serfling (2000), with a few restrictions for technical convenience. After a brief consideration of its performance in the location problem for elliptical distributions, we define a multivariate rank vector using this. We discuss several aspects of its performance in estimating components of the covariance matrix in the data-generating elliptical distribution: its eigenvectors, eigenvalues and the covariance matrix itself. Several simulation studies and data examples outline the utility of these methods, and we also discuss their implementation in Sufficient Dimension Reduction (Adragni and Cook, 2009) and functional outlier detection.

Chapter 5 discusses another application of the idea of data depth-based inverse ranking, this time in regularized regression. We propose a new class of nonconvex penalty functions in the paradigm of multitask sparse penalized regression using penalties based on data depth. Focusing on a one-step sparse estimator of the coefficient matrix using local linear approximation of the penalty function, we derive its theoretical properties and provide the algorithm for its computation. For orthogonal design and independent responses, the resulting thresholding rule enjoys near-minimax optimal risk performance, similar to the adaptive lasso (Zou, 2006). A simulation study as well as real data analysis demonstrate its effectiveness compared to some of the present methods that provide sparse solutions in multivariate regression.

Chapter 2

Generalized Model Discovery using Statistical Evaluation Maps

2.1 Introduction

In a typical statistical or data science exercise, both *data* and a *statistical model* are involved. While there is often little or no ambiguity about data, there can be many alternatives about how to analyze such data, and how to interpret the results. This broadly constitute the realm of statistical models. In this chapter, we interpret the term *statistical model* very broadly. We recognize that various possible transformations of the data, different model fitting algorithms, practical safeguards put in place to ensure robustness and sensitivity balance in the results, different methods of data analysis, different statistical paradigms of interpretation of results, as all equally deserving to be considered as crucial components of a statistical model. The example below illustrates this idea.

Example 2.1.1 (Tree data)**.** Consider the data contained in `data(trees)` in the statistical software `R`. There are 31 observations on girth, height and volume. Observed data for these variables are (X_{1i}, X_{2i}, Y_i) respectively, for for $i = 1, \ldots, n$. We denote $p = 2$ for the two explanatory variables X_1 and X_2, used to explain the properties of the response variable Y.

Define the Box-Cox transformation (Box and Cox, 1964) on the response variable as $C(y, \lambda) = \log(y)\mathbb{I}_{\lambda=0} + y^{\lambda}\mathbb{I}_{\lambda\neq 0}$. We assume that Y_i's in the data are related to the other variables according to the statistical relation

$$C(Y_i, \lambda) = \beta_0 + \beta \log(X_{1i}) + \beta \log(X_{2i}) + e_i \tag{2.1.1}$$

Here $\{e_i\}$ is a sequence of random variables, and we assume that $\mathbb{E}e_i = 0$ and $\mathbb{E}e_i^2 = \sigma_i^2 < \infty$. The parameters in this system are $\boldsymbol{\theta}_n = (\lambda, \beta_0, \beta_1, \beta_2, \sigma_1^2, \ldots, \sigma_n^2) \in \mathbb{R}^{p_n}$ where $p_n = n + 4$.

Even in this rather simple framework, we can imagine several *statistical models* as being *per se* equally interesting or important. These can include (i) the Gauss-Markov linear regression model with $\lambda = 0$, (ii) linear regression with any other fixed, non-random λ, (iii) a model where λ is estimated form data but then a linear regression model used for the rest of the analysis ignoring the randomness in the estimated λ, (iv) using a fixed λ value like 0 or 1, then using *ordinary least squares* (OLS) method to estimate regression parameters, followed by inference based on the residual bootstrap (see Efron (1979); Efron and Tibshirani (1993); Shao and Tu (1995)), (v) using robustness-driven M-estimation techniques for simultaneous estimation of $(\lambda, \beta_0, \beta_1, \beta_2)$, followed by a *wild bootstrap* resampling scheme for statistical inference (Wu, 1986; Mammen, 1993), which provides robustness against heteroscedasticity.

We submit that these are all plausible models, important from one or more considerations. Some like (iii) reflect tradition, others like (v) reflect desirable caution coupled with modern computational power. The above list of possible models is far from exhaustive (e.g. in (iv) each alternative resampling scheme may be called a separate model), but serves to illustrate the fact that statistical models arise in most of the standard procedures of data analysis, be it from classical Statistics, robustness considerations, Bayesian paradigm, risk management perspective, Occam's razor, or

combinations thereof. Such models typically differ from each other in many ways, and not just in the number of covariates, or number of parameters to estimate. Often, as in the case of the heteroscedastic model coupled with resampling-based inference above, a very classical approach towards modeling or model selection, or a selection based only on a superficial reading of parsimony, can lead to leaving out greatly versatile models on both robustness and efficiency counts. In this chapter, we address this problem of elicitation of suitable models for analyzing data in a very general framework. We consider candidate models that need not be nested, or philosophically or otherwise compatible with each other.

Our primary goal is a clear separation of the candidate models into two groups: those that adequately explain some user-defined characteristics exhibited in the data, which we designate *adequate models*, and those that do not (inadequate models). The first subsection in Section 2.2 contains notations and a technical definition of model adequacy, as well as a generic description of a baseline model, which we call the *preferred model*. This may be the most complex candidate model (e.g. the model with all covariates in regression), a model in popular or current use, a hypothesized model, or a model with known parsimony or computational advantage. As we shall see, this formulation of statistical models is broader than the traditional definition of correct or wrong models in model selection literature (e.g. Shao (1993, 1996)). Each candidate model has its own set of unknown parameters, which are estimated using a model-specific optimization framework. The next subsection outlines this estimation method. Following this, all model parameter estimates are mapped to a common Euclidean reference frame $\mathbb{R}^{d_n}, d_n \in \mathbb{N}$ through user-defined transformation functions for ease of comparison between models.

We focus on this transformed model space $\mathbb{R}^{d_n}$, and propose using a function called the *evaluation map* in Section 2.3, which compares each candidate model against the preferred model. An evaluation map typically compares a point in the parameter

space of any candidate model with the distribution of estimated parameters in the preferred model, and data depth functions are special cases of functions that may act as an evaluation map. After this we introduce a quantity called the e-value, which we define as a non-negative summary statistic for the evaluation map distribution corresponding to a candidate model. The model e-value is a measure of how well a candidate model explains the interesting features of the data, which is based on a user-specified function. Under very general theoretical conditions we show that population e-values for theoretical models asymptotically tends to zero, while for adequate models they tend to the e-value of the preferred model. Thus we allow the possibility that none of the candidate models, including the preferred model, adequately explain the properties of the data at hand. In such cases, only the preferred model will have a high score. Our proposal thus includes the provision for triggering a re-evaluation of models and data based on scientific caution, when only the preferred model achieves a significantly non-zero score.

We adopt a fairly general resampling-based procedure to approximate the distribution of evaluation maps for a candidate model, and in Section 2.4 establish consistency of the resampling procedure adopted in this chapter, when one or more models are considered simultaneously. Following this we show that under certain conditions on the resampling schemes, population e-values for both adequate and inadequate models can be consistently recovered. Thus, we formulate a unified system where resampling elicits both the e-value of a model, along with the joint sampling distribution of all its parameter estimators. This allows for automatic inference and prediction with any model.

Additionally, in Section 2.3 and Section 2.4 we allow several quantities, like number of parameters in each candidate model or the number of characteristics of interest from the data on which the evaluation map is computed, to tend to infinity with sample size. This *dimension asymptotics* approach allows any candidate model to have

increasing parameter dimensionality with sample size, which imitates the reality of the scientific discovery process where additional data is often used in conjunction with more fine-tuned or insightful models. Similarly, allowing the number of characteristics used for comparing models to grow with the sample size reflects the scientific process. Throughout these sections, for theoretical purposes we adopt a framework involving a triangular array of models and parameters, where various parameter values and dimensions and even estimation and model evaluation procedures are allowed to change with sample size. This is partially for the same reason of being in tune with the reality of scientific discovery process, but also for additional theoretical advantages that such a framework offers, and for the purpose of being inclusive of techniques like local asymptotics, uniform convergence and several others that will form part of our future work.

Our proposal thus far involves four choices: that of (a) a preferred model, (b) a map from the parameter space to $\mathbb{R}^{d_n}$ for each candidate model, (c) an evaluation map, which is a function defined on $\mathbb{R}^{d_n}$ and probability distributions on it to compare each model to the preferred model, (d) a resampling strategy. In Section 2.5 we demonstrate how all of these come together in tackling the traditional model selection problem of identifying necessary covariates in a regression-like setup. In such problems, there is a maximum number of parameters p_n to consider, and various candidate models consider subsets of a common set of p_n parameters. The candidate models can be arranged in a lattice, with the supremum being the *least parsimonious* or complete model that involves all p_n parameters. There are 2^{p_n} such models, and a full evaluation of all such models is an NP-Hard problem (Natarajan, 1995). For this reason various algorithms to reduce computations by evaluating far fewer models (Schwarz, 1978; Konishi and Kitagawa, 1996), as well as sparsity-based approaches (Tibshirani, 1996; Fan and Li, 2001; Zou, 2006) have been proposed, which compromise optimality and other properties of the model selection procedure.

In this context, we use data depths as evaluation functions, allowing us to establish a preference ordering among the adequate models. Subsequently we are able to propose a very fast algorithm which has the following simple and generic steps:

1. Start from the model with all covariates, i.e. the full model and compute its e-value using resampling;

2. Take the marginal models by dropping each covariate, compute their corresponding e-values;

3. Collect covariates that cause a decrease in e-value compared to the full model.

As evident from the above steps, this recipe only requires computation of the full model. Coupled with the fact that a fast and parallelizable generalized bootstrap procedure (Chatterjee and Bose, 2005) based on Monte-Carlo simulation can be used as the resampling method of choice, we end up with an extremely fast covariate selection method. This procedure is able to tackle with ease tricky modelling situations like linear mixed models and robust regression, and also provides asymptotic model selection consistency owing to the machinery developed in Section 2.3 and Section 2.4.

In Section 2.6 we present two illustrative examples on how our fast algorithm is implemented, and its relative performance in covariate selection problems. One of the examples in this section involves random effects, to illustrate the breadth of applicability of the proposed methodology. Finally, in Section 2.7 we discuss the scope and implications of this framework, future research plans, caveats and end with some concluding comments. Regarding real data applications of the e-values procedure, we have performed substantial amounts of them in diverse modelling situations: this we are going to defer to Chapter 3.

Before proceeding to the next section we state some necessary notations. For any

function $\mathbf{h}$ of the parameters in any model, we will often simplify notations by using

$$\begin{aligned}\mathbf{h} &\equiv \mathbf{h}_{sn} \equiv \mathbf{h}\left(\boldsymbol{\theta}_{sn}\right), \\ \widehat{\mathbf{h}} &\equiv \widehat{\mathbf{h}}_{sn} \equiv \mathbf{h}\left(\widehat{\boldsymbol{\theta}}_{sn}\right), \\ \widehat{\mathbf{h}}_r &\equiv \widehat{\mathbf{h}}_{rsn} \equiv \mathbf{h}\left(\widehat{\boldsymbol{\theta}}_{rsn}\right).\end{aligned}$$

The notation $a_n \asymp b_n$ implies that $a_n = O(b_n)$ as well as $b_n = O(a_n)$. The notation $\mathbf{R}$, typically with various subscripts like $\mathbf{R}_n, \mathbf{R}_{sn}, \mathbf{R}_{rsn}$ and so, are used as generic for remainder terms, which contribute asymptotically negligible terms in our results. While we include all necessary algebraic details, often the tedious algebra behind moment calculations and probabilistic bound computations is omitted to contain this chapter to a reasonable length and preserve clarity. However, our technical conditions are always comprehensive and explicit, and such algebraic computations can be easily carried out without much intellectual effort. In designing the technical conditions for the theoretical properties in this chapter, we have striven for simplicity and not on minimal requirements. Thus, the various assumptions made in this chapter are often sufficient conditions, rather than necessary ones, for the theoretical results.

2.2 The general framework

2.2.1 The frame of models

In any statistical model, each parameter has an assigned role. A parameter may be a constant related to the scientific process, tuning constant related to a computational procedure or a prediction algorithm, or may perform some other function. Examples of the former in Example 2.1.1 are the regression slope parameters β_1 and β_2, which quantify how the volume of wood in a tree changes with its height or girth. An exam-

ple of the latter in the same context can be the parameter λ, or a tolerance or iteration limits of an iterative model fitting procedure. Parameters can have similar roles in many models, for example, the regression coefficients β_1 and β_2 in Example 2.1.1 are used in all the listed models in that example. We use these general facts to describe *frame of models* that we use in this paper.

In this chapter, we consider a context where the union of all parameters from all candidate models forms a countable set. Naturally, problems where the number of parameters are finite, as in a majority of statistical applications, are included in our framework. We exclude all constants that are invariant across candidate models from this count, or any unknown quantity that is not estimated in any model and is not used subsequently. The parameters across all models are laid out in any arbitrary but fixed fashion indexed by the set of integers $\{1, 2, \ldots\}$. For example, in 2.1.1 we may consider $p_n = n + 4$ as the maximum number of parameters in the system, and denote the p_n-dimensional vector of parameters with the generic notation

$$\boldsymbol{\theta}_n = \left(\lambda, \beta_0, \beta_1, \beta_2, \sigma_1^2, \ldots, \sigma_n^2\right) = \left(\theta_{n,1}, \theta_{n,2}, \ldots, \theta_{n,p_n}\right) \text{ notationally.}$$

We now associate a candidate model $\mathcal{M}_n$, either from a scientific discovery process or a hypothesis testing process, with two quantities:

(a) The set $\mathcal{S}_n = \{j_1, \ldots, j_{p_{sn}}\} \subseteq \{1, 2, \ldots\}$ of indices where the parameter values are unknown and estimated from the data; and
(b) An ordered vector of known constants $\mathbf{c}_n = (c_{nj} : j \notin \mathcal{S}_n)$ for parameters not indexed by $\mathcal{S}_n$.

For any n the sets $\mathcal{S}_n$ are finite, thus each model may include only a finite number of unknown real-valued constants.

The generic parameter vector corresponding to this model, say $\boldsymbol{\theta}(\mathcal{M}_n) \in \boldsymbol{\Theta}_{mn} \subseteq$

$\boldsymbol{\Theta}_n = \times_j \boldsymbol{\Theta}_{n,j}$, will thus have the structure

$$\theta_{mnj} = \begin{cases} \text{Unknown } \theta_{mnj} & \text{for } j \in \mathcal{S}_n; \\ \text{Known } c_{nj} & \text{for } j \notin \mathcal{S}_n. \end{cases}$$

Each $\theta_{mnj} \subseteq \mathbb{R}$, thus all parameters are real-valued. It may be noted that in most cases, simple re-parametrization can be used to define models in a way such that the known constants in $\mathbf{c}_n$ are all zero.

We assume that at stage n there is have a *preferred model*, which we denote by $\mathcal{M}_{*n}$: and is identified by the set of indices $\mathcal{S}_{*n} \subseteq \{1, 2, \ldots\}$ having p_{*n} elements, and known constants $\mathbf{c}_{*n}$. We also designate a a fixed element of $\mathcal{M}_{*n}$ as the *preferred parameter vector*, say $\boldsymbol{\theta}_{0n}$. Depending on the context, the preferred model may relate to a hypothesized model, or the most complex or the most simple model, or relate to the current state of the art, a 'gold standard', or be 'preferred' by some other predefined criteria; whereas the preferred parameter vector is generally indicative of the data generating process. Note that the preferred model is just one of the candidate models, and its usage will shortly be clear.

2.2.2 Transformation to a common platform

Suppose $\mathbf{G}_{mn} : \boldsymbol{\Theta}_n \to \mathbb{R}^{d_n}$ is a known transformation to map parameters from model $\mathcal{M}_n$ to $\mathbb{R}^{d_n}$. While the candidate models may be very diverse and may relate to different physical realities, theories or hypotheses, computational or data analytic choices, the Euclidean space $\mathbb{R}^{d_n}$ is a common ground where all models may be compared. We use the notation $\mathbf{G}_{*n}$ for the transformation of the preferred model. In principle, each $\mathbf{G}_{mn}$ can also be designed to map to some proper subset $\mathcal{G}_n$ of $\mathbb{R}^{d_n}$. However, in such cases we would have to address technical issues relating to topological, measure-theoretic and geometric or algebraic properties of $\mathcal{G}_n$ while studying

theoretical results, which may be considered avoidable since the statistician gets to choose the maps $\mathbf{G}_{mn}$. Consequently, we assume that the co-domain of each map $\mathbf{G}_{mn}$ is $\mathbb{R}^{d_n}$ in this paper, and avoid unnecessary mathematical complications.

The choice of $\mathbf{G}_{mn}$ may depend on the purpose for building the scientific model, and the way we interpret the model. This transformation allows us to consider the *science case* where the actual parameter values and their interpretation is subject to scrutiny, or *use cases* like prediction and classification problems.

Example 2.2.1 (Example 2.1.1 continued)**.** In Example 2.1.1 consider the three types of models:

1. Linear regression model: $Y_i = X_{i1}\beta_1 + X_{i2}\beta_2 + \epsilon_i, \epsilon_i \sim N(0, \sigma^2)$ with $\sigma > 0$ for $i = 1, \ldots, n$;

2. Semiparametric regression model: $Y_i = X_{i1}\beta_1 + g(X_{i2}) + \epsilon_i$, for an unknown function g;

3. Semiparametric single index model: $Y_i = h(X_{i1}\beta_1 + X_{i2}\beta_2) + \epsilon_i$ for some unknown function h.

If we consider only the linear regression model and are interested in the estimated linear effects on the covariates, any candidate model $\mathcal{M}_n$ shall correspond to $\mathcal{S}_n \subseteq \{1, 2\}$ and $\mathbf{c}_n \in \mathbb{R}^{2-|\mathcal{S}_n|}$. Consequently an identity transformation for all models is enough to put them in a comparable platform. However, when all three types of models above are considered together, comparing and choosing between them becomes tricky. While it is certainly possible consider all modelling methods as special cases of a general model: $Y_i = h(X_{i1}\beta_1 + g(X_{i2})) + \epsilon_i$ in presence of suitable technical conditions, restrict $h(.)$ in (3) and $g(.)$ in (2) as linear combinations of elements in some B-spline basis, and represent a model as a collection of elements in the space of the combined set of spline basis coefficients: it makes their interpretation less

intuitive. A more interpretable platform in this scenario can be the predicted value of responses, and one can simply take as $\mathbf{G}_{mn}$ the vector of fitted values obtained in each method.

We now define an important concept for use in the rest of this paper. Each candidate model corresponds to a subspace of the full parameter space $\mathbf{\Theta}_n$. For any given model $\mathcal{M}_n$, entries of its corresponding subspace $\mathbf{\Theta}_{mn}$ are specified by elements from $\mathbf{\Theta}_j$ for indices $j \in \mathcal{S}_n$, and entries from $\mathbf{c}_n$ when $j \notin \mathcal{S}_n$. Consequently, we define their versions in the transformed space $\mathcal{G}_n$:

$$\begin{aligned} \mathcal{G}_{mn} &:= \{\mathbf{G}_{mn}(\boldsymbol{\theta}(\mathcal{M}_n)) : \boldsymbol{\theta}(\mathcal{M}_n) \in \mathbf{\Theta}_{mn}\} \\ \mathcal{G}_{*n} &:= \{\mathbf{G}_{*n}(\boldsymbol{\theta}(\mathcal{M}_{*n})) : \boldsymbol{\theta}(\mathcal{M}_{*n}) \in \mathbf{\Theta}_{*n}\} \end{aligned}$$

In this framework,

Definition 2.2.2. For $\mathbf{g} \in \mathbb{R}^{d_n}$ and $\mathcal{G}'_n \subseteq \mathbb{R}^{d_n}$, we define the following:

$$d(\mathbf{g}, \mathcal{G}'_n) := \inf_{\mathbf{g}' \in \mathcal{G}'_n} \|\mathbf{g} - \mathbf{g}'\|$$

where $\|.\|$ is the Euclidean norm. Then

(a) For two sequences of models, say $\{\mathcal{M}_{1n}\}$ and $\{\mathcal{M}_{2n}\}$, we say $\{\mathcal{M}_{1n}\}$ *is nested within* $\{\mathcal{M}_{2n}\}$ if, for all sequences $\{\mathbf{g}_{1n} : \mathbf{g}_{1n} \in \mathcal{G}_{1n}\}$ we have

$$\lim_{n\to\infty} d(\mathbf{g}_{1n}, \mathcal{G}_{2n}) = 0 \tag{2.2.1}$$

(b) A sequence of models $\{\mathcal{M}_n\}$ is called *adequate* if the model $\mathcal{M}_{0n}$ corresponding to the singleton set $\mathbf{\Theta}_{0n} = \{\boldsymbol{\theta}_{0n}\}$, i.e. when $\mathcal{S}_{0n} = \varnothing$ and $\mathbf{c}_{0n} = \boldsymbol{\theta}_{0n}$, is nested within $\mathcal{M}_n$.

(c) A model that is not adequate is an *inadequate* model.

This notion of adequacy of a model depends on the choice of the preferred parameter vector, as well as the transformation maps $\mathbf{G}_{mn}$. The preferred model is always adequate, as is $\mathcal{M}_{0n}$, so the set of adequate models is non-empty by construction. Since the notion of parsimony is important in this context, we define the *minimal adequate* model as the adequate model that has the smallest number of parameters estimated from the data. Our framework ensures that there is always a minimal adequate model ($\mathcal{M}_{0n}$), though in general, its uniqueness is not guaranteed.

In classical model selection problems, as in linear regression where a subset of covariates $\mathbf{X}_s$ is used in fitting the expression $Y = \mathbf{X}_s\boldsymbol{\beta}_s + \epsilon$, this concept of model adequacy captures standard notions of model 'correctness'. Given a full-rank covariate matrix $\mathbf{X} \in \mathbb{R}^{k_n \times p}$, candidate models are fully specified by the set $\mathcal{S} \in \{1, \ldots, p\}$ of non-zero indices in $\boldsymbol{\beta}$, and for obvious choices of $\{\mathbf{G}_{mn}\}$, the condition for model adequacy reduces to $\mathbb{E}Y - \mathbf{X}_s\boldsymbol{\beta}_s = 0$. Thus the concept of the minimal adequate model merges with that of a 'true model' used in many studies.

We elicit the above broader definition to capture the limiting cases that arise in such situations. For instance, in the above example consider $p = 2$ and the triangular data generating model to be $Y_{ni} = X_{1i}\beta_{01} + X_{2i}\delta_n + \epsilon$ for some $\beta_{01} \in \mathbb{R}, \delta_n = o(1)$ and $i = 1, \ldots, k_n$. In our framework, given that the model with all covariates is the preferred model, the sequence of models $\mathcal{M}_n$ so that $\boldsymbol{\Theta}_{mn} = \{(\beta_1, 0)^T : \beta_1 \in \mathbb{R}\}$ shall be considered an adequate model. Such models frequently arise from prior choices in bayesian variable selection techniques (e.g. Narisetty and He (2014); Ročkovà and George (2016)).

2.2.3 Method of estimation

Since some or all the parameter values are unknown in a typical scientific problem, they have to be *estimated* from empirical observations. Suppose at stage n, the empirical data we have at hand is denoted by the set $\mathcal{B}_n = \{B_{n1}, \ldots, B_{nk_n}\}$, where we

do not restrict either the dimension of any of the A_{ni}'s, or declare any properties or restrictions on them. In particular, each B_{ni} may be infinite dimensional element, or a finite dimensional vector. The size of $\mathcal{B}_n$, which we call the *sample size* and denote by k_n is assumed to be a non-decreasing sequence of integers that tends to infinity as $n \to \infty$.

We consider here a known triangular array of functions, say $\Psi_{mni}(\cdot)$, for which the following equation has a unique minimizer in $\boldsymbol{\Theta}_{mn}$:

$$\Psi_{mn}(\boldsymbol{\theta}) = \mathbb{E} \sum_{i=1}^{k_n} \Psi_{mni}\big(\boldsymbol{\theta}, B_{ni}\big) \tag{2.2.2}$$

for any candidate model $\mathcal{M}_n$. Suppose this minimizer is $\boldsymbol{\theta}_{mn}$. We borrow the terminology *energy function* from optimization and other literature to denote such functions. They functions have also been called *contrast functions*, (see Pfanzagl (1969); Michel and Pfanzagl (1971); Bose and Chatterjee (2003)). The estimator $\hat{\boldsymbol{\theta}}_{mn}$ of $\boldsymbol{\theta}_{mn}$ is obtained as a minimizer of the sample analog of the above, i.e.

$$\hat{\boldsymbol{\theta}}_{mn} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}_{mn}}{\arg\min} \sum_{i=1}^{k_n} \Psi_{mni}\big(\boldsymbol{\theta}, B_{ni}\big) \tag{2.2.3}$$

The *preferred model estimate*, say $\hat{\boldsymbol{\theta}}_{*n}$ is described in an identical way. Thus

$$\hat{\boldsymbol{\theta}}_{*n} = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}_{*n}}{\arg\min} \sum_{i=1}^{k_n} \Psi_{*ni}\big(\boldsymbol{\theta}, B_{ni}\big) \tag{2.2.4}$$

where $\Psi_{*ni}(\cdot)$ are a known triangular array of functions.

Naturally, only the unknown elements of the generic model vector $\boldsymbol{\theta}$, say $\boldsymbol{\theta}(\mathcal{S}_n)$, and their sample equivalents are relevant for the above minimization problems. Hence for ease of exposition we shall assume that $\Psi_{mni}(\boldsymbol{\theta}, .) \equiv \Psi_{sni}(\boldsymbol{\theta}(\mathcal{S}_n), .)$ for $i = 1, \ldots, k_n$, i.e. the estimating functionals depend and operate only on the index sets to be

estimated.

We designate the subvector of $\boldsymbol{\theta}_{mn}$ at indices $\mathcal{S}_n$ by $\boldsymbol{\theta}_{sn}$, and assume the following very general conditions on this estimation process:

(S0) For inadequate models, the model corresponding to the singleton set $\{\boldsymbol{\theta}_{mn}\} \subseteq \boldsymbol{\Theta}_n$ is inadequate.

(S1) Define the Hilbert space $\ell_2 = \{\{x_n, n = 1, 2, \ldots\} : x_n \in \mathbb{R}, \sum_{n \geqslant 1} x_n^2 < \infty\}$, and embed $\mathbb{R}^{p_{sn}}$ in it as and when necessary, as the first $p_{sn} = |\mathcal{S}_n|$ elements of ℓ_2. Denote by $[\boldsymbol{\theta}]$ the probability distribution of the random variable $\boldsymbol{\theta}$. Then for any candidate model $\mathcal{M}_n$ there exists a tight sequence of probability measures $\mathbb{T}_{sn}$ on ℓ_2 with weak limit $\mathbb{T}_{s,\infty} \in \tilde{\ell}_2$, which is the set of probability measures on ℓ_2, and positive real numbers $a_{sn} \asymp a_{*n}$ such that

(a) For all n, $\left[a_{sn}\left(\hat{\boldsymbol{\theta}}_{sn} - \boldsymbol{\theta}_{sn}\right)\right]$ is the distribution of the marginal of $\mathbb{T}_{sn}$ under the first p_{sn} coordinates;

(b) For the preferred model solution $\boldsymbol{\theta}_{*n}$, $\|a_{*n}(\mathbf{G}_{*n} - \mathbf{G}_{0n})\| \to 0$ as $n \to \infty$.

Because of the definition of inadequate models, we need (S0) to ensure that the sequence of solutions for inadequate models do not actually end up converging to the preferred model vector $\boldsymbol{\theta}_{0n}$. We need (S1) to prove the population-level results in the next section, covering potentially biased estimation methods with bias going to 0 as n grows. A few technical conditions will eventually get added to this in Section 2.4 to establish consistency results of the resampling scheme used.

2.3 Statistical evaluation maps and e-values

2.3.1 A general evaluation map

We now introduce another function, the *statistical evaluation function*:

$$E_n : \mathbb{R}^{d_n} \times \tilde{\mathbb{R}}^{d_n} \to [0, \infty)$$

which takes as arguments a point from $\mathbb{R}^{d_n}$ and a probability measure from $\tilde{\mathbb{R}}^{d_n}$, and maps that pair into a non-negative real number. Roughly, the quantity $E_n(\mathbf{y}, [\mathbf{Y}])$ is a measure of where exactly the point $\mathbf{y}$ sits with respect to the distribution of the random variable $\mathbf{Y} \in \mathbb{R}^{d_n}$.

The exact nature of the evaluation function, which will make this rough notion precise, depends on the context. We shall discuss this in detail shortly. Good examples of evaluation functions are probabilities of sets like $A_\delta = \{x : |x| < \delta\}$ under $N(0, \sigma^2)$ distribution for $\sigma > 0$, unimodal probability density functions that uniformly decrease away from the mode in any direction, and various *data depth* functions. In fact, depths offer a very rich collection of relevant functions: although their properties are somewhat more restrictive than those our evaluation map requires initially. While we later use halfspace depth (Tukey, 1975) as our choice of evaluation map in model selection, for the majority of our theoretical analysis we do not restrict the evaluation maps to be only depth functions in order to avoid some of technical assumptions on traditional depth functions that are not required in our context until Section 2.5.

2.3.2 The e-value of models

We now associate with each model $\mathcal{M}_n$ a functional of the evaluation map E_n: which we call the e-value. An example of e-value is the mean evaluation map function:

$$e_n(\mathcal{M}_n) = \mathbb{E}E_n\left(\hat{\mathbf{G}}_{mn}, [\hat{\mathbf{G}}_{*n}]\right) \tag{2.3.1}$$

which we concentrate on for the rest of the paper. However, any other functional of $E_n(\hat{\mathbf{G}}_{mn}, [\hat{\mathbf{G}}_{*n}])$ may also be used here, and a large proportion of our theoretical discussion in the rest of the paper is applicable to any smooth functional of the distribution of $E_n(\hat{\mathbf{G}}_{mn}, [\hat{\mathbf{G}}_{*n}])$. Furthermore, the distribution of $E_n(\hat{\mathbf{G}}_{mn}, [\hat{\mathbf{G}}_{*n}])$ is itself informative, and has an important role to play in the study of uniform convergence. We defer all this discussion and analysis to future research.

Remark. From a hypothesis testing perespective, e-values generalize the concept of p-values. Consider the problem of finding out the right tail probability with respect to a null distribution, say $[T_{0n}]$, for a test statistic $\hat{T}_{mn}$. Here we can designate the model corresponding to the null hypothesis as the preferred model, take the smooth transformation as $\mathbf{G}_{mn} \equiv \hat{T}_{mn}$ and given the evaluation map $E_n(\hat{T}_n, [T_{0n}]) = \mathbb{I}_{\hat{T}_n > T_{0n}}$ the e-value is calculated as $P(\hat{T}_n > T_{0n})$. A higher e-value (or p-value) indicates a high degree of similarity between the null and alternate model, or in other words the alternate model is 'adequate' for the null model. However, in terms of usefulness the inadequate models will be the useful ones in this context.

There are two random quantities involved in the expression of $e(\mathcal{M}_n)$ above, namely $\hat{\mathbf{G}}_{mn}$ and $\hat{\mathbf{G}}_{*n}$. Typically, the distribution of either of these random quantities are not known, and have to be elicited from data. We shall use resampling methods for this purpose, the details of which will be outlined in later sections.

2.3.3 Model adequacy and e-values

We now present our first result on the model elicitation process, which as claimed earlier, separates the inadequate models from the adequate ones.

For this, we first assume two conditions on the transformation $\mathbf{G}_{mn}$. Note that the j-th element of the function $\mathbf{G}_{mn}$, denoted by $G_{mnj}(\cdot) \equiv G_j(\cdot)$, is a map from a subset of $\mathbb{R}^{p_n}$ to $\mathbb{R}$, for $j = 1, \ldots, d_n$. Here we assume that

(G1) $d_n = o(\min_{\mathcal{S}_n}\{a_{sn}, a_{*n}\})$;
(G2) The functions $G_j(\cdot)$ are smooth functions in a neighborhood of $\boldsymbol{\theta}_{mn} \equiv \boldsymbol{\theta}$. Specifically, there exists a $\delta > 0$ such that for $\mathbf{x} = \boldsymbol{\theta} + \mathbf{t}$ with $\|\mathbf{t}\| < \delta$, we have the following expansion

$$G_j(\mathbf{x}) = G_j(\boldsymbol{\theta}) + \mathbf{G}_{1j}^T(\boldsymbol{\theta})\mathbf{t} + 2^{-1}\mathbf{t}^T\mathbf{R}_j(\boldsymbol{\theta} + c\mathbf{t})\mathbf{t} \tag{2.3.2}$$

for some $c \in (0, 1)$. We assume that there is a positive definite matrix $\mathbf{M}_j$ such that

$$\sup_{\mathbf{t}:\|t\|<\delta} \mathbf{R}_j(\boldsymbol{\theta} + c\mathbf{t}) < \mathbf{M}_j; \quad \lambda_{max}(\mathbf{M}_j) < \infty \tag{2.3.3}$$

Also, the technical conditions assumed on the sequence of evaluation maps are as follows:

(E1) Each E_n is invariant to location and scale transformations, i.e. for any $a \in \mathbb{R}, \mathbf{b} \in \mathbb{R}^{d_n}$ and random variable $\mathbb{G}$ having distribution $\mathbb{G} \in \tilde{\mathbb{R}}^{d_n}$,

$$E_n(\mathbf{x}, \mathbb{G}) = E_n(a\mathbf{x} + \mathbf{b}, [a\mathbf{G} + \mathbf{b}]) \tag{2.3.4}$$

(E2) Each E_n is Lipschitz continuous in the first argument, i.e. there exists an

$\alpha_n > 0$, possibly depending on the measure $\mathbb{G} \in \tilde{\mathcal{G}}_n$ such that

$$|E_n(\mathbf{x}, \mathbb{G}) - E_n(\mathbf{y}, \mathbb{G})| < \|\mathbf{x} - \mathbf{y}\|^{\alpha_n} \tag{2.3.5}$$

(E3) Suppose $\{\mathbb{Y}_n\}$ is a tight sequence of probability measures on ℓ_2, with weak limit $\mathbb{Y}_\infty$. Further assume that $\mathbf{Y}_n \in \mathbb{R}^{d_n}$ is a random variable that follows the marginal distribution of the first d_n co-ordinates under $\mathbb{Y}_n$. Also suppose $E_\infty : \ell_2 \times \tilde{\ell}_2 \to [0, \infty)$ be a map such that $\mathbb{E}E_\infty(\mathbf{y}, \mathbb{Y}_\infty) < \infty$, and when restricted to the first d_n co-ordinates, E_∞ matches E_n. Then we assume that

$$\lim_{n\to\infty} \mathbb{E}E_n(\mathbf{Y}_n, [\mathbf{Y}_n]) = \mathbb{E}E_\infty(\mathbf{y}, \mathbb{Y}_\infty) \tag{2.3.6}$$

(E4) Now suppose that $\mathbf{Z}_n \in \mathbb{R}^{d_n}$ is another sequence of random variables. Then, if $\|\mathbf{Z}_n\| \xrightarrow{P} \infty$, we assume the following condition as $n \to \infty$:

$$E_n(\mathbf{Z}_n, [\mathbf{Y}_n]) \xrightarrow{P} 0 \tag{2.3.7}$$

Clearly, these properties are not mutually exclusive, and some may be derived from others, but we present these together for ease of verification. Additionally some properties like Lipschitz continuity and (E4) are simply for technical convenience, while we only require the condition (E3) that is weaker than uniform convergence.

We are now at a stage to present our population-level result that forms the foundation of all the following analysis.

Theorem 2.3.1. *Consider a sequence of evaluation functions E_n satisfying properties (E1)-(E4). Then as $n \to \infty$:*

1. *For the preferred model $\mathcal{M}_{*n}$, $e_n(\mathcal{M}_{*n}) \to e_\infty < \infty$;*
2. *When $\mathcal{M}_n$ is an adequate model, $|e_n(\mathcal{M}_n) - e_n(\mathcal{M}_{*n})| \to 0$;*

3. When $\mathcal{M}_n$ is an inadequate model, $e_n(\mathcal{M}_n) \to 0$.

This result ensures that for large enough n, it is possible to find some threshold $\epsilon_n \leqslant e_n(\mathcal{M}_{*n})$ such that all inadequate models have e-values less than the threshold, while e-values for all adequate models fall above it. The choice of ϵ_n, of course, depends on several factors like the evaluation map, estimation technique used and sample size: some cases of which we shall pursue later (Section 2.5 and Section 3.3).

2.4 Estimation of e-values through resampling

In this section we shall use a resampling scheme to estimate the distributions corresponding to the smooth functionals of candidate model parameters we consider, i.e. $\hat{\mathbf{G}}_{mn}$, and discuss consistency of the resulting procedure in estimating model e-values through imposing certain necessary conditions on the resampling scheme used.

A special case of the family of resampling methods that we use in this chapter is the m-out-of-n bootstrap, which we abbreviate as *moon bootstrap*. There are numerous problems where the moon-bootstrap provides consistent approximation to the distribution of statistics of interest (e.g. Shao (1996); Chatterjee and Bose (2005)). Since all such cases are too numerous to list and review of resampling consistency is not central to this chapter, we only demonstrate the properties of our resampling procedure in some interesting frameworks.

Recall that in (2.2.3) and (2.2.4) we obtain the estimator $\hat{\boldsymbol{\theta}}_{mn}$ by minimizing the *energy functional* or *estimating functional* $\hat{\Psi}_{sn}(\boldsymbol{\theta}) = \sum_{i=1}^{k_n} \Psi_{sni}(\boldsymbol{\theta}, B_{ni})$. The parameter $\boldsymbol{\theta}_{mn}$ is the unique minimizer of the expectation of the above over all $\boldsymbol{\theta} \in \boldsymbol{\Theta}_{mn}$. In this section, we occasionally drop the subscript $_s$ and $_*$ when there is no scope for confusion for notational simplicity, since the developments presented in the rest of this section are applicable to any model. We often drop the second argument from estimating functionals, thus for example $\Psi_{ni}(\boldsymbol{\theta}) \equiv \Psi_{sni}(\boldsymbol{\theta}, B_{ni})$. Any other notational

simplifications in various contexts of this section will be presented as related contexts arise.

2.4.1 Smooth estimating functional models

We shall consider the case is where the functions $\Psi_{sni}(\cdot,\cdot)$ is smooth in the first argument, which covers a vast number of models routinely considered in statistics.

In a neighborhood of $\boldsymbol{\theta}_{sn}$, which is the solution of $\Psi_{sn} = \sum_i \Psi_{sni}$, we assume that the functions Ψ_{sni} are thrice continuously differentiable in the first argument, with the successive derivatives being denoted by Ψ_{ksni}, $k = 0, 1, 2$. That is, there exists a $\delta > 0$ such that for any $\boldsymbol{\theta} = \boldsymbol{\theta}_{sn} + \mathbf{t}$ satisfying $\|\mathbf{t}\| < \delta$ we have

$$\frac{d}{d\boldsymbol{\theta}}\Psi_{sni}(\boldsymbol{\theta}) = \Psi_{0sni}(\boldsymbol{\theta}) \in \mathbb{R}^{p_{sn}} \tag{2.4.1}$$

where $p_{sn} = |\mathcal{S}_n|$. For the a^{th} element of $\Psi_{0sni}(\boldsymbol{\theta}), a = 1, \ldots p_{sn}$, denoted by $\Psi_{0sni(a)}(\boldsymbol{\theta})$, we have

$$\Psi_{0sni(a)}(\boldsymbol{\theta}) = \Psi_{0sni(a)}(\boldsymbol{\theta}_{sn}) + \Psi_{1sni(a)}(\boldsymbol{\theta}_{sn})t + 2^{-1}\mathbf{t}^T\Psi_{2sni(a)}(\boldsymbol{\theta}_{sn} + c\mathbf{t})\mathbf{t} \tag{2.4.2}$$

for some $c \in (0, 1)$ possibly depending on a. We assume that for each $\mathcal{S}_n$ and n, there is a sequence of σ-fields $\mathcal{F}_{sn1} \subset \mathcal{F}_{sn2} \ldots \mathcal{F}_{snk_n}$ such that $\{\sum_{i=1}^{j} \Psi_{0sni}(\boldsymbol{\theta}_{sn}), \mathcal{F}_{snj}\}$ is a martingale.

Also, let the spectral decomposition of the matrix $\boldsymbol{\Gamma}_{0sn} = \sum_{i=1}^{k_n} \mathbb{E}\Psi_{0sni}(\boldsymbol{\theta}_{sn})\Psi_{0sni}^T(\boldsymbol{\theta}_{sn})$ be given by

$$\boldsymbol{\Gamma}_{0sn} = \mathbf{P}_{0sn}\boldsymbol{\Lambda}_{0sn}\mathbf{P}_{0sn}^T \tag{2.4.3}$$

where $\mathbf{P}_{0sn} \in \mathbb{R}^{p_{sn}} \times \mathbb{R}^{p_{sn}}$ is an orthogonal matrix whose columns contain the eigenvectors, and $\boldsymbol{\Lambda}_{0sn}$ is a diagonal matrix contining the eigenvalues of $\boldsymbol{\Gamma}_{0sn}$. We assume

that $\mathbf{\Gamma}_{0sn}$ is positive definite, that is, all the diagonal entries of $\mathbf{\Lambda}_{0sn}$ are positive numbers. We assume that there is a constant $\delta_{0s} > 0$ such that $\lambda_{min}(\mathbf{\Gamma}_{0sn}) > \delta_{0sn}$ for all sufficiently large n. The matrices $\mathbf{\Lambda}_{0sn}^{c}$ for various real numbers c are defined in the obvious way, that is, these are diagonal matrices where the j-th diagonal entry is raised to the power c.

Let $\mathbf{\Gamma}_{1sni}(\boldsymbol{\theta}_{sn})$ be the $p_{sn} \times p_{sn}$ matrix whose a-th row is $\mathbb{E}\Psi_{1sni(a)}(\boldsymbol{\theta}_{sn})$; we assume this expectation exists. Define

$$\mathbf{\Gamma}_{1sn}(\boldsymbol{\theta}_{sn}) = \sum_{i=1}^{k_n} \mathbf{\Gamma}_{1sni}(\boldsymbol{\theta}_{sn}) \tag{2.4.4}$$

We assume that $\mathbf{\Gamma}_{1sn} \equiv \mathbf{\Gamma}_{1sn}(\boldsymbol{\theta}_{sn})$ is nonsingular for each $\mathcal{M}_n$ and n. Suppose the singular value decomposition of $\mathbf{\Gamma}_{1sn}$ is given by

$$\mathbf{\Gamma}_{1sn} = \mathbf{P}_{1sn}\mathbf{\Lambda}_{1sn}\mathbf{Q}_{1sn}^{T} \tag{2.4.5}$$

where $\mathbf{P}_{1sn}, \mathbf{Q}_{1sn} \in \mathbb{R}^{p_{sn}} \times \mathbb{R}^{p_{sn}}$ are orthogonal matrices, and $\mathbf{\Lambda}_{1sn}$ is a diagonal matrix. We assume that the diagonal entries of $\mathbf{\Lambda}_{1sn}$ are all positive, which implies that in the parameter space the energy functional Ψ_{sn} actually achieves a minimal value at $\boldsymbol{\theta}_{sn}$, the solution of the optimization problem. We define the matrices $\mathbf{\Lambda}_{1sn}^{c}$ for various real numbers c as diagonal matrices where the j-th diagonal entry is raised to the power c. Correspondingly, we define $\mathbf{\Gamma}_{1sn}^{c} = \mathbf{P}_{1sn}\mathbf{\Lambda}_{1sn}^{c}\mathbf{Q}_{1sn}^{T}$, and assume that there is a constant $\delta_{1sn} > 0$ such that $\lambda_{min}(\mathbf{\Gamma}_{1sn}^{T}\mathbf{\Gamma}_{1sn}) > \delta_{1sn}$ for all sufficiently large n.

We now define the matrix

$$\mathbf{A}_{sn} := \mathbf{\Gamma}_{0sn}^{-1/2}\mathbf{\Gamma}_{1sn}. \tag{2.4.6}$$

and assume the following conditions:

(S2) The minimum eigenvalue of $\mathbf{A}_{sn}^T\mathbf{A}_{sn}$ tends to infinity. Specifically, for the sequence $a_{sn} \uparrow \infty$ from condition (S1), we have

$$\lambda_{\min}\left(\mathbf{\Gamma}_{1sn}\mathbf{\Gamma}_{0sn}^{-1}\mathbf{\Gamma}_{1sn}^T\right) \asymp a_{sn}^2 \tag{2.4.7}$$

(S3) Also there exists a sequence $\tau_{sn} \uparrow \infty, \tau_{sn}^{-1} = O(1)$ as $n \to \infty$ such that

$$\lambda_{\max}\left(\mathbf{\Gamma}_{1sn}^{-1}\mathbf{\Gamma}_{0sn}^2\mathbf{\Gamma}_{1sn}^{-1T}\right) = o(\tau_{sn}^{-2}) \tag{2.4.8}$$

as $n \to \infty$ for any $\mathcal{S}_n$, and

$$\mathbb{E}\left\|\mathbf{A}_{sn}^{-1}\left(\sum_{i=1}^{k_n}\Psi_{1sni} - \mathbf{\Gamma}_{1sn}\right)\mathbf{A}_{sn}^{-1}\right\|_F^2 = o(p_{sn}\tau_{sn}^{-2}) \tag{2.4.9}$$

where $\|\mathbf{A}\|_F$ denotes the Frobenium norm of matrix $\mathbf{A}$.

(S4) For the symmetric matrix $\Psi_{2sni(a)}(\boldsymbol{\theta})$ and for some $\delta_0 > 0$, there exists a symmetric matrix $\mathbf{M}_{2sni(a)}$ such that

$$\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_{mn}|<\delta_0} \Psi_{2sni(a)}(\boldsymbol{\theta}) < \mathbf{M}_{2sni(a)}, \tag{2.4.10}$$

satisfying

$$\sum_{a=1}^{p_{sn}}\sum_{i=1}^{k_n}\mathbb{E}\lambda_{max}^2\left(\mathbf{M}_{2sni(a)}\right) = o\left(a_{sn}^6 n^{-1} p_{sn}\tau_{sn}^{-2}\right) \tag{2.4.11}$$

For any vector $\mathbf{c} \in \mathbb{R}^{p_{sn}}$ with $\|\mathbf{c}\| = 1$, define $\mathbf{Z}_{mni} = -\mathbf{c}^T\mathbf{\Gamma}_{0sn}^{-1/2}\Psi_{0sni}$ for $i = 1, \ldots k_n$. We assume that

$$\sum_{i=1}^{k_n}\mathbf{Z}_{sni}^2 \xrightarrow{P} 1, \text{ and } \mathbb{E}\left[\max_i|\mathbf{Z}_{sni}|\right] \to 0. \tag{2.4.12}$$

from hereon using $\Psi_{ksni} \equiv \Psi_{ksni}(\boldsymbol{\theta}_{sn})$, for $k = 0, 1, 2$.

We now consider an array of resampling weights $\mathbb{W}_{rsni}$, which for any n may be collected together in the vector $\mathbf{W}_{rsn} = (\mathbb{W}_{rsn1}, \ldots, \mathbb{W}_{rsnk_n})^T \in \mathbb{R}^{k_n}$. We assume that this is an exchangeable array of non-negative random variables, independent of the data. The index r denotes that these are related to the resampling procedure. The actual implementation of the resampling procedure is carried out by generating independent copies $\{\mathbf{W}_{1sn}, \ldots, \mathbf{W}_{Rsn}\}$ for some sufficiently large integer R, and using them in a Monte Carlo procedure, where for any $r = 1, \ldots, R$, we minimize

$$\sum_{i=1}^{k_n} \mathbb{W}_{rsni} \Psi_{mni}\left(\theta, B_{ni}\right) \tag{2.4.13}$$

to obtain the resampling version of the estimator $\hat{\boldsymbol{\theta}}_{rsn} \in \mathbb{R}^{p_{sn}}$. This is the *generalized bootstrap* (Chatterjee and Bose, 2005).

We assume that for each $i = 1, \ldots, k_n$, $\mathbb{E}\mathbb{W}_{rsni} = \mu_{sn}$ and $\mathbb{V}\mathbb{W}_{rsni} = \tau_{sn}^2$, and write the centered and scaled resampling weights as

$$W_{rsni} = \tau_{sn}^{-1}\left(\mathbb{W}_{rsni} - \mu_{sn}\right), \tag{2.4.14}$$

thus $\mathbb{W}_{rsni} = \mu_{sn} + \tau_{sn} W_{rsni}$. Since $\mathbb{W}_{rsni} \geqslant 0$ almost surely and is non-degenerate, we have $\mu_{sn} > 0$, and assume that $\mu_{sn} + \tau_{sn}^2 = O(\tau_{sn}^2)$. Our analysis below suggests that the properties of the resampling procedure depend only on the *coefficient of variation* ratio τ_{sn}/μ_{sn}, so without loss of generality we can set $\mu_{sn} = 1$ for all s and n.

We assume the following conditions on the resampling weights as $n \to \infty$:

$$\mathbb{E}\mathbb{W}_{rsn1} = \mu_{sn}, \tag{2.4.15}$$

$$\mathbb{V}\mathbb{W}_{rsn1} = \tau_{sn}^2 \uparrow \infty, \tag{2.4.16}$$

$$\tau_{sn}^2 = o(a_{sn}^2), \tag{2.4.17}$$

$$\mathbb{E}W_{rsn1}W_{rsn2} = O(k_n^{-1}), \tag{2.4.18}$$

$$\mathbb{E}W_{rsn1}^2 W_{rsn2}^2 \to 1, \tag{2.4.19}$$

$$\mathbb{E}W_{rsn1}^4 < \infty. \tag{2.4.20}$$

Example 2.4.1 (The m-out-of-n (moon) bootstrap)**.** In our framework, the *moon*-bootstrap is identified with $\mathcal{W}_{rsn}$ having a Multinomial distribution with parameters m and probabilities $k_n^{-1}(1,\ldots,1) \in \mathbb{R}^{k_n}$, by a factor of k_n/m. Thus we have $\mathbb{E}\mathbb{W}_{rsn1} = \mu_{sn} = (m^{-1}k_n)(m/k_n) = 1$, and $\mathbb{V}\mathbb{W}_{rsn1} = \tau_{sn}^2 = (m^{-1}k_n)^2(mk_n^{-1}(1 - k_n^{-1}) = O(m^{-1}k_n)$. In typical applications of the *moon*-bootstrap, as in its application in this chapter, we require that $m \to \infty$ and $m/k_n \to 0$ as $n \to \infty$. Thus we have $\tau_{sn}^2 \to \infty$ as $n \to \infty$, thus the *scale* factor of the resampling weights $\mathbb{W}_{rsni}$ tend to infinity with n. We use the term *scale-enhanced* resampling for schemes like the *moon*-bootstrap where the variance of (properly centered) resampling weights tends to infinity with n.

Example 2.4.2 (The scale-enhanced Bayesian bootstrap)**.** A version of Bayesian bootstrap may be constructed by choosing $\mathbb{W}_{rsni}$ to be independent and identically distributed Gamma random variables, with mean $\mu_{sn} = 1$ and variance $\tau_{sn}^2 \to \infty$ as $n \to \infty$. The functionality of this resampling scheme and Bayesian interpretation remain similar to the standard Bayesian bootstrap, however some convenient properties like conjugacy are lost.

Theorem 2.4.3. *Assume conditions (S0)-(S5) and that* $p_{sn}^2 k_n^{-1} \to 0$ *as* $n \to \infty$*. Addi-*

tionally, assume that the resampling weights $\mathbb{W}_{rsni}$ *are exchangeable random variables satisfying the conditions (2.4.15)-(2.4.20). Define* $\hat{\mathbf{B}}_{sn} := \tau_{sn}^{-1}\hat{\mathbf{\Gamma}}_{0sn}^{1/2}\hat{\mathbf{\Gamma}}_{1sn}^{-1}$*, where* $\hat{\mathbf{\Gamma}}_{0sn}$ *and* $\hat{\mathbf{\Gamma}}_{1sn}$ *are sample equivalents of* $\mathbf{\Gamma}_{0sn}$ *and* $\mathbf{\Gamma}_{1sn}$*, respectively. Then* $\mathbf{A}_{sn}(\hat{\boldsymbol{\theta}}_{sn} - \boldsymbol{\theta}_{sn})$ *converges weakly to the standard Normal distribution in* p_{sn} *dimension, and conditional on the data,* $\hat{\mathbf{B}}_{sn}(\hat{\boldsymbol{\theta}}_{rsn} - \hat{\boldsymbol{\theta}}_{sn})$ *also converges weakly to the same distribution in probability.*

2.4.2 Bootstrap estimation of e-values

We now consider the sample equivalent of the e-value and prove that it consistently estimates the population e-value for certain resampling schemes. We use a resampling scheme satisfying conditions in the previous subsection, and two independent bootstrap samples, indexed by $_{r\cdot}$ and $_{r_1\cdot}$, from the preferred model. We use the first set of samples to generate coefficient vectors $\hat{\boldsymbol{\theta}}_{rmn}$ corresponding to the model $\mathcal{M}_n$:

$$\hat{\theta}_{rmnj} = \begin{cases} \text{Unknown } \hat{\theta}_{rsnj} & \text{for } j \in \mathcal{S}_n; \\ \text{Known } c_{nj} & \text{for } j \notin \mathcal{S}_n. \end{cases} \tag{2.4.21}$$

and the second set of samples to get bootstrap approximation of $[\hat{\mathbf{G}}_{*n}]$. Given that $\mathbf{G}_{mn}(\hat{\boldsymbol{\theta}}_{rmn}) \equiv \hat{\mathbf{G}}_{rmn}$ and $\mathbf{G}_{*n}(\hat{\boldsymbol{\theta}}_{r_1*n}) \equiv \hat{\mathbf{G}}_{r_1*n}$, we define the sample e-value as

$$\hat{e}(\mathcal{M}_n) := \mathbb{E}_r E_n(\hat{\mathbf{G}}_{rmn}, [\hat{\mathbf{G}}_{r_1*n}]) \tag{2.4.22}$$

The expectation above is taken on the first set of bootstrap samples.

Theorem 2.4.4. *Consider a resampling scheme satisfying technical conditions in the previous subsection, and an evaluation map* E_n *satisfying the assumptions (E1)-(E4). Define* $b_{sn} = a_{sn}/\tau_{sn}$*, and assume that (a)* $b_{sn} \asymp b_{*n}$*, (b)* $d_n = o(\min_{\mathcal{S}_n}(\{b_{sn}, b_{*n}\})$*. Then as* $n \to \infty$*,*

1. *For any adequate model* $\mathcal{M}_n$ *we have* $|\hat{e}_n(\mathcal{M}_n) - \hat{e}_n(\mathcal{M}_{*n})| \xrightarrow{P_n} o_P(1)$*;*

2. *For any inadequate model* $\mathcal{M}_n$ *we have* $\hat{e}_n(\mathcal{M}_n) \xrightarrow{P_n} o_P(1)$.

where $s_n \xrightarrow{P_n} t_n$ *means* s_n *converges in probability to* t_n *conditional on the data.*

Proving the above theorem requires largely similar arguments used in the proof of its population counterpart, i.e. theorem 2.3.1. Interestingly, as shown in the proof, we do not actually require τ_n to go to infinity to achieve convergence for adequate models: only $b_{sn} \to \infty$ is good enough. The slower rate is only required to separate out e-value estimates of inadequate models from those of the adequate models. In practice when dealing with $\sqrt{n}$-consistent estimators (i.e. $a_{sn} = a_{*n} = \sqrt{n}$ for all $\mathcal{M}_n$, this would mean choosing the variance parameter $\tau_{sn}^2 = \tau_n^2$ of the resampling weight distribution $\mathbb{W}_{sn} \equiv \mathbb{W}_n$ such that $\tau_n^2 \to \infty$ and $\tau_n^2/n \to 0$ as $n \to \infty$.The bootstrap model selection criterion by Shao (1996) had used the same specification of bootstrap weights to obtain a criterion that achieves asymptotic model selection consistency: albeit in a very specific setup compared to our formulation. Also, numerous examples exist in model selection literature of using similar quantities explicitly as a penalty term in model selection criteria (Schwarz, 1978; Konishi and Kitagawa, 1996) or the loss function (Zou, 2006).

2.5 Fast variable selection using data depth

The traditional application domain for statistical model selection has been in *covariate selection*: for regression, mixed effect models, time series and other problems. Also, in many instances, the number of parameters does not grow significantly faster than the sample size. In such situations, it is feasible to consider the least parsimonious model as the preferred model. This is routinely done in practice, for example in classical

model selection techniques (Konishi and Kitagawa, 1996; Claeskens and Hjort, 2008), and the fence method (Jiang et al., 2008).

From now on we assume that the least parsimonious model has $p_n = p$ parameters for all n, and thus drop $_n$ in all subscripts that depend on p_n, e.g. in $\mathcal{M}_n, \boldsymbol{\theta}_{mn}, \mathbf{G}_{mn}$, as well as $_*$ in all subscripts corresponding to the preferred model. Although we still keep the subscript in e_n because it is calculated based on the estimators $\hat{\boldsymbol{\theta}}_m$ that depends on a size n-sample. We shall consider as preferred model the least parsimonious model with all covariates estimated from the data, and refer to it as the 'full model' from now on. In a typical variable selection problem, all candidate models are sub-models of this model, in the sense that one or more of the parameters are set to zero instead of being estimated from the data. An example is that of linear regression with total p covariates, and different candidate models are obtained by setting subsets of regression coefficients to zero. In such models, obtaining the most parsimonious model that fits the data, for example by using the Bayesian Information Criterion (BIC) (Schwarz, 1978), a full-scale analysis would require analyzing all 2^p possible candidate models. This is an NP-Hard problem (Natarajan, 1995), and becomes computationally intractible even for moderate data dimensions ($n \simeq 100, p \simeq 50$). Several *ad-hoc* techniques that are in use do not guarantee, in the absence of stringent conditions, that the probability of selecting the most parsimonious model that fits the data tends to one as sample size increases. In this section we shall devise a fast and scalable algorithm to tackle this problem, i.e. detect variables with non-zero coefficients, through implementing our generic e-values framework.

2.5.1 A plugin parameter estimate

We are going fit only the full model in the process of performing covariate selection. We first obtain a consistent estimator $\hat{\boldsymbol{\theta}}_* = (\hat{\theta}_{*1}, \ldots, \hat{\theta}_{*p})^T$ of the full coefficient vector for this. For a general model $\mathcal{M}$ specified by the set $\mathcal{S} = \{j_1, \ldots, j_{p_s}\} \subseteq \{1, 2, \ldots, p\}$

and the vector of potentially non-zero constants $\mathbf{c}$, we define the parameter estimates to be

$$\hat{\theta}_{mj} = \begin{cases} \text{Unknown } \hat{\theta}_{*j} & \text{for } j \in \mathcal{S}; \\ \text{Known } c_j & \text{for } j \notin \mathcal{S}. \end{cases} \tag{2.5.1}$$

Thus, we do not fit the model $\mathcal{M}$ separately, but simply plug-in estimators from the full model at the indices in $\mathcal{S}$. Following (2.4.21), bootstrapped model estimates are obtained as

$$\hat{\theta}_{rmj} = \begin{cases} \text{Unknown } \hat{\theta}_{r*j} & \text{for } j \in \mathcal{S}; \\ \text{Known } c_n & \text{for } j \notin \mathcal{S} \end{cases} \tag{2.5.2}$$

The logic behind this is simple: for a candidate model $\mathcal{M}$, a joint distribution of the estimator of its parameters, i.e. $[\hat{\boldsymbol{\theta}}_s]$, can actually be obtained from the marginal of $[\hat{\boldsymbol{\theta}}_*]$ at indices $\mathcal{S}$. This makes it easy to guarantee that the distribution of parameter estimates for any selected model is consistently approximated through the corresponding sampling distributions by our method. We conjecture that this logic may be applied in the context of several other model selection methods also, but do not pursue that line of study in this paper.

The above plug-in step has two more major advantages. First, we do not separately analyze each candidate model, and instead use resampling, implying significant computational savings. Second, this approach leads to an easier comparison of any candidate model to the preferred model.

2.5.2 Simplifications

At this stage we make a few simplifying assumptions that will allow us to obtain specialized results relevant in the context. First of all we assume $\mathbf{G}_m$ to be the

identity function, i.e. $\mathbf{G}_m(\boldsymbol{\theta}) = \boldsymbol{\theta}$ for any $\mathcal{M}$ and $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. This vastly simplifies the definition of nested models and model adequacy: we now consider a model $\mathcal{M}_1$ to be nested in $\mathcal{M}_2$ if $\mathcal{S}_1 \subseteq \mathcal{S}_2$ and $\mathbf{c}_2$ is a subvector of $\mathbf{c}_1$. Also a model $\mathcal{M}$ is adequate simply if the preferred parameter vector $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}(\mathcal{M})$.

For the evaluation functions, we take a single map $E : \mathbb{R}^p \times \tilde{\mathbb{R}}^p \to [0, \infty)$ for all n that satisfies the following properties:

(D1) The map E is invariant to affine transformations, i.e. for any non-singular matrix $\mathbf{A} \in \mathbb{R}^{p\times p}$, and $\mathbf{b} \in \mathbb{R}^p$ and random variable $\mathbf{G}$ having distribution $\mathbb{G} \in \tilde{\mathbb{R}}^p$, the set of probability measures on $\mathbb{R}^p$,

$$E(\mathbf{x}, \mathbb{G}) = E(\mathbf{A}\mathbf{x} + \mathbf{b}, [\mathbf{A}\mathbf{G} + \mathbf{b}]) \tag{2.5.3}$$

(D2) The map E is Lipschitz continuous in the first argument, i.e. there exists an $\alpha > 0$, possibly depending on the measure $\mathbb{G} \in \tilde{\mathbb{R}}^p$ such that for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$,

$$|E(\mathbf{x}, \mathbb{G}) - E(\mathbf{y}, \mathbb{G})| < \|\mathbf{x} - \mathbf{y}\|^\alpha \tag{2.5.4}$$

(D3) Assume that $\mathbf{Y}_n \in \mathbb{R}^p$ is a sequence of random variables converging in distribution to some $\mathbb{Y} \in \tilde{\mathbb{R}}^p$. Then $E(\mathbf{y}, [\mathbf{Y}_n])$ converges uniformly to $E(\mathbf{y}, \mathbb{Y})$.

(D4) For any $\mathbb{G} \in \tilde{\mathbb{R}}^p$, $\lim_{\|\mathbf{x}\|\to\infty} E(\mathbf{x}, \mathbb{G}) = 0$.

(D5) For any $\mathbb{G} \in \mathbb{R}^p$ with a point of symmetry $\boldsymbol{\mu}(\mathbb{G}) \in \mathbb{R}^p$, we have for any $t \in (0, 1)$ and any $\mathbf{x} \in \mathbb{R}^p$

$$E(\mathbf{x}, \mathbb{G}) < E(\boldsymbol{\mu}(\mathbb{G}) + t(\mathbf{x} - \boldsymbol{\mu}(\mathbb{G})), \mathbb{G}) < E(\boldsymbol{\mu}(\mathbb{G}), \mathbb{G}) = \sup_{\mathbf{x}\in\mathbb{R}^p} E(\mathbf{x}, \mathbb{G}) < \infty \tag{2.5.5}$$

That is, the evaluation takes a maximum value at $\boldsymbol{\mu}(\mathbb{G})$, and is strictly decreasing

along any ray connecting $\boldsymbol{\mu}(\mathbb{G})$ to any point $\mathbf{x} \in \mathbb{R}^p$.

The second property is a restatement of (E2) assuming a common evaluation map for all n. The first, third and fourth properties are stronger versions of (E1), (E3) and (E4). (D5) will be essential in proving the theoretical results that follow. Also note that (D1), (D3), (D4) and (D5) have traditionally been used for depth functions, and lipschitz continuity and uniform convergence arises implicitly for many implementations of data depth (see Chapter 1). Coupled with the fact that we shall be using depth functions as evaluation maps in numerical sections that follow shortly, from hereon we shall use the notation $D(\mathbf{x}, \mathbb{G})$ in place of $E(\mathbf{x}, \mathbb{G})$ for clarity.

We shall assume elliptical asymptotic distributions for full model estimators $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_*$. Following Fang et al. (1990), elliptical distributions can be formally defined using their characteristic function:

Definition 2.5.1. A p-dimensional random vector $\mathbf{X}$ is said to elliptically distributed if and only if there exist a vector $\boldsymbol{\mu} \in \mathbb{R}^p$, a positive semi-definite matrix $\boldsymbol{\Omega} \equiv \boldsymbol{\Sigma}^{-1} \in \mathbb{R}^{p \times p}$ and a function $\phi : \mathbb{R}_+ \to \mathbb{R}$ such that the characteristic function $\mathbf{t} \mapsto \phi_{\mathbf{X}-\boldsymbol{\mu}}(\mathbf{t})$ of $\mathbf{X} - \boldsymbol{\mu}$ corresponds to $\mathbf{t} \mapsto \phi(\mathbf{t}^T \boldsymbol{\Sigma} \mathbf{t}), \mathbf{t} \in \mathbb{R}^p$.

The density function of an elliptically distributed random variable takes the form:

$$h(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = |\boldsymbol{\Omega}|^{1/2} g((\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Omega} (\mathbf{x} - \boldsymbol{\mu}))$$

where g is a non-negative scalar-valued density function that is continuous and strictly increasing, and is called the *density generator* of the elliptical distribution. We denote such an elliptical distribution by $\mathcal{E}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$. For the asymptotic parameter distribution we also assume the following conditions:

(S1a) The limiting distribution $\mathbb{T}$ of the full model estimate, i.e. $a_n(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0), (a_{sn} \equiv a_n)$ is distributed as $\mathcal{E}(\mathbf{0}_p, \mathbf{V}, g)$, for some positive-definite matrix $\mathbf{V}$ and density generator

function g;

(S1b) For almost every data sequence $\mathcal{B}$, There exists a sequence of positive definite matrices $\mathbf{V}_n$ such that $\text{plim}_{n\to\infty}\mathbf{V}_n = \mathbf{V}$.

In practice we mostly deal with Gaussian limiting distributions, which naturally satisfy (S1a), while (S1b) is standard for such methods of estimation.

2.5.3 Derivation of the algorithm

We are now at a stage to present a result that forms the foundation of our fast algorithm.

Theorem 2.5.2. *Consider a depth function $D : \mathbb{R}^p \times \tilde{\mathbb{R}}^p \mapsto [0, \infty)$ satisfying properties (D1)-(D5), and an estimator $\boldsymbol{\theta}$ that satisfies (S0), (S1a) and (S1b). Then, given a (finite) sequence of nested correct models, say $\mathcal{M}_1, \ldots, \mathcal{M}_k$ where a model is nested under all the models with higher indices, we shall have*

$$e_n(\mathcal{M}_1) > \ldots > e_n(\mathcal{M}_k)$$

for large enough n.

This above theorem is still rather general in nature, considering a generic nested structure for adequate models in which the constant part of coefficient vector can take any value. To use this framework for statistical model selection, we shall elicit the following result:

Corollary 2.5.3. *Consider the subcollection of candidate models $\mathbb{M}_0 = \{\mathcal{M} : c_j = 0 \quad \forall \quad j \notin \mathcal{S}\}$. Suppose $\mathcal{M}_0 \in \mathbb{M}_0$ is an adequate model such that its associated index set $\mathcal{S}_0 = \{j : \theta_{0j} \neq 0\}$, i.e. it estimates all non-zero indices in the preferred coefficient*

vector $\boldsymbol{\theta}_0$. *Then there exists a positive integer* N *so that for all* $n_1 > N$,

$$\mathcal{M}_0 = \arg\max_{\mathcal{M}\in\mathbb{M}_0} \left[e_{n_1}(\mathcal{M})\right] \tag{2.5.6}$$

For the purpose of statistical model selection, we assume that the data is generated using a 'true' vector of parameters, and only a subset of parameters influence the outcome. Here we take our preferred vector of parameters, i.e. $\boldsymbol{\theta}_0$ as this true parameter vector. Restricting our attention to the subcollection of models in the above corollary is necessary because of the objective being covariate selection, and the second condition guarantees uniqueness of the minimal adequate model $\mathcal{M}_0$. Also notice that we can now fully specify candidate models by the index set $\mathcal{S}$, and since we perform all subsequent analysis in this restricted setup, from now on we shall refer the candidate model by $\mathcal{S}$ instead of $\mathcal{M}$. This will carry over to corresponding subscripts as well (e.g. $\boldsymbol{\theta}_s$ in place of $\boldsymbol{\theta}_m$ etc.).

At this point the total number of candidate models being considered is 2^p. However, in the e-values framework, to determine the minimal adequate model $\mathcal{S}_0$ one does not need to sift through all possible subsets or employ *ad-hoc* search strategies like forward selection/ backward deletion. We show that checking e-values at only p marginal models is sufficient for this purpose. In order to do this, we further restrict our attention to those candidate models where only a single parameter set to zero. That is, for such models $p_s = p - 1$. This collection of marginal sub-models can be studied in parallel: e.g. computations for these can be done on separate processors or computers.

The following result offers an alternate representation of the minimal adequate model using this much smaller set of models, after which the fast selection algorithm will be immediate.

Corollary 2.5.4. *Consider the models* $\mathcal{S}_{-j} = \{1, \ldots, p\}\backslash\{j\}$ *for* $j = 1, \ldots, p$. *Then*

for the same conditions and positive integer N as in corollary 2.5.3 we shall have

$$\mathcal{S}_0 = \{j : e_{n_1}(\mathcal{S}_{-j}) < e_{n_1}(\mathcal{S}_*)\} \tag{2.5.7}$$

for any positive integer $n_1 > N$.

In short, this happens because dropping an essential predictor makes the model inadequate, which has very small e-value for large enough sample size, whereas dropping a non-essential predictor increases the e-value: thus simply collecting those predictors that cause decrease in the e-value on dropping them from the model suffices for variable selection.

Thus, our fast algorithm for the evaluation of models shall consist of only 3 steps: (a) fit the full model and estimate its e-value, (b) replace each covariate by 0 and compute e-value of all such reduced models, and (c) collect covariates dropping which causes the e-value to go down. A safer version of this recipe can be to keep dropping one covariate at each step until no sub-model achieves a lower e-value. In numeric studies we conducted we did not find substantial difference between selecting covariates directly based on whether $e_n(\mathcal{S}_{-j}) < e_n(\mathcal{S}_*)$, and this backward deletion method. Also in an empirical data-analytic setup, the performance of our algorithm is dependent on several factors, like sample size, signal-to-noise ratio, the estimation model and the resampling technique used: although we later show that our method in general performs better than the state-of-the-art across multiple modelling situations that take the above into account.

2.5.4 Bootstrap implementation

A sample version of the above variable selection recipe that incorporates bootstrap to estimate the sampling distributions $[\hat{\boldsymbol{\theta}}]$, $[\hat{\boldsymbol{\theta}}_s]$ is the following:

1. Generate two independent set of bootstrap weights, of size R and R_1, and obtain the corresponding approximations to the full model sampling distribution, say $[\hat{\boldsymbol{\theta}}_r]$ and $[\hat{\boldsymbol{\theta}}_{r_1}]$;

2. For $j = 1, 2, \ldots p$, estimate the e-value of $\mathcal{S}_{-j}$ as

$$\hat{e}_n(\mathcal{S}_{-j}) = \mathbb{E}_r D(\hat{\boldsymbol{\theta}}_{r,-j}, [\hat{\boldsymbol{\theta}}_{r_1}]) \tag{2.5.8}$$

   with $\hat{\boldsymbol{\theta}}_{r,-j}$ obtained from $\hat{\boldsymbol{\theta}}_r$ by replacing the j-th coordinate with 0;

3. Estimate the set of non-zero covariates as $\hat{\mathcal{S}}_0 = \{j : \hat{e}_n(\mathcal{S}_{-j}) < \hat{e}_n(\mathcal{S}_*)\}$

To make the sample e-values appropriately mimic the population level quantities, the bootstrap method used must adhere to the guidelines in Section 2.4. Subsequently theorem 2.4.4 shall ensure asymptotic model selection consistency, i.e. $\mathbb{P}_n(\hat{\mathcal{S}}_0 = \mathcal{S}_0) \xrightarrow{P} 1$ as $n \to \infty$, with the probability $\mathbb{P}_n$ being calculated over the second resampling scheme conditional on the given data.

Example 2.5.5 (Generalized linear models (GLM))**.** In the GLM setup: $\mathbf{Y} = g^{-1}(\mathbf{X}\boldsymbol{\beta}) + \boldsymbol{\epsilon}; \boldsymbol{\epsilon} \sim N(0, \sigma^2\mathbf{I}_p)$ and g being the link function, we can obtain bootstrapped copies of $\hat{\boldsymbol{\theta}}$ using the moon bootstrap (example 2.4.1) or the scale-enhanced Gamma bootstrap (example 2.4.2). For moon bootstrap the resampling sample size m is the variance of the multinomial distribution from which the iid bootstrap weights are drawn; while in the bayesian Gamma bootstrap $\mathbb{W}_r$ follow a Gamma distribution, so that its scale parameter is the variance. To obtain asymptotic model selection consistency, an intermediate rate of this bootstrap variance $\tau_{sn}^2 \equiv \tau_n^2$ is required as per theorem 2.4.4. We achieve this by taking functions of the sample size as τ_n^2, e.g. $\tau_n^2 = n^\gamma; 0 < \gamma < 1$ or $\tau_n^2 = \log(n)$. For moon bootstrap, this means drawing larger with-replacement samples with increasing n, say of size size m_n, ensuring that $m_n \to \infty, m_n/n \to 0$ as $n \to \infty$.

Example 2.5.6 (Linear Mixed models)**.** Consider a random intercept-only model:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$$

There are m independent groups of observations with k observations in each groups, with $\mathbf{Z}_{n\times k}$ the within-group random effects design matrix. Also $\boldsymbol{\gamma}$ is a k-dimensional random effect vector ($k \leqslant n$), with $\boldsymbol{\gamma} \sim \mathcal{N}_k(\mathbf{0}_k, \Delta)$, Δ being positive definite. Here we use the generalized bootstrap scheme of Chatterjee and Bose (2005), taking equal resampling weights $w_{ri} \sim \text{Gamma}(1,1)-1$ inside a group. Given the original estimates $\hat{\boldsymbol{\beta}}, \hat{\sigma}^2, \hat{\Delta}$, and τ_n satisfying similar conditions as last example, a simple relationship exists between $\hat{\boldsymbol{\beta}}$ and its bootstrap counterpart $\hat{\boldsymbol{\beta}}_r$:

$$\hat{\boldsymbol{\beta}}_r = \hat{\boldsymbol{\beta}} + \frac{\tau_n}{\sqrt{n}}(\mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{W}_r\mathbf{X}^T\hat{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) + \mathbf{R}_{rn} \tag{2.5.9}$$

with $\mathbb{E}_r\|\mathbf{R}_{rn}\|^2 = o_P(1), \mathbf{W}_r = \text{diag}(w_{r1}\mathbf{I}_k, \ldots, w_{rm}\mathbf{I}_k)$ and $\hat{\mathbf{V}} = \hat{\sigma}^2\mathbf{I}_p + \mathbf{Z}\hat{\Delta}\mathbf{Z}^T$. This is immediate from theorem 3.2 in Chatterjee and Bose (2005). Depending on the structure of the matrix Δ, the calculation of $\hat{\boldsymbol{\beta}}_r$ repeatedly can be computation-intensive, and the above parametric procedure effectively bypasses it by approximating $\hat{\boldsymbol{\beta}}_r$ through dropping the last term in (2.5.9) above. Although a similar approach can certainly be used for GLMs as well, computationally they are much more effective here.

2.6 Simulation studies

We now present the results of two simulation studies to compare the performance of our proposed fast variable selection method using model e-values, with the model selection procedures obtained from backward deletion and all subset regression versions that aim to minimize the Akaike Information Criterion (AIC: Akaike (1970)) or

the BIC for linear model, and sparse regularization-based methods for linear mixed models. In both examples below, we assume that the expectation of the response Y is a linear function of a few covariates, and the model selection problem is the classical one of identifying the set of covariates which have a non-zero effect on $\mathbb{E}Y$.

2.6.1 Selecting covariates in linear regression

For the first simulation, we use the first $p = 10$ columns of a simulated dataset from Prof. Charles Geyer's website (`http://www.stat.umn.edu/geyer/5102/data/ex6-8.txt`) and $n = 100$ randomly chosen rows, and arrange then in a $n \times p$ covariate matrix $\mathbf{X}$. Each non-zero regression slope parameter takes the value 1, and we add independent standard Normal noise to generate the response vector, thus obtaining the framework $Y = \mathbf{X}\boldsymbol{\beta} + \epsilon$.

We generate data under different choices of the size of the minimal adequate model:b by first selecting $k \in \{2, 4, 6, 8\}$, then setting the first k coefficients of the regression slope $\boldsymbol{\beta}$ to be 1, and the rest $p - k$ slope parameters to be zero. The values of $\tau = \tau_n/\sqrt{n}$, the standard deviation of the resampling weights scaled by $\sqrt{n}$, is selected on a grid between 1 and 10 in 0.1 length intervals. We use a resampling Monte Carlo size $R = R_1 = 1000$ for use in the sample version of (2.5.8), Finally the entire exercise is repeated 1000 times independently. We report here the results on the proportion of times out of this 1000 replications of the study when the minimal adequate model is selected. This is the numeric approximation of the 'probability of selecting the true model'.

We use the backward deletion and all-subset regression search strategies while using AIC and BIC as the model selection criterion. We use the leaps-and-bound algorithm, implemented in the `R` package `leaps`, for all-subset search. We display the results of this study in Figure 2.1 for the moon bootstrap, in Figure 2.2 for the gamma bootstrap and Figure 2.3 for a wild bootstrap (Mammen, 1993) version of

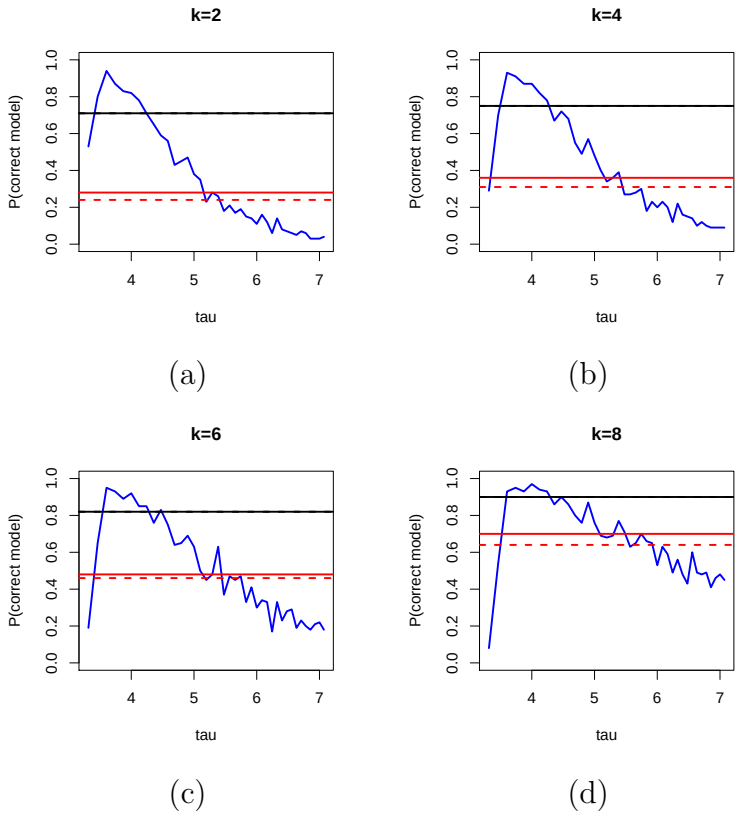


Figure 2.1: Empirical probabilities of selecting the correct model through moon bootstrap for several levels of sparsity: The e-values method- blue solid, AIC backward deletion- red dotted, AIC all subset- red solid, BIC backward deletion- black dotted, BIC all subset- black solid

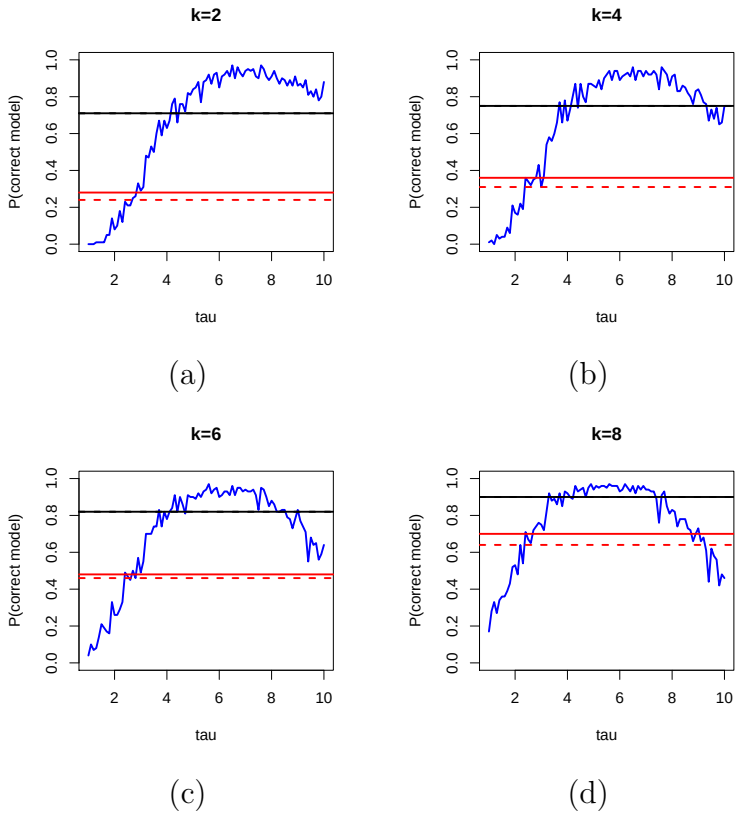


Figure 2.2: Empirical probabilities of selecting the correct model through gamma bootstrap for several levels of sparsity: The e-values method- blue solid, AIC backward deletion- red dotted, AIC all subset- red solid, BIC backward deletion- black dotted, BIC all subset- black solid

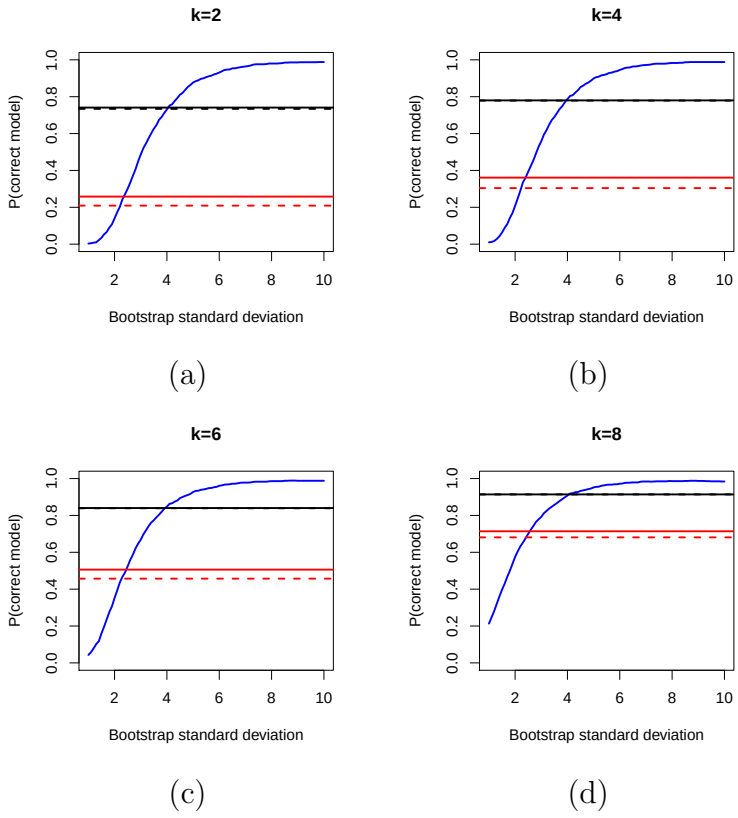


Figure 2.3: Empirical probabilities of selecting the correct model through wild bootstrap for several levels of sparsity: The e-values method- blue solid, AIC backward deletion- red dotted, AIC all subset- red solid, BIC backward deletion- black dotted, BIC all subset- black solid

the linear regression equivalent of (2.5.9) with i.i.d. $N(0,1)$ weights. In all three methods, and for all of $k \in \{2,4,6,8\}$ the proposed e-value based method performs better than AIC or BIC, as long as τ_n^2 is not too small or too large. This is entirely as expected. The parametric wild bootstrap procedure has the best performance among the three, giving almost perfect detection for even very large values of τ_n. We experimented with other choices of n, p, R_1, R_2, and it seems considering $\tau \in (4,8)$ in this problem ensures exact minimal adequate model selection with high chance, and typically better performance than BIC in this regard. As long as R and R_1 are of the order of a few hundreds or higher, the variation from the resampling Monte Carlo step seems ignorable.

2.6.2 Model selection in the presence of random effects

Here we use the repeated measures simulation setup from Peng and Lu (2012), which has 9 fixed effects and 4 random effects, with true $\boldsymbol{\beta} = (0,1,1,0,0,0,0,0,0)$ and random effect covariance matrix:

$$D = \begin{pmatrix} 9 & & & \\ 4.8 & 4 & & \\ 0.6 & 1 & 1 & \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

The error variance σ^2 is set at 1. The goal is to select the covariates of the fixed effect, thus essentially identify the covariates corresponding to the entries where $\boldsymbol{\beta}$ is non-zero. We use two scenarios for our study: one where the number of subjects considered is $m = 30$, and the number of observations per subject is $n_i = 5$, and another with 60 subjects and 10 observations per subject.

We consider $\tau \in \{1, \ldots, 15\}$ here, and use the approximation in (2.5.9) to calculate

Method	Tuning	FPR%	FNR%	Model size	FPR%	FNR%	Model size
		$n_i = 5, m = 30$			$n_i = 10, m = 60$		
e-value based	$\tau = 1$	59.9	0.0	5.61	44.3	0.0	4.43
	$\tau = 2$	33.0	0.0	3.45	15.5	0.0	2.54
	$\tau = 3$	15.9	0.0	2.59	5.2	0.0	2.17
	$\tau = 4$	8.0	0.0	2.28	2.8	0.0	2.09
	$\tau = 5$	5.2	0.0	2.18	2.0	0.0	2.06
	$\tau = 6$	2.7	0.0	2.09	0.7	0.0	2.02
	$\tau = 7$	2.2	0.0	2.07	0.3	0.0	2.01
	$\tau = 8$	1.5	0.0	2.05	0.3	0.0	2.01
	$\tau = 9$	1.0	0.0	2.03	0.3	0.0	2.01
	$\tau = 10$	0.7	0.0	2.02	0.3	0.0	2.01
	$\tau = 12$	0.7	0.0	2.02	0.0	0.0	2.00
	$\tau = 15$	0.7	0.0	2.02	0.0	0.0	2.00
Peng and Lu (2012)	BIC	21.5	9.9	2.26	1.5	1.9	2.10
	AIC	17	11.0	2.43	1.5	3.3	2.20
	GCV	20.5	6	2.30	1.5	3	2.18
	$\sqrt{\log n/n}$	21	15.6	2.67	1.5	4.1	2.26

Table 2.1: Comparison between our method and that proposed by Peng and Lu (2012) through average false positive percentage, false negative percentage and model size

the bootstrapped coefficients. We consider multiple characteristics of the model that obtains the highest e-value, including the number of parameters it involves, the proportion of times the minimal adequate model is obtained, the proportion of times a zero-valued (non-zero-valued) element of beta was identified as non-zero (zero), that is, the proportion of false positives (negatives), and so on.

In the method proposed by Peng and Lu (2012), the tuning parameter can be selected using several different criteria. We present the false positive percentage (FPR%), false negative percentage (FNR%) and model sizes corresponding to four such criteria. Our results are presented in Table 2.1. It can be seem the e-value based method handsomely outperforms the method proposed by Peng and Lu (2012), especially in smaller sample sizes, as long as $\tau \geqslant 4$.

We also compare the percentages of times the correct model was identified, and these results are presented in Table 2.2, along with the corresponding results from two other papers. The proposed e-value based procedure performs best here for $\tau \geqslant 6$ for the smaller sample setting, and for $\tau \geqslant 12$ for larger sample setting.

Method		Setting 1	Setting 2
e-value based	$\tau = 1$	3	14
	$\tau = 2$	30	60
	$\tau = 3$	61	86
	$\tau = 4$	79	92
	$\tau = 5$	87	94
	$\tau = 6$	93	98
	$\tau = 7$	94	99
	$\tau = 8$	96	99
	$\tau = 9$	97	99
	$\tau = 10$	98	99
	$\tau = 12$	98	100
	$\tau = 15$	98	100
Bondell et al. (2010)		73	83
Peng and Lu (2012)		49	86
Fan and Li (2012)		90	100

Table 2.2: Comparison of our method and three sparsity-based methods of mixed effect model selection through accuracy of selecting correct fixed effects

2.7 Discussion and conclusion

In the above sections we present an expansive framework and principle, where the definition of a statistical model is very broad, and estimation procedures and resampling algorithms very general. In such a scenario, we propose a scheme of simultaneous model selection and resampling-based inference, using the newly defined e-value. An extremely fast algorithm, based on using data depth as evaluation function, obtains consistent true model selection through fitting a single model. Simulation results show that the procedure performs better than traditional methods in two illustrative examples. Last but not least we provide a number of theoretical results for characterization of our method in both pupulation and sample setting.

While the above framework is extremely open-ended, multiple details require cautious approach and more detailed studies. The choice of the resampling algorithm, and the method of choosing the tuning parameter τ_n associated with it should be subject to further scrutiny. Our results suggest excellent *asymptotic* properties that seem to be borne out in our simulation experiments, but finite-sample performance of our procedure needs further study. We have remarked earlier that uniform convergence, local asymptotics and detailed asymptotic studies are needed to understand

the workings of our proposal more thoroughly. The current framework includes *dimension asymptotics* where the parameter dimensions are allowed to grow with the sample size, but we do not include extremely high-dimensional parameters in our study. The sensitivity of the results to the choice of the evaluation maps, and the way $E_n(\mathbf{y}, [\mathbf{Y}])$ is summarized to obtain the e-value deserve further attention. A further, perhaps philosophical, issue to look into is the sensitivity of the results to the choice of the preferred model. While in practice this may not matter much, the choice of the preferred model reflects a choice of paradigms and scientific principles.

In recent times, there is a growing concern about statistical inference after the implementation of a model selection step. Discussions and several interesting results relating to this matter may be found in Yang (2005); Leeb and Potscher (2005); Chang et al. (2014); Tibshirani et al. (2015, 2016) and several references therein. The general principle discussed in this chapter advocates obtaining consistent resampling-based distributions of the estimators of *all* parameters from *all* candidate models. Thus in our framework, statistical inference is not the usual two-step procedure where the first step involves selection of a model, and the second step of actual inference somehow adjusts for the uncertainties of the first step. Our proposal is one of a *joint selection and inference* procedure, where the consistent resampling-based approximations of the sampling distributions of any collection of models are simultaneously used for inference, as well as establishing an e-value of a model, which may be used to preferentially treat a subset of models.

A study of the research on post model selection inference reveals that some of the issues there may be addressed using *uniform convergence* and related ideas. Based on the concepts and tools presented in this chapter, we have the ingredients at hand to conduct such studies. Additionally, current studies essentially conclude that the goal of identifying the true data-generating model with probability tending to one, under the assumption that it is already one of the candidate models. This is not

immediately compatible with several other goals of optimal statistical inference. Note that the problem of identification of one of the candidate models as a 'true model' has not been a goal of this chapter, although our theoretical and numeric results establish that such identification is achieved easily if such a situation were to arise. We also note that traditional 'true statistical model' considered in some related literature typically do not consider the domain scientific knowledge or background, and are solely based on a limited version of parsimony. Keeping this in mind, we plan to investigate the application of e-values to achieve multiple targets of optimal inference.

2.8 Proofs

Proof of theorem 2.3.1. Part 1 follows directly from assumption (E3).

Part 2. Assuming now that $\mathcal{M}_n$ is an adequate model, we again use the location invariance property of E_n:

$$E_n(\hat{\mathbf{G}}_{mn}, [\hat{\mathbf{G}}_n]) = E_n\left(\hat{\mathbf{G}}_{mn} - \mathbf{G}_{*n}, \left[\hat{\mathbf{G}}_{*n} - \mathbf{G}_{*n}\right]\right) \tag{2.8.1}$$

and decompose the first argument

$$\hat{\mathbf{G}}_{mn} - \mathbf{G}_{*n} = (\hat{\mathbf{G}}_{mn} - \hat{\mathbf{G}}_{*n}) + (\hat{\mathbf{G}}_{*n} - \mathbf{G}_{*n}) \tag{2.8.2}$$

Now we have, for any $\mathcal{M}_n$,

$$\hat{\boldsymbol{\theta}}_{mn} \equiv \hat{\boldsymbol{\theta}} = \boldsymbol{\theta} + a_n^{-1}\mathbf{T}_n \equiv \boldsymbol{\theta}_{mn} + a_{sn}^{-1}\mathbf{T}_{mn}$$

where $\mathbf{T}_{mn}$ is distributed as $\mathbb{T}_{sn}$ in $\mathcal{S}_n$ indices and fixed to 0 in other indices. In terms

of these, we can write the j-th element of $\mathbf{G}_{mn}(.) \equiv \mathbf{G}(.)$ as

$$G_j(\hat{\boldsymbol{\theta}}) = G_j(\boldsymbol{\theta}) + a_n^{-1}\mathbf{G}_{1j}^T(\boldsymbol{\theta})\mathbf{T}_n + 2a_n^{-2}\mathbf{T}_n^T\mathbf{R}_j(\hat{\boldsymbol{\theta}}, \mathbf{T}_n)\mathbf{T}_n$$

Our technical conditions are sufficient to ensure that for any $\mathbf{c} \in \mathbb{R}^{d_n}$ with $\|\mathbf{c}\| = 1$

$$\mathbb{E}\left(\sum_{j=1}^{d_n} c_j \mathbf{T}_n^T \mathbf{R}_j(\hat{\boldsymbol{\theta}}, \mathbf{T}_n)\mathbf{T}_n\right)^2 = O(a_n d_n)$$

we omit the details of the algebra here.

Thus we have that $a_n(\hat{\mathbf{G}} - \mathbf{G}) = \mathbf{G}_1^T\mathbf{T}_n + \mathbf{R}_n$, with $\mathbb{E}\|\mathbf{R}_n^2\| = o(1)$. Coming back to the first summand of the right-hand side in (2.8.2) we get

$$\hat{\mathbf{G}}_{mn} - \hat{\mathbf{G}}_{*n} = \mathbf{G}_{mn} - \mathbf{G}_{*n} + O_P(\min\{a_{sn}, a_{*n}\}^{-1}) \tag{2.8.3}$$

Since $\mathcal{M}_n$ is an adequate model, $\mathbf{G}_{mn} - \mathbf{G}_{0n} = o(n)$. Also $\mathbf{G}_{*n} - \mathbf{G}_{0n} = o(n)$. Thus, substituting the above right-hand side in (2.8.2) we get

$$\begin{aligned}\left|E_n\left(\hat{\mathbf{G}}_{mn} - \mathbf{G}_{*n}, \left[\hat{\mathbf{G}}_{*n} - \mathbf{G}_{*n}\right]\right) - E_n\left(\hat{\mathbf{G}}_{*n} - \mathbf{G}_{*n}, \left[\hat{\mathbf{G}}_{*n} - \mathbf{G}_{*n}\right]\right)\right| \\ = o_P(\min\{a_{sn}, a_{*n}, n\})\end{aligned} \tag{2.8.4}$$

from of Lipschitz continuity of E_n. Adding back $\mathbf{G}_{*n}$ everywhere and applying (E1) again,

$$|E_n(\hat{\mathbf{G}}_{mn}, [\hat{\mathbf{G}}_{*n}]) - E_n(\hat{\mathbf{G}}_{*n}, [\hat{\mathbf{G}}_{*n}])| = o_P(\min\{a_{sn}, a_{*n}, n\}) \tag{2.8.5}$$

the proof of part 2 is immediate now.

Part 3. Since the evaluation map E_n is invariant under location and scale trans-

formations, we have

$$E_n(\hat{\mathbf{G}}_{mn}, [\hat{\mathbf{G}}_n]) = E_n\left(a_{*n}(\hat{\mathbf{G}}_{mn} - \mathbf{G}_{*n}), \left[a_{*n}(\hat{\mathbf{G}}_{*n} - \mathbf{G}_{*n})\right]\right) \tag{2.8.6}$$

Decomposing the first argument,

$$a_{*n}(\hat{\mathbf{G}}_{mn} - \mathbf{G}_{*n}) = \frac{a_{*n}}{a_{sn}}.a_{sn}(\hat{\mathbf{G}}_{mn} - \mathbf{G}_{mn}) + a_{*n}(\mathbf{G}_{mn} - \mathbf{G}_{0n}) + a_{*n}(\mathbf{G}_{0n} - \mathbf{G}_{*n}) \tag{2.8.7}$$

Since $\mathcal{M}_n$ is inadequate, given $\delta > 0$ there exists a subsequence indexed by $\{k_n\}$ such that $\|\mathbf{G}_{mk_n} - \mathbf{G}_{0k_n}\| > \delta$. Since $a_{*n} \uparrow \infty$, this implies $a_{*n}\|\mathbf{G}_{mn} - \mathbf{G}_{0n}\| \to \infty$. Finally $a_{*n}(\hat{\mathbf{G}}_{mn} - \mathbf{G}_{mn}) = O_P(1)$ using similar arguments as in proof of part 2 above, $a_{sn} \asymp a_{*n}$, and norm of the third part goes to 0 by part b of assumption (S1). We now get the needed by assumption (E4). □

Proof of Theorem 2.4.3. We consider a generic point $\boldsymbol{\theta} = \boldsymbol{\theta}_{sn} + \mathbf{A}_{sn}^{-1}\mathbf{t}$. From the Taylor series expansion, we have

$$\Psi_{0sni(a)}(\boldsymbol{\theta}) = \Psi_{0sni(a)}(\boldsymbol{\theta}_{sn}) + \Psi_{1sni(a)}(\boldsymbol{\theta}_{sn})\mathbf{A}_{sn}^{-1}\mathbf{t} + 2^{-1}\mathbf{t}^T\mathbf{A}_{sn}^{-1T}\Psi_{2sni(a)}(\tilde{\boldsymbol{\theta}}_{sn})\mathbf{A}_{sn}^{-1}\mathbf{t} \tag{2.8.8}$$

for $a = 1, \ldots, p_{sn}$, and $\tilde{\boldsymbol{\theta}}_{sn} = \boldsymbol{\theta}_{sn} + c\mathbf{A}_{sn}^{-1}\mathbf{t}$ for some $c \in (0, 1)$.

Recall our convention that for any function $\mathbf{h}(\boldsymbol{\theta})$ evaluated at the true parameter value $\boldsymbol{\theta}_{sn}$, we use the notation $\mathbf{h} \equiv \mathbf{h}(\boldsymbol{\theta}_{sn})$. Also define the p_{sn} dimensional vector $\mathbf{R}_{sn}(\tilde{\boldsymbol{\theta}}_{sn}, \mathbf{t})$ whose a-th element is given by

$$\mathbf{R}_{sn(a)}(\tilde{\boldsymbol{\theta}}_{sn}, \mathbf{t}) = \mathbf{t}^T\mathbf{A}_{sn}^{-1T}\sum_{i=1}^{k_n}\Psi_{2sni(a)}(\tilde{\boldsymbol{\theta}}_{sn})\mathbf{A}_{sn}^{-1}\mathbf{t} \tag{2.8.9}$$

Thus we have

$$
\begin{aligned}
&p_{sn}^{-1/2}\mathbf{A}_{sn}^{-1}\sum_{i=1}^{k_n}\Psi_{0sni}(\boldsymbol{\theta}_{sn}+\mathbf{A}_{sn}^{-1}\mathbf{t}) \\
&= p_{sn}^{-1/2}\mathbf{A}_{sn}^{-1}\sum_{i=1}^{k_n}\Psi_{0sni} + p_{sn}^{-1/2}\mathbf{A}_{sn}^{-1}\sum_{i=1}^{k_n}\Psi_{1sni}\mathbf{A}_{sn}^{-1}\mathbf{t} + 2^{-1}p_{sn}^{-1/2}\mathbf{A}_{sn}^{-1}\mathbf{R}_{sn}(\tilde{\boldsymbol{\theta}}_{sn},\mathbf{t}) \\
&= p_{sn}^{-1/2}\mathbf{A}_{sn}^{-1}\sum_{i=1}^{k_n}\Psi_{0sni} + p_{sn}^{-1/2}\mathbf{A}_{sn}^{-1}\boldsymbol{\Gamma}_{1sn}\mathbf{A}_{sn}^{-1}\mathbf{t} + p_{sn}^{-1/2}\mathbf{A}_{sn}^{-1}\big(\sum_{i=1}^{k_n}\Psi_{1sni}-\boldsymbol{\Gamma}_{1sn}\big)\mathbf{A}_{sn}^{-1}\mathbf{t} \\
&\quad + 2^{-1}p_{sn}^{-1/2}\mathbf{A}_{sn}^{-1}\mathbf{R}_{sn}(\tilde{\boldsymbol{\theta}}_{sn},\mathbf{t}) \qquad (2.8.10)
\end{aligned}
$$

Fix $\epsilon > 0$. We first show that there exists a $C_0 > 0$ such that

$$
\mathbb{P}\Big[\,\|p_{sn}^{-1/2}\mathbf{A}_{sn}^{-1}\sum_{i=1}^{k_n}\Psi_{0sni}\| > C_0\Big] < \epsilon/2. \qquad (2.8.11)
$$

For this, we compute

$$
\begin{aligned}
p_{sn}^{-1}\mathbb{E}\|\mathbf{A}_{sn}^{-1}\sum_{i=1}^{k_n}\Psi_{0sni}\|^2 &= p_{sn}^{-1}\mathbb{E}\sum_{i,j=1}^{k_n}\Psi_{0sni}^T\mathbf{A}_{sn}^{-1T}\mathbf{A}_{sn}^{-1}\Psi_{0snj} \\
&= p_{sn}^{-1}\,\mathrm{Tr}\left[\mathbf{A}_{sn}^{-1T}\mathbf{A}_{sn}^{-1}\right]\mathbb{E}\sum_{i=1}^{k_n}\Psi_{0sni}\Psi_{0sni}^T \\
&= p_{sn}^{-1}\,\mathrm{Tr}\left[\mathbf{A}_{sn}^{-1T}\mathbf{A}_{sn}^{-1}\boldsymbol{\Gamma}_{0sn}\right] \\
&= O(1) \qquad (2.8.12)
\end{aligned}
$$

from assumption (2.4.8).

Now define

$$
\mathbf{S}_{sn}(\mathbf{t}) = p_{sn}^{-1/2}\mathbf{A}_{sn}^{-1}\big(\sum_{i=1}^{k_n}\Psi_{0sni}(\boldsymbol{\theta}_{sn}+\mathbf{A}_{sn}^{-1}\mathbf{t}) - \sum_{i=1}^{k_n}\Psi_{0sni}\big) - p_{sn}^{-1/2}\boldsymbol{\Gamma}_{1sn}^{-1}\boldsymbol{\Gamma}_{0sn}\mathbf{t} \quad (2.8.13)
$$

We next show that for any $C > 0$, for all sufficiently large n, we have

$$\mathbb{E}\Big[\sup_{\|\mathbf{t}\|\leqslant C}\|\mathbf{S}_{sn}(\mathbf{t})\|\Big]^2 = o(1) \tag{2.8.14}$$

This follows from (2.4.9) and (2.4.11).

Note that

$$\mathbf{S}_{sn}(\mathbf{t}) = p_{sn}^{-1/2}\mathbf{A}_{sn}^{-1}\big(\sum_{i=1}^{k_n}\Psi_{1sni} - \boldsymbol{\Gamma}_{1sn}\big)\mathbf{A}_{sn}^{-1}\mathbf{t} + 2^{-1}p_{sn}^{-1/2}\mathbf{A}_{sn}^{-1}\mathbf{R}_{sn}(\tilde{\boldsymbol{\theta}}_{sn},\mathbf{t}) \tag{2.8.15}$$

Thus

$$\begin{aligned}&\sup_{\|\mathbf{t}\|\leqslant C}\|\mathbf{S}_{sn}(\mathbf{t})\| \leqslant \\ &p_{sn}^{-1/2}\sup_{\|\mathbf{t}\|\leqslant C}\|\mathbf{A}_{sn}^{-1}\big(\sum_{i=1}^{k_n}\Psi_{1sni} - \boldsymbol{\Gamma}_{1sn}\big)\mathbf{A}_{sn}^{-1}\mathbf{t}\| + 2^{-1}p_{sn}^{-1/2}\sup_{\|\mathbf{t}\|\leqslant C}\|\mathbf{A}_{sn}^{-1}\mathbf{R}_{sn}(\tilde{\boldsymbol{\theta}}_{sn},\mathbf{t})\|\end{aligned} \tag{2.8.16}$$

We consider each of these terms separately.

For any matrix $\mathbf{M} \in \mathbb{R}^{p\times p}$, we have

$$\begin{aligned}\sup_{|\mathbf{t}|\leqslant C}\|\mathbf{M}\mathbf{t}\| &= \sup_{\|\mathbf{t}\|\leqslant C}\Big[\sum_{i=1}^{p}\big(\sum_{j=1}^{p}m_{ij}t_j\big)^2\Big]^{1/2} \\ &\leqslant \sup_{\|\mathbf{t}\|\leqslant C}\Big[\sum_{i=1}^{p}\sum_{j=1}^{p}m_{ij}^2\sum_{j=1}^{p}t_j^2\Big]^{1/2} \\ &= \|\mathbf{M}\|_F\sup_{\|\mathbf{t}\|\leqslant C}\|\mathbf{t}\| \\ &= C\|\mathbf{M}\|_F\end{aligned} \tag{2.8.17}$$

Using $\mathbf{M} = \mathbf{A}_{sn}^{-1}\big(\sum_{i=1}^{k_n}\Psi_{1sni} - \boldsymbol{\Gamma}_{1sn}\big)\mathbf{A}_{sn}^{-1}$ and (2.4.9), we get one part of the result.

For the other term, we similarly have

$$\Big[\sup_{\|\mathbf{t}\|\leqslant C}\|p_{sn}^{-1/2}\mathbf{A}_{sn}^{-1}\mathbf{R}_{sn}(\tilde{\boldsymbol{\theta}}_{sn},\mathbf{t})\|\Big]^2 = p_{sn}^{-1}\sup_{\|\mathbf{t}\|\leqslant C}\Big[\|\mathbf{A}_{sn}^{-1}\mathbf{R}_{sn}(\tilde{\boldsymbol{\theta}}_{sn},\mathbf{t})\|\Big]^2$$
$$\leqslant p_{sn}^{-1}\lambda_{max}\big(\mathbf{A}_{sn}^{-1T}\mathbf{A}_{sn}^{-1}\big)\sup_{\|\mathbf{t}\|\leqslant C}\|\mathbf{R}_{sn}(\tilde{\boldsymbol{\theta}}_{sn},\mathbf{t})\|^2$$
$$\leqslant p_{sn}^{-1}\lambda_{max}\big(\mathbf{A}_{sn}^{-1}\mathbf{A}_{sn}^{-1T}\big)\sup_{\|\mathbf{t}\|\leqslant C}\|\mathbf{R}_{sn}(\tilde{\boldsymbol{\theta}}_{sn},\mathbf{t})\|^2$$
$$\leqslant p_{sn}^{-1}a_{sn}^{-2}\sup_{\|\mathbf{t}\|\leqslant C}\|\mathbf{R}_{sn}(\tilde{\boldsymbol{\theta}}_{sn},\mathbf{t})\|^2 \tag{2.8.18}$$

Note that

$$\big(\sup_{\|\mathbf{t}\|\leqslant C}\|\mathbf{R}_{sn}(\tilde{\boldsymbol{\theta}}_{sn},\mathbf{t})\|\big)^2 = \sup_{\|\mathbf{t}\|\leqslant C}\|\mathbf{R}_{sn}(\tilde{\boldsymbol{\theta}}_{sn},\mathbf{t})\|^2 \tag{2.8.19}$$

Now

$$\|\mathbf{R}_{sn}(\tilde{\boldsymbol{\theta}}_{sn},\mathbf{t})\|^2 = \sum_{a=1}^{p_{sn}}\big(\mathbf{R}_{sn(a)}(\tilde{\boldsymbol{\theta}}_{sn},\mathbf{t})\big)^2$$
$$= \sum_{a=1}^{p_{sn}}\big(\mathbf{t}^T\mathbf{A}_{sn}^{-1T}\sum_{i=1}^{k_n}\Psi_{2sni(a)}(\tilde{\boldsymbol{\theta}}_{sn})\mathbf{A}_{sn}^{-1}\mathbf{t}\big)^2$$
$$= \sum_{a=1}^{p_{sn}}\sum_{i,j=1}^{k_n}\mathbf{t}^T\mathbf{A}_{sn}^{-1T}\Psi_{2sni(a)}(\tilde{\boldsymbol{\theta}}_{sn})\mathbf{A}_{sn}^{-1}\mathbf{t}.\mathbf{t}^T\mathbf{A}_{sn}^{-1T}\Psi_{2snj(a)}(\tilde{\boldsymbol{\theta}}_{sn})\mathbf{A}_{sn}^{-1}\mathbf{t} \tag{2.8.20}$$

Based on this, we have

$$\begin{aligned}
\sup_{\|\mathbf{t}\|\leqslant C}\|\mathbf{R}_{sn}(\tilde{\boldsymbol{\theta}}_{sn},\mathbf{t})\|^2 &= \sup_{\|\mathbf{t}\|\leqslant C}\sum_{a=1}^{p_{sn}}\sum_{i,j=1}^{k_n}\mathbf{t}^T\mathbf{A}_{sn}^{-1T}\Psi_{2sni(a)}(\tilde{\boldsymbol{\theta}}_{sn})\mathbf{A}_{sn}^{-1}\mathbf{t}.\mathbf{t}^T\mathbf{A}_{sn}^{-1T}\Psi_{2snj(a)}(\tilde{\boldsymbol{\theta}}_{sn})\mathbf{A}_{sn}^{-1}\mathbf{t}\\
&\leqslant \sup_{\|\mathbf{t}\|\leqslant C}\sum_{a=1}^{p_{sn}}\sum_{i,j=1}^{k_n}\mathbf{t}^T\mathbf{A}_{sn}^{-1T}\mathbf{M}_{2sni(a)}\mathbf{A}_{sn}^{-1}\mathbf{t}.\mathbf{t}^T\mathbf{A}_{sn}^{-1T}\mathbf{M}_{2snj(a)}\mathbf{A}_{sn}^{-1}\mathbf{t}\\
&\leqslant \sup_{\|\mathbf{t}\|\leqslant C}\|\mathbf{A}_{sn}^{-1}\mathbf{t}\|^4\sum_{a=1}^{p_{sn}}\Big(\sum_{i=1}^{k_n}\lambda_{max}\big(\mathbf{M}_{2sni(a)}\big)\Big)^2\\
&\leqslant C^4 n\lambda_{max}^2\big(\mathbf{A}_{sn}^{-1T}\mathbf{A}_{sn}^{-1}\big)\sum_{a=1}^{p_{sn}}\sum_{i=1}^{k_n}\lambda_{max}^2\big(\mathbf{M}_{2sni(a)}\big) \qquad (2.8.21)
\end{aligned}$$

Putting all these together, we have

$$\begin{aligned}
\mathbb{E}\Big[\sup_{\|\mathbf{t}\|\leqslant C}\|p_{sn}^{-1/2}\mathbf{A}_{sn}^{-1}\mathbf{R}_{sn}(\tilde{\boldsymbol{\theta}}_{sn},\mathbf{t})\|\Big]^2 &= p_{sn}^{-1}\mathbb{E}\Big[\sup_{\|\mathbf{t}\|\leqslant C}\mathbf{A}_{sn}^{-1}\mathbf{R}_{sn}(\tilde{\boldsymbol{\theta}}_{sn},\mathbf{t})\Big]^2\\
&\leqslant p_{sn}^{-1}\mathbf{A}_{sn}^{-2}\mathbb{E}\Big[\sup_{\|\mathbf{t}\|\leqslant C}\|\mathbf{R}_{sn}(\tilde{\boldsymbol{\theta}}_{sn},\mathbf{t})\|\Big]^2\\
&= O\big(p_{sn}^{-1}a_{sn}^{-2}\big)\mathbb{E}\Big[\sup_{\|\mathbf{t}\|\leqslant C}\|\mathbf{R}_{sn}(\tilde{\boldsymbol{\theta}}_{sn},\mathbf{t})\|\Big]^2\\
&= O\big(p_{sn}^{-1}na_{sn}^{-6}\big)\sum_{a=1}^{p_{sn}}\sum_{i=1}^{k_n}\mathbb{E}\lambda_{max}^2\big(\mathbf{M}_{2sni(a)}\big)\\
&= o(1) \qquad (2.8.22)
\end{aligned}$$

using (2.4.11).

Now define

$$\mathbf{S}_{sn}(\mathbf{t}) = p_{sn}^{-1/2}\mathbf{A}_{sn}^{-1}\big(\sum_{i=1}^{k_n}\Psi_{0sni}(\boldsymbol{\theta}_{sn}+\mathbf{A}_{sn}^{-1}\mathbf{t})-\sum_{i=1}^{k_n}\Psi_{0sni}\big)-p_{sn}^{-1/2}\boldsymbol{\Gamma}_{1sn}^{-1}\boldsymbol{\Gamma}_{0sn}\mathbf{t} \quad (2.8.23)$$

hence

$$p_{sn}^{-1/2}\mathbf{A}_{sn}^{-1}\sum_{i=1}^{k_n}\Psi_{0sni}(\boldsymbol{\theta}_{sn}+p_{sn}^{1/2}\mathbf{A}_{sn}^{-1}\mathbf{t})=\mathbf{S}_{sn}(\mathbf{t})+p_{sn}^{-1/2}\mathbf{A}_{sn}^{-1}\sum_{i=1}^{k_n}\Psi_{0sni}+\mathbf{A}_{sn}^{-1}\boldsymbol{\Gamma}_{1sn}\mathbf{A}_{sn}^{-1}\mathbf{t} \tag{2.8.24}$$

and thus

$$\begin{aligned}
&\inf_{\|\mathbf{t}\|=C}\Big\{p_{sn}^{-1/2}\mathbf{t}^T\boldsymbol{\Gamma}_{1sn}\mathbf{A}_{sn}^{-1}\sum_{i=1}^{k_n}\Psi_{0sni}(\boldsymbol{\theta}_{sn}+p_{sn}^{1/2}\mathbf{A}_{sn}^{-1}\mathbf{t})\Big\}\\
&=\inf_{\|\mathbf{t}\|=C}\Big\{\mathbf{t}^T\boldsymbol{\Gamma}_{1sn}\mathbf{S}_{sn}(\mathbf{t})+p_{sn}^{-1/2}\mathbf{t}^T\boldsymbol{\Gamma}_{1sn}\mathbf{A}_{sn}^{-1}\sum_{i=1}^{k_n}\Psi_{0sni}+\mathbf{t}^T\boldsymbol{\Gamma}_{1sn}\mathbf{A}_{sn}^{-1}\boldsymbol{\Gamma}_{1sn}\mathbf{A}_{sn}^{-1}\mathbf{t}\Big\}\\
&\geqslant\inf_{\|\mathbf{t}\|=C}\mathbf{t}^T\boldsymbol{\Gamma}_{1sn}\mathbf{S}_{sn}(\mathbf{t})+p_{sn}^{-1/2}\inf_{\|\mathbf{t}\|=C}\mathbf{t}^T\boldsymbol{\Gamma}_{1sn}\mathbf{A}_{sn}^{-1}\sum_{i=1}^{k_n}\Psi_{0sni}+\inf_{\|\mathbf{t}\|=C}\mathbf{t}^T\boldsymbol{\Gamma}_{1sn}\mathbf{A}_{sn}^{-1}\boldsymbol{\Gamma}_{1sn}\mathbf{A}_{sn}^{-1}\mathbf{t}\\
&\geqslant-C\delta_{1s}\sup_{\|\mathbf{t}\|=C}\|\mathbf{S}_{sn}(\mathbf{t})\|-C\delta_{1s}p_{sn}^{-1/2}\|\mathbf{A}_{sn}^{-1}\sum_{i=1}^{k_n}\Psi_{0sni}\|+C^2\delta_{0s}
\end{aligned} \tag{2.8.25}$$

The last step above utilizes facts like $\mathbf{a}^T\mathbf{b}\geqslant-\|a\|\|b\|$.

Consequently, defining $C_1=C\delta_{0s}/\delta_{1s}$, we have

$$\begin{aligned}
&\mathbb{P}\Big[\inf_{\|\mathbf{t}\|=C}\Big\{\mathbf{t}^T\boldsymbol{\Gamma}_{1sn}\mathbf{A}_{sn}^{-1}\sum_{i=1}^{k_n}\Psi_{0sni}(\boldsymbol{\theta}_{sn}+p_{sn}^{1/2}\mathbf{A}_{sn}^{-1}\mathbf{t})\Big\}<0\Big]\\
\leqslant&\mathbb{P}\Big[\sup_{\|\mathbf{t}\|=C}\|\mathbf{S}_{sn}(\mathbf{t})\|+\|\mathbf{A}_{sn}^{-1}\sum_{i=1}^{k_n}\Psi_{0sni}\|>C_1\Big]\\
\leqslant&\mathbb{P}\Big[\sup_{\|\mathbf{t}\|=C}\|\mathbf{S}_{sn}(\mathbf{t})\|>C_1/2\Big]+\mathbb{P}\Big[\|\mathbf{A}_{sn}^{-1}\sum_{i=1}^{k_n}\Psi_{0sni}\|>C_1/2\Big]\\
<&\,\epsilon
\end{aligned} \tag{2.8.26}$$

for all sufficiently large n, using (2.8.11) and (2.8.14).

This implies that with a probability greater than $1-\epsilon$ there is a root $\mathbf{T}_{sn}$ of the equations $\sum_{i=1}^{k_n}\Psi_{0sni}(\boldsymbol{\theta}_{sn}+\mathbf{A}_{sn}^{-1}\mathbf{t})$ in the ball $\{\|\mathbf{t}\|<C\}$, for some $C>0$ and

all sufficiently large n. Defining $\hat{\boldsymbol{\theta}}_{sn} = \boldsymbol{\theta}_{sn} + \mathbf{A}_{sn}^{-1}\mathbf{T}_{sn}$, we obtain the desired result. Issues like dependence on ϵ and other technical details are handled using standard arguments, see Chatterjee and Bose (2005) for related arguments.

Since we have

$$\sup_{\|\mathbf{t}\|<C} \|\mathbf{S}_{sn}(\mathbf{t})\| = o_P(1) \tag{2.8.27}$$

and $\mathbf{T}_{sn}$ lies in the set $\|\mathbf{t}\| < C$, define $-\mathbf{R}_{sn} = \mathbf{S}_{sn}(\mathbf{T}_{sn}) = o_P(1)$. We consequently have

$$\begin{aligned}
-\mathbf{R}_{sn} &= \mathbf{S}_{sn}(\mathbf{T}_{sn}) \\
&\quad p_{sn}^{-1/2}\mathbf{A}_{sn}^{-1}\big(\sum_{i=1}^{k_n}\Psi_{0sni}(\boldsymbol{\theta}_{sn} + \mathbf{A}_{sn}^{-1}\mathbf{T}_{sn}) - \sum_{i=1}^{k_n}\Psi_{0sni}\big) - p_{sn}^{-1/2}\boldsymbol{\Gamma}_{1sn}^{-1}\boldsymbol{\Gamma}_{0sn}\mathbf{T}_{sn} \\
&\quad p_{sn}^{-1/2}\mathbf{A}_{sn}^{-1}\sum_{i=1}^{k_n}\Psi_{0sni} - p_{sn}^{-1/2}\boldsymbol{\Gamma}_{1sn}^{-1}\mathbf{T}_{sn}
\end{aligned} \tag{2.8.28}$$

Thus,

$$\begin{aligned}
\mathbf{T}_{sn} &= -\boldsymbol{\Gamma}_{0sn}^{-1}\boldsymbol{\Gamma}_{1sn}\mathbf{A}_{sn}^{-1}\sum_{i=1}^{k_n}\Psi_{0sni} + p^{1/2}\boldsymbol{\Gamma}_{0sn}^{-1}\boldsymbol{\Gamma}_{1sn}\mathbf{R}_{sn} \\
&= -\boldsymbol{\Gamma}_{0sn}^{-1/2}\sum_{i=1}^{k_n}\Psi_{0sni} + p^{1/2}\boldsymbol{\Gamma}_{0sn}^{-1}\boldsymbol{\Gamma}_{1sn}\mathbf{R}_{sn}
\end{aligned} \tag{2.8.29}$$

Note that our conditions imply that for any $\mathbf{c}$ with $\|\mathbf{c}\| = 1$, we have that $\mathbf{c}^T\mathbf{T}_{sn}$ has two terms, where $\mathbb{V}\big(-\mathbf{c}^T\boldsymbol{\Gamma}_{0sn}^{-1/2}\sum_{i=1}^{k_n}\Psi_{0sni}\big) = 1$ and

$$\mathbb{E}\big[p^{1/2}\mathbf{c}^T\boldsymbol{\Gamma}_{0sn}^{-1}\boldsymbol{\Gamma}_{1sn}\mathbf{R}_{sn}\big]^2 = O(1) \tag{2.8.30}$$

using (2.4.8). Using (2.4.12) we also have that for any $\mathbf{c}$ with $\|\mathbf{c}\| = 1$, $\mathbf{c}^T\mathbf{T}_{sn} \rightsquigarrow N(0,1)$.

Define

$$\hat{\mathbf{A}}_{sn} = \mu_{sn}\tau_{sn}^{-1}\hat{\mathbf{\Gamma}}_{0sn}^{1/2}\hat{\mathbf{\Gamma}}_{1sn}^{-1}. \tag{2.8.31}$$

We now follow steps that are very similar to the above, but for the resampling procedure as implemented in (2.4.13) to obtain the result. We omit the details.

□

Proof of theorem 2.4.4. Part 1. Taking a similar approach as in the proof of theorem 2.3.1, we get

$$\begin{aligned} G_j(\hat{\boldsymbol{\theta}}_r) &= G_j(\hat{\boldsymbol{\theta}}) + b_n^{-1}\mathbf{G}_{1j}^T(\hat{\boldsymbol{\theta}})\mathbf{T}_{rn} + 2b_n^{-2}\mathbf{T}_n^T\mathbf{R}_j(\hat{\boldsymbol{\theta}}, \mathbf{T}_{rn})\mathbf{T}_{rn}, \\ &= G_j(\hat{\boldsymbol{\theta}}) + b_n^{-1}\mathbf{G}_{1j}^T(\boldsymbol{\theta})\mathbf{T}_{rn} + b_n^{-1}(\hat{\mathbf{G}}_{1j} - \mathbf{G}_{1j})^T\mathbf{T}_{rn} + 2b_n^{-2}\mathbf{T}_{rn}^T\mathbf{R}_j(\hat{\boldsymbol{\theta}}, \mathbf{T}_{rn})\mathbf{T}_n, \\ &= G_j(\hat{\boldsymbol{\theta}}) + b_n^{-1}\mathbf{G}_{1j}^T(\boldsymbol{\theta})\mathbf{T}_{rn} + b_n^{-1}\mathbf{R}_{rnj1} + b_n^{-2}\mathbf{R}_{rnj2} \end{aligned}$$

Our technical conditions are sufficient to ensure that for any $\mathbf{c} \in \mathbb{R}^{d_n}$ with $\|\mathbf{c}\| = 1$

$$\mathbb{E}_r\left(\sum_{j=1}^{d_n} c_j\mathbf{R}_{rnj1}\right)^2 = o_P(b_n^{-1}d_n); \quad \mathbb{E}_r\left(\sum_{j=1}^{d_n} c_j\mathbf{R}_{rnj2}\right)^2 = O_P(b_n d_n),$$

we omit the details of the algebra here. Thus we get $b_n(\hat{\mathbf{G}}_r - \hat{\mathbf{G}}) = \mathbf{G}_1^T\mathbf{T}_{rn} + \mathbf{R}_{rn}$ with $\mathbb{E}_r\|\mathbf{R}_{rn}\|^2 = o_P(1)$. Hence

$$\hat{\mathbf{G}}_{rmn} - \hat{\mathbf{G}}_{r*n} = \hat{\mathbf{G}}_{mn} - \hat{\mathbf{G}}_{*n} + O_{P_n}(\min\{b_{sn}, b_{*n}\}^{-1}) + o_P(1) \tag{2.8.32}$$

where $s_n = o_{P_n}(t_n)$ means $s_n/t_n \to 0$ in probability conditional on the data.

Now following assumption (E1),

$$E_n(\hat{\mathbf{G}}_{rmn}, [\hat{\mathbf{G}}_{r_1*n}]) = E_n((\hat{\mathbf{G}}_{rmn} - \hat{\mathbf{G}}_{*n}), [\hat{\mathbf{G}}_{r_1*n} - \hat{\mathbf{G}}_{*n}]) \tag{2.8.33}$$

Expanding first argument of the right-hand side

$$\hat{\mathbf{G}}_{rmn} - \hat{\mathbf{G}}_{*n} = (\hat{\mathbf{G}}_{rmn} - \hat{\mathbf{G}}_{r*n}) + (\hat{\mathbf{G}}_{r*n} - \hat{\mathbf{G}}_{*n})$$

We now apply (2.8.32) and then (2.8.3) to the first summand on the right side. Scalar invariance and lipschitz continuity of the evaluation map E_n implies

$$\begin{aligned}&\left|E_n(\hat{\mathbf{G}}_{rmn}, [\hat{\mathbf{G}}_{r_1*n}]) - E_n(\hat{\mathbf{G}}_{r*n}, [\hat{\mathbf{G}}_{r_1*n}])\right| \\ &= O_{P_n}(\min\{b_{sn}, b_{*n}\}^{-1}) + O_P(\min\{a_{sn}, a_{*n}\}^{-1})\end{aligned}$$

Finally from theorem 2.4.3 and assumption (G2), $b_{sn}(\hat{\mathbf{G}}_{r*n} - \hat{\mathbf{G}}_{*n})$ and $a_{sn}(\hat{\mathbf{G}}_{*n} - \mathbf{G}_{*n})$ converge to the same limiting distribution for almost every data sequence; thus

$$\mathbb{E}_r E_n(\hat{\mathbf{G}}_{r*n}, [\hat{\mathbf{G}}_{r_1*n}]) = \mathbb{E} E_n(\hat{\mathbf{G}}_{*n}, [\hat{\mathbf{G}}_{*n}]) + o_{P_n}(1)$$

The proof follows since $\tau_{sn} = o(a_{sn}) \Rightarrow b_{sn} \to \infty$.

Part 2. Continuing from (2.8.33) and applying scale invariance,

$$E_n(\hat{\mathbf{G}}_{rmn}, [\hat{\mathbf{G}}_{r_1*n}]) = E_n(b_{*n}(\hat{\mathbf{G}}_{rmn} - \hat{\mathbf{G}}_{*n}), [b_{*n}(\hat{\mathbf{G}}_{r_1*n} - \hat{\mathbf{G}}_{*n})])$$

and then

$$\begin{aligned}b_{*n}(\hat{\mathbf{G}}_{rmn} - \hat{\mathbf{G}}_{*n}) = \frac{b_{*n}}{b_{sn}}.b_{sn}(\hat{\mathbf{G}}_{rmn} - \hat{\mathbf{G}}_{mn}) + \frac{b_{*n}}{a_{sn}}.a_{sn}(\hat{\mathbf{G}}_{mn} - \mathbf{G}_{mn}) \\ - \frac{b_{*n}}{a_{*n}}.a_{*n}(\hat{\mathbf{G}}_{*n} - \mathbf{G}_{*n}) + b_{*n}(\mathbf{G}_{mn} - \mathbf{G}_{*n})\end{aligned}$$

Since $b_{*n} = a_{*n}/\tau_{*n}, \tau_{*n} = o(a_{*n}), a_{*n} \asymp a_{sn}$, lipschitz continuity of E_n ensures

that

$$\begin{aligned}
&\mathbb{E}_r E_n(b_{*n}(\hat{\mathbf{G}}_{rmn} - \hat{\mathbf{G}}_{*n}), [b_{*n}(\hat{\mathbf{G}}_{r_1*n} - \hat{\mathbf{G}}_{*n})]) \\
&= \mathbb{E}_r E_n \left(\frac{b_{*n}}{b_{sn}}.b_{sn}(\hat{\mathbf{G}}_{rmn} - \hat{\mathbf{G}}_{mn}) + b_{*n}(\mathbf{G}_{mn} - \mathbf{G}_{*n}), [b_{*n}(\hat{\mathbf{G}}_{r_1*n} - \hat{\mathbf{G}}_{*n})] \right) \\
&\quad + o_P(1)
\end{aligned}$$

From theorem 2.4.3 and assumption (G2) $b_{sn}(\hat{\mathbf{G}}_{rmn} - \hat{\mathbf{G}}_{mn})$ and $a_{sn}(\hat{\mathbf{G}}_{mn} - \mathbf{G}_{mn})$ converge to the same limiting distribution for almost every data sequence; and $b_{*n} \uparrow \infty$ implies $\|b_{*n}(\mathbf{G}_{mn} - \mathbf{G}_{*n})\| \to \infty$. Also by assumption $b_{*n} \asymp b_{sn}$. The proof now follows from assumption (E4). □

Proof of theorem 2.5.2. Since we are dealing with a finite sequence of nested models, it is enough to prove that $e_n(\mathcal{M}_1) > e_n(\mathcal{M}_2)$ for large enough n.

Suppose $\mathbb{T}_0 = \mathcal{E}(\mathbf{0}_p, \mathbf{I}_p, g)$. Affine invariance implies invariant to rotational transformations, and since depth decreases along any ray from the origin, $D(\boldsymbol{\theta}, \mathbb{T}_0)$ is a monotonocally decreasing function of $\boldsymbol{\theta}^T\boldsymbol{\theta}$ for any $\boldsymbol{\theta} \in \mathbb{R}^p$. Now consider the models $\mathcal{M}_{10}, \mathcal{M}_{20}$ that have 0 in all indices outside $\mathcal{S}_1$ and $\mathcal{S}_2$, respectively. Take some $\boldsymbol{\theta}_{10} \in \boldsymbol{\Theta}_{10}$, which is the parameter space corresponding to $\mathcal{M}_{10}$, and replace its (zero) entries at indices $j \in \mathcal{S}_2 \backslash \mathcal{S}_1$ by some non-zero $\boldsymbol{\delta} \in \mathbb{R}^{p-|\mathcal{S}_2 \backslash \mathcal{S}_1|}$. Denote it by $\boldsymbol{\theta}_{1\boldsymbol{\delta}}$. Then we shall have

$$\begin{aligned}
\boldsymbol{\theta}_{1\boldsymbol{\delta}}^T\boldsymbol{\theta}_{1\boldsymbol{\delta}} > \boldsymbol{\theta}_{10}^T\boldsymbol{\theta}_{10} \quad &\Rightarrow \quad D(\boldsymbol{\theta}_{10}, \mathbb{T}_0) > D(\boldsymbol{\theta}_{1\boldsymbol{\delta}}, \mathbb{T}_0) \\
&\Rightarrow \quad \mathbb{E}_{s1} D(\boldsymbol{\theta}_{10}, \mathbb{T}_0) > \mathbb{E}_{s1} D(\boldsymbol{\theta}_{1\boldsymbol{\delta}}, \mathbb{T}_0)
\end{aligned}$$

where $\mathbb{E}_s$ denotes the expectation taken over the marginal of the distributional argument $\mathbb{T}_0$ at indices $\mathcal{S}_1$. Notice now that by consitruction $\boldsymbol{\theta}_{1\boldsymbol{\delta}} \in \boldsymbol{\Theta}_{20}$, the parameter space corresponding to $\mathcal{M}_{20}$, and since the above holds for all possible $\boldsymbol{\delta}$, we can take

expectation over indices $\mathcal{S}_2 \backslash \mathcal{S}_1$ in both sides to obtain $\mathbb{E}_{s1} D(\boldsymbol{\theta}_{10}, \mathbb{T}_0) > \mathbb{E}_{s2} D(\boldsymbol{\theta}_{20}, \mathbb{T}_0)$, with $\boldsymbol{\theta}_{20}$ denoting a general element in $\boldsymbol{\Theta}_{20}$.

Now combining (S1a) and (S1b) we get $a_n \mathbf{V}_n^{-1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \rightsquigarrow \mathbb{T}_0$. Suppose $\mathbb{T}_n := [a_n \mathbf{V}_n^{-1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)]$. Now choose a positive $\epsilon < (\mathbb{E}_{s1} D(\boldsymbol{\theta}_{10}, \mathbb{T}_0) - \mathbb{E}_{s2} D(\boldsymbol{\theta}_{20}, \mathbb{T}_0))/2$. Then, for large enough n we shall have

$$|D(\boldsymbol{\theta}_{10}, \mathbb{T}_n) - D(\boldsymbol{\theta}_{10}, \mathbb{T}_0)| < \epsilon \quad \Rightarrow \quad |\mathbb{E}_{s1} D(\boldsymbol{\theta}_{10}, \mathbb{T}_n) - \mathbb{E}_{s1} D(\boldsymbol{\theta}_{10}, \mathbb{T}_0)| < \epsilon$$

following condition (D4). Similarly we have $|\mathbb{E}_{s2} D(\boldsymbol{\theta}_{20}, \mathbb{T}_n) - \mathbb{E}_{s2} D(\boldsymbol{\theta}_{20}, \mathbb{T}_0)| < \epsilon$ for the same n for which the above holds. This implies $\mathbb{E}_{s1} D(\boldsymbol{\theta}_{10}, \mathbb{T}_n) > \mathbb{E}_{s2} D(\boldsymbol{\theta}_{20}, \mathbb{T}_n)$.

Now apply the affine transformation $\mathbf{t}(\boldsymbol{\theta}) = \mathbf{V}_n^{1/2} \boldsymbol{\theta}/a_n + \boldsymbol{\theta}_0$ to both arguments of the depth function above. This will keep the depths constant following affine invariance, i.e. $D(\mathbf{t}(\boldsymbol{\theta}_{10}), [\hat{\boldsymbol{\theta}}]) = D(\boldsymbol{\theta}_{10}, \mathbb{T}_n)$ and $D(\mathbf{t}(\boldsymbol{\theta}_{20}), [\hat{\boldsymbol{\theta}}]) = D(\boldsymbol{\theta}_{20}, \mathbb{T}_n)$. Since this transformation maps $\boldsymbol{\Theta}_{10}$ to $\boldsymbol{\Theta}_1$, the parameter space corresonding to $\mathcal{M}_1$, we get $\mathbb{E}_{s1} D(\mathbf{t}(\boldsymbol{\theta}_{10}), [\hat{\boldsymbol{\theta}}]) > \mathbb{E}_{s2} D(\mathbf{t}(\boldsymbol{\theta}_{20}), [\hat{\boldsymbol{\theta}}])$, i.e. $e_n(\mathcal{M}_1) > e_n(\mathcal{M}_2)$.

□

Proof of corollary 2.5.3. By construction, $\mathcal{M}_0$ is the unique minimal adequate model in $\mathbb{M}_0$, and should be nested in all other adequate models therein. Hence theorem 2.5.2 implies $e_n(\mathcal{M}_0) > e_n(\mathcal{M}^c)$ for any adequate model $\mathcal{M}^c \in \mathbb{M}_0$ and large enough n.

For an inadequate model $\mathcal{M}^w$, suppose $N(\mathcal{M}^w)$ is the integer such that $e_{n_1}(\mathcal{M}^w) < e_{n_1}(\mathcal{M}_*)$ for all $n_1 > N(\mathcal{M}^w)$. Part 3 of theorem 2.3.1 ensures that such an integer exists for every inadequate model. Now define $N = \max_{\mathcal{M}^w \in \mathbb{M}_0} N(\mathcal{M}^w)$: we can do this since $\mathbb{M}_0$ has countably finite elements. Thus $e_{n_1}(\mathcal{M}_0)$ is larger than e-values of all inadequate models in $\mathbb{M}_0$. □

Proof of corollary 2.5.4. Consider $j \in \mathcal{S}_0$. Then $\boldsymbol{\theta}_0 \notin \mathcal{S}_{-j}$, hence $\mathcal{S}_{-j}$ is inadequate. By choice of n_1, e-values of all inadequate models are less than that of $\mathcal{S}_*$, hence

$e_{n_1}(\mathcal{S}_{-j}) < e_{n_1}(\mathcal{S}_*)$.

On the other hand, suppose there exists a j such that $e_{n_1}(\mathcal{S}_{-j}) \leqslant e_{n_1}(\mathcal{S}_*)$ but $j \notin \mathcal{S}_0$. Now $j \notin \mathcal{S}_0$ means that $\mathcal{S}_{-j}$ is an adequate model. Since $\mathcal{S}_{-j}$ is nested within $\mathcal{S}_*$ for any j, and the full model is always adequate, we have $e_{n_1}(\mathcal{S}_{-j}) > e_{n_1}(\mathcal{S}_*)$ by theorem 2.5.2: leading to a contradiction and thus completing the proof. □

Chapter 3

Applications of the Evaluation Maps Framework

3.1 Identifying Driving Factors Behind Indian Monsoon Precipitation

Various studies indicate that our knowledge about the physical drivers of precipitation in India is incomplete; this is in addition to the known difficulties in modeling precipitation itself (Knutti et al., 2010; Trenberth et al., 2003; Wang et al., 2005; Trenberth, 2011). For example, Gosswami (2005) discovered an upward trend in frequency and magnitude of extreme rain events, using daily central Indian rainfall data on a $10^\circ \times 12^\circ$ grid, but a similar study on a $1^\circ \times 1^\circ$ gridded data by Ghosh et al. (2016) suggested that there are both increasing and decreasing trends of extreme rainfall events, depending on the location. Additionally, Krishnamurthy et al. (2009) reported increasing trends in exceedances of the 99th percentile of daily rainfall; however, there is also a decreasing trend for exceedances of the 90th percentile data in many parts of India. Significant spatial and temporal variabilities at various scales have also been discovered for Indian Monsoon (Dietz and Chatterjee, 2014, 2015).

Here we attempt to identify the driving factors behind precipitation during the

Indian monsoon season using our e-value based model selection technique. Data is obtained from the repositories of the National Climatic Data Center (NCDC) and National Oceanic and Atmospheric Administration (NOAA), for the years 1978-2012. We obtained data 35 potential covariates of the Indian summer precipitation:

(A) Station-specific: (from 36 weather stations across India) Latitude, longitude, elevation, maximum and minimum temperature, tropospheric temperature difference (ΔTT), Indian Dipole Mode Index (DMI), Niño 3.4 anomaly;

(B) Global:

- u-wind and v-wind at 200, 600 and 850 mb;
- 10 indices of Madden-Julian Oscillations: 20E, 70E, 80E, 100E, 120E, 140E, 160E, 120W, 40W, 10W;
- Teleconnections: North Atlantic Oscillation (NAO), East Atlantic (EA), West Pacific (WP), East Pacific-North Pacific (EPNP), Pacific/North American (PNA), East Atlantic/Western Russia (EAWR), Scandinavia (SCA), Tropical/Northern Hemisphere (TNH), Polar/Eurasia (POL);
- Solar Flux;
- Land-Ocean Temperature Anomaly (TA).

These covariates are all based on existing knowledge and conjectures from the actual Physics driving Indian summer precipitations. The references provided earlier in this section, and multiple references contained therein may be used for background knowledge on the physical processes related to Indian monsoon rainfall, which after decades of study remains one of the most challenging problems in climate science.

As a modeling step, we consider the annual medians of all the above covariates as fixed effects, the log yearly rainfall at a weather station as response variable,

Variable dropped	$\hat{e}_n(\mathcal{S}_{-j})$
- Tmax	0.1490772
- X120W	0.2190159
- ELEVATION	0.2288938
- X120E	0.2290021
- ΔTT_Deg_Celsius	0.2371846
- X80E	0.2449195
- LATITUDE	0.2468698
- TNH	0.2538924
- Nino34	0.2541503
- X10W	0.2558397
- LONGITUDE	0.2563105
- X100E	0.2565388
- EAWR	0.2565687
- X70E	0.2596766
- v_wind_850	0.2604214
- X140E	0.2609039
- X40W	0.261159
- SolarFlux	0.2624313
- X160E	0.2626321
- EPNP	0.2630901
- TempAnomaly	0.2633658
- u_wind_850	0.2649837
- WP	0.2660394
<none>	0.2663496
- POL	0.2677756
- Tmin	0.268231
- X20E	0.2687891
- EA	0.2690791
- u_wind_200	0.2692731
- u_wind_600	0.2695297
- SCA	0.2700276
- DMI	0.2700579
- PNA	0.2715089
- v_wind_200	0.2731708
- v_wind_600	0.2748239
- NAO	0.2764488

Table 3.1: Ordered values of $\hat{e}_n(\mathcal{S}_{-j})$ after dropping the j-th variable from the full model in the Indian summer precipitation data

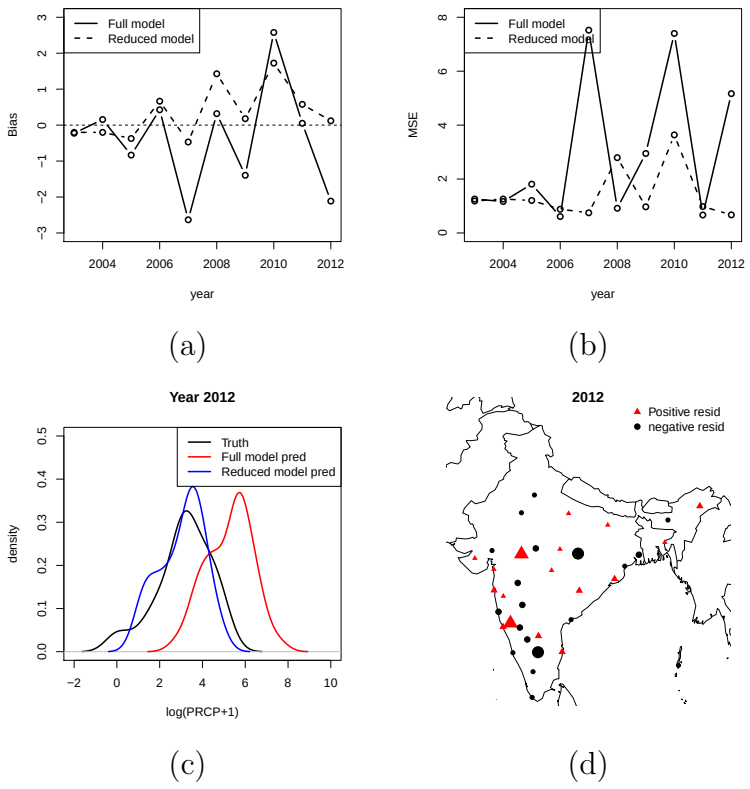


Figure 3.1: Comparing full model rolling predictions with reduced models: (a) Bias across years, (b) MSE across years, (c) density plots for 2012, (d) stationwise residuals for 2012

and include year-specific random intercepts. Table 3.1 lists the estimated $\hat{e}(\mathcal{S}_{-j})$ values in increasing order for the full model as well as all 35 models where a single variable is dropped. We implement the gamma bootstrap with Monte Carlo resample sizes $R = R_1 = 1000$. We use data until 2002 as training data, which contains $n = 897$ samples. The mixed effects model trained on this data is evaluated for $\tau_n = n^\gamma; \gamma = 0.01, 0.02, \ldots, 0.16$. We take the covariate set corresponding to the tuning parameter which minimizes future fixed effect prediction errors on the testing data, i.e. for the period 2003-2012. Outputs of the fast e-value procedure for this covariate set are listed in Table 3.1. The variables listed above *none* in this table are considered relevant by our e-value criterion.

All the variables selected by our procedure have documented effects on Indian monsoon (Krishnamurthy and Kinter III, 2003; Moon et al., 2012). The single largest contributor is *maximum temperature*, whose relation to precipitation is based on the Clausius-Clapeyron relation is now classical knowledge in Physics. It seems that wind velocities high up in the atmosphere are not significant contributors, and the fact that many covariates are selected in the process highlights the complexity of the system.

To check out-of-sample prediction performance of the estimated minimal adequate, we use a rolling validation scheme. For each of the 10 test years: 2003–2012, we select important variables from the model built on past 25 year's data (i.e. use data from 1978–2002 for 2003, 1979-2003 for 2004 and so on), build a model using them and compare predictions on test year obtained from this model with those from the full model. Figure 3.1 summarizes results obtained through this process. Across all testing years, reduced model predictions have less bias as well as are more stable (panels a and b, respectively). The better approximations of truth by reduced models is also evident from the density plot for 2012 in panel c, and there does not seem to be any spatial patterns in its residuals as well (panel d).

3.2 Spatio-temporal Dependence Analysis in fMRI data

In a second application, we apply our proposed method of model selection to analyze brain activity data obtained using functional Magnetic Resonance Imaging (fMRI). Typically, the brain is divided by a grid into three-dimensional array elements called voxels, and activity is measured at each voxel. More specifically, a series of three-dimensional images are obtained by measuring Blood Oxygen Level Dependent (BOLD) signals for a time interval as the subject performs several tasks at specific time points. A single fMRI image typically consists of voxels in the order of 10^5, which makes even fitting the simplest of statistical models computationally intensive when it is repeated for all voxels to generate inference, e.g. investigating the differential activation of brain region in response to a task.

The dataset we work with comes from a recent study involving 19 test subjects and two types of visual tasks (Wakeman and Henson, 2015). Each subject went through 9 runs, in which they were showed faces or scrambled faces at specific time points. In each run 210 images were recorded in 2 second intervals, and each 3D image was of the dimension of $64 \times 64 \times 33$, which means there were 135168 voxels. Here we use the data from a single run on subject 1, and perform a voxelwise analysis to find out the effect of time lags and BOLD responses at neighboring voxels on the BOLD response at a voxel. Formally we consider two models at voxel $i \in \{1, 2, ..., V\}$ at a time point $t \in \{1, 2, ..., T\}$.

3.2.1 Temporal model

The first model we consider is a K-th order autoregressive model in which we try to determine the effect of time lag upto 5 past frames on the BOLD response in voxel i

through the coefficients $(\delta_{i1}, ..., \delta_{i5})$:

$$y_i(t) = x_{ia}(t)\beta_{ia} + x_{ib}(t)\beta_{ib} + \sum_{l=1}^{q} t^{l-1}\gamma_{il} + \sum_{K=1}^{5} y_i(t-k)\delta_{i,t-k} + \epsilon_i(t)$$

Here $x_{ia}(t)$ and $x_{ib}(t)$ are stimulus values corresponding to the two tasks at time t and $\sum_{l=1}^{q} t^{l-1}\gamma_{il}$ is the polynomial drift terms to account for background noise. The stimulus values are calculated through a deterministic equation given the exact time points a face (stimulus a) or scrambled image (stimulus b) is shown (Eloyan et al., 2014).

In this analysis we consider $K = 5$ and $q = 2$, i.e. an AR(5) model with quadratic drift. With this specification, a very small fraction of voxels had any neighbors selected with any autoregressive effects (less than 1%), and most of them was in empty areas, indicating noise.

3.2.2 Spatial model

Our second model is a spatial regression model which tries to determine the amount of spatial dependence that exists between neighboring voxels. For this, apart from the two stimulus term and two drift terms, we consider BOLD responses at all the immediate neighbors of a voxel as potential predictors:

$$y_i(t) = x_{ia}(t)\beta_{ia} + x_{ib}(t)\beta_{ib} + \sum_{l=1}^{q} t^{l-1}\gamma_{il} + \sum_{n \in N_i} y_n(t)\delta_{i,n} + \epsilon_i(t)$$

Here N_i is the set of neighbors of voxel i, $\delta_{i,n}$ is the coefficient corresponding to the effect of neighbor n of voxel i. We consider only immediate neighbors of a voxel. In 3-dimensional space there are 26 such neighbors for voxel not at the periphery of the grid, so the total number of predictors in the voxelwise model in this case is 30. We exclude any voxel on the periphery of the $64 \times 64 \times 33$ grid from the analysis. We also

consider the drift term to be quadratic as before. Further, since a very small fraction of voxels were positive for lag terms in the previous temporal model, we decided not to include any autoregressive term here.

Clubbing together the stimuli, drift terms and neighbor terms into a combined design matrix $\tilde{\mathbf{X}} = (\tilde{\mathbf{x}}(1)^T, ..., \tilde{\mathbf{x}}(T)^T)^T$ and coefficient vector $\boldsymbol{\theta}_i$, we can write $y_i(t) = \tilde{\mathbf{x}}(t)^T\boldsymbol{\theta}_i + \epsilon_i(t)$. We now estimate the set of non-zero coefficients in $\boldsymbol{\theta}_i$ using our method. Suppose this set is R_i, and its subsets containing coefficient corresponding to neighbor and non-neighbor (i.e. stimuli and drift) terms are S_i and T_i, respectively. To quantify the effect of neighbors we now calculate the corresponding F-statistic:

$$F_i = \frac{(\sum_{n\in S_i} \tilde{x}_{i,n}\hat{\theta}_{i,n})^2}{(y_i(t) - \sum_{n\in T_i} \tilde{x}_{i,n}\hat{\theta}_{i,n})^2} \frac{|n - T_i|}{|S_i|}$$

and obtain its p-value, i.e. $P(F_i \geqslant F_{|S_i|,|n-T_i|})$.

Figure 3.2 shows plots of the voxels with a significant p-value from the above F-test. Both left and right visual cortex areas show high spatial dependence, although this is much higher on the left side. Signals from the right eye are processed by the left visual cortex, and high spatial dependence among voxels in both these areas suggest that the right eye was more involved in processing visual signals for this specific subject. We also notice activity in cerebellum, the role of which in visual perception is well-documented (Calhoun et al., 2010; Kirschen et al., 2010).

In terms of future work, we aim to expand on the encouraging findings of this study and repeat the procedure on other individuals in the study. An interesting direction here might be including subject specific random effects and correlating their clinical outcomes (if any) to the observed spatial dependency patterns in their brain.

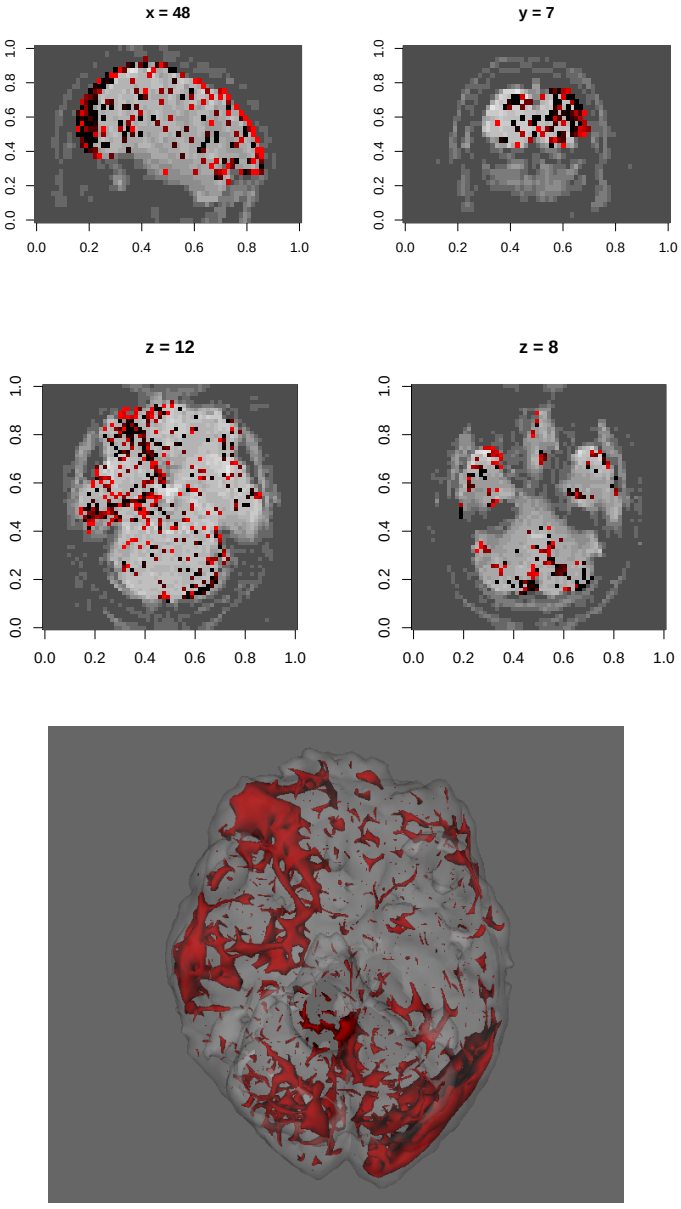


Figure 3.2: (Top) Plot of significant p-values at 95% confidence level at the specified cross-sections; (bottom) a smoothed surface obtained from the p-values clearly shows high spatial dependence in right optic nerve, auditory nerves, auditory cortex and left visual cortex areas

3.3 Selection of Important Single Nucleotide Polymorphisms behind behavioral traits from Familial Genome Wide Association Studies data

3.3.1 Motivation

Genome Wide Association Studies (GWAS), where genetic variants across the full human genome are analyzed, are becoming more and more relevant in recent years for the purpose of determining which of the variants are associated behind the expression of complex traits. The advent of efficient and economical genotyping technology enables researchers to scan the genome at hundreds of thousands of Single Nucleotide Polymorphisms (SNPs), and improvements in computational speed in the past few decades have helped in feasible analysis of the huge amount of data collected in order to detect significant associations (Visscher et al., 2012). One major challenge in such studies is the small effects individual SNPs have: detecting which requires large sample sizes (Manolio et al., 2009). For quantitative behavioral traits, for example alcohol dependence, drug abuse, Anorexia and depression, this problem is amplified because of the additional noise introduced by variation due to the environment the subject grew up in. This is one of the motivations of performing GWAS on families (GWAF) instead of unrelated individuals, through which the environmental variation can be reduced: so as to require smaller samples to detect the same magnitude of SNP effect. Another major reason, of course, of performing GWAS on familial data is to detect gene-environment interactions associated with development of behavioral traits. The data analyzed in such families typically consist of trait information and genotypes of from parents and their children, who can be either identical twins, non-identical twins or adpoted.

Single-marker tests, i.e. analyzing the effect of multiple SNPs separately on the

quantitative phenotype and then selecting a group of SNPs by setting suitable thresholds on the resulting p-values, often after correction for multiple testing, is the most commonly used method to detect SNPs associated with the phenotype being studied. Although simultaneously estimating the fixed effect of a single SNP as well as the stratified population variance covariance matrix reflecting the familial structure, and repeating this for a large number of SNPs is a computationally prohibitive task, several fast approximation methods exist in the literature that tackle this while maintaining moderately high power. The GRAMMAR method of Aulchenko et al. (2007) and the association test of Chen and Abecasis (2007) are examples of this. While these two methods are able to efficiently analyze GWAF data, they assume that phenotypic similarity within families is entirely due to their genetic similarity and ignore the effect of shared environment. In GWAF data from nuclear (i.e. unrelated) families, the proportion of phenotypic variation explained by the shared environmental effects is often substantial, sometimes as high as 51% (McGue et al., 2013) or 74% (De Neve et al., 2013): in which case such methods shall not be able to account for this added variation. To remedy this, Li et al. (2011) proposed a rapid method (RFGLS) that computes p-values corresponding to each SNP through a rapid approximation of the single-SNP generalized least squares model taking into account genetic and environmental sources of familial similarity.

A major issue with all such methods of single-marker analysis is that they are not always effective for detecting functionally relevant SNPs or regions in the genome. A single SNP is sometimes not enough to capture the extent of association (Yang et al., 2012; Ke, 2012). This includes cases when there are multiple causal SNPs closely located inside a gene in high Linkage Disequilibrium (LD) with one another. The causal SNP may even not be genotyped if its variants are unlikely to be present in the sample population (e.g. the variant of the SNP rs671 responsible for low alcohol tolerance in asians is rare in caucasians), and other SNPs highly correlated with it

are genotyped instead.

Here we propose to tackle this through fitting mixed effect models with the behavioral trait phenotype as respone and a group of SNPs (e.g. SNPs inside a single gene) as fixed effect predictors, and selecting important SNPs through a model selection approach. Although the major impediment of applying model selection techniques in GWAS setup is the high computational cost, some fast methods have been proposed that are able to perform SNP seletion from a multi-SNP model on GWAS data from *unrelated individuals* (Zhang et al., 2014; Frommelet et al., 2012). However, these methods still rely on fitting models corresponding to multiple predictor sets. This makes them unsuitable to be adapted to the GWAF setup because of the much higher computational costs associated with training multiple mixed effect models that take into account within-family correlation between individuals, as compared to models that assume independent observations.

We shall use our e-values framework to provide a solution to this situation. As showed in the last chapter, our variable selection technique based on e-values requires only fitting the 'full model': which makes it suitable to be utilized here.

3.3.2 The MCTFR data

The familial GWAS dataset collected and studied by Minnesota Center for Twin and Family Research (MCTFR)(Li et al., 2011; Miller et al., 2012; McGue et al., 2013) consists of samples from three longitudinal studies conducted by the MCTFR: (1) the Minnesota Twin Family Study (MTFS: Iacono et al. (1999)) that covers twins and their parent, (2) the Sibling Interaction and Behavior Study (SIBS: McGue et al. (2007)) that includes adopted and biological sibling pairs and their parents, and (3) the enrichment study (ES: Keyes et al. (2009)) that extended the MTFS by oversampling 11 year old twins who are highly likely to develop substance abuse. While 9827 individuals completed the initial assessments for participation in the study, af-

ter several steps of screening the final sample consisted of 7188 caucasian individuals clustered in 2300 nuclear families.

DNA samples collected from the subjects were analyzed using Illuminas Human660W-Quad Array for 561,490 non-intensity SNP markers. After several data cleaning steps for quality control, 527,829 SNPs were retained. Covariates for each sample included age, sex, birth year, generation (parent or offspring), as well as two-way interactions between generation and other three covariates each. As for the quantitative phenotypes, five of them were studied in this GWAS: (1) Nicotine dependence, (2) Alcohol consumption, (3) Alcohol dependence, (4) Illegal drug usage, and (5) Behavioral disinhibition. The response variables corresponding to these phenotypes were derived from questionnaires using a hierarchical approach based on factor analysis (Hicks et al., 2011).

A more detailed description of the data is available in Miller et al. (2012). Several studies have been performed that focus on different aspects of this dataset. Li et al. (2011) used RFGLS to single out causal SNPs behind the height of participants, while McGue et al. (2013) used the same method to study SNPs behind the development of all five indicators of behavioral disinhibition mentioned above. Irons (2012) focused on the effect of several factors affecting alcohol use in the study population, namely the effects of polymorphisms in the ALDH2 gene and the GABA system genes, as well as the effect of early exposure to alcohols as adolescents to adult outcomes. Finally Coombes (2016) used a bootstrap-based combination test and a sequential score test to evaluate gene-environment interactions behind phenotypic outcomes in the data.

3.3.3 Statistical model

We shall demonstrate the use of e-values in this context using a Linear Mixed Model (LMM) framework. We assume that the families modeled are unrelated to one another, i.e. they are nuclear pedigrees. We stick to this structure for ease of represen-

tation, although as discussed shortly, the model fitting process remains unchanged for larger pedigrees.

We assume there are a total of m families, with the i-th pedigree containing n_i individuals. Denote by $\mathbf{y}_i = (y_{i1}, \ldots, y_{in_i})^T$ the quantitative trait values for individuals in that pedigree, while the matrix $\mathbf{G}_i \in \mathbb{R}^{n_i \times p_s}$ containing their genotypes for a bunch of SNPs. Let $\mathbf{C}_i \in \mathbb{R}^{n_i \times p}$ denote the data on p covariates for individuals in the pedigree i. Given these, we consider the following model.

$$\mathbf{Y}_i = \alpha + \mathbf{G}_i\boldsymbol{\beta}_g + \mathbf{C}_i\boldsymbol{\beta}_c + \boldsymbol{\epsilon}_i \tag{3.3.1}$$

with α the intercept term, $\boldsymbol{\beta}_g$ and $\boldsymbol{\beta}_c$ fixed coefficient terms corresponding to the multiple SNPs and covariates, respectively, and $\boldsymbol{\epsilon}_i \sim \mathcal{N}_{n_i}(\mathbf{0}, \mathbf{V_i})$ the random error term. To account for the within-family dependency structure, we break up the random error variance into three independent components:

$$\mathbf{V}_i = \sigma_a^2\boldsymbol{\Phi}_i + \sigma_c^2\mathbf{1}\mathbf{1}^T + \sigma_e^2\mathbf{I}_{n_i} \tag{3.3.2}$$

The first part represents a within-family random effect term to account for effects of other SNPs. The matrix $\boldsymbol{\Phi}_i$ is the relationship matrix within the i-th pedigree. Its (s,t)-th element represents two times the kinship coefficient, which is the probability that given that a random gene is drawn each from individuals s and t in pedigree i, these genes are 'identical by descent', i.e. come from same common ancestor. The second part accounts for shared environmental effect within the family, while the third term finally quantifies other sources of variation unique to an individual.

Following basic probability, the kinship coefficient of a parent-child pair is 1/4, a full sibling pair or non-identical (or dizygous = DZ) twins is 1/4, and for identical (or monozygous = MZ) twins is 1/2 in a nuclear pedigree. Following this, we can

construct the $\mathbf{\Phi}_i$ matrices for different types of families:

$$\mathbf{\Phi}_{MZ} = \begin{bmatrix} 1 & 0 & 1/2 & 1/2 \\ 0 & 1 & 1/2 & 1/2 \\ 1/2 & 1/2 & 1 & 1 \\ 1/2 & 1/2 & 1 & 1 \end{bmatrix}, \mathbf{\Phi}_{DZ} = \begin{bmatrix} 1 & 0 & 1/2 & 1/2 \\ 0 & 1 & 1/2 & 1/2 \\ 1/2 & 1/2 & 1 & 1/2 \\ 1/2 & 1/2 & 1/2 & 1 \end{bmatrix}, \mathbf{\Phi}_{Adopted} = \mathbf{I}_4$$

for families with parents (indices 1 and 2) and MZ twins, DZ twins, or two adapted children (indices 3 and 4), respectively.

We use the R package `regress` to fit the above model with additive error structure. The package requires specifying the dependency structure of all samples in the data. For ease of representation, we only consider nuclear pedigrees with MZ twins in our simulation study and data analysis, which simplifies the overall relationship matrix $\mathbf{\Phi} = \text{diag}(\mathbf{\Phi}_1, \ldots, \mathbf{\Phi}_m)$ as $\mathbf{I}_m \otimes \mathbf{\Phi}_{MZ}$. Note that, situations in which the pedigree structure is not nuclear can be readily handled in this situation by supplying the overall $\mathbf{\Phi}$ matrix. The second overall structural component will be $\mathbf{I}_m \otimes \mathbf{1}\mathbf{1}^T$. The `regress` procedure includes the third structure in (3.3.2) by default.

3.3.4 A conditional e-value

We now take a closer look at the evaluation map distributions corresponding to reduced model coefficient estimates, in order to better detect the weak signals we are dealing with here and reduce the number of false positives. Carrying over notations from Chapter 2, recall that the model formed by dropping the j-the index is denoted by $\mathcal{M}_{-j}$, and we are going to inspect $\mathbb{D}_{-j}$, the distribution of $D(\hat{\boldsymbol{\beta}}_{-j}, [\hat{\boldsymbol{\beta}}])$ by comparing it with $\mathbb{D} :=$ distribution of $D(\hat{\boldsymbol{\beta}}, [\hat{\boldsymbol{\beta}}])$. Also define by $\mu(.)$ the mean operator on the corresponding distributions.

Recall from Chapter 2 that we approximate the above distributions and the final

e-value through resampling. The quality of approximation depends on the variance parameter τ_n^2, and as seen in the simulation section, on the type of bootstrap scheme used (moon/ gamma/ generalized). Because of the high-computational overhead of the `regress` procedure, we shall use the parametric generalized bootstrap scheme here. We would also like to emphasize that all observations in this subsection are entirely empirical and from controlled simulation setups, and further studies are warranted to theoretically characterize such behavior.

We denote by $\hat{\mathbb{D}}_{-j}(\tau)$ and $\hat{\mathbb{D}}(\tau)$ the approximations of $\mathbb{D}_{-j}$ and $\mathbb{D}$, respectively, using a generalized bootstrap scheme with standard deviation $\tau_n \equiv \tau$. According to theorem 2.4.4, in case $\mathbb{D}_{-j}$ is an inadequate model distribution, the mean of $\hat{\mathbb{D}}_{-j}(\tau)$ goes to 0 in probability for an intermediate rate of the bootstrap standard deviation. As τ increases, all reduced model distributions approach $\hat{\mathbb{D}}(\tau)$. However, depending on the magnitude of signals at the non-zero indices, we observe that this behavior follows two different regimes. We demonstrate this using the simulation setup we elaborate on later in the chapter. In the plots below, h represents the relative magnitude of non-zero entries in the coefficient vector: for which we consider two choices, $h = 5$ and $h = 0.05$.

(a) Large signal regime ($h = 5$: Figure 3.3) When $\mathbb{D}_{-j}$ corresponds to an inadequate model, i.e. the j-th coefficient of the true parameter vector is non-zero, for small values of τ we can clearly distinguish this distribution from that of $\hat{\mathbb{D}}(\tau)$ in their density plots. As τ increases, the inadequate model distributions seem to have more and more positive bias. However, when j is a non-essential covariate, the reduced distributions are close to $\hat{\mathbb{D}}(\tau)$ for all values of τ.

(b) Small signal regime ($h = 0.05$: Figure 3.4) When the actual signal in β_j is weak, the inadequate reduced model distributions still approach $\hat{\mathbb{D}}(\tau)$ as τ goes up

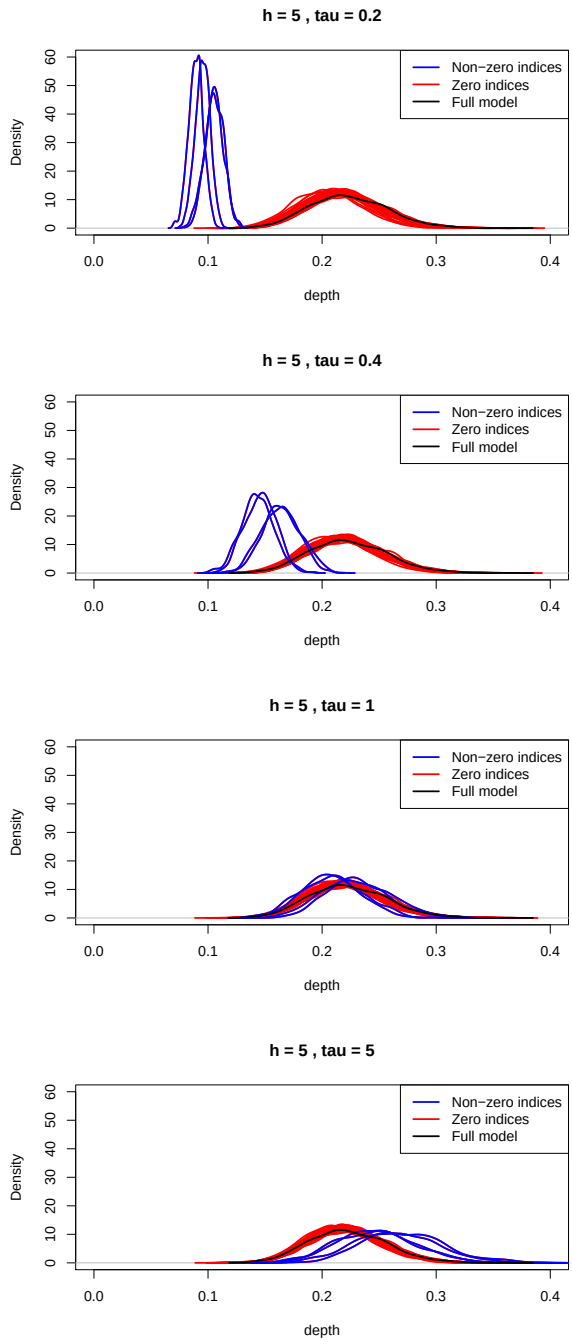


Figure 3.3: Density plots for $\hat{\mathbb{D}}(\tau)$ and $\hat{\mathbb{D}}_{-j}(\tau)$ for all j in simulation setup, with signal parameter $h = 5$ and bootstrap standard deviations $\tau = 0.2, 0.4, 1, 5$

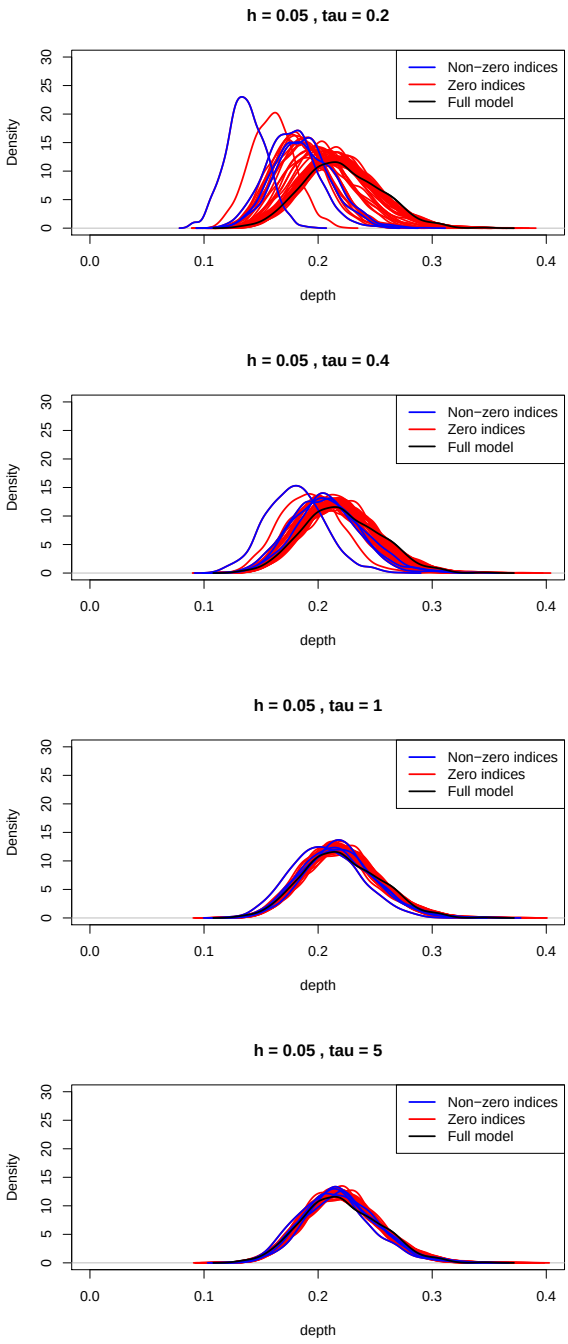


Figure 3.4: Density plots for $\hat{\mathbb{D}}(\tau)$ and $\hat{\mathbb{D}}_{-j}(\tau)$ for all j in simulation setup, with signal parameter $h = 0.05$ and bootstrap standard deviations $\tau = 0.2, 0.4, 1, 5$

but stabilize at the full model distribution instead of passing it for very large τ ($\tau = 5$ here). However the adequate model distributions seem to exhibit a similar behavior: albeit staying to the left of inadequate model density plots in general.

This increased ambiguity of reduced model distributions for small signals make it difficult to distinguish between two types of model distributions using the mean operator, which ends up being very conservative in the second case. For this reason we consider the usage of a different summarizing function that will be able to capture the differentiate between the two types of reduced model distributions across a broader range of the signal-to-noise ratio, specifically by setting a lower detection threshold than the same operator on the full model distribution. Here we focus on a specific alternate formulation of the e-value that is based on a tail quantile of $\hat{\mathbb{D}}_{-j}(\tau)$:

$$e_q(\mathcal{M}_{-j}|\tau) = q\text{-th quantile of } \hat{\mathbb{D}}_{-j}(\tau) \tag{3.3.3}$$

for some fixed $q \in (0, 1)$. Notice that for any q, the quantity $e_q(\mathcal{M}_{-j}|\tau)$ is conditional on the bootstrap standard deviation parameter τ.

The motivation behind this is the observation that the inadequate and adequate model distributions have different tail behaviors for intermediate values of τ, and setting an appropriate upper threshold to tail probabilities for a suitable fixed quantile of these distributions with respect to $\hat{\mathbb{D}}(\tau)$ can possibly separate out the two types of distributions. The choice of threshold potentially depends on several factors such as the value of q, the statistical model used, degree of sparsity of parameters in the data generating process. In the following section we shall experiment with different thresholds to illustrate this. Also note that we still retain the main flavor of the e-values method, by training only the full model and then use Monte Carlo resampling to compute $e_q(\mathcal{M}_{-j}|\tau)$ for all j and a range of τ.

3.3.5 Simulation

We now compare the performance of the above formulation of quantile e-values in a simulation setup. For this, consider the model in (3.3.1) with no environmental covariates. We consider familes with MZ twins and first generate the covariate matrices $\mathbf{G}_i$. We take a total of $p_g = 50$ SNPs, and to simulate correlation among SNPs in the genome generate them in correlated blocks of 6, 4 ,6, 4 and 30. We set the correlation between two SNPs inside a block at 0.7, and consider the blocks to be uncorrelated. For each parent we generate two independent vectors of length 50 with the above correlation structure, and entries within each block being 0 or 1 following Bernoulli distributions with probabilities 0.2, 0.4, 0.4, 0.25 and 0.25 (Minor Allele Frequency or MAF) for SNPs in the 5 blocks, respectively. The genotype of a person is then determined by taking the sum of these two vectors: thus entries in $\mathbf{G}_i$ can take the values 0, 1 or 2. Finally we set the common genotype of the twins by randomly choosing one allele vector from each of the parents and taking their sum.

We repeat the above process for $m = 250$ families. In GWAS there are generally a small number of causal SNPs, each explaining small proportions of the overall variability in response variable. To reflect this in our simulation setup, we assume that the first entries in each of the first four blocks above are causal, and each of them explains $h/(\sigma_a^2 + \sigma_c^2 + \sigma_e^2)\%$ of the overall variability. The term h is known as the *heritability* of the corresponding SNP (and can of course vary across SNPs). The value of the non-zero coefficient in k-th group: $k = 1, ..., 4$, say β_k is calculated using the formula:

$$\beta_k = \sqrt{\frac{h}{(\sigma_a^2 + \sigma_c^2 + \sigma_e^2).2\mathrm{MAF}_k(1 - \mathrm{MAF}_k)}} \tag{3.3.4}$$

We fix the following values for the error variance components: $\sigma_a^2 = 4, \sigma_c^2 = 1, \sigma_e^2 = 1$, and generate pedigree-wise response vectors $\mathbf{y}_1, \ldots, \mathbf{y}_{250}$ using the above setup. To

consider different SNP effect sizes, we repeat the above setup for $h \in \{10, 5, 2, 1, 0\}$, generating 1000 datasets for each value of h.

Methods and metrics

For this simulated data, we compare our e-value based approach with two other methods:

(1) Model selection on linear model: Here we ignore the dependency structure within families by training linear models on the simulated data and selecting SNPs with non-zero effects by backward deletion using a modification of the BIC called mBIC2. This has been showed to give better results than single-marker analysis in GWAS for unrelated individuals (Frommelet et al., 2012) and provides approximate False Discovery Rate (FDR) control at level 0.05 (Bogdan et al., 2011).

(2) Single-marker mixed model: We train single-SNP versions of (3.3.1) using a fast approximation of the Generalized Least Squares procedure (named Rapid Feasible Generalized Least Squares or RFGLS: Li et al. (2011)), obtain marginal p-values from corresponding t-tests and use the Benjamini-Hochberg (BH) procedure to select significant SNPs at FDR = 0.05.

We compute the e-values by setting projection depth (Zuo, 2003) as the evaluation function. With the e-value being the q-th quantile of the evaluation map distribution, we set the detection threshold value at the t-th multiple of q for some $0 < t < 1$. This means all indices j such that q-th quantile of the bootstrap approximation of $\hat{\mathbb{D}}_{-j}(\tau)$ is less than the tq-th quantile of $\hat{\mathbb{D}}(\tau)$ will get selected as the set of active predictors. We repeat the e-value procedure for different values of the bootstrap standard deviation $s \in \{0.3, 0.35, \ldots, 0.95, 2\}$. Consequently, we take as the final estimated set of SNPs the SNP set $\hat{\mathcal{S}}(\tau)$ that minimizes fixed effect prediction error

(PE) on an independently generated test dataset $\{(\mathbf{y}_{test,i}, \mathbf{G}_{test,i}), i = 1, \ldots, 250\}$ from the same setup above:

$$\mathrm{PE}(\tau|q,t) = \sum_{i=1}^{250}\sum_{j=1}^{4}\left(y_{test,ij} - \mathbf{g}_{test,ij}^{T}\hat{\boldsymbol{\beta}}_{\hat{\mathcal{S}}(\tau)}\right)^2;$$

$$\hat{\mathcal{S}}_0(q,t) = \arg\min_{\tau} \mathrm{PE}(\tau|q,t)$$

The metrics to evaluate each method we implement are:

1. True Positive (TP): proportion of causal SNPs detected;
2. True Negative (TN): proportion of non-causal SNPs undetected;
3. Relaxed True Positive (TPR): proportion of detecting any SNP in each of the 4 blocks with causal SNPs, i.e. for the selected index set $\hat{\mathcal{S}}_0(q,t)$,

$$\mathrm{TPR}(\hat{\mathcal{S}}_0(q,t)) = \frac{1}{4}\sum_{i=1}^{4}\mathbb{I}(\text{Block } i \cap \hat{\mathcal{S}}_0(q,t) \neq \varnothing)$$

4. Relaxed True Negative (TNR): proportion of SNPs in block 5 undetected.

We consider the third and fourth metrics to cover situations in which the causal SNP is not detected itself, but highly correlated SNPs with the causal SNP are. This is common in GWAS. Finally, we average all the above proportions over 1000 replications, and repeat the process for $q \in \{0.9, 0.5, 0.2, 0.1\}; t \in \{0.8, 0.7, 0.6, 0.5\}$.

Results

We present the simulation results in Table 3.2 and Table 3.3. Applying BIC on linear models performs poorly compared RFGLS and then correction for multiple testing on marginal LMMs for all heritability values: possibly because the linear models

6x Heritability	mBIC2	RFGLS +BH	quantile e-values				
			q	$t = 0.8$	$t = 0.7$	$t = 0.6$	$t = 0.5$
$h = 10$	0.79/0.99	0.95/0.92	0.9	0.95/0.97	0.95/0.97	0.95/0.98	0.94/0.98
			0.5	0.96/0.97	0.96/0.98	0.95/0.98	0.94/0.98
			0.2	0.96/0.94	0.96/0.97	0.95/0.97	0.95/0.98
$h = 5$	0.41/0.99	0.62/0.97	0.9	0.72/0.95	0.7/0.96	0.69/0.96	0.66/0.97
			0.5	0.78/0.94	0.75/0.94	0.72/0.95	0.71/0.96
			0.2	0.83/0.91	0.78/0.94	0.75/0.95	0.73/0.95
$h = 2$	0.11/0.99	0.14/0.99	0.9	0.26/0.97	0.24/0.97	0.23/0.98	0.21/0.98
			0.5	0.34/0.95	0.28/0.96	0.27/0.97	0.26/0.97
			0.2	0.46/0.91	0.34/0.95	0.3/0.96	0.27/0.96
$h = 1$	0.05/0.99	0.04/0.99	0.9	0.12/0.98	0.1/0.98	0.09/0.99	0.08/0.99
			0.5	0.16/0.96	0.13/0.97	0.12/0.97	0.11/0.98
			0.2	0.25/0.93	0.16/0.96	0.13/0.97	0.13/0.97
$h = 0$	–/0.99	–/0.99	0.9	–/0.99	–/0.99	–/0.99	–/0.99
			0.5	–/0.98	–/0.98	–/0.99	–/0.99
			0.2	–/0.94	–/0.98	–/0.98	–/0.99

Table 3.2: Average True Positive (TP) and True Negative (TN) proportions over 1000 replications for all three methods

6x Heritability	mBIC2	RFGLS +BH	quantile e-values				
			q	$t = 0.8$	$t = 0.7$	$t = 0.6$	$t = 0.5$
$h = 10$	0.84/0.99	0.96/0.99	0.9	0.96/0.97	0.96/0.97	0.95/0.98	0.94/0.98
			0.5	0.96/0.97	0.96/0.97	0.95/0.98	0.95/0.98
			0.2	0.97/0.95	0.96/0.97	0.96/0.97	0.95/0.98
$h = 5$	0.48/0.99	0.64/0.99	0.9	0.73/0.95	0.71/0.95	0.7/0.96	0.67/0.97
			0.5	0.79/0.93	0.76/0.94	0.73/0.95	0.72/0.95
			0.2	0.85/0.91	0.79/0.93	0.76/0.94	0.74/0.95
$h = 2$	0.16/0.99	0.16/0.99	0.9	0.29/0.96	0.27/0.97	0.25/0.98	0.23/0.98
			0.5	0.37/0.95	0.31/0.96	0.3/0.96	0.29/0.97
			0.2	0.53/0.91	0.38/0.95	0.33/0.95	0.3/0.96
$h = 1$	0.08/0.99	0.05/0.99	0.9	0.15/0.97	0.13/0.98	0.12/0.98	0.1/0.99
			0.5	0.2/0.96	0.17/0.97	0.15/0.97	0.13/0.98
			0.2	0.35/0.93	0.21/0.96	0.17/0.97	0.16/0.97
$h = 0$	–/0.98	–/0.99	0.9	–/0.97	–/0.98	–/0.98	–/0.99
			0.5	–/0.95	–/0.97	–/0.97	–/0.98
			0.2	–/0.90	–/0.95	–/0.97	–/0.97

Table 3.3: Average Relaxed True Positive (TPR) and Relaxed True Negative (TNR) proportions over 1000 replications for all three methods

are trained on a smaller amount of data and ignore the variation due to shared environment in the parents.

Our proposed e-values work better than the two methods for detecting true signals across different values of h: the average TP rate going down slowly than other methods across the majority of choices for (q, t). Both mBIC2 and RFGLS+BH have very high true negative detection rates, which is matched by our method for higher values of q. Since all reduced model distributions reside on the left of the full model distribution, we expect the variable selection process to turn more conservative at higher values of t.This effect is more noticeable for lower q. This indicates that the right tails of evaluation map distributions are more useful for this purpose. Finally for $h = 0$, we report only TN or TNR values since no signals should ideally be detected: in terms of this a value of $q = 0.9$ or $q = 0.5$ leads to the same TN and TNR performance as RFGLS+BH for all choices of t. Finally, TPR performances for all methods are better than the corresponding TPTN performances. However, for mBIC2 this seems to be due to detecting SNPs in the first four blocks by chance since for $h = 0$ its TNR is less than TN.

Considering that when analyzing a large number of SNPs false positives need to be minimized, setting $q = 0.9, t = 0.5$ is a safe choice choice for e-values in this simulation setup. Note here that the previous model selection algorithm using e-values depended on comparing the mean of the evaluation map distribution $\mathbb{D}_{-j}$ with that of $\mathbb{D}$. Compared to that here we end up comparing a tail quantile of $\mathbb{D}$, and set the detection threshold at a smaller value than the same quantile of $\mathbb{D}$.

3.3.6 Analysis of the Minnesota Twin Studies data

We now apply the above technique on genes from a familial GWAS dataset collected and studied by Minnesota Center for Twin and Family Research (Miller et al., 2012; McGue et al., 2013; Li et al., 2011). Here a total of 7188 Caucasian individuals, who

come from ~ 2300 families, have been genotyped. A detailed description of the data is available at Miller et al. (2012).

In total, five quantitative phenotypes were studied in this GWAS: (1) Nicotine dependence, (2) Alcohol consumption, (3) Alcohol dependence, (4) Illegal drug usage, and (5) Behavioral disinhibition. The response variables corresponding to these phenotypes were derived from questionnaires using a hierarchical approach based on factor analysis (Hicks et al., 2011). SNP genotype data were collected from the sample using Illuminas Human660W-Quad Array, and 529828 SNPs were retained in the dataset after a screening for quality control.

We assume a nuclear pedigree structure, and for simplicity only analyze pedigrees with MZ twins only. After adjusting for missing data, here we have 682 such 4-member families. For the response variable, we look at the effect of genetic factors behind alcohol consumption, which has previously been found to be highly heritable in this dataset (McGue et al., 2013). As a first pass we decide to analyze SNPs inside some of the most-studied genes with respect to alcohol abuse: GABRA2, ADH1B, ADH1C, SLC6A3, SLC6A4, OPRM1, CYP2E1, DRD2, ALDH2, and COMT (Coombes, 2016) through separate gene-level models. The ADH genes did not contain many SNPs individually, so we decided to club all existing ADH genes (ADH1-ADH7) together in our analysis.

For each gene, We train the LMM in (3.3.1) on 75% of randomly selected families, perform our conditional e-values procedure for $\tau = 0.2, 0.4, \ldots, 2.8, 3$; and select the predictor set $\hat{\mathcal{S}}_0(\tau)$ that minimizes fixed effect prediction error on the data from the other 25% of families. To enforce a stricter control on which SNPs get selected, we use $q = 0.9$ and $t = 0.5$ here based on results in the simulation setup, and use projection depth as the evalution function.

We show the results of our gene-specific analyses in Figure 3.5, Figure 3.6 and Figure 3.7. The exon locations are obtained from annotation data extracted from

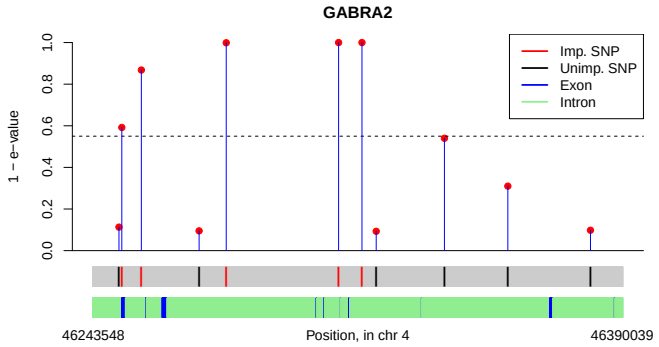
Gene	Total/detected SNP	Non-zero SNPs ordered per position in genome
GABRA2	11/5	rs572227(-), rs534459(+), rs502038(-), rs1808851(+), rs279856(-)
ADH	21/5	rs17027523(-), rs13103626(+), rs10516430(+), rs12503056(+), rs2004316(-)
SLC6A3	18/4	rs2042449(+), rs464049(-), rs460700(-), rs460000(+)
SLC6A4	5/0	None
OPRM1	46/29	rs9371718(-), rs1937600(+), rs9397637(+), rs12662873(-), rs1316368(+), rs1937587(-), rs6921403(-), rs1937580(+), rs1937645(+), rs1892361(-), rs1937633(+), rs1937631(-), rs12527197(-), rs1892360(-), rs1892356(+), rs1937619(-), rs1332849(-), rs9371749(+), rs9285539(+), rs9322439(-), rs11752884(+), rs4870241(-), rs689219(-), rs9371761(+), rs12199858(+), rs9371762(-), rs612450(+), rs9384159(+), rs6938958(-)
CYP2E1	9/5	rs9419702(-), rs9419624(+), rs7906770(-), rs9419569(+), rs9419629(+)
DRD2	17/0	None
ALDH2	5/5	rs7398343(+), rs7297186(+), rs3803167(+), rs10219736(-), rs3742004(-)
COMT	15/9	rs4646312(-), rs165656(-), rs165722(+), rs2239393(-), rs4680(+), rs174699(-), rs165728(+), rs5993891(+), rs2239395(-)

Table 3.4: Table of analyzed genes and detected SNPs in them. Positive/ negative sign indicates type of association found.

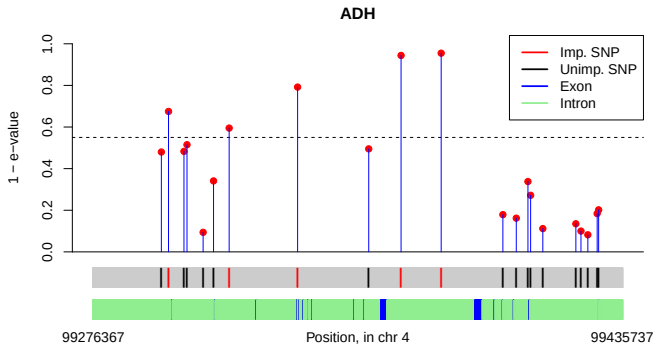
the UCSC Genome Browser database (Rosenbloom et al., 2015). Also Table 3.4 summarizes the selected SNPs for each gene. In general, SNPs tend to get selected in groups with neighboring SNPs, which suggests high Linkage Disequilibrium (LD). Also most of the selected SNPs either overlap or in close proximity to the coding regions of genes, i.e. exons, which underline their functional relevance.

Finally, below are some gene-specific observations:

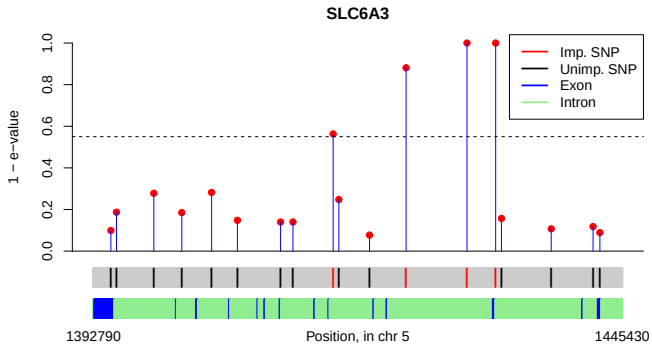
GABRA2: As seen in the plots, the first two SNPs detected are close to two separate exons. The 4th and 5th detected SNPs, rs1808851 and rs279856, are at perfect LD with rs279858 in the larger 7188-individual dataset (Irons, 2012). This SNP had not been genotyped in our sample, but is the marker in GABRA2 most frequently associated in the literature with alcohol abuse (Cui et al., 2012). Interestingly, a



(a)

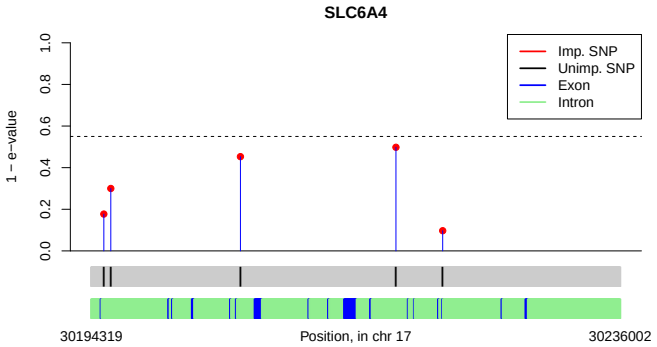


(b)

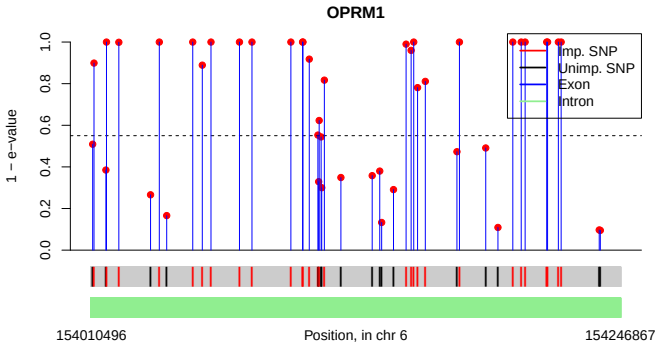


(c)

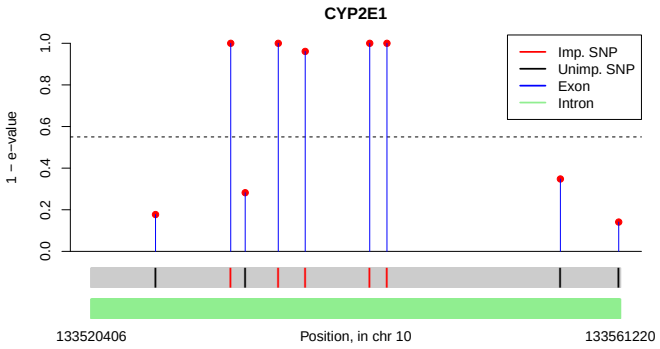
Figure 3.5: Plot of e-values for genes analyzed: (a) GABRA2, (b) ADH1 to ADH7, (c) SLC6A3



(d)

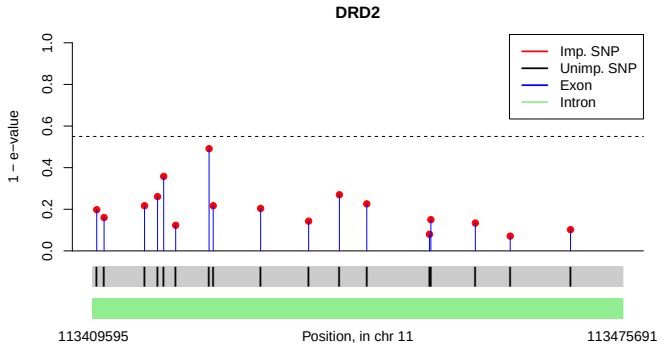


(e)

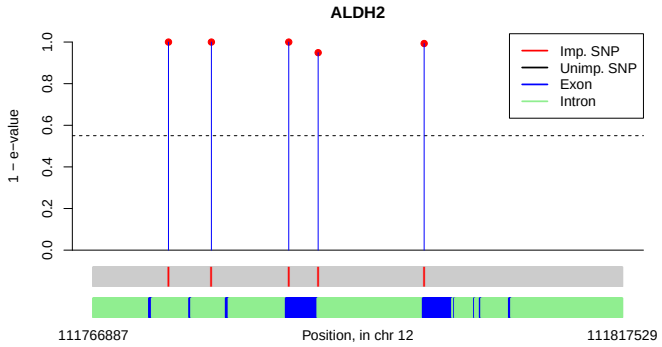


(f)

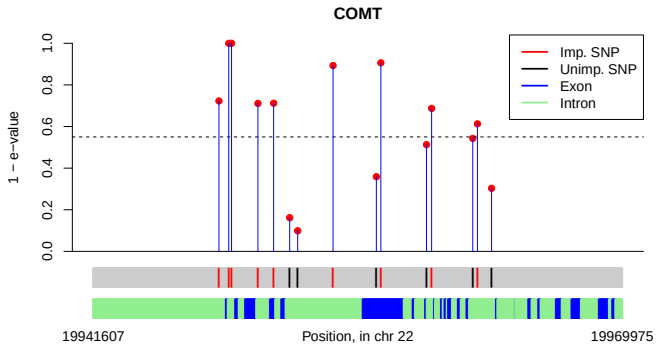
Figure 3.6: Plot of e-values for genes analyzed: (d) SLC6A4, (e) OPRM1, (f) CYP2E1



(g)



(h)



(i)

Figure 3.7: Plot of e-values for genes analyzed: (g) DRD2, (h) ALDH2, (i) COMT

single SNP RFGLS analysis of the same twin studies data that used Bonferroni correction on marginal p-values to detect SNPs had missed these SNPs (Irons, 2012). This highlights the advantage of our approach.

ADH genes: Multiple studies have associated rs1229984 in the ADH1B gene (position 99318162 of chromosome 4) with alcohol dependence (`https://www.snpedia.com/index.php/Rs1229984`), which as seen in the plot of ADH2 is close to an exon region. Our data does not contain this marker, but detects rs13103626 and rs10516430 at positions 99317251 and 99337881 respectively. The SNP rs17027523 is interesting: they reside in the uncharacterized long non-coding RNA gene LOC100507053. One previous study (Gelernter et al., 2014; Xu et al., 2015) found significant associations for 5 SNPs in this gene with alcohol consumption for African American population through single-SNP analysis on non-familial GWAS data. Notably, their analysis found a much stronger evidence of the association in African-American part of the sample than the European American part, while our findings are entirely from a Caucasian sample.

SLC6A3: Our analysis does not detect rs27072, which has been associated with alcohol withdrawal symptoms (`https://www.snpedia.com/index.php/Rs27072`). Two of the four neighboring SNPs we detect are in intron regions, while the other two very close to exons.

OPRM1: The minor allele of the SNP rs1799971 (chr 6, position 154039662) has been associated with stronger alcohol cravings (`https://www.snpedia.com/index.php/Rs1799971`). We detect rs12662873 that resides within 1 kb of this SNP. There are 28 more SNPs detected by our procedure, which seem to reside in 3 clusters.

CYP2E1: Five of the 9 SNPs studied are detected through our analysis. Four of them are within 10 kb of one another (base pairs 133534822 to 133543210 in chr 10). Although CYP2E1 produces one of the three major enzymes required in alcohol metabolism, effect of SNPs in this gene on alcohol dependence is sparse. In the analysis of Lind et al. (2012) rs4646976 at 133534223 position was most associated with a measure of breath alcohol concentration: this is within our detected region. This study had also detected rs4838767 in the promoter region of CYP2E1 (position 133520114) associated with multiple alcohol consumption measures, but we did not detect the closest SNP to this as having non-zero effect on our response.

ALDH2: All 5 SNPs we study are close to exons, and get picked up by the e-value procedure. While all five are at a lesser base pair position than the well-known SNP rs671 (`https://www.snpedia.com/index.php/Rs671`, position 111803962), one of the SNPs we analyze (rs3742004) is within 5 kb of this SNP.

COMT: The SNP rs4680 has long been associated with schizophrenia and substance abuse, including alcoholism. We detect this SNP with an e-value of 0.144, as well as 8 other SNPs. Interestingly, a previous case-control study (Voisey et al., 2011) associated rs4680 and rs165774 with alcohol dependence through a SNP-wise chi-squared test, and had these two SNPs in high LD in their study population. Compared to this, in our simultaneous model of all COMT polymorphisms, rs165774 is one of the two SNPs with very high e-value.

SLCA6A4 and DRD2: Our analysis did not detect any of the SNPs in these genes having non-zero effect on alcohol consumption. Variants of these two genes have known interaction effects behind alcohol withdrawal-induced seizure (Karpyak et al., 2010) and bipolar disorder (Wang et al., 2014). For this reason we also ran the e-values procedure on the combined set of SNPs from these genes, but did not detect

any signal there as well for our sample.

3.3.7 Future work: incorporating group selection for GWAS

To expand the above approach to the full GWAS data, we need to incorporate strategies for dealing with the hierarchical structure of SNPs: there are a large number of genes in the human genome, and each of them contains a number of SNPs. Since our method requires the number of predictors to be less than number of sample size, it is plausible to start with an initial screening step to eliminate genes that are evident not relevant. Methods like the grouped Sure Independent Screening (Li et al., 2012) and min-P test (Westfall and Young, 1993) will be relevant here. Following this, in a multi-gene predictor set, there are several possible strategies to select important genes *and* important SNPs in them:

1. **Two-step e-values:** First construct multi-SNP models for each gene, trained on SNP data inside that gene and a common behavioral trait response and select SNPs in each model using e-values. Now train a modle using selected SNPs from all genes, and run group selection procedure in this model using e-values. This means dropping *groups* of predictors from the full model and checking the reduced model e-values. The one-step e-values method outlined in Chapter 2 will work here because of the same logic, and setting the groups as the collection of SNPs corresponding to a gene should be able to achieve our objective here.

2. **SNP-level e-values only:** First select important genes using an aggregation method of SNP-trait associations (e.g. Lamparter et al. (2016)) and run e-value based SNP selection on the set of SNPs within these genes.

3. **Gene-level e-values only:** Train separate models for each gene (after initial screening), select SNPs within those models using a fast screening method (e.g.

RFGLS) and run group-level or SNP-level e-value selection in that full set of SNPs.

We plan to study merits and demerits of these strategies and the computational issues associated with them in detail through synthetic studies as well as in the GWAS data from MCTFR.

Chapter 4

Signed Peripherality Functions in Multivariate Analysis

4.1 Introduction

Consider a real separable Hilbert space $\mathcal{H}$, and the following two functions. Firstly the *sign function* $S : \mathcal{H} \times \mathcal{H} \to \mathcal{H}$, which is defined as

$$S(x; \mu_x) = \frac{x - \mu_x}{\|x - \mu_x\|} \mathbb{I}_{x \neq \mu_x} \tag{4.1.1}$$

with respect to the *location parameter* $\mu_x \in \mathcal{H}$, and the norm $\|\cdot\|$ used above is the norm of the underlying Hilbert space. This is a direct generalization of the real-valued case of the indicator of whether the point x is to the right, left or at μ_x. This function had first been introduced by Möttönen and Oja (1995), and has seen widespread application in robust statistics across the past two decades (Locantore et al., 1999; Oja, 2010; Wang et al., 2015).

Next we describe the *peripherality function*, for which some mathematical preliminaries are necessary for easier exposition. Let $(\Omega, \mathcal{A}, \alpha)$ be a probability space, and let $\mathcal{B}$ be the Borel σ-algebra generated by the norm topology of $\mathcal{H}$. A $\mathcal{H}$-valued random variable is a mapping $X : \Omega \mapsto \mathcal{H}$ such that for every $B \in \mathcal{B}$, $X^{-1}(B) \in \mathcal{A}$. It

is easy to see that $\alpha_x = \alpha(X^{-1}(\cdot))$ is a probability measure on the measurable space $(\mathcal{H}, \mathcal{B})$. Mathematical details about such probability measures on Hilbert spaces are available from a number of places (Segal, 1958; Gross, 1967).

Let $\mathcal{M}$ be a set of probability measures on $\mathcal{H}$. A *peripherality function* $P : \mathcal{H} \times \mathcal{M} \to \mathbb{R}$, is a function that satisfies the following condition:
For every probability measure $F \in \mathcal{M}$, there exists a constant $\mu_F \in \mathcal{H}$ such that for every $t \in [0, 1]$ and every $x \in \mathcal{H}$

$$P\Big(\mu_F; F\Big) \leqslant P\Big(\mu_F + t(x - \mu_F); F\Big).$$

That is, for every fixed F, the peripherality function achieves a minimum at μ_F, and is non-decreasing in every direction away from μ_F. If we impose the practical restriction that $\inf_x P(x; F)$ is finite and bounded below, then we may as well impose without loss of generality $P(\mu_F; F) = 0$ and consequently $P(x; F) \geqslant 0$ for all $x \in \mathcal{H}$ and $F \in \mathcal{M}$. In many cases of interest, $P(\cdot; \cdot)$ is uniformly bounded above as well.

The peripherality function quantifies whether the point x is near or far from μ_F. We will impose additional conditions on this function as we proceed, but it can be seem immediately that any distance measure between x and μ_F satisfies the bare minimum requirement mentioned above.

In this chapter, we demonstrate interesting applications arising from composing the sign function and the peripherality function together, to form the *signed-peripherality function*. We define this function with three parameters $\mu_x \in \mathcal{H}$, $F \in \mathcal{M}$ and $\mu_y \in \mathcal{H}$, argument $x \in \mathcal{H}$ and range $\mathcal{H}$. More precisely, we use two functions $\kappa_s : \mathcal{H} \to \mathcal{H}$, $\kappa_p : \mathcal{H} \to \mathcal{H}$ that are respectively composed with the sign transformation and the peripherality function, and then multiplied together to obtain the

function $\kappa : \mathcal{H} \times \mathcal{H} \times \mathcal{M} \times \mathcal{H} \times \mathcal{H} \to \mathcal{H}$ defined as

$$\kappa(x; \mu_x, F, \mu_y) = \kappa_s(S(x; \mu_x))\kappa_p(P(x; F)) + \mu_y. \tag{4.1.2}$$

We have deliberately set the location parameters μ_x, μ_F, μ_y to be potentially non-identical, this additional flexibility has some advantage for robust data analysis. In many applications, the value of these three parameters may be identical, which leads to no conflict in our framework.

We are going to elaborate on the case when $\mathcal{H}$ is the p-dimensional Euclidean space $\mathbb{R}^p$ in (4.1.2) above, for some positive integer p. In this situation, a whole class of peripherality functions can be defined from *Data depth* functions. Peripherality functions can be defined as some inverse ranking based on data depth, and the concept of *outlyingness* associated with data depth (see Zuo and Serfling (2000)) is essentially same as what we use in this paper. Coming back to (4.1.2), we fix $\kappa_s(x) = x, \mu_y = \mathbf{0}_p$, and shall consider two separate choices of κ_p. In Section 4.2 we show that when κ_p is a monotonically decreasing function of its argument, it leads to favorable asymptotic and finite sample efficiency results in robust multivariate location estimation and high-dimensional testing. On the other hand, an opposite characterization of $\kappa_p(.)$, i.e. when it is an monotonically *increasing* function, results in better performance compared to the spatial sign-based principal component analysis (PCA) in Section 4.3, as well as robustification of Sufficient Dimension Reduction (Adragni and Cook, 2009) in Section 4.4 and functional PCA in Section 4.5.

4.2 The robust location problem

Consider an elliptic distribution in $\mathbb{R}^p$, denoted by $\mathcal{E}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$, for which we take the characterization of Fang et al. (1990) as given in Chapter 2. In this section we

focus on the problem of estimation and testing for the location parameter $\boldsymbol{\mu}$ in this distribution using data-dependent weights on the spatial sign vectors:

$$\mathbf{X}_w = w(\mathbf{X})\mathbf{S}(\mathbf{X})$$

where $\mathbf{S}(\mathbf{x}) \equiv \mathbf{S}(\mathbf{x}; \mathbf{0}_p) = \|\mathbf{x}\|^{-1}\mathbf{x}.\mathbb{I}_{\mathbf{x}\neq\mathbf{0}_p}$, adapting the definition of spatial signs in (4.1.1) for $\mathbb{R}^p$. For now the only condition we impose on these weights, say $w(.)$, is that they need to be scalar-valued affine invariant and square-integrable functions of $\mathbf{X}$, or equivalently of the norm of the standardized random variable $\mathbf{Z} \equiv \boldsymbol{\Sigma}^{-1/2}(\mathbf{X}-\boldsymbol{\mu})$. In other words, it is possible to write $w(\mathbf{X})$ as $f(r)$, with $r = \|\mathbf{Z}\|$. Our theoretical analysis in this section assumes this general weighs structure. The role of peripherality functions *vis-á-vis* the characterization in (4.1.2) comes in the form of empirical evidence, where we demonstrate better performance compared to spatial sign-based procedures when $f(r)$ is taken as a decreasing function of r.

The simplest use of weighted signs in the location problem would be to construct an outlier-robust alternative to the Hotelling's T^2 test using their sample mean vector and covariance matrix. Formally, given a size-n sample $\mathbb{X}_n = (\mathbf{X}_1, \ldots, \ldots \mathbf{X}_n)^T$ of independent and identically distributed (as $\mathbf{X}$) random variables, this means testing for $H_0 : \boldsymbol{\mu} = \mathbf{0}_p$ vs. $H_1 : \boldsymbol{\mu} \neq \mathbf{0}_p$ based on the test statistic:

$$T_{n,w} = n\bar{\mathbf{X}}_w^T(Cov(\mathbf{X}_w))^{-1}\bar{\mathbf{X}}_w$$

with $\bar{\mathbf{X}}_w = \sum_{i=1}^n \mathbf{X}_{w,i}/n$ and $\mathbf{X}_{w,i} = w(\mathbf{X}_i)\mathbf{S}(\mathbf{X}_i)$ for $i = 1, 2, ..., n$. However, the following holds true for this weighted sign test:

Proposition 4.2.1. *Consider n random variables $\mathbb{Z}_n = (\mathbf{Z}_1, ..., \mathbf{Z}_n)^T$ distributed independently and identically as $\mathcal{E}(\boldsymbol{\mu}, k\mathbf{I}_p, g); k > 0$, and the class of hypothesis tests defined above. Then, given any $\alpha \in (0, 1)$, local power at $\boldsymbol{\mu} \neq \mathbf{0}_p$ for the level-α test*

based on $T_{n,w}$ is maximum when $w(\mathbf{Z}_1) = c$, a constant independent of $\mathbf{Z}_1$.

This essentially means that power-wise the (unweighted) spatial sign test (Oja, 2010) is optimal in the given class of hypothesis tests when the data comes from a spherically symmetric distribution. Our simulations show that this empirically holds for non-spherical elliptic distributions as well.

4.2.1 The weighted spatial median

In order to explore usage of weighted spatial signs in the location problem that improve upon the state-of-the-art, we now concentrate on the following optimization problem:

$$\boldsymbol{\mu}_w = \arg\min_{\boldsymbol{\mu}_0 \in \mathbb{R}^p} \mathbb{E}(w(\mathbf{X})|\mathbf{X} - \boldsymbol{\mu}_0|) \tag{4.2.1}$$

This can be seen as a generalization of the Fermat-Weber location problem, which has the spatial median (Brown, 1983; Chaudhuri, 1996) as its solution, using data-dependent weights. Using affine invariant weights in (4.2.1) ensures that the weights are independent of $\boldsymbol{\mu}_0$, which allows the optimization problem to have a unique solution. We call this solution the *weighted spatial median* of F, and denote it by $\boldsymbol{\mu}_w$. In a sample setup it is estimated by iteratively solving the equation $\sum_{i=1}^n w(\mathbf{X}_i)\mathbf{S}(\mathbf{X}_i - \hat{\boldsymbol{\mu}}_w)/n = \mathbf{0}_p$.

The sample weighted spatial median $\hat{\boldsymbol{\mu}}_w$ is a $\sqrt{n}$-consistent estimator of $\boldsymbol{\mu}_w$, and gives its asymptotic distribution:

Theorem 4.2.2. *Let $\mathbf{A}_w, \mathbf{B}_w$ be two matrices, dependent on the weight function w such that*

$$\mathbf{A}_w = \mathbb{E}\left[\frac{w(\boldsymbol{\epsilon})}{\|\boldsymbol{\epsilon}\|}\left(1 - \frac{\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T}{\|\boldsymbol{\epsilon}\|^2}\right)\right]; \quad \mathbf{B}_w = \mathbb{E}\left[\frac{(w(\boldsymbol{\epsilon}))^2\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T}{\|\boldsymbol{\epsilon}\|^2}\right]$$

	t_3	t_5	t_{10}	t_{20}	Normal
$p = 5$	1.28	1.20	1.16	1.14	1.13
$p = 10$	1.15	1.10	1.07	1.07	1.06
$p = 20$	1.09	1.05	1.04	1.03	1.03
$p = 50$	1.05	1.02	1.01	1.01	1.01

Table 4.1: Table of $ARE(\boldsymbol{\mu}_w; \boldsymbol{\mu}_s)$ for different spherical distributions

where $\boldsymbol{\epsilon} \sim \mathcal{E}(\mathbf{0}_p, \boldsymbol{\Sigma}, g)$. *Then*

$$\sqrt{n}(\hat{\boldsymbol{\mu}}_w - \boldsymbol{\mu}_w) \rightsquigarrow N_p(\mathbf{0}_p, \mathbf{A}_w^{-1}\mathbf{B}_w\mathbf{A}_w^{-1}) \tag{4.2.2}$$

The above theorem generalizes equivalent results for the spatial median (Oja, 2010), and can be proved in a similar fashion. Note that setting $w(\boldsymbol{\epsilon}) = 1$ above yields the asymptotic covariance matrix for the spatial median. Following this, the asymptotic relative efficiency (ARE) of $\boldsymbol{\mu}_w$ corresponding to some non-uniform weight function with respect to the spatial median, say $\boldsymbol{\mu}_s$ will be:

$$ARE(\boldsymbol{\mu}_w, \boldsymbol{\mu}_s) = \left[\frac{\det(\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1})}{\det(\mathbf{A}_w^{-1}\mathbf{B}_w\mathbf{A}_w^{-1})}\right]^{1/p} \tag{4.2.3}$$

with $\mathbf{A} = \mathbb{E}[1/\|\boldsymbol{\epsilon}\|(\mathbf{I}_p - \boldsymbol{\epsilon}\boldsymbol{\epsilon}^T/\|\boldsymbol{\epsilon}\|^2)]$ and $\mathbf{B} = \mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T/\|\boldsymbol{\epsilon}\|^2]$. This is further simplified under spherical symmetry:

Corollary 4.2.3. *For a spherical distribution* $\mathcal{E}(\boldsymbol{\mu}, k\mathbf{I}_p, g); k \in \mathbb{R}, \boldsymbol{\mu} \in \mathbb{R}^p$, *we have*

$$ARE(\boldsymbol{\mu}_w, \boldsymbol{\mu}_s) = \frac{\left[\mathbb{E}\left(\frac{f(r)}{r}\right)\right]^2}{\mathbb{E}f^2(r)\left[\mathbb{E}\left(\frac{1}{r}\right)\right]^2}$$

At this point, choices of weights that are decreasing functions of r lead ARE values larger than 1. For example, Table 4.1 summarizes the AREs for several families of elliptic distributions, numerically calculated using 10,000 random samples, and taking $f(r) = 1/(1 + r)$. It is evident from the table that the weighted spatial median out-

performs its unweighted counterpart for all data dimensions and distribution families considered. While the performance is much better for small values of p, weighting the signs seems to have less and less effect as p grows larger. Assuming a first order autoregressive (AR1) covariance structure, i.e. $\mathbf{\Sigma}_{ij} = \rho^{|i-j|}, \rho \in (0,1)$ also results in largely similar ARE values as those obtained in Table 4.1 which assume $\mathbf{\Sigma} = \mathbf{I}_p$.

4.2.2 A high-dimensional test of location

It is possible to take an alternative approach to the location testing problem by using the covariance-type U-statistic $C_{n,w} = \sum_{i=1}^{n} \sum_{j=1}^{i-1} \mathbf{X}_{w,i}^T \mathbf{X}_{w,j}$. This class of test statistics are especially attractive since they are readily generalized to cover high-dimensional situations, i.e. when $p > n$. The Chen and Qin (CQ) high-dimensional test of location for multivariate normal $\mathbf{X}_i$ (Chen and Qin, 2010) is a special case of this test that uses the statistic $C_n = \sum_{i=1}^{n} \sum_{j=1}^{i-1} \mathbf{X}_i^T \mathbf{X}_j$, and a recent paper (Wang et al. (2015), from here on referred to as WPL test) shows that one can improve upon the power of the CQ test for non-gaussian elliptical distributions by using spatial signs $\mathbf{S}(\mathbf{X}_i)$ in place of the actual variables.

Given these, and some mild regularity conditions, the following holds for our generalized test statistic $C_{n,w}$ under H_0 as $n, p \to \infty$:

$$\frac{C_{n,w}}{\sqrt{\frac{n(n-1)}{2}\mathrm{Tr}(\mathbf{B}_w^2)}} \rightsquigarrow N(0,1) \tag{4.2.4}$$

and under contiguous alternatives $H_1 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$,

$$\frac{C_{n,w} - \frac{n(n-1)}{2}\boldsymbol{\mu}_0^T \mathbf{A}_w^2 \boldsymbol{\mu}_0 (1 + o(1))}{\sqrt{\frac{n(n-1)}{2}\mathrm{Tr}(\mathbf{B}_w^2)}} \rightsquigarrow N(0,1) \tag{4.2.5}$$

we provide the details behind deriving these two results in the supplementary material,

$\boldsymbol{\mu} = \text{rep}(.15, p)$				
p	n	CQ	WPL	$C_{n,w}$
500	20	0.051	0.376	0.418
500	50	0.060	0.832	0.866
1000	20	0.044	0.541	0.584
1000	50	0.039	0.973	0.987
$\boldsymbol{\mu} = \text{rep}(0, p)$				
p	n	CQ	WPL	$C_{n,w}$
500	20	0.049	0.061	0.063
500	50	0.039	0.061	0.064
1000	20	0.042	0.060	0.063
1000	50	0.043	0.050	0.050

Table 4.2: Table of empirical powers of level-0.05 tests for the Chen and Qin (CQ), WPL and $C_{n,w}$ statistics

which involve modified regularity conditions and sketches of proofs along the lines of Wang et al. (2015).

The ARE of this test statistic with respect to its unweighted version, i.e. the WPL statistic, is expressed as:

$$ARE(C_{n,w}, \text{WPL}; \boldsymbol{\mu}_0) = \frac{\boldsymbol{\mu}_0^T \mathbf{A}_w^2 \boldsymbol{\mu}_0}{\boldsymbol{\mu}_0^T \mathbf{A}^2 \boldsymbol{\mu}_0} \sqrt{\frac{\text{Tr}(\mathbf{B}^2)}{\text{Tr}(\mathbf{B}_w^2)}} (1 + o(1))$$

when $\boldsymbol{\Sigma} = k\mathbf{I}_p$ and $f(r) = 1/(1+r)$, this again simplifies to $\mathbb{E}^2(f(r)/r)/[\mathbb{E}f^2(r).\mathbb{E}^2(1/r)]$. The ARE values will be exactly same as those in Table 4.1, which indicates that for large data dimension the WPL test and that based on $C_{n,w}$ are almost equivalent.

However, in a practical high-dimensional setup one almost always has to work with a low sample size. For this reason, comparing the the two tests with respect to their *finite sample* efficiencies instead should give a better idea of their practical utility. We do this in Table 4.2, which lists empirical powers calculated from 1000 replications of each setup under an AR1 covariance structure (with $\rho = 0.8$). While under $H_0 : \boldsymbol{\mu} = \mathbf{0}_p$ all tests have similar performance, $C_{n,w}$ beats the other two under deviations from H_0.

4.3 Depth-based rank covariance matrix

We shall now focus on scatter functionals of the weighted sign vectors $\mathbf{X}_w$ defined in the previous section. For this purpose, given a measure of data depth $D(., F)$ we take the weights to be any monotonically decreasing transformation on that depth function which takes values in $[0, M]$ for some $M < \infty$. We call this an *inverse depth* function, and denote it by $D^-(\mathbf{x}, F)$ for $\mathbf{x} \in \mathbb{R}^p$. With respect to the formulation of (4.1.2) this corresponds to an affine invariant peripherality function paired with a nonnegative-valued monotonically *increasing* κ_p that is bounded above. Some examples of inverse depth functions include but are not limited to $D^-(\mathbf{x}, F) := \max_{\mathbf{x}} D(\mathbf{x}, F) - D(\mathbf{x}, F)$ and $D^-(\mathbf{x}, F) := \exp(-D(\mathbf{x}, F))$.

In the analysis that follows, we shall assume the max definition of $D^-(\mathbf{x}, F)$ above, i.e. $D^-(\mathbf{x}, F) = \max_{\mathbf{x}} D(\mathbf{x}, F)$ for ease of representation, although all the analysis go through in exactly the same fashion with other definitions. Also we slightly tweak the notations to make things this section onwards easier to follow.Data depth is as much a property of a vector-valued random variable $\mathbf{X} \in \mathbb{R}^p$ as it is of the underlying distribution $F \equiv [\mathbf{X}]$, so from now on we shall be using $D(\mathbf{x}, [\mathbf{X}])$ and $D(\mathbf{x}, F)$ intermittently to denote the depth of a point $\mathbf{x}$. We expand this notation to inverse depths as well (i.e. $D^-(\mathbf{x}, [\mathbf{X}]) \equiv D^-(\mathbf{x}, F)$ etc.).

Now, given the weights $w(\mathbf{x}) = D^-(\mathbf{x}, [\mathbf{X}])$, we can write the transformation of any point $\mathbf{x} \in \mathbb{R}^p$ as:

$$\tilde{\mathbf{x}} = D^-(\mathbf{x}, [\mathbf{X}])\mathbf{S}(\mathbf{x} - \boldsymbol{\mu}) \tag{4.3.1}$$

with $\mathbf{S}(.)$ being the spatial sign functional. The transformed random variable, say $\tilde{\mathbf{X}}$, can be seen as the multivariate rank corresponding to $\mathbf{X}$ (e.g. Serfling (2006)). The notion of multivariate ranks goes back to Puri and Sen (1971), where they take the vector consisting of marginal univariate ranks as multivariate rank vector. Subsequent

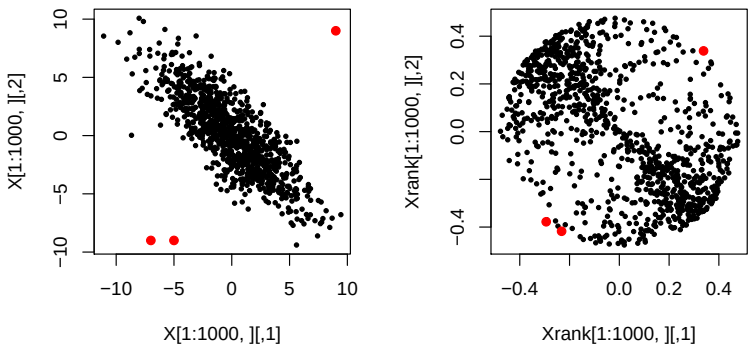


Figure 4.1: (Left) 1000 points randomly drawn from $\mathcal{N}_2\left((0,0)^T, \left(\begin{smallmatrix} 5 & -4 \\ -4 & 5 \end{smallmatrix}\right)\right)$ and (Right) their multivariate ranks based on halfspace depth

definitions of multivariate ranks were proposed by Möttönen and Oja (1995); Hallin and Paindaveine (2002) and Chernozhukov et al. (2017).

Compared to these previous formulations, our definition of multivariate ranks provides more intuitive representation of the transformation applied on the data. Figure 4.1 gives an idea of how our rank vector $\tilde{\mathbf{X}}$ is distributed when $\mathbf{X}$ has a bivariate normal distribution. Compared to the spatial sign, which are distributed on the surface of p-dimensional unit ball centered at $\boldsymbol{\mu}$, these spatial ranks have the same direction as original data and reside *inside* the p-dimensional ball around $\boldsymbol{\mu}$ with a finite radius (the choice of the radius depends on the inverse transformation used: e.g. for the case of max transformation, this radius is $\max_{\mathbf{x}} D(\mathbf{x}, [\mathbf{X}])$). As a result, the rank transformation preserves the shape of the data more effectively.

Now consider the spectral decomposition for the covariance matrix of F: $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^T$, $\boldsymbol{\Gamma}$ being orthogonal and $\boldsymbol{\Lambda}$ diagonal with positive diagonal elements. Also normalize the original random variable as $\mathbf{z} = \boldsymbol{\Gamma}^T\boldsymbol{\Lambda}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$. In this setup, we can

represent the transformed random variable as

$$\begin{aligned}\tilde{\mathbf{x}} &= \tilde{D}^{-}(\mathbf{x},[\mathbf{X}])\mathbf{S}(\mathbf{x}-\boldsymbol{\mu}) \\ &= D^{-}(\boldsymbol{\Gamma}\boldsymbol{\Lambda}^{1/2}\mathbf{z}+\boldsymbol{\mu},[\boldsymbol{\Gamma}\boldsymbol{\Lambda}^{1/2}\mathbf{Z}+\boldsymbol{\mu}]).\mathbf{S}(\boldsymbol{\Gamma}\boldsymbol{\Lambda}^{1/2}\mathbf{z}) \\ &= D^{-}(\mathbf{z},[\mathbf{Z}]).\boldsymbol{\Gamma}\mathbf{S}(\boldsymbol{\Lambda}^{1/2}\mathbf{z}) \\ &= \left[\boldsymbol{\Gamma}\boldsymbol{\Lambda}^{1/2}\frac{\|\mathbf{z}\|}{\|\boldsymbol{\Lambda}^{1/2}\mathbf{z}\|}\right].D^{-}(\mathbf{z},[\mathbf{Z}])\mathbf{S}(\mathbf{z}) \qquad (4.3.2)\end{aligned}$$

following affine invariance of D hence D^{-}. Now $D^{-}(\mathbf{z},[\mathbf{Z}])$ is an even function in $\mathbf{z}$ because of affine invariance, as is $\|\mathbf{z}\|/\|\boldsymbol{\Lambda}^{1/2}\mathbf{z}\|$. Since $\mathbf{S}(\mathbf{z})$ is odd in $\mathbf{z}$ for spherically symmetric $\mathbf{z}$, it follows that $\mathbb{E}\tilde{\mathbf{X}}=\mathbf{0}_p$. Consequently we obtain an expression for the covariance matrix of $\tilde{\mathbf{X}}$:

Theorem 4.3.1. *Let the random variable $\mathbf{X}\in\mathbb{R}^p$ follow an elliptical distribution with center $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}=\boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^T$, its spectral decomposition. Then, given a depth function $D(.)$, the covariance matrix of the transformed random variable $\tilde{\mathbf{X}}$ is*

$$\mathbb{V}(\tilde{\mathbf{X}})=\boldsymbol{\Gamma}\tilde{\boldsymbol{\Lambda}}\boldsymbol{\Gamma}^T, \quad \textit{with} \quad \tilde{\boldsymbol{\Lambda}}=\mathbb{E}_{\mathbf{Z}}\left[(D^{-}(\mathbf{z},[\mathbf{Z}]))^2\frac{\boldsymbol{\Lambda}^{1/2}\mathbf{z}\mathbf{z}^T\boldsymbol{\Lambda}^{1/2}}{\mathbf{z}^T\boldsymbol{\Lambda}\mathbf{z}}\right] \qquad (4.3.3)$$

where $\mathbf{Z}=(Z_1,...,Z_p)^T\sim\mathcal{N}(\mathbf{0},\mathbf{I}_p)$, so that $\tilde{\boldsymbol{\Lambda}}$ is a diagonal matrix with diagonal entries

$$\tilde{\lambda}_i=\mathbb{E}_{\mathbf{Z}}\left[\frac{(D^{-}(\mathbf{z},[\mathbf{Z}]))^2\lambda_i z_i^2}{\sum_{j=1}^{p}\lambda_j z_j^2}\right] \qquad (4.3.4)$$

We call $\tilde{\boldsymbol{\Sigma}}:=\mathbb{V}(\tilde{\mathbf{X}})$ the Depth Covariance Matrix (DCM). Notice that the matrix of eigenvectors of the covariance matrix of $\mathbf{X}$, i.e. $\boldsymbol{\Gamma}$, remains unchanged in the transformation $\mathbf{X}\mapsto\tilde{\mathbf{X}}$. As a result, the multivariate rank vectors can be used for robust principal component analysis, which we are going to discuss shortly. However,

as one can see in the above expression, the diagonal entries of $\tilde{\mathbf{\Lambda}}$ do not change if a scale change is done on all entries of $\mathbf{\Lambda}$, meaning the $\tilde{\mathbf{\Lambda}}$ matrices corresponding to F and cF for some $c \neq 0$ will be same. Thus the DCM is not equivariant under affine transformations.

We need to follow the general framework of M-estimation with data-dependent weights in Huber (1981) to construct an affine equivariant counterpart of the DCM. Specifically, we implicitly define the Affine-equivariant Depth Covariance Matrix (ADCM) as

$$\tilde{\mathbf{\Sigma}}_{\mathrm{o}} = \frac{1}{\mathbb{V}(\tilde{Z}_1)}\mathbb{E}\left[\frac{(D^{-}(\mathbf{x},[\mathbf{X}]))^2(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T}{(\mathbf{x}-\boldsymbol{\mu})^T\tilde{\mathbf{\Sigma}}_{\mathrm{o}}^{-1}(\mathbf{x}-\boldsymbol{\mu})}\right] \tag{4.3.5}$$

Its affine equivariance follows from the fact that the weights $D^{-}(\mathbf{x},[\mathbf{X}])$ depend only on the standardized quantities $\mathbf{z}$ that come from the underlying spherical distribution G. We solve (4.3.5) iteratively by obtaining a sequence of positive definite matrices $\tilde{\mathbf{\Sigma}}_{\mathrm{o}}^{(k)}$ until convergence:

$$\tilde{\mathbf{\Sigma}}_{\mathrm{o}}^{(k+1)} = \frac{1}{\mathbb{V}(\tilde{Z}_1)}\mathbb{E}\left[\frac{(D^{-}(\mathbf{x},[\mathbf{X}]))^2(\tilde{\mathbf{\Sigma}}_{\mathrm{o}}^{(k)})^{1/2}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T(\tilde{\mathbf{\Sigma}}_{\mathrm{o}}^{(k)})^{1/2}}{(\mathbf{x}-\boldsymbol{\mu})^T(\tilde{\mathbf{\Sigma}}_{\mathrm{o}}^{(k)})^{-1}(\mathbf{x}-\boldsymbol{\mu})}\right]$$

To ensure existence and uniqueness of this estimator, let us consider the class of scatter estimators $\mathbf{\Sigma}_M$ that are obtained as solutions of the following equation:

$$\mathbb{E}_{\mathbf{Z}_M}\left[u(\|\mathbf{z}_M\|)\frac{\mathbf{z}_M\mathbf{z}_M^T}{\|\mathbf{z}_M\|^2} - v(\|\mathbf{z}_M\|)\mathbf{I}_p\right] = 0 \tag{4.3.6}$$

with $\mathbf{z}_M = \mathbf{\Sigma}_M^{-1/2}(\mathbf{x}-\boldsymbol{\mu})$. The above equation produces a unique solution under the following assumptions on the scalar-valued functions u and v (Huber, 1981):

(M1) The function $u(r)/r^2$ is monotone decreasing, and $u(r) > 0$ for $r > 0$;

(M2) The function $v(r)$ is monotone decreasing, and $v(r) > 0$ for $r > 0$;

(M3) Both $u(r)$ and $v(r)$ are bounded and continuous;

(M4) $u(0)/v(0) < p$;

(M5) For any hyperplane in the sample space $\mathcal{X}$, (i) $P(H) = \mathbb{E}_{\mathbf{X}}\mathbb{I}_{\mathbf{x}\in H} < 1 - pv(\infty)/u(\infty)$ and (ii) $P(H) \leqslant 1/p$.

In our case we take $v(r) = \mathbb{V}(\tilde{Z}_1)$, i.e. a constant, thus (M2) and (M3) are trivially satisfied. As for u, we notice that most well-known depth functions can be expressed as simple functions of the norm of the standardized random variable. For example, $PD(\mathbf{z}, [\mathbf{Z}]) = 1 - F_{01}(\|\mathbf{z}\|); MhD(\mathbf{z}, [\mathbf{Z}]) = (1 + \|\mathbf{z}\|^2)^{-1}; HD(\mathbf{z}, [\mathbf{Z}]) = (1 + \|\mathbf{z}\|/MAD(F_{01}))^{-1}$ etc., where $F_{01} \equiv \mathcal{E}(0, 1, g)$, and MAD is median absolute deviation. Thus we can take u as square of the corresponding inverse depth functions:

$$u_{PD}(r) = F_{01}^2(r); \quad u_{MhD}(r) = \frac{r^4}{(1+r^2)^2}; \quad u_{HSD}(r) = \frac{r^2}{(1 + r/MAD(F_{01}))^2}$$

It is easy to verify that the above choices of u satisfy (M1) and (M3). To check (M4) and (M5), first notice that $\mathbf{Z}$ has a spherically symmetric distribution, so that its norm and sign are independent. Since $D(\mathbf{z}, [\mathbf{Z}])$ depends only on $\|\mathbf{z}\|$, we have

$$\mathbb{V}(\tilde{Z}_1) = \mathbb{V}\left(D^-(\mathbf{Z}, [\mathbf{Z}])\frac{Z_1}{\|\mathbf{Z}\|}\right) = \mathbb{V}(D^-(\mathbf{Z}, [\mathbf{Z}]))\mathbb{V}(S_1(\mathbf{Z})) = \frac{1}{p}\mathbb{V}(D^-(\mathbf{Z}, [\mathbf{Z}]))$$

as $\mathbb{V}(\mathbf{S}(\mathbf{Z})) = \mathbb{V}((S_1(\mathbf{Z}), S_2(\mathbf{Z}), ..., S_p(\mathbf{Z}))^T) = \mathbf{I}_p/p$. Now for MhD and HD $u(\infty) = 1, u(0) = 0$, so (M4) and (M5) are immediate. To achieve this for PD, we only need to replace $u_{PD}(r)$ with $u^*_{PD}(r) = F_{01}^2(r) - 1/4$.

4.3.1 Calculating the sample DCM and ADCM

Let us denote $\mathbb{S}(\mathbf{x}; \boldsymbol{\mu}) = \mathbf{S}(\mathbf{x} - \boldsymbol{\mu})\mathbf{S}(\mathbf{x} - \boldsymbol{\mu})^T$. Then, given the depth function and known location center $\boldsymbol{\mu}$, one can show that the vectorized form of $\sqrt{n}$-times the

sample DCM, i.e. $(1/\sqrt{n})\sum_{i=1}^{n}(D^{-}(\mathbf{x}_i,[\mathbf{X}]))^2\mathbb{S}(\mathbf{x}_i;\boldsymbol{\mu})$ has an asymptotic multivariate normal distribution with mean $\sqrt{n}.\,\mathrm{vec}(\mathbb{E}[(D^{-}(\mathbf{X},[\mathbf{X}]))^2\mathbb{S}(\mathbf{X};\boldsymbol{\mu})])$ and a certain covariance matrix by straightforward application of the Central Limit Theorem (CLT). But in practice the population depth function $D(\mathbf{x},[\mathbf{X}])$ is estimated by the depth function based on the empirical distribution function $[\mathbb{X}_n]$, i.e. by $D(\mathbf{x},[\mathbb{X}_n])$ (recall from Section 4.2 that $\mathbb{X}_n = (\mathbf{X}_1,\ldots\mathbf{X}_n)^T$). Here we make the following assumption regarding how $D^{-}(\mathbf{x},[\mathbf{X}])$ is approximated by its sample counterpart:

(P5) *Uniform convergence*: $\sup_{\mathbf{x}\in\mathbb{R}^p}|D(\mathbf{x},[\mathbb{X}_n]) - D(\mathbf{x},[\mathbf{X}])| \to 0$ as $n\to\infty$.

The assumption that empirical depths converge uniformly at all points $\mathbf{x}$ to their population versions holds under very mild conditions for several well known depth functions: for example projection depth (Zuo, 2003) and simplicial depth (Dumbgen, 1992). One also needs to replace the known location parameter $\boldsymbol{\mu}$ by some estimator $\hat{\boldsymbol{\mu}}_n$. Examples of robust estimators of location that are relevant here include the spatial median (Haldane, 1948; Brown, 1983), Oja median (Oja, 1983), projection median (Zuo, 2003) etc. Now, given $D(.,[\mathbb{X}_n])$ and $\hat{\boldsymbol{\mu}}_n$, to plug them into the sample DCM and still go through with the CLT we need the following result:

Lemma 4.3.2. *Consider a random variable $\mathbf{X}\in\mathbb{R}^p$ having a continuous and symmetric distribution with location center $\boldsymbol{\mu}$ such that $\mathbb{E}\|\mathbf{x}-\boldsymbol{\mu}\|^{-3/2} < \infty$. Given n random samples from this distribution, suppose $\hat{\boldsymbol{\mu}}_n$ is an estimator of $\boldsymbol{\mu}$ so that $\sqrt{n}(\hat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}) = O_P(1)$. Then with the above notations, and given the assumption (D5) we have*

$$\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n}(D^{-}(\mathbf{x}_i,[\mathbb{X}_n]))^2\mathbb{S}(\mathbf{x}_i;\hat{\boldsymbol{\mu}}_n) - \frac{1}{n}\sum_{i=1}^{n}(D^{-}(\mathbf{x}_i,[\mathbf{X}]))^2\mathbb{S}(\mathbf{x}_i;\boldsymbol{\mu})\right] \xrightarrow{P} 0$$

Following this, we are now in a position to state the result for consistency of the sample DCM:

Theorem 4.3.3. *Consider n iid samples from the distribution in Lemma 4.3.2. Then, given a depth function $D(., [\mathbf{X}])$ and an estimate of center $\hat{\boldsymbol{\mu}}_n$ so that $\sqrt{n}(\hat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}) = O_P(1)$,*

$$\sqrt{n}\left[\text{vec}\left\{\frac{1}{n}\sum_{i=1}^{n}(D^{-}(\mathbf{x}, [\mathbb{X}_n]))^2\mathbb{S}(\mathbf{x}_i; \hat{\boldsymbol{\mu}}_n)\right\} - \mathbb{E}\left[vec\left\{(D^{-}(\mathbf{x}, [\mathbf{X}]))^2\mathbb{S}(\mathbf{x}; \boldsymbol{\mu})\right\}\right]\right] \quad (4.3.7)$$
$$\rightsquigarrow \mathcal{N}_{p^2}(\mathbf{0}, \tilde{\mathbf{V}}(F))$$

with $\tilde{\mathbf{V}}(F) = \mathbb{V}\left[\text{vec}\left\{(D^{-}(\mathbf{x}_i, [\mathbf{X}]))^2\mathbb{S}(\mathbf{x}; \boldsymbol{\mu})\right\}\right]$.

This holds for any general non-degenerate F. In case F is elliptical, an elaborate form of the covariance matrix $\tilde{\mathbf{V}}_F$ explicitly specifying each of its elements (more directly those of its $\boldsymbol{\Gamma}^T$-rotated version) can be obtained, which is given in Subection 4.7.1. This form is useful when deriving limiting distributions of eigenvectors and eigenvalues of the sample DCM.

In contrast to the DCM, the issue of estimating $\boldsymbol{\mu}$ to plug into the ADCM is easily handled by simultaneously solving for the location and scatter functionals $(\boldsymbol{\mu}_{\text{o}}, \tilde{\boldsymbol{\Sigma}}_{\text{o}})$:

$$\mathbb{E}\left[\frac{\tilde{\boldsymbol{\Sigma}}_{\text{o}}^{-1/2}(\mathbf{x} - \boldsymbol{\mu}_{\text{o}})}{\|\tilde{\boldsymbol{\Sigma}}_{\text{o}}^{-1/2}(\mathbf{x} - \boldsymbol{\mu}_{\text{o}})\|}\right] = \mathbf{0}_p \quad (4.3.8)$$

$$\mathbb{E}\left[\frac{(D^{-}(\mathbf{x}, [\mathbf{X}]))^2\tilde{\boldsymbol{\Sigma}}_{\text{o}}^{-1/2}(\mathbf{x} - \boldsymbol{\mu}_{\text{o}})(\mathbf{x} - \boldsymbol{\mu}_{\text{o}})^T\tilde{\boldsymbol{\Sigma}}_{\text{o}}^{-1/2}}{(\mathbf{x} - \boldsymbol{\mu}_{\text{o}})^T\tilde{\boldsymbol{\Sigma}}_{\text{o}}^{-1}(\mathbf{x} - \boldsymbol{\mu}_{\text{o}})}\right] = \mathbb{V}(\tilde{Z}_1)\mathbf{I}_p \quad (4.3.9)$$

In the general framework of (4.3.5), for any fixed $\boldsymbol{\Sigma}_M$ there exists a unique and fixed solution of the location problem $\mathbb{E}_{\mathbf{Z}_M}(w(\|\mathbf{z}_M\|\mathbf{z}_M) = \mathbf{0}_p$ under the following condition:

(M6) The function $w(r)r$ is monotone increasing for $r > 0$.

This condition is easy to verify for our choice of the weights: $w(\|\mathbf{z}_M\|) = \tilde{D}(\mathbf{z}_M, [\mathbf{Z}_M])/\|\mathbf{z}_M\|$. Consequently, uniqueness of any simultaneous fixed point solution of (4.3.8) and (4.3.9) is guaranteed when $\mathbf{X}$ has a symmetric distribution (Huber, 1981).

In practice it is difficult to calculate the scale multiple $\mathbb{V}(\tilde{Z}_1)$ analytically for known depth functions and an arbitrary F. Hence we instead calculate the standardized version if the ADCM: $\tilde{\mathbf{\Sigma}}_{\mathrm{o}}^{*} = \tilde{\mathbf{\Sigma}}_{\mathrm{o}}/\mathbb{V}(\tilde{Z}_1)$ (so that the determinant equals 1), alongwith $\boldsymbol{\mu}_{\mathrm{o}}$ using the following iterative algorithm:

1. Start from some initial estimates $(\boldsymbol{\mu}_{\mathrm{o}}^{(0)}, \tilde{\mathbf{\Sigma}}_{\mathrm{o}}^{(0)})$. Set $t = 0$;

2. Calculate the standardized observations $\mathbf{z}_i^{(t)} = (\tilde{\mathbf{\Sigma}}_{\mathrm{o}}^{(t)})^{-1/2}(\mathbf{x}_i - \boldsymbol{\mu}_{\mathrm{o}}^{(t)})$;

3. Update the location estimate:

$$\boldsymbol{\mu}_{\mathrm{o}}^{(t+1)} = \frac{\sum_{i=1}^{n} \tilde{\mathbf{x}}_i / \|\mathbf{z}_i^{(t)}\|}{\sum_{i=1}^{n} 1/\|\mathbf{z}_i^{(t)}\|}$$

4. Update the scatter estimate:

$$\begin{aligned} \tilde{\mathbf{\Sigma}}_{\mathrm{o}}^{*(t+1)} &= \frac{1}{n}\sum_{i=1}^{n} \frac{(D^{-}(\mathbf{x}_i, [\mathbb{X}_n]))^2(\mathbf{x}_i - \boldsymbol{\mu}_{\mathrm{o}}^{(t+1)})(\mathbf{x}_i - \boldsymbol{\mu}_{\mathrm{o}}^{(t+1)})^T}{\|\mathbf{z}_i^{(t)}\|^2} \\ \tilde{\mathbf{\Sigma}}_{\mathrm{o}}^{*(t+1)} &\leftarrow \frac{\tilde{\mathbf{\Sigma}}_{\mathrm{o}}^{*(t+1)}}{\det(\tilde{\mathbf{\Sigma}}_{\mathrm{o}}^{*(t+1)})^{1/p}} \end{aligned}$$

5. Continue until convergence.

Notice that owing to the uniform convergence property we can safely replace $D^{-}(\mathbf{x}_i, [\mathbf{X}])$ with its sample version and use the iterative algorithm above to obtain a consistent estimate of the of the solution of (4.3.9).

4.3.2 Robust PCA using eigenvectors of DCM

We shall now elaborate on using the DCM for robust principal components analysis. From now on we assume that the eigenvalues of $\mathbf{\Sigma}$ are distinct: $\lambda_1 > \lambda_2 > ... > \lambda_p$ to obtain asymptotic distributions of its eigenvectors. In case any of the eigenvalues have

multiplicity larger than 1, limiting distributions of the corresponding eigenprojection matrices can be obtained analogous to those of the sign covariance matrix (Magyar and Tyler, 2014).

Influence functions

We start by deriving the influence functions for eigenvectors of the DCM and ADCM. This will help in demonstrating the robustness of their estimates, as well as deriving their asymptotic efficiencies. Influence functions of the DCM as well as its eigenvectors and eigenvalues, which are essential to understand how much influence a sample point, especially an infinitesimal contamination, has on any functional on the distribution (Hampel et al., 1986). Given any probability distribution F, the influence function of any point $\mathbf{x}_0$ in the sample space $\mathcal{X}$ for some functional $T(F)$ on the distribution is defined as

$$IF(\mathbf{x}_0; T, F) = \lim_{\epsilon \to 0} \frac{1}{\epsilon}(T(F_\epsilon) - T(F))$$

where F_ϵ is F with an additional mass of ϵ at $\mathbf{x}_0$, i.e. $F_\epsilon = (1-\epsilon)F + \epsilon\delta_{\mathbf{x}_0}$; $\delta_{\mathbf{x}_0}$ being the distribution with point mass at $\mathbf{x}_0$. When $T(F) = \mathbb{E}_F g$ for some F-integrable function g, $IF(\mathbf{x}_0; T, F) = g(\mathbf{x}_0) - T(F)$. It now follows that for the DCM,

$$IF(\mathbf{x}_0; \tilde{\boldsymbol{\Sigma}}, F) = (D^-(\mathbf{x}_0, [\mathbf{X}]))^2 \mathbb{S}(\mathbf{x}_0; \boldsymbol{\mu}) - \tilde{\boldsymbol{\Sigma}}$$

Following Croux and Haesbroeck (2000), we now get the influence function of the

i^{th} eigenvector of $\tilde{\boldsymbol{\Sigma}}$, say $\tilde{\boldsymbol{\Gamma}} = (\tilde{\boldsymbol{\gamma}}_1, \ldots, \tilde{\boldsymbol{\gamma}}_p); i = 1, ..., p$:

$$\begin{aligned} IF(\mathbf{x}_0; \tilde{\boldsymbol{\gamma}}_i, F) &= \sum_{k=1;k\neq i}^{p} \frac{1}{\tilde{\lambda}_i - \tilde{\lambda}_k} \left\{ \boldsymbol{\gamma}_k^T IF(\mathbf{x}_0; \tilde{\boldsymbol{\Sigma}}, \boldsymbol{\gamma}_i) \right\} \boldsymbol{\gamma}_k \\ &= \sum_{k=1;k\neq i}^{p} \frac{1}{\tilde{\lambda}_i - \tilde{\lambda}_k} \left\{ \boldsymbol{\gamma}_k^T (D^-(\mathbf{x}_0, [\mathbf{X}]))^2 \mathbb{S}(\mathbf{x}_0; \boldsymbol{\mu}) \boldsymbol{\gamma}_i - \lambda_i \boldsymbol{\gamma}_k^T \boldsymbol{\gamma}_i \right\} \boldsymbol{\gamma}_k \\ &= \sum_{k=1;k\neq i}^{p} \frac{\sqrt{\lambda_i \lambda_k} z_{0i} z_{0k}}{\lambda_i - \lambda_k} . \frac{(D^-(\mathbf{z}_0, [\mathbf{Z}]))^2}{\mathbf{z}_0^T \boldsymbol{\Lambda} \mathbf{z}_0} \boldsymbol{\gamma}_k \end{aligned} \tag{4.3.10}$$

where $\boldsymbol{\Gamma}^T \boldsymbol{\Lambda}^{-1/2} (\mathbf{x}_0 - \boldsymbol{\mu}) = \mathbf{z}_0 = (z_{01}, ..., z_{0p})^T$. Clearly this influence function will be bounded, which indicates good robustness properties of principal components.

For the ADCM, we first notice that the influence function of any affine equivariant estimate of scatter $\mathbf{C}$ can be expressed as

$$IF(\mathbf{x}_0, \mathbf{C}, F) = \alpha_{\mathbf{C}}(\|\mathbf{z}_0\|) \frac{\mathbf{z}_0 \mathbf{z}_0^T}{\mathbf{z}_0^T \mathbf{z}_0} - \beta_{\mathbf{C}}(\|\mathbf{z}_0\|) \mathbf{C}$$

for scalar valued functions $\alpha_{\mathbf{C}}, \beta_{\mathbf{C}}$ (Hampel et al., 1986). Following this, the influence function of an eigenvector $\boldsymbol{\gamma}_{\mathbf{C},i}$ of $\mathbf{C}$ is derived:

$$IF(\mathbf{x}_0, \boldsymbol{\gamma}_{\mathbf{C},i}, F) = \alpha_{\mathbf{C}}(\|\mathbf{z}_0\|) \sum_{k=1,k\neq i}^{p} \frac{\sqrt{\lambda_i \lambda_k}}{\lambda_i - \lambda_k} . \frac{z_{0i} z_{0k}}{\mathbf{z}_0^T \mathbf{z}_0} \boldsymbol{\gamma}_k$$

When $\mathbf{C} = \boldsymbol{\Sigma}_M$, i.e. the solution to (4.3.5), then Huber (1981) shows that

$$\alpha_{\mathbf{C}}(\|\mathbf{z}_0\|) = \frac{p(p+2)u(\|\mathbf{z}_0\|)}{\mathbb{E}_G \left[p u(\|\mathbf{y}\|) + u'(\|\mathbf{y}\|) \|\mathbf{y}\| \right]}$$

Setting $u(\|\mathbf{z}_0\|) = (D^-(\mathbf{z}_0, [\mathbf{Z}]))^2$ ensures that the influence function of eigenvectors of the ADCM is bounded as well as increasing in magnitude with $\|\mathbf{z}_0\|$.

In Figure 4.2 we consider first eigenvectors of our scatter estimates, as well as two well-known robust estimates of scatter: the Sign Covariance Matrix (SCM) (Taskinen

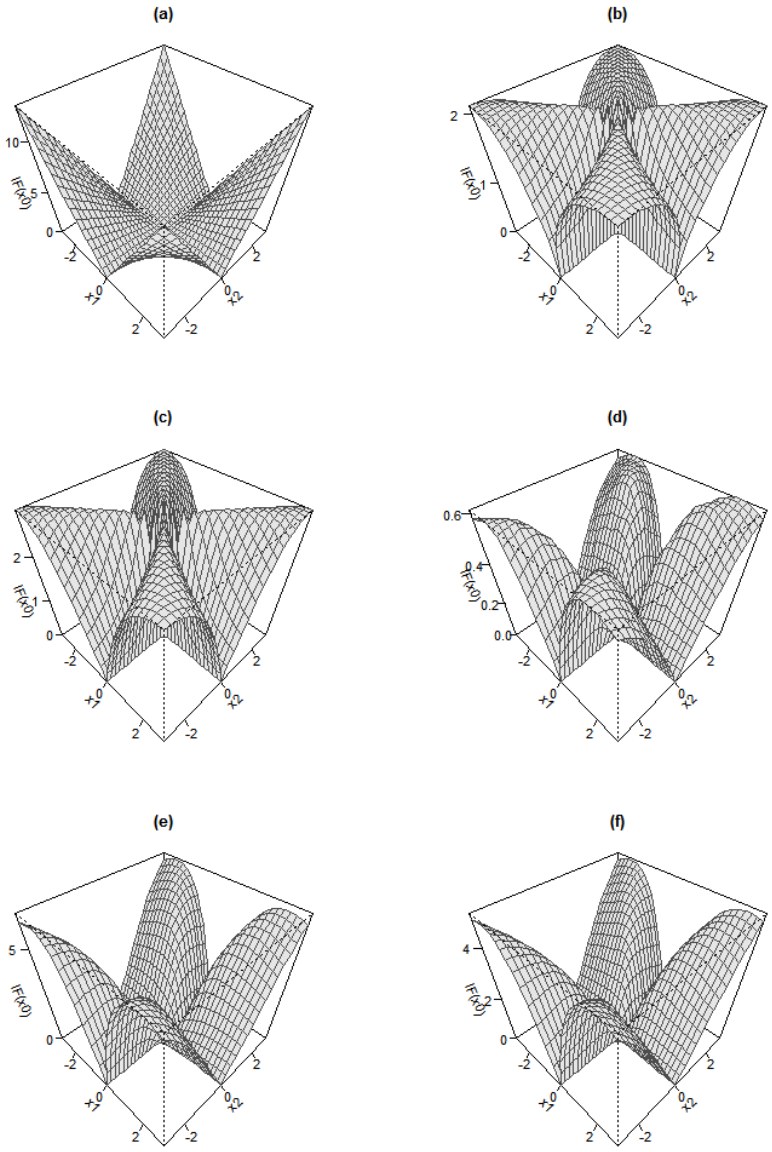


Figure 4.2: Plot of the norm of influence function for first eigenvector of (a) sample covariance matrix, (b) SCM, (c) Tyler's scatter matrix and DCMs for (d) Halfspace depth, (e) Mahalanobis depth, (f) Projection depth for a bivariate normal distribution with $\boldsymbol{\mu} = \mathbf{0}, \boldsymbol{\Sigma} = \text{diag}(2, 1)$

et al., 2012) and Tyler's shape matrix (Tyler, 1987), for the $\mathcal{N}_2((0,0)^T, \text{diag}(2,1))$ and plot norms of these influence functions for different values of $\mathbf{x}_0$. Influence function for the i^{th} eigenvectors of these two matrices (say $\boldsymbol{\gamma}_{S,i}$ and $\boldsymbol{\gamma}_{T,i}$, respectively) are as follows:

$$\begin{aligned} IF(\mathbf{x}_0; \boldsymbol{\gamma}_{S,i}, F) &= \sum_{k=1;k\neq i}^{p} \frac{\sqrt{\lambda_i \lambda_k}}{\lambda_{S,i} - \lambda_{S,k}} . \frac{z_{0i} z_{0k}}{\mathbf{z}_0^T \boldsymbol{\Lambda} \mathbf{z}_0} \boldsymbol{\gamma}_k, \text{ with } \lambda_{S,i} = \mathbb{E}_{\mathbf{Z}} \left(\frac{\lambda_i z_i^2}{\sum_{j=1}^{p} \lambda_j z_j^2} \right) \\ IF(\mathbf{x}_0; \boldsymbol{\gamma}_{T,i}, F) &= (p+2) \sum_{k=1;k\neq i}^{p} \frac{\sqrt{\lambda_i \lambda_k}}{\lambda_i - \lambda_k} . \frac{z_{0i} z_{0k}}{\mathbf{z}_0^T \mathbf{z}_0} \boldsymbol{\gamma}_k \end{aligned}$$

Their corresponding plots demonstrate the 'inlier effect', i.e. points close to symmetry center and the center itself having high influence, which results in loss of efficiency. The influence function for the sample covariance matrix is obtained by replacing $(p+2)$ by $\|\mathbf{z}_0\|^2$ in the expression of $IF(\mathbf{x}_0; \boldsymbol{\gamma}_{T,i}, F)$ above, hence is unbounded and the corresponding eigenvector estimators are not robust. In comparison, all three DCMs considered here have a bounded influence function as well as small values of the influence function at 'deep' points.

Asymptotic and finite-sample efficiencies

Suppose $\hat{\mathbf{V}}$ is a $\sqrt{n}$-consistent estimator of a scatter functional $\mathbf{V}$. Then the asymptotic variance of its eigenvectors are (Anderson, 2003)

$$A\mathbb{V}(\sqrt{n}\hat{\boldsymbol{\gamma}}_{v,i}) = \sum_{k=1;k\neq i}^{p} \frac{\lambda_{v,i}\lambda_{v,k}}{(\lambda_{v,i} - \lambda_{v,k})^2} \boldsymbol{\gamma}_{v,k} \boldsymbol{\gamma}_{v,k}^T \tag{4.3.11}$$

On the other hand, asymptotic variances of eigenvectors of the DCM can be derived using an approach similar to Taskinen et al. (2012):

$$A\mathbb{V}(\sqrt{n}\hat{\tilde{\boldsymbol{\gamma}}}_i) = \sum_{k=1;k\neq i}^{p} \frac{1}{(\tilde{\lambda}_k - \tilde{\lambda}_i)^2} \mathbb{E}\left[\frac{(D^-(\mathbf{z}, [\mathbf{Z}]))^4 \lambda_i \lambda_k z_i^2 z_k^2}{(\mathbf{z}^T \Lambda \mathbf{z})^2} \right] \boldsymbol{\gamma}_k \boldsymbol{\gamma}_k^T \tag{4.3.12}$$

We discuss this in detail in Subection 4.7.2. Following the above, we can now derive the asymptotic relative efficiencies of eigenvectors from the sample DCM with respect to the sample covariance matrix:

$$\begin{aligned} ARE(\hat{\tilde{\boldsymbol{\gamma}}}_i, \hat{\boldsymbol{\gamma}}_i; F) &= \frac{\mathrm{Tr}(A\mathbb{V}(\sqrt{n}\hat{\boldsymbol{\gamma}}_i))}{\mathrm{Tr}(A\mathbb{V}(\sqrt{n}\hat{\tilde{\boldsymbol{\gamma}}}_i))} \\ &= \left[\sum_{k=1;k\neq i}^{p} \frac{\lambda_i\lambda_k}{(\lambda_i-\lambda_k)^2}\right]\left[\sum_{k=1;k\neq i}^{p} \frac{\lambda_i\lambda_k}{(\tilde{\lambda}_i-\tilde{\lambda}_k)^2}\mathbb{E}\left(\frac{(D^-(\mathbf{z},[\mathbf{Z}]))^4 z_i^2 z_k^2}{(\mathbf{z}^T\boldsymbol{\Lambda}\mathbf{z})^2}\right)\right]^{-1} \end{aligned}$$

Obtaining ARE of the ADCM is, in comparison to DCM, more straightforward. The asymptotic covariance matrix of an eigenvector of the affine equivariant scatter functional $\mathbf{C}$ is given by:

$$A\mathbb{V}(\sqrt{n}\hat{\boldsymbol{\gamma}}_{\mathbf{C},j}) = A\mathbb{V}(c_{12}, F_0)\sum_{k=1,k\neq i}^{p}\frac{\lambda_i\lambda_k}{\lambda_i-\lambda_k}.\boldsymbol{\gamma}_i\boldsymbol{\gamma}_k^T$$

where $A\mathbb{V}(c_{12}, F_0)$ is the asymptotic variance of an off-diagonal element of $\mathbf{C}$ when the underlying distribution is $F_0 \equiv \mathcal{E}(\mathbf{0}_p, \mathbf{I}_p, g)$. Following Croux and Haesbroeck (2000) this equals

$$A\mathbb{V}(\mathbf{c}_{12}, F_0) = \mathbb{E}_{F_0}\left[\alpha_{\mathbf{C}}(\|\mathbf{z}\|)^2(S_1(\mathbf{z})S_2(\mathbf{z}))^2\right] = \mathbb{E}_{F_0}\alpha_{\mathbf{C}}(\|\mathbf{z}\|)^2.\mathbb{E}_{F_0}(S_1(\mathbf{z})S_2(\mathbf{z}))^2$$

again using the fact that $\|\mathbf{Z}\|$ and $\mathbf{S}(\mathbf{Z})$ are independent when $\mathbf{Z} \sim F_0$. It now follows that

$$ARE(\hat{\boldsymbol{\gamma}}_{\mathbf{C},i}, \hat{\boldsymbol{\gamma}}_{\boldsymbol{\Sigma},i}; F) = \frac{\mathbb{E}_{F_0}\alpha_{\boldsymbol{\Sigma}}(\|\mathbf{z}\|)^2}{\mathbb{E}_{F_0}\alpha_{\mathbf{C}}(\|\mathbf{z}\|)^2} = \frac{\mathbb{E}_{F_0}\|\mathbf{z}\|^4.\left[\mathbb{E}_{F_0}(pu\|\mathbf{z}\| + u'(\|\mathbf{z}\|)\|\mathbf{z}\|)\right]^2}{\mathbb{E}_{F_0}(u(\|\mathbf{z}\|))^2} \quad (4.3.13)$$

Table 4.3 considers 6 different elliptic distributions (namely, bivariate t with df = 5, 6, 10, 15, 25 and bivariate normal) and summarizes ARE for first eigenvectors for ADCMs corresponding to projection depth (PD-ACM) and halfspace depth (HD-

Distribution	PD-ACM				HD-ACM			
	$p = 2$	$p = 5$	$p = 10$	$p = 20$	$p = 2$	$p = 5$	$p = 10$	$p = 20$
t_5	4.73	3.99	3.46	3.26	4.18	3.63	3.36	3.15
t_6	2.97	3.28	2.49	2.36	2.59	2.45	2.37	2.32
t_{10}	1.45	1.47	1.49	1.52	1.30	1.37	1.43	1.49
t_{15}	1.15	1.19	1.23	1.27	1.01	1.10	1.17	1.24
t_{25}	0.97	1.02	1.07	1.11	0.85	0.94	1.02	1.08
MVN	0.77	0.84	0.89	0.93	0.68	0.77	0.84	0.91

Table 4.3: Table of AREs of the ADCM for different choices of p and data-generating distributions, and two choices of depth functions

ACM). Due to difficulty of analytically obtain the AREs, we calculate them using Monte-Carlo simulation of 10^6 samples and subsequent numerical integration. The ADCM seems to be particularly efficient in lower dimensions for distributions with heavier tails (t_5 and t_6), while for distributions with lighter tails, the AREs increase with data dimension. At higher values of p the ADCM is almost as efficient as the sample covarnace matrix when the data comes from multivariate normal distribution.

We now obtain finite sample efficiencies of the three DCMs as well as their depth-weighted affine equivariant counterparts by a simulation study, and compare them with the same from the SCM and Tyler's scatter matrix. We consider the same 6 elliptical distributions considered in ARE calculations above, and from every distribution draw 10,000 samples each for sample sizes $n = 20, 50, 100, 300, 500$. All distributions are centered at $\mathbf{0}_p$, and have covariance matrix $\mathbf{\Sigma} = \text{diag}(p, p-1, ...1)$. We consider 3 choices of p: 2, 3 and 4.

We use the concept of principal angles (Miao and Ben-Israel, 1992) to find out error estimates for the first eigenvector of a scatter matrix. In our case, the first eigenvector will be

$$\boldsymbol{\gamma}_1 = (1, \overbrace{0, ..., 0}^{p-1})^T$$

For an estimate of the eigenvector, say $\hat{\boldsymbol{\gamma}}_1$, error in prediction is measured by the smallest angle between the two lines, i.e. $\cos^{-1} |\hat{\boldsymbol{\gamma}}_1^T \hat{\boldsymbol{\gamma}}_1|$. A smaller absolute value of

this angle is equivalent to better prediction. We repeat this 10000 times and calculate the *Mean Squared Prediction Angle*:

$$MSPA(\hat{\boldsymbol{\gamma}}_1) = \frac{1}{10000}\sum_{m=1}^{10000}\left(\cos^{-1}\left|\boldsymbol{\gamma}_1^T\hat{\boldsymbol{\gamma}}_1^{(m)}\right|\right)^2$$

Finally, the finite sample efficiency of some eigenvector estimate $\hat{\boldsymbol{\gamma}}_{e,1}$ relative to that obtained from the sample covariance matrix, say $\hat{\boldsymbol{\gamma}}_{\boldsymbol{\Sigma},1}$ is obtained as:

$$FSE(\hat{\boldsymbol{\gamma}}_{e,1}, \hat{\boldsymbol{\gamma}}_{\boldsymbol{\Sigma},1}) = \frac{MSPA(\hat{\boldsymbol{\gamma}}_{\boldsymbol{\Sigma},1})}{MSPA(\hat{\boldsymbol{\gamma}}_{e,1})}$$

Table 4.4, Table 4.5 and Table 4.6 give these FSE values for $p = 2, 3, 4$, respectively. In general, all the efficiencies increase as the dimension p goes up. DCM-based estimators (columns 3-5 in each table) outperform SCM and Tyler's scatter matrix, and among the 3 depths considered, projection depth seems to give the best results. Its finite sample performances are better than Tyler's and Huber's M-estimators of scatter as well as their symmetrized counterparts (Table 4 in Sirkiä et al. (2007)), and quite close to the affine equivariant spatial sign covariance matrix (Table 2 in Ollilia et al. (2003)). The depth-weighted iterated versions of these 3 SCMs (columns 6-8 in each table) seem to further better the performance of their corresponding orthogonal equivariant counterparts.

Robust estimation of eigenvalues, and a plug-in estimator of $\boldsymbol{\Sigma}$

As we have seen in theorem 4.3.1, eigenvalues of the DCM are not same as the population eigenvalues, whereas the ADCM only gives back standardized eigenvalues. However, it is possible to robustly estimate the original eigenvalues by working with the individual columns of the robust score matrix. We do this using the following steps:

1. Randomly divide the sample indices $\{1, 2, ..., n\}$ into k disjoint groups $\{G_1, ..., G_k\}$ of size $\lfloor n/k \rfloor$ each;

2. Assume the data is centered. Transform the data matrix: $\mathbb{S}_n = \hat{\hat{\mathbf{\Gamma}}}^T \mathbb{X}_n$;

3. Calculate coordinate-wise variances for each group of indices G_j:

$$\hat{\lambda}_{i,j} = \frac{1}{|G_j|} \sum_{l \in G_j} (s_{li} - \bar{s}_{G_j,i})^2; \quad i = 1, ..., p; j = 1, ..., k$$

   where $\bar{\mathbf{s}}_{G_j} = (\bar{s}_{G_j,1}, ..., \bar{s}_{G_j,p})^T$ is the vector of column-wise means of S_{G_j}, the submatrix of $\mathbb{S}_n$ with row indices in G_j.

4. Obtain estimates of eigenvalues by taking coordinate-wise medians of these variances:

$$\hat{\lambda}_i = \text{median}(\hat{\lambda}_{i,1}, ..., \hat{\lambda}_{i,k}); \quad i = 1, ..., p$$

The number of subgroups used to calculate this median-of-small-variances estimator can be determined following Minsker (2015). After this, we construct a consistent plug-in estimator of the population covariance matrix $\mathbf{\Sigma}$:

Theorem 4.3.4. *Consider the estimates $\hat{\lambda}_i$ obtained from the above algorithm, and the matrix of eigenvectors $\hat{\hat{\mathbf{\Gamma}}}$ estimated using the sample DCM. Define $\hat{\mathbf{\Sigma}} = \hat{\hat{\mathbf{\Gamma}}} \hat{\mathbf{\Lambda}} \hat{\hat{\mathbf{\Gamma}}}$; $\hat{\mathbf{\Lambda}} = diag(\hat{\lambda}_1, \ldots, \hat{\lambda}_p)$. Then as $n \to \infty$,*

$$\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_F \xrightarrow{P} 0$$

$\|.\|_F$ being the Frobenius norm.

Given that we already have the eigenvector estimates of the DCM, the estimates $\hat{\lambda}_i$ are easy to compute, and finite-sample error bounds for them can be obtained as a special case of the general results provided in Minsker (2015).

$t_5, p = 2$	SCM	Tyler	HDCM	MHDCM	PDCM	HD-wCM	MhD-wCM	PD-wCM
n=20	0.80	0.83	0.95	0.95	0.89	1.00	0.96	0.89
n=50	0.86	0.90	1.25	1.10	1.21	1.32	1.13	1.25
n=100	1.02	1.04	1.58	1.20	1.54	1.67	1.24	1.63
n=300	1.24	1.28	1.81	1.36	1.82	1.93	1.44	1.95
n=500	1.25	1.29	1.80	1.33	1.84	1.91	1.39	1.97
$t_6, p = 2$	SCM	Tyler	HDCM	MHDCM	PDCM	HD-wCM	MhD-wCM	PD-wCM
n=20	0.77	0.79	0.92	0.92	0.86	0.96	0.92	0.85
n=50	0.76	0.78	1.11	1.00	1.08	1.17	1.03	1.13
n=100	0.78	0.79	1.27	1.06	1.33	1.35	1.11	1.41
n=300	0.88	0.91	1.29	1.09	1.35	1.38	1.15	1.45
n=500	0.93	0.96	1.37	1.13	1.40	1.44	1.19	1.48
$t_{10}, p = 2$	SCM	Tyler	HDCM	MHDCM	PDCM	HD-wCM	MhD-wCM	PD-wCM
n=20	0.70	0.72	0.83	0.84	0.77	0.89	0.87	0.79
n=50	0.58	0.60	0.90	0.84	0.86	0.95	0.88	0.91
n=100	0.57	0.59	0.92	0.87	0.97	0.98	0.90	1.03
n=300	0.62	0.64	0.93	0.85	0.99	0.99	0.91	1.06
n=500	0.62	0.65	0.93	0.86	1.00	1.00	0.92	1.08
$t_{15}, p = 2$	SCM	Tyler	HDCM	MHDCM	PDCM	HD-wCM	MhD-wCM	PD-wCM
n=20	0.63	0.66	0.76	0.78	0.72	0.81	0.81	0.73
n=50	0.52	0.52	0.79	0.75	0.80	0.84	0.79	0.85
n=100	0.51	0.52	0.83	0.77	0.88	0.88	0.81	0.94
n=300	0.55	0.56	0.84	0.79	0.91	0.89	0.84	0.98
n=500	0.56	0.59	0.85	0.80	0.93	0.91	0.86	0.99
$t_{25}, p = 2$	SCM	Tyler	HDCM	MHDCM	PDCM	HD-wCM	MhD-wCM	PD-wCM
n=20	0.63	0.65	0.77	0.79	0.74	0.80	0.81	0.74
n=50	0.49	0.50	0.73	0.71	0.76	0.78	0.75	0.80
n=100	0.45	0.46	0.73	0.69	0.81	0.78	0.73	0.87
n=300	0.51	0.52	0.78	0.75	0.87	0.83	0.79	0.94
n=500	0.53	0.55	0.79	0.75	0.87	0.84	0.80	0.94
$\mathcal{N}_p$, $p = 2$	SCM	Tyler	HDCM	MHDCM	PDCM	HD-wCM	MhD-wCM	PD-wCM
n=20	0.56	0.60	0.69	0.71	0.67	0.73	0.74	0.68
n=50	0.42	0.43	0.66	0.66	0.70	0.71	0.69	0.75
n=100	0.42	0.43	0.69	0.66	0.77	0.74	0.71	0.83
n=300	0.47	0.49	0.71	0.69	0.82	0.76	0.73	0.88
n=500	0.48	0.50	0.73	0.71	0.83	0.78	0.76	0.89

Table 4.4: Finite sample efficiencies of several scatter matrices: $p = 2$, t_v is t-distribution with v degrees of freedom, $\mathcal{N}_p$ is p-variate normal

$t_5, p = 3$	SCM	Tyler	HDCM	MHDCM	PDCM	HD-wCM	MhD-wCM	PD-wCM
n=20	0.96	0.97	1.06	1.03	0.99	1.07	1.06	0.97
n=50	1.07	1.08	1.28	1.20	1.18	1.33	1.23	1.20
n=100	1.12	1.15	1.49	1.31	1.40	1.57	1.38	1.48
n=300	1.49	1.54	2.09	1.82	2.07	2.19	1.93	2.18
n=500	1.60	1.66	2.18	1.87	2.21	2.27	1.95	2.30
$t_6, p = 3$	SCM	Tyler	HDCM	MHDCM	PDCM	HD-wCM	MhD-wCM	PD-wCM
n=20	0.90	0.92	1.00	0.99	0.95	1.02	1.01	0.94
n=50	0.95	0.96	1.16	1.09	1.09	1.21	1.14	1.11
n=100	0.98	0.99	1.32	1.22	1.25	1.38	1.27	1.29
n=300	1.10	1.14	1.57	1.40	1.58	1.62	1.47	1.64
n=500	1.17	1.20	1.57	1.43	1.60	1.63	1.51	1.67
$t_{10}, p = 3$	SCM	Tyler	HDCM	MHDCM	PDCM	HD-wCM	MhD-wCM	PD-wCM
n=20	0.87	0.88	0.95	0.94	0.90	0.97	0.98	0.89
n=50	0.77	0.79	0.96	0.92	0.94	0.99	0.96	0.95
n=100	0.75	0.76	1.02	0.95	1.01	1.06	1.00	1.05
n=300	0.73	0.75	1.03	0.98	1.10	1.08	1.03	1.15
n=500	0.73	0.76	1.02	0.98	1.09	1.06	1.02	1.14
$t_{15}, p = 3$	SCM	Tyler	HDCM	MHDCM	PDCM	HD-wCM	MhD-wCM	PD-wCM
n=20	0.84	0.86	0.92	0.92	0.89	0.94	0.94	0.87
n=50	0.75	0.76	0.92	0.90	0.90	0.96	0.94	0.93
n=100	0.66	0.67	0.91	0.87	0.95	0.96	0.92	1.00
n=300	0.61	0.64	0.90	0.87	1.00	0.93	0.91	1.04
n=500	0.65	0.67	0.89	0.87	0.99	0.93	0.91	1.03
$t_{25}, p = 3$	SCM	Tyler	HDCM	MHDCM	PDCM	HD-wCM	MhD-wCM	PD-wCM
n=20	0.78	0.79	0.87	0.89	0.87	0.89	0.92	0.86
n=50	0.70	0.71	0.88	0.86	0.88	0.91	0.90	0.90
n=100	0.61	0.63	0.86	0.83	0.89	0.90	0.88	0.94
n=300	0.58	0.59	0.83	0.80	0.92	0.87	0.85	0.98
n=500	0.62	0.64	0.83	0.82	0.94	0.88	0.87	0.99
$\mathcal{N}_p$, $p = 3$	SCM	Tyler	HDCM	MHDCM	PDCM	HD-wCM	MhD-wCM	PD-wCM
n=20	0.76	0.78	0.85	0.87	0.84	0.87	0.90	0.83
n=50	0.66	0.67	0.82	0.81	0.84	0.86	0.86	0.86
n=100	0.56	0.58	0.77	0.75	0.83	0.82	0.79	0.87
n=300	0.53	0.55	0.75	0.74	0.85	0.79	0.78	0.90
n=500	0.56	0.58	0.76	0.76	0.87	0.80	0.80	0.92

Table 4.5: Finite sample efficiencies of several scatter matrices: $p = 2$, t_v is t-distribution with v degrees of freedom, $\mathcal{N}_p$ is p-variate normal

$t_5, p = 4$	SCM	Tyler	HDCM	MHDCM	PDCM	HD-wCM	MhD-wCM	PD-wCM
n=20	1.04	1.02	1.10	1.07	1.02	1.09	1.07	0.98
n=50	1.08	1.08	1.16	1.16	1.13	1.19	1.19	1.13
n=100	1.31	1.31	1.42	1.38	1.36	1.46	1.44	1.36
n=300	1.46	1.54	1.81	1.76	1.95	1.88	1.88	1.95
n=500	1.92	1.93	2.23	2.03	2.31	2.35	2.19	2.39
$t_6, p = 4$	SCM	Tyler	HDCM	MHDCM	PDCM	HD-wCM	MhD-wCM	PD-wCM
n=20	1.00	1.05	1.03	1.05	1.00	1.04	1.04	0.95
n=50	1.03	1.01	1.13	1.12	1.11	1.19	1.17	1.10
n=100	1.08	1.12	1.25	1.23	1.27	1.24	1.25	1.22
n=300	1.34	1.36	1.64	1.52	1.60	1.67	1.61	1.68
n=500	1.26	1.34	1.55	1.49	1.60	1.65	1.61	1.69
$t_{10}, p = 4$	SCM	Tyler	HDCM	MHDCM	PDCM	HD-wCM	MhD-wCM	PD-wCM
n=20	0.90	0.89	0.95	0.98	0.98	0.96	1.01	0.95
n=50	0.90	0.91	1.01	0.98	0.98	1.03	1.04	0.99
n=100	0.87	0.87	0.93	0.95	1.01	0.99	1.01	1.05
n=300	0.87	0.87	1.09	1.09	1.17	1.14	1.16	1.23
n=500	0.88	0.92	1.10	1.10	1.23	1.19	1.22	1.29
$t_{15}, p = 4$	SCM	Tyler	HDCM	MhDCM	PDCM	HD-wCM	MhD-wCM	PD-wCM
n=20	0.92	0.90	0.94	0.94	0.96	0.95	0.97	0.89
n=50	0.82	0.83	0.88	0.91	0.93	0.88	0.93	0.93
n=100	0.84	0.87	0.92	0.95	1.00	0.93	0.96	1.00
n=300	0.73	0.75	0.96	0.99	1.10	1.00	1.06	1.12
n=500	0.73	0.76	0.95	0.96	1.06	0.94	0.97	1.06
$t_{25}, p = 4$	SCM	Tyler	HDCM	MhDCM	PDCM	HD-wCM	MhD-wCM	PD-wCM
n=20	0.89	0.92	0.92	0.92	0.90	0.96	0.95	0.89
n=50	0.82	0.84	0.89	0.90	0.91	0.93	0.96	0.92
n=100	0.77	0.76	0.90	0.90	0.96	0.94	0.98	1.04
n=300	0.73	0.77	0.93	0.91	0.98	1.00	0.98	1.03
n=500	0.67	0.71	0.83	0.83	0.96	0.88	0.90	1.00
$\mathcal{N}_p$, $p = 4$	SCM	Tyler	HDCM	MhDCM	PDCM	HD-wCM	MhD-wCM	PD-wCM
n=20	0.82	0.84	0.87	0.90	0.91	0.89	0.93	0.89
n=50	0.80	0.81	0.87	0.88	0.88	0.88	0.92	0.88
n=100	0.68	0.71	0.80	0.85	0.91	0.82	0.86	0.92
n=300	0.61	0.63	0.82	0.85	0.93	0.86	0.91	0.96
n=500	0.60	0.64	0.77	0.80	0.90	0.82	0.86	0.96

Table 4.6: Finite sample efficiencies of several scatter matrices: $p = 2$, t_v is t-distribution with v degrees of freedom, $\mathcal{N}_p$ is p-variate normal

4.4 Robust PCA and supervised models

In the presence of a vector of univariate responses, say $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)^T$, there is substantial literature devoted to utilizing the subspace generated by the basis of $\mathbb{V}(\mathbf{X}) = \Sigma$ in modelling $\mathbb{E}(Y|\mathbf{X})$. This ranges from the simple Principal Components Regression (PCR) to Partial Least Squares (PLS) and Envelope methods (Cook et al., 2010). Here we concentrate on robust inference using Sufficient Dimension Reduction (SDR) (Adragni and Cook, 2009), mainly because it provides a general framework for reducing dimensionality of data directly using top eigenvectors of the covariance matrix of $\mathbf{X}$ (albeit in a different manner than PCR) or an appropriate affine transformation of it.

SDR attempts to find out a linear transformation R on $\mathbf{X}$ such that $E(Y|\mathbf{X}) = \mathbb{E}(Y|R(\mathbf{X}))$. Assuming that $R(\mathbf{X})$ takes values in $\mathbb{R}^d, d \leqslant \min(n, p)$, this can be achieved through an inverse regression model:

$$\mathbf{X}_y = \bar{\boldsymbol{\mu}} + \boldsymbol{\Gamma}\mathbf{v}_y + \boldsymbol{\epsilon} \tag{4.4.1}$$

where $\mathbf{X}_y = \mathbf{X}|Y = y, \bar{\boldsymbol{\mu}} = \mathbb{E}\mathbf{X}$, $\boldsymbol{\Gamma}$ is a $p \times d$ semi-orthogonal basis for $\mathcal{S}_{\boldsymbol{\Gamma}}$, the spanning subspace of $\{\mathbb{E}\mathbf{X}_y - \bar{\boldsymbol{\mu}}|y \in S_Y\}$ (S_y is sample space of Y) and $\mathbf{v}_y = (\boldsymbol{\Gamma}^T\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T(\mathbb{E}\mathbf{X}_y - \bar{\boldsymbol{\mu}}) \in \mathbb{R}^d$. The random error term $\boldsymbol{\epsilon}$ follows a multivariate normal distribution with mean $\mathbf{0}_p$ and covariance matrix Δ for some positive definite $\Delta \in \mathbb{R}^{p\times p}$. This formulation is straightforward to implement when Y is categorical, while for continuous responses, the vector $\mathbf{y}$ is divided into a number of slices.

Under this model the minimal sufficient transformation is $R(\mathbf{X}) = \boldsymbol{\Gamma}^T\Delta^{-1}\mathbf{X}$. The simplest case of this model is when $\Delta = \sigma^2\mathbf{I}_p$, for which the maximum likelihood estimator of $R(\mathbf{X})$ turns out to be the first d principal components of $\hat{\boldsymbol{\Sigma}}$, the sample covariance matrix. Let us denote the matrix of these PC estimates by $\hat{\boldsymbol{\Gamma}}_d \in \mathbb{R}^{p\times d}$. Now taking $\hat{\mathbb{E}}\mathbf{X}_y = \bar{\mathbf{X}}_y$ and $\hat{\bar{\boldsymbol{\mu}}} = \bar{\mathbf{X}}$, one can now estimate σ^2 as: $\hat{\sigma}^2 = \sum_{j=1}^{p} s_{ii}/p$, where

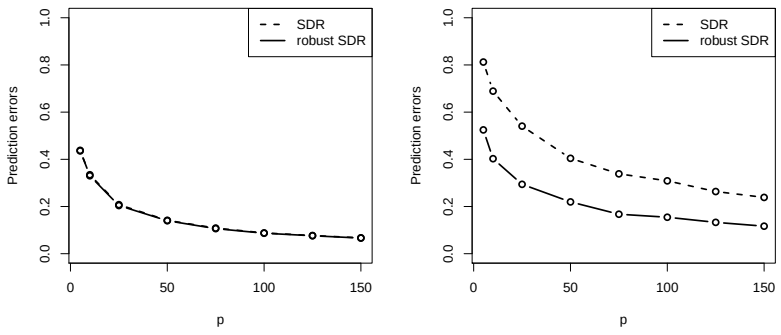


Figure 4.3: Plot of the norm of influence function for first eigenvector of (a) sample covariance matrix, (b) SCM, (c) Tyler's scatter matrix and DCMs for (d) Halfspace depth, (e) Mahalanobis depth, (f) Projection depth for a bivariate normal distribution with $\boldsymbol{\mu} = \mathbf{0}, \Sigma = \text{diag}(2, 1)$

s_{jj} is the j^{th} diagonal element of the estimated inverse regression residual covariance matrix $\hat{\mathbb{V}}_Y(\mathbf{X}_Y - \bar{\mathbf{X}} - \hat{\mathbf{\Gamma}}_d \hat{\mathbf{v}}_Y)$.

Following this, predictions for a new observation $\mathbf{x}$ is obtained as a weighted sum of the responses:

$$\hat{\mathbb{E}}(Y|\mathbf{X} = \mathbf{x}) = \frac{\sum_{i=1}^n w_i Y_i}{\sum_{i=1}^n w_i}; \quad w_i = \exp\left[-\frac{1}{\hat{\sigma}^2}\|\hat{\mathbf{\Gamma}}_d^T(\mathbf{x} - \mathbf{X}_i)\|^2\right]$$

We formulate a robust version of the above procedure by estimating the quantities $\mathbf{\Gamma}, \bar{\boldsymbol{\mu}}, \boldsymbol{\mu}_y, \sigma^2$ by robust methods. Specifically, we take the following as their estimates:

- $\hat{\tilde{\mathbf{\Gamma}}}_d$ = first d eigenvectors of the sample DCM;
- $\hat{\boldsymbol{\mu}}_s$ = spatial median of the rows of the data matrix $\mathbb{X}_n$;
- $\hat{\boldsymbol{\mu}}_{s,y}$ = spatial median of the rows of $\mathbb{X}_y = \mathbb{X}_n|Y = y$, for all $y \in S_Y$;
- $\tilde{\sigma}^2 = \sum_{j=1}^p [\widehat{\text{MAD}}_Y(\mathbb{X}_{Y,j} - \hat{\mu}_{s,j} - \hat{\tilde{\boldsymbol{\gamma}}}_{d,j}^T \tilde{\mathbf{v}}_Y)]^2/p$, with $\hat{\tilde{\mathbf{\Gamma}}}_d = (\hat{\tilde{\boldsymbol{\gamma}}}_{d,1}, ..., \hat{\tilde{\boldsymbol{\gamma}}}_{d,p})^T$, and MAD being the median absolute deviation.

The following simulation study using the same setup as in (Adragni and Cook, 2009) compares the performance of our robust SDR with the original method with or without the presence of bad leverage points in the covariate matrix $\mathbb{X}_n$. For a fixed dimension p, we take $n = 200, d = 1$, generate the responses Y as independent standard normal, and the predictors as $\mathbb{X}_Y = \boldsymbol{\gamma}^* v_Y^* + \boldsymbol{\epsilon}$, with $\boldsymbol{\gamma}^*_{p\times 1} = (1, ..., 1)^T, v_Y = Y + Y^2 + Y^3$ and $\mathbb{V}(\boldsymbol{\epsilon}) = 25\mathbf{I}_p$. We measure performance of both SDR models by their mean squared prediction error on another set of 200 observations generated similarly, and taking the average of these errors on 100 such training-test pair of datasets. Finally we repeat the whole setup for different choices of $p = 5, 10, 25, 50, 75, 100, 125, 150$.

Panel (a) of Figure 4.3 compares prediction errors using robust and maximum likelihood SDR estimates when the covariate matrix contains no outliers, and the two methods are virtually indistinguishable. We now introduce outliers in each of the 100 datasets by adding 100 to first $p/5$ coordinates of the first 10 samples in $\mathbb{X}_n$, and repeat the analysis. Panel (b) of the figure shows that although our robust method performs slightly worse than the case when there were no outliers, it remains more accurate in predicting out-of-sample observations for all values of p.

4.5 Robust inference with functional data

Detection of anomalous observations is of importance in real-life problems involving functional data analysis, and functional PCA is a widely used tool in this setting. In this section we use robust principal components from the DCM for this purpose. We shall use the approach of Boente and Salibian-Barrera (2015) for performing robust PCA on functional data using the estimated PCs from the DCM. Here we have a data matrix $\mathbf{H}$, that stores the values of a set of n curves, say $\mathcal{F} = \{f_1, \ldots, f_n\} \in L^2[0, 1]$, each observed at a set of common design points $\{t_1, ..., t_m\}$. We model each of these functions as a linear combination of p mutually orthogonal B-spline basis functions

$\mathcal{D} = \{\delta_1, ..., \delta_p\}$. Following this, we map data for each of the functions onto the coordinate system formed by the spline basis:

$$T(\mathbf{H}; \mathcal{F}, \mathcal{D})_{ij} = \sum_{l=2}^{m} f_i(t_l)\delta_j(t_l)(t_l - t_{l-1}); \quad 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant p \tag{4.5.1}$$

We now do depth-based PCA on the transformed $n \times p$ data matrix $T(\mathbf{H}; \mathcal{F}, \mathcal{D}) \equiv T(\mathbf{H})$, and obtain the rank-$q$ approximation ($q \leqslant p$) of the i^{th} observation using the robust $p \times q$ loading matrix $\tilde{\mathbf{P}}$ and robust $q \times 1$ score vector $\tilde{\mathbf{s}}_i$:

$$\widehat{T}(\mathbf{H})_i = \hat{\boldsymbol{\mu}}_s + \tilde{\mathbf{P}}\tilde{\mathbf{s}}_i$$

with $\hat{\boldsymbol{\mu}}_s$ being the spatial median of $T(\mathbf{H})$. Then we transform this approximation back to the original coordinates: $\hat{f}_i(t_l) = \sum_{j=1}^{p} \widehat{T}(\mathbf{H})_{ij}\delta_j(t_l)$.

We shall demonstrate the utility of our robust method for detecting functional outliers through two data examples. For any method of PCA with k components on a dataset of n observations and p variables, the *score distance* (SD) and *orthogonal distance* (OD) for i^{th} observation ($i = 1, 2, ..., n$) are defined as:

$$SD_i = \sqrt{\sum_{j=1}^{k} \frac{s_{ij}^2}{\lambda_j}}; \quad OD_i = \|\mathbf{x}_i - \mathbf{P}\mathbf{s}_i^T\|$$

where $\mathbf{S} = (\mathbf{s}_1, \ldots, \mathbf{s}_n)^T \in \mathbb{R}^{n\times k}$ is the score matrix, $\mathbf{P} \in \mathbb{R}^{p\times k}$ the loading matrix, and $\lambda_1, \ldots, \lambda_k$ are eigenvalues obtained from the PCA, and $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are the n observation vectors. From a practical standpoint, SD_i can be interpreted as a weighted norm of the projection of the i^{th} point on the hyperplane formed by first k principal components, and OD_i the orthogonal distance of point i from that hyperplane. For outlier detection, following Hubert et al. (2005) we set the upper cutoff values for score distances at $\sqrt{\chi^2_{2,.975}}$ and orthogonal distances at $[\text{median}(OD^{2/3}) +$

$\text{MAD}(OD^{2/3})\Phi^{-1}(0.975)]^{3/2}$, where $\Phi(.)$ is the standard normal cumulative distribution function.

We consider the El-Niño data, which is part of a larger dataset on potential factors behind El-Niño oscillations in the tropical pacific available in `http://www.cpc.ncep.noaa.gov/data/indices/`, as the first test case for outlier detection using our robust functional PCA. This records monthly average Sea Surface Temperatures from June 1970 to May 2004, and the yearly oscillations follow more or less the same pattern (see panel a of Figure 4.4). Using a cubic spline basis with knots at alternate months starting in June gives a close approximation of the yearly time series data (panel c), and performing depth-based PCA with $q = 1$ results in two points having their SD and OD larger than cutoff (panel e). These points correspond to the time periods June 1982 to May 1983 and June 1997 to May 1998 are marked by black curves in panels a and c), and pinpoint the two seasons with strongest El-Niño events.

Our second application is on the Octane data, which consists of 226 variables and 39 observations (Esbensen et al., 1994). Each sample is a gasoline compound with a certain octane number, and has its NIR absorbance spectra measured in 2 nm intervals between 1100 - 1550 nm. There are 6 outliers here: compounds 25, 26 and 36-39, which contain alcohol. We use the same basis structure as the one in El-Niño data here, and again the top robust PC turns out to be sufficient in identifying all 6 outliers (panels b, d and f of Figure 4.4).

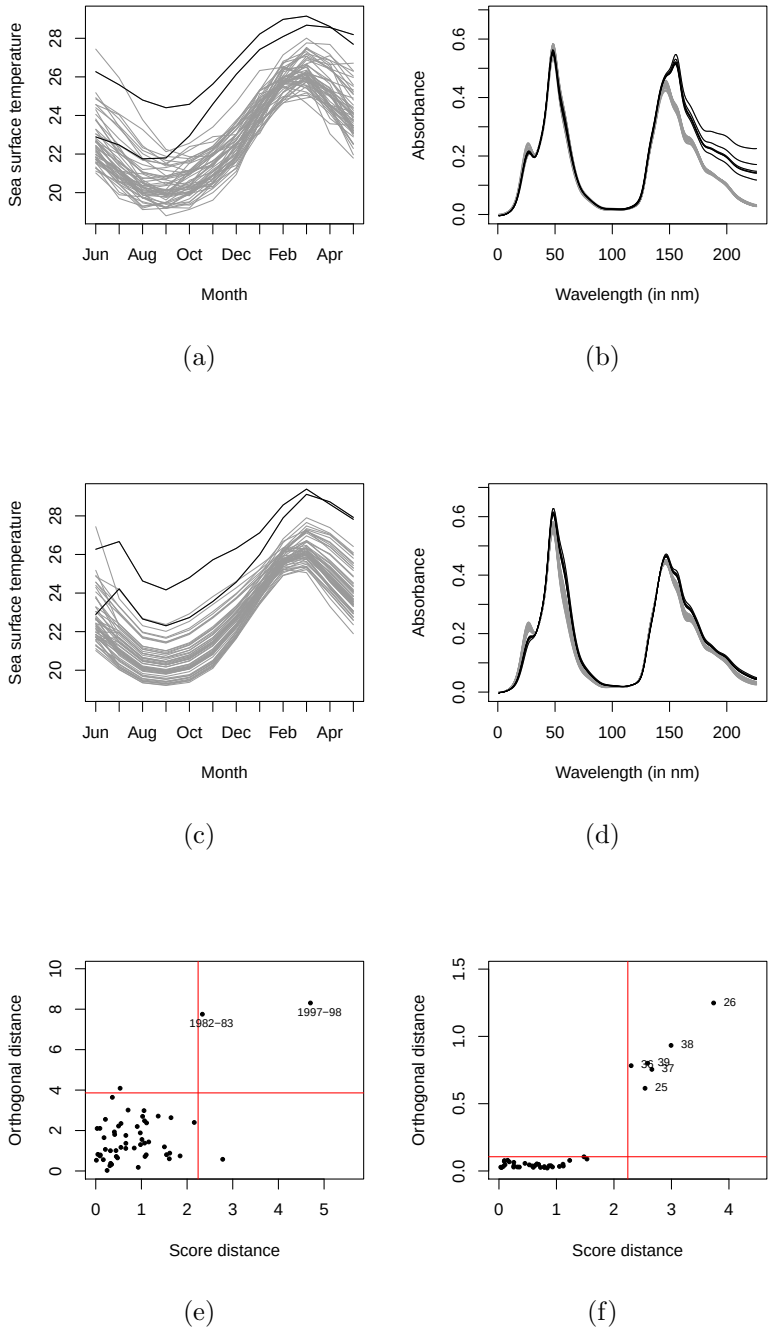


Figure 4.4: Actual sample curves, their spline approximations and diagnostic plots respectively for El-Niño (a,c,e) and Octane (b,d,f) datasets

4.6 Conclusion

In the above sections we elaborate on a proposed transformation based on the idea of combining sign functions in an inner product space and certain transformations of general peripherality functions defined using probability measures on the same space. Based on the conditions we impose in course of the chapter, we essentially end up using data depths as scalar multiples of the spatial sign in estimation and testing problems for the location parameter of an elliptical distribution in $\mathbb{R}^p$, and using inverse depth functions for robust estimation of different components of its covariance matrix. As demonstrated by several simulation studies and data examples, in all these situations the use of this composite transformation vector brings about efficiency gains, as well as favorable robustness properties in terms of bounded influence functions and deviations from Gaussianity.

Regarding the multivariate rank transformation we propose, to be noted is the fact that the mapping $\mathbf{X} \mapsto \tilde{\mathbf{X}}$ is in fact one-to-one for elliptical distributions. Thus such rank vectors can possibly be used for inference based on transformation-retransformation type techniques (e.g. Chakraborty and Chaudhuri (1996); Chakraborty et al. (1998)). We defer this to future research. Finally, while these rank vectors have excellent intuitive appeal in preserving the shape of a multivariate data cloud, it would be interesting to study the properties of transformations similar to (4.1.2) in general Hilbert spaces, as well as explore their applications in different data-analytic domains.

4.7 Appendix

4.7.1 Form of $\tilde{\mathbf{V}}(F)$

First observe that for F having covariance matrix $\Sigma = \mathbf{\Gamma\Lambda\Gamma}^T$,

$$\tilde{\mathbf{V}}(F) = (\mathbf{\Gamma} \otimes \mathbf{\Gamma})\tilde{\mathbf{V}}(F_{\mathbf{\Lambda}})(\mathbf{\Gamma} \otimes \mathbf{\Gamma})^T$$

where $F_{\mathbf{\Lambda}}$ has the same elliptic distribution as F, but with covariance matrix $\mathbf{\Lambda}$. Now,

$$\begin{aligned}
\tilde{\mathbf{V}}(F_{\mathbf{\Lambda}}) &= \mathbb{E}_{\mathbf{Z}}\left[\text{vec}\left\{\frac{(D^-(\mathbf{z},[\mathbf{Z}]))^2\mathbf{\Lambda}^{1/2}\mathbf{z}\mathbf{z}^T\mathbf{\Lambda}^{1/2}}{\mathbf{z}^T\mathbf{\Lambda}\mathbf{z}} - \tilde{\mathbf{\Lambda}}\right\}\text{vec}^T\left\{\frac{(D^-(\mathbf{z},[\mathbf{Z}]))^2\mathbf{\Lambda}^{1/2}\mathbf{z}\mathbf{z}^T\mathbf{\Lambda}^{1/2}}{\mathbf{z}^T\mathbf{\Lambda}\mathbf{z}} - \tilde{\mathbf{\Lambda}}\right\}\right] \\
&= \mathbb{E}\left[\text{vec}\left\{(D^-(\mathbf{z},[\mathbf{Z}]))^2\mathbb{S}(\mathbf{\Lambda}^{1/2}\mathbf{z};\mathbf{0})\right\}\text{vec}^T\left\{(D^-(\mathbf{z},[\mathbf{Z}]))^2\mathbb{S}(\mathbf{\Lambda}^{1/2}\mathbf{z};\mathbf{0})\right\}\right] \\
&- \quad \text{vec}(\tilde{\mathbf{\Lambda}})\text{vec}^T(\tilde{\mathbf{\Lambda}})
\end{aligned}$$

The matrix $\text{vec}(\tilde{\mathbf{\Lambda}})\,\text{vec}^T(\tilde{\mathbf{\Lambda}})$ consists of elements $\lambda_i\lambda_j$ at $(i,j)^{\text{th}}$ position of the $(i,j)^{\text{th}}$ block, and 0 otherwise. These positions correspond to variance and covariance components of on-diagonal elements. For the expectation matrix, all its elements are of the form $\mathbb{E}[\sqrt{\lambda_a\lambda_b\lambda_c\lambda_d}z_az_bz_cz_d.(D^-(\mathbf{z},[\mathbf{Z}]))^4/(\mathbf{z}^T\mathbf{\Lambda}\mathbf{z})^2]$, with $1 \leqslant a,b,c,d \leqslant p$. Since $(D^-(\mathbf{z},[\mathbf{Z}]))^4/(\mathbf{z}^T\mathbf{\Lambda}\mathbf{z})^2$ is even in $\mathbf{z}$, which has a circularly symmetric distribution, all such expectations will be 0 unless $a=b=c=d$, or they are pairwise equal. Following a similar derivation for spatial sign covariance matrices in Magyar and Tyler (2014), we collect the non-zero elements and write the matrix of expectations:

$$(\mathbf{I}_{p^2}+\mathbf{K}_{p,p})\left\{\sum_{a=1}^{p}\sum_{b=1}^{p}\tilde{\gamma}_{ab}(\mathbf{e}_a\mathbf{e}_a^T\otimes\mathbf{e}_b\mathbf{e}_b^T) - \sum_{a=1}^{p}\tilde{\gamma}_{aa}(\mathbf{e}_a\mathbf{e}_a^T\otimes\mathbf{e}_a\mathbf{e}_a^T)\right\} + \sum_{a=1}^{p}\sum_{b=1}^{p}\tilde{\gamma}_{ab}(\mathbf{e}_a\mathbf{e}_b^T\otimes\mathbf{e}_a\mathbf{e}_b^T)$$

where $\mathbf{I}_k = (\mathbf{e}_1,...,\mathbf{e}_k), \mathbf{K}_{m,n} = \sum_{i=1}^{m}\sum_{j=1}^{n}\mathbf{J}_{ij}\otimes\mathbf{J}_{ij}^T$ with $\mathbf{J}_{ij}\in\mathbb{R}^{m\times n}$ having 1 as (i,j)-th element and 0 elsewhere, and $\tilde{\gamma}_{mn} = \mathbb{E}[\lambda_m\lambda_n z_m^2 z_n^2.(D^-(\mathbf{z},[\mathbf{Z}]))^4/(\mathbf{z}^T\mathbf{\Lambda}\mathbf{z})^2]; 1 \leqslant m,n \leqslant p$.

Putting everything together, denote $\hat{\tilde{\mathbb{S}}}(F_{\mathbf{\Lambda}}) = \sum_{i=1}^{n}(D^{-}(\mathbf{z}_i, [\mathbb{Z}_n]))^2\mathbb{S}(\mathbf{\Lambda}^{1/2}\mathbf{z}_i; \hat{\boldsymbol{\mu}}_n)/n$. Then the different types of elements in the matrix $\tilde{\mathbf{V}}(F_{\mathbf{\Lambda}})$ are as given below ($1 \leqslant a, b, c, d \leqslant p$):

- Variance of on-diagonal elements

$$A\mathbb{V}(\sqrt{n}\hat{\tilde{s}}_{aa}(F_{\mathbf{\Lambda}})) = \mathbb{E}\left[\frac{(\tilde{D}^{-}(\mathbf{z}, [\mathbf{Z}]))^4\lambda_a^2 z_a^4}{(\mathbf{z}^T\mathbf{\Lambda}\mathbf{z})^2}\right] - \tilde{\lambda}_a^2$$

- Variance of off-diagonal elements ($a \neq b$)

$$A\mathbb{V}(\sqrt{n}\hat{\tilde{s}}_{ab}(F_{\mathbf{\Lambda}})) = \mathbb{E}\left[\frac{(D^{-}(\mathbf{z}, [\mathbf{Z}]))^4\lambda_a\lambda_b z_a^2 z_b^2}{(\mathbf{z}^T\mathbf{\Lambda}\mathbf{z})^2}\right]$$

- Covariance of two on-diagonal elements ($a \neq b$)

$$A\mathbb{V}(\sqrt{n}\hat{\tilde{s}}_{aa}(F_{\mathbf{\Lambda}}), \sqrt{n}\hat{\tilde{s}}_{bb}(F_{\mathbf{\Lambda}})) = \mathbb{E}\left[\frac{(D^{-}(\mathbf{z}, [\mathbf{Z}]))^4\lambda_a\lambda_b z_a^2 z_b^2}{(\mathbf{z}^T\mathbf{\Lambda}\mathbf{z})^2}\right] - \tilde{\lambda}_a\tilde{\lambda}_b$$

- Covariance of two off-diagonal elements ($a \neq b \neq c \neq d$)

$$A\mathbb{V}(\sqrt{n}\hat{\tilde{s}}_{ab}(F_{\mathbf{\Lambda}}), \sqrt{n}\hat{\tilde{s}}_{cd}(F_{\mathbf{\Lambda}})) = 0$$

- Covariance of one off-diagonal and one on-diagonal element ($a \neq b \neq c$)

$$A\mathbb{V}(\sqrt{n}\hat{\tilde{s}}_{ab}(F_{\mathbf{\Lambda}}), \sqrt{n}\hat{\tilde{s}}_{cc}(F_{\mathbf{\Lambda}})) = 0$$

4.7.2 Asymptotics of eigenvectors and eigenvalues

The following result allows us to obtain asymptotic joint distributions of eigenvectors and eigenvalues of the sample DCM, provided we know the limiting distribution of

the sample DCM itself:

Theorem 4.7.1. *(Taskinen et al., 2012) Let $F_{\mathbf{\Lambda}}$ be an elliptical distribution with a diagonal covariance matrix $\mathbf{\Lambda}$, and $\hat{\mathbf{C}}$ be any positive definite symmetric $p \times p$ matrix such that at $F_{\mathbf{\Lambda}}$ the limiting distribution of $\sqrt{n}\,\mathrm{vec}(\hat{\mathbf{C}} - \mathbf{\Lambda})$ is a p^2-variate (singular) normal distribution with mean zero. Write the spectral decomposition of $\hat{\mathbf{C}}$ as $\hat{\mathbf{C}} = \hat{\mathbf{P}}\hat{\mathbf{\Lambda}}\hat{\mathbf{P}}^T$. Then the limiting distributions of $\sqrt{n}\,\mathrm{vec}(\hat{\mathbf{P}} - \mathbf{I}_p)$ and $\sqrt{n}\,\mathrm{vec}(\hat{\mathbf{\Lambda}} - \mathbf{\Lambda})$ are multivariate (singular) normal and*

$$\sqrt{n}\,\mathrm{vec}(\hat{\mathbf{C}} - \mathbf{\Lambda}) = \left[(\mathbf{\Lambda} \otimes \mathbf{I}_p) - (\mathbf{I}_p \otimes \mathbf{\Lambda})\right] \sqrt{n}\,\mathrm{vec}(\hat{\mathbf{P}} - \mathbf{I}_p) + \sqrt{n}\,\mathrm{vec}(\hat{\mathbf{\Lambda}} - \mathbf{\Lambda}) + o_P(1) \quad (4.7.1)$$

The first matrix picks only off-diagonal elements of the LHS and the second one only diagonal elements. We shall now use this as well as the form of the asymptotic covariance matrix of the vec of sample DCM, i.e. $\tilde{\mathbf{V}}(F)$ to obtain limiting variance and covariances of eigenvalues and eigenvectors.

Corollary 4.7.2. *Consider the sample DCM $\tilde{\mathbb{S}}(F) = \sum_{i=1}^{n} (D^-(\mathbf{x}_i, [\mathbb{X}_n]))^2 \mathbb{S}(\mathbf{x}_i; \hat{\boldsymbol{\mu}}_n)/n$ and its spectral decomposition $\tilde{\mathbb{S}}(F) = \hat{\tilde{\mathbf{\Gamma}}}\hat{\tilde{\mathbf{\Lambda}}}\hat{\tilde{\mathbf{\Gamma}}}^T$. Then the matrices $\mathbf{G} = \sqrt{n}(\hat{\tilde{\mathbf{\Gamma}}} - \mathbf{\Gamma})$ and $\mathbf{L} = \sqrt{n}(\hat{\tilde{\mathbf{\Lambda}}} - \tilde{\mathbf{\Lambda}})$ have independent distributions. The random variable $\mathrm{vec}(\mathbf{G})$ asymptotically has a p^2-variate normal distribution with mean $\mathbf{0}_{p^2}$, and the asymptotic variance and covariance of different columns of $\mathbf{G} = (\mathbf{g}_1, ..., \mathbf{g}_p)$ are as follows:*

$$A\mathbb{V}(\mathbf{g}_i) = \sum_{k=1;k\neq i}^{p} \frac{1}{(\tilde{\lambda}_k - \tilde{\lambda}_i)^2} \mathbb{E}\left[\frac{(D^-(\mathbf{z}, [\mathbf{Z}]))^4 \lambda_i \lambda_k z_i^2 z_k^2}{(\mathbf{z}^T \mathbf{\Lambda} \mathbf{z})^2}\right] \boldsymbol{\gamma}_k \boldsymbol{\gamma}_k^T \quad (4.7.2)$$

$$A\mathbb{V}(\mathbf{g}_i, \mathbf{g}_j) = -\frac{1}{(\tilde{\lambda}_i - \tilde{\lambda}_j)^2} \mathbb{E}\left[\frac{(D^-(\mathbf{z}, [\mathbf{Z}]))^4 \lambda_i \lambda_j z_i^2 z_j^2}{(\mathbf{z}^T \mathbf{\Lambda} \mathbf{z})^2}\right] \boldsymbol{\gamma}_j \boldsymbol{\gamma}_i^T; \quad i \neq j \quad (4.7.3)$$

where $\mathbf{\Gamma} = (\boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_p)$. The vector consisting of diagonal elements of bfL, say

$\mathbf{l} = (l_1, ..., l_p)^T$ *asymptotically has a p-variate normal distribution with mean* $\mathbf{0}_p$ *and variance-covariance elements:*

$$A\mathbb{V}(l_i) = \mathbb{E}\left[\frac{(D^-(\mathbf{z}, [\mathbf{Z}]))^4 \lambda_i^2 z_i^4}{(\mathbf{z}^T \Lambda \mathbf{z})^2}\right] - \tilde{\lambda}_i^2 \tag{4.7.4}$$

$$A\mathbb{V}(l_i, l_j) = \mathbb{E}\left[\frac{(D^-(\mathbf{z}, [\mathbf{Z}]))^4 \lambda_i \lambda_j z_i^2 z_j^2}{(\mathbf{z}^T \Lambda \mathbf{z})^2}\right] - \tilde{\lambda}_i \tilde{\lambda}_j; \quad i \neq j \tag{4.7.5}$$

Proof of Corollary 4.7.2. In spirit, this corollary is similar to Theorem 13.5.1 in Anderson (2003). Due to the decomposition (4.7.1) we have, for the distribution $F_{\mathbf{\Lambda}}$, the following relation between any off-diagonal element of $\tilde{\mathbb{S}}(F_{\mathbf{\Lambda}})$ and the corresponding element in the estimate of eigenvectors $\hat{\tilde{\mathbf{\Gamma}}}(F_{\mathbf{\Lambda}})$:

$$\sqrt{n}\hat{\tilde{\gamma}}_{ij}(F_{\mathbf{\Lambda}}) = \sqrt{n}\frac{\tilde{\mathbb{S}}_{ij}(F_{\mathbf{\Lambda}})}{\tilde{\lambda}_i - \tilde{\lambda}_j}; \quad i \neq j$$

So that for eigenvector estimates of the original F we have

$$\sqrt{n}(\hat{\tilde{\boldsymbol{\gamma}}}_i - \boldsymbol{\gamma}_i) = \sqrt{n}\mathbf{\Gamma}(\hat{\tilde{\boldsymbol{\gamma}}}_i(F_{\mathbf{\Lambda}}) - \mathbf{e}_i) = \sqrt{n}\left[\sum_{k=1;k\neq i}^{p} \hat{\tilde{\gamma}}_{ik}(F_{\mathbf{\Lambda}})\boldsymbol{\gamma}_k + (\hat{\tilde{\gamma}}_{ii}(F_{\mathbf{\Lambda}}) - 1)\boldsymbol{\gamma}_i\right] \tag{4.7.6}$$

Now $\sqrt{n}(\hat{\tilde{\gamma}}_{ii}(F_{\mathbf{\Lambda}}) - 1) = o_P(1)$ and $A\mathbb{V}(\sqrt{n}\tilde{\mathbb{S}}_{ik}(F_{\mathbf{\Lambda}}), \sqrt{n}\tilde{\mathbb{S}}_{il}(F_{\mathbf{\Lambda}})) = 0$ for $k \neq l$, so the above equation implies

$$A\mathbb{V}(\mathbf{g}_i) = AVar(\sqrt{n}(\hat{\tilde{\boldsymbol{\gamma}}}_i - \boldsymbol{\gamma}_i)) = \sum_{k=1;k\neq i}^{p} \frac{A\mathbb{V}(\sqrt{n}\tilde{\mathbb{S}}_{ik}(F_{\mathbf{\Lambda}}))}{(\tilde{\lambda}_i - \tilde{\lambda}_k)^2}\boldsymbol{\gamma}_k\boldsymbol{\gamma}_k^T$$

For the covariance terms, from (4.7.6) we get, for $i \neq j$,

$$\begin{aligned} A\mathbb{V}(\mathbf{g}_i, \mathbf{g}_j) &= A\mathbb{V}(\sqrt{n}(\hat{\tilde{\boldsymbol{\gamma}}}_i - \boldsymbol{\gamma}_i), \sqrt{n}(\hat{\tilde{\boldsymbol{\gamma}}}_j - \boldsymbol{\gamma}_j)) \\ &= A\mathbb{V}\left(\sum_{k=1;k\neq i}^{p} \sqrt{n}\hat{\tilde{\gamma}}_{ik}(F_{\boldsymbol{\Lambda}})\boldsymbol{\gamma}_k, \sum_{k=1;k\neq j}^{p} \sqrt{n}\hat{\tilde{\gamma}}_{jk}(F_{\boldsymbol{\Lambda}})\boldsymbol{\gamma}_k\right) \\ &= A\mathbb{V}\left(\sqrt{n}\hat{\tilde{\gamma}}_{ij}(F_{\boldsymbol{\Lambda}})\boldsymbol{\gamma}_j, \sqrt{n}\hat{\tilde{\gamma}}_{ji}(F_{\boldsymbol{\Lambda}})\boldsymbol{\gamma}_i\right) \\ &= -\frac{A\mathbb{V}(\sqrt{n}\tilde{\mathbb{S}}_{ij}(\Lambda))}{(\tilde{\lambda}_i - \tilde{\lambda}_j)^2}\boldsymbol{\gamma}_j\boldsymbol{\gamma}_i^T \end{aligned}$$

The exact forms given in the statement of the corollary now follows from the Form of $\tilde{\mathbf{V}}(F)$ in Subection 4.7.1.

For the on-diagonal elements of $\tilde{\mathbb{S}}(F_{\boldsymbol{\Lambda}})$ Theorem 4.7.1 gives us $\sqrt{n}\hat{\tilde{\lambda}}_i(F_{\boldsymbol{\Lambda}}) = \sqrt{n}\tilde{\mathbb{S}}_{ii}(F_{\boldsymbol{\Lambda}})$ for $i = 1, ..., p$. Hence

$$\begin{aligned} A\mathbb{V}(l_i) &= A\mathbb{V}(\sqrt{n}\hat{\tilde{\lambda}}_i - \sqrt{n}\tilde{\lambda}_i) \\ &= A\mathbb{V}(\sqrt{n}\hat{\tilde{\lambda}}_i(F_{\boldsymbol{\Lambda}}) - \sqrt{n}\tilde{\lambda}_i(F_{\boldsymbol{\Lambda}})) \\ &= A\mathbb{V}(\sqrt{n}\tilde{\mathbb{S}}_{ii}(F_{\boldsymbol{\Lambda}})) \end{aligned}$$

A similar derivation gives the expression for $A\mathbb{V}(l_i, l_j); i \neq j$. Finally, since the asymptotic covariance between an on-diagonal and an off-diagonal element of $\tilde{\mathbb{S}}(F_{\boldsymbol{\Lambda}})$, it follows that the elements of $\mathbf{G}$ and diagonal elements of $\mathbf{L}$ are independent. □

4.7.3 Proofs

Proof of Proposition 4.2.1. Under contiguous alternatives $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$, the weighted sign test statistic $T_{n,w}$ has mean $\mathbb{E}(w(\mathbf{Z})\mathbf{S}(\mathbf{Z}))$. For spherically symmetric $\mathbf{Z}$, $w(\mathbf{Z})$ depends on $\mathbf{Z}$ only through its norm. Since $\|\mathbf{Z}\|$ and $\mathbf{S}(\mathbf{Z})$ are independent, we get

$\mathbb{E}(w(\mathbf{Z})\mathbf{S}(\mathbf{Z})) = \mathbb{E}w(\mathbf{Z}).\mathbb{E}\mathbf{S}(\mathbf{Z})$. The same kind of decomposition holds for $\mathbb{V}(w(\mathbf{Z})\mathbf{S}(\mathbf{Z}))$.

We can now simplify the approximate local power $\beta_{n,w}$ of the level-α $(0 < \alpha < 1)$ test based on $T_{n,w}$:

$$\begin{aligned}\beta_{n,w} &= K_p\left(\chi^2_{p,\alpha} + n\times\right.\\ &\quad \left.(\mathbb{E}(w(\mathbf{Z})\mathbf{S}(\mathbf{Z}))^T[\mathbb{E}(w^2(\mathbf{Z})\mathbf{S}(\mathbf{Z})\mathbf{S}(\mathbf{Z})^T)]^{-1}(\mathbb{E}(w(\mathbf{Z})\mathbf{S}(\mathbf{Z}))\right)\\ &= K_p\left(\chi^2_{p,\alpha} + \frac{\mathbb{E}^2 w(\mathbf{Z})}{\mathbb{E}w^2(\mathbf{Z})}.\mathbb{E}\mathbf{S}(\mathbf{Z})^T[\mathbb{V}\mathbf{S}(\mathbf{Z})]^{-1}\mathbb{E}\mathbf{S}(\mathbf{Z})\right)\end{aligned}$$

where K_p and $\chi^2_{p,\alpha}$ are distribution function and upper-α cutoff of a χ^2_p distribution, respectively. Since $\mathbb{E}^2 w(\mathbf{Z}) \leqslant \mathbb{E}w(\mathbf{Z})$, $\beta_{n,w}$ the largest possible value of $\beta_{n,w}$ is $K_p(\chi^2_{p,\alpha} + \mathbb{E}\mathbf{S}(\mathbf{Z})^T[\mathbb{V}\mathbf{S}(\mathbf{Z})]^{-1}\mathbb{E}\mathbf{S}(\mathbf{Z}))$, the approximate power of the unweighted sign test statistic. Equality is of course achieved when $w(\mathbf{Z})$ is a constant independent of $\mathbf{Z}$. □

Proof of corollary 4.2.3. Since $\boldsymbol{\epsilon} \sim \mathcal{E}(\boldsymbol{\mu}, k\mathbf{I}_p, g)$, $r = \|\boldsymbol{\epsilon}\|$ and $\mathbf{S}(\boldsymbol{\epsilon})$ are independent. Also $w(\boldsymbol{\epsilon}) = f(r)$. Thus

$$\begin{aligned}\mathbf{A}_w &= \mathbb{E}\left(\frac{f(r)}{r}\right)\mathbb{E}(\mathbf{I}_p - \mathbf{S}(\boldsymbol{\epsilon})\mathbf{S}(\boldsymbol{\epsilon})^T);\\ \mathbf{B}_w &= \mathbb{E}f^2(r)\mathbb{E}(\mathbf{S}(\boldsymbol{\epsilon})\mathbf{S}(\boldsymbol{\epsilon})^T)\end{aligned}$$

We conclude by substituting these in (4.2.3). □

Sketch of proofs for equations (4.2.4) and (4.2.5). A first step to obtain asymptotic normality for the high-dimensional location test statistic $C_{n,w}$ is obtaining an equivalent result of Lemma 2.1 in Wang et al. (2015):

Lemma 4.7.3. *Under the conditions*
(C1) $Tr(\boldsymbol{\Sigma}^4) = o(Tr^2(\boldsymbol{\Sigma}^2))$,
(C2) $Tr^4(\boldsymbol{\Sigma})/Tr^2(\boldsymbol{\Sigma}^2)\exp[-Tr^2(\boldsymbol{\Sigma})/128p\lambda^2_{\max}(\boldsymbol{\Sigma})] = o(1)$

when H_0 is true we have

$$\mathbb{E}[(\boldsymbol{\epsilon}_{w1}^T\boldsymbol{\epsilon}_{w2})^4] = O(1)\mathbb{E}^2[(\boldsymbol{\epsilon}_{w1}^T\boldsymbol{\epsilon}_{w2})^2] \quad (4.7.7)$$

$$\mathbb{E}[(\boldsymbol{\epsilon}_{w1}^T B_w\boldsymbol{\epsilon}_{w1})^2] = O(1)\mathbb{E}^2[(\boldsymbol{\epsilon}_{w1}^T B_w\boldsymbol{\epsilon}_{w1})^2] \quad (4.7.8)$$

$$\mathbb{E}[(\boldsymbol{\epsilon}_{w1}^T B_w\boldsymbol{\epsilon}_{w2})^2] = o(1)\mathbb{E}^2[(\boldsymbol{\epsilon}_{w1}^T B_w\boldsymbol{\epsilon}_{w1})^2] \quad (4.7.9)$$

with $\boldsymbol{\epsilon} \sim \mathbb{E}(\mathbf{0}_p, \boldsymbol{\Lambda}, g)$ and $\boldsymbol{\epsilon}_w = w(\boldsymbol{\epsilon})\mathbf{S}(\boldsymbol{\epsilon})$.

A proof of this lemma is derived using results in section 3 of El Karoui (2009), noticing that any-scalar valued 1-Lipschitz function of $\boldsymbol{\epsilon}_w$ is a M_w-Lipschitz function of $\mathbf{S}(\boldsymbol{\epsilon})$, with $M_w = \sup_{\boldsymbol{\epsilon}} w(\boldsymbol{\epsilon})$. Same steps as in the proof of Theorem 2.2 in Wang et al. (2015) follow now, using the lemma above in place of Lemma 2.1 therein, to establish asymptotic normality of $C_{n,w}$ under H_0.

To derive the asymptotic distribution under contiguous alternatives we need the conditions (C3)-(C6) in Wang et al. (2015), as well as slightly modified versions of Lemmas A.4 and A.5:

Lemma 4.7.4. *Given that condition (C3) holds, we have $\lambda_{\max}(\mathbf{B}_w) \leqslant 2\frac{\lambda_{\max}}{Tr(\boldsymbol{\Sigma})}(1+o(1))$.*

Lemma 4.7.5. *Define $\mathbf{D}_w = \mathbb{E}\left[\frac{w^2(\boldsymbol{\epsilon})}{\|\boldsymbol{\epsilon}\|^2}(\mathbf{I}_p - \mathbf{S}(\boldsymbol{\epsilon})\mathbf{S}(\boldsymbol{\epsilon})^T)\right]$. Then $\lambda_{\max}(\mathbf{A}_w) \leqslant \mathbb{E}(w(\boldsymbol{\epsilon})/\|\boldsymbol{\epsilon}\|)$ and $\lambda_{\max}(\mathbf{D}_w) \leqslant \mathbb{E}(w(\boldsymbol{\epsilon})/\|\boldsymbol{\epsilon}\|)^2$. Further, if (C3) and (C4) hold then $\lambda_{\min}(\mathbf{A}_w) \geqslant \mathbb{E}(w(\boldsymbol{\epsilon})/\|\boldsymbol{\epsilon}\|)(1+o(1))/\sqrt{3}$.*

The proof now exactly follows steps in the proof of theorem 2.3 in Wang et al. (2015), replacing vector signs by weighted signs, using the fact that $w(\boldsymbol{\epsilon})$ is bounded above by M_w while applying conditions (C5)-(C6) and lemmas A.1, A.2, A.3, and finally using the above two lemmas in place of lemmas A.4 and A.5 respectively. □

Proof of Theorem 4.3.1. The proof follows directly from writing out the expression

of $\mathbb{V}(\tilde{\mathbf{X}})$:

$$\begin{aligned}\mathbb{V}(\tilde{\mathbf{X}}) &= \mathbb{E}(\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T) - \mathbb{E}\tilde{\mathbf{X}}\mathbb{E}\tilde{\mathbf{X}}^T \\ &= \mathbf{\Gamma}.\mathbb{E}\left[(D^-(\mathbf{z},[\mathbf{Z}]))^2 \frac{\|\mathbf{z}\|^2}{\|\mathbf{\Lambda}^{1/2}\mathbf{z}\|}\mathbf{\Lambda}^{1/2}\mathbf{S}(\mathbf{z})\mathbf{S}(\mathbf{z})^T\mathbf{\Lambda}^{1/2}\right]\mathbf{\Gamma}^T - \mathbf{0}_p\mathbf{0}_p^T \\ &= \mathbf{\Gamma}.\mathbb{E}\left[(D^-(\mathbf{z},[\mathbf{Z}]))^2 \frac{\mathbf{\Lambda}^{1/2}\mathbf{z}\mathbf{z}^T\mathbf{\Lambda}^{1/2}}{\mathbf{z}^T\mathbf{\Lambda}\mathbf{z}}\right]\mathbf{\Gamma}^T\end{aligned}$$

□

Proof of Lemma 4.3.2. For two positive definite matrices $\mathbf{A}, \mathbf{B}$, we denote by $\mathbf{A} > \mathbf{B}$ that $\mathbf{A} - \mathbf{B}$ is positive definite. Also, denote

$$\mathbb{S}_n = \sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n}\left|(D^-(\mathbf{x}_i,[\mathbb{X}_n]))^2 - (D^-(\mathbf{x}_i,[\mathbf{X}]))^2\right|\mathbb{S}(\mathbf{x}_i;\hat{\boldsymbol{\mu}}_n)\right]$$

Due to the assumption of uniform convergence, given $\epsilon > 0$ we can find $N \in \mathbb{N}$ such that

$$|(D^-(\mathbf{x}_i,[\mathbb{X}_{n_1}]))^2 - (D^-(\mathbf{x}_i,[\mathbf{X}]))^2| < \epsilon \tag{4.7.10}$$

for all $n_1 \geqslant N; i = 1, 2, ..., n_1$. This implies

$$\begin{aligned}\mathbb{S}_{n_1} &< \epsilon\sqrt{n_1}\left[\frac{1}{n_1}\sum_{i=1}^{n_1}\mathbb{S}(\mathbf{x}_i;\hat{\boldsymbol{\mu}}_{n_1})\right] \\ &= \epsilon\sqrt{n_1}\left[\frac{1}{n_1}\sum_{i=1}^{n_1}\left\{\mathbb{S}(\mathbf{x}_i;\hat{\boldsymbol{\mu}}_{n_1}) - \mathbb{S}(\mathbf{x}_i;\boldsymbol{\mu})\right\} + \frac{1}{n_1}\sum_{i=1}^{n_1}\mathbb{S}(\mathbf{x}_i;\boldsymbol{\mu})\right] \qquad (4.7.11)\end{aligned}$$

We now construct a sequence of positive definite matrices $\{a_k(\mathbf{B}_k + \mathbf{C}_k) : k \in \mathbb{N}\}$ so that

$$a_k = \frac{1}{k}, \quad \mathbf{B}_k = \sqrt{N_k}\left[\frac{1}{N_k}\sum_{i=1}^{N_k}\left\{\mathbb{S}(\mathbf{x}_i;\hat{\boldsymbol{\mu}}_{N_k}) - \mathbb{S}(\mathbf{x}_i;\boldsymbol{\mu})\right\}\right]$$

$$\mathbf{C}_k = \sqrt{N_k}\left[\frac{1}{N_k}\sum_{i=1}^{N_k}\mathbb{S}(\mathbf{x}_i;\boldsymbol{\mu})\right]$$

where $N_k \in \mathbb{N}$ gives the relation (4.7.10) in place of N when we take $\epsilon = 1/k$. Under conditions $\mathbb{E}\|\mathbf{x}-\boldsymbol{\mu}\|^{-3/2} < \infty$ and $\sqrt{n}(\hat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}) = O_P(1)$, the sample SCM with unknown location parameter $\hat{\boldsymbol{\mu}}_n$ has the same asymptotic distribution as the SCM with known location $\boldsymbol{\mu}$ (Dürre et al., 2014), hence $\mathbf{B}_k = o_P(1)$. Also $\mathbf{C}_k = O_P(1)$, thus $a_k(\mathbf{B}_k + \mathbf{C}_k) \xrightarrow{P} 0$ as $k \to \infty$.

Now (4.7.11) implies that for any $\epsilon_1 > 0$, $\mathbb{S}_{N_k} > \epsilon_1 \Rightarrow a_k(\mathbf{B}_k + \mathbf{C}_k) > \epsilon_1$, which means $P(\mathbb{S}_{N_k} > \epsilon_1) < P(a_k(\mathbf{B}_k + \mathbf{C}_k) > \epsilon_1)$. Hence the subsequence $\{\mathbb{S}_{N_k}\} \xrightarrow{P} 0$. Since the main sequence $\{\mathbb{S}_n\}$ is bounded below by 0, this implies $\{\mathbb{S}_n\} \xrightarrow{P} 0$. Finally, we have that

$$\begin{aligned}\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n}(D^-(\mathbf{x}_i,[\mathbb{X}_n]))^2\mathbb{S}(\mathbf{x}_i;\hat{\boldsymbol{\mu}}_n) - \frac{1}{n}\sum_{i=1}^{n}(D^-(\mathbf{x}_i,[\mathbf{X}]))^2\mathbb{S}(\mathbf{x}_i;\boldsymbol{\mu})\right] \;\leqslant\; \\ \mathbb{S}_n + \max_{\mathbf{x}}(D(\mathbf{x},[\mathbf{X}]))^2.\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n}\{\mathbb{S}(\mathbf{x}_i;\hat{\boldsymbol{\mu}}_n) - \mathbb{S}(\mathbf{x}_i;\boldsymbol{\mu})\}\right] \qquad (4.7.12)\end{aligned}$$

The second summand on the right hand side is $o_P(1)$ due to Dürre et al. (2014) as mentioned before, so we have the needed. □

Proof of Theorem 4.3.3. The quantity in the statement of the theorem can be broken down as:

$$\begin{aligned}\sqrt{n}\left[\operatorname{vec}\left\{\frac{1}{n}\sum_{i=1}^{n}(D^-(\mathbf{x}_i,[\mathbf{X}]))^2\mathbb{S}(\mathbf{x}_i;\hat{\boldsymbol{\mu}}_n)\right\} - \operatorname{vec}\left\{\frac{1}{n}\sum_{i=1}^{n}(D^-(\mathbf{x}_i,[\mathbf{X}]))^2\mathbb{S}(\mathbf{x}_i;\boldsymbol{\mu})\right\}\right] + \\ \sqrt{n}\left[\operatorname{vec}\left\{\frac{1}{n}\sum_{i=1}^{n}(D^-(\mathbf{x}_i,[\mathbf{X}]))^2\mathbb{S}(\mathbf{x}_i;\boldsymbol{\mu})\right\} - \mathbb{E}\left[\operatorname{vec}\left\{(D^-(\mathbf{x},[\mathbf{X}]))^2\mathbb{S}(\mathbf{x};\boldsymbol{\mu})\right\}\right]\right]\end{aligned}$$

The first part goes to 0 in probability by Lemma 4.3.2, and applying Slutsky's theorem we get the required convergence. □

Proof of Theorem 4.3.4. We are going to prove the following:

1. $\|\hat{\tilde{\mathbf{\Gamma}}} - \mathbf{\Gamma}\|_F \xrightarrow{P} 0$, and
2. $\|\hat{\mathbf{\Lambda}} - \mathbf{\Lambda}\|_F \xrightarrow{P} 0$

as $n \to \infty$. For (1), we notice $\sqrt{n}\,\text{vec}(\hat{\tilde{\mathbf{\Gamma}}} - \mathbf{\Gamma})$ asymptotically has a (singular) multivariate normal distribution following Corollary 4.7.2, so that $\|\hat{\tilde{\mathbf{\Gamma}}} - \mathbf{\Gamma}\|_F = O_P(1/\sqrt{n})$ using Prokhorov's theorem.

It is now enough to prove convergence in probability of the individual eigenvalue estimates $\hat{\lambda}_i; i = 1, ..., p$. For this, define estimates $\hat{\lambda}_i^*$ as median-of-small-variances estimator of the *true* score vectors $\mathbf{\Gamma}^T \mathbb{X}_n$. For this we have

$$|\hat{\lambda}_i^* - \lambda_i| \xrightarrow{P} 0 \tag{4.7.13}$$

using Theorem 3.1 of Minsker (2015), with $\mu = \lambda_i$. Now $\hat{\lambda}_i = \text{med}_j(\mathbb{V}(\mathbb{X}_{G_j}^T \hat{\tilde{\boldsymbol{\gamma}}}_i))$ and $\hat{\lambda}_i^* = \text{med}_j(\mathbb{V}(\mathbb{X}_{G_j}^T \boldsymbol{\gamma}_i))$, so that

$$\begin{aligned} |\hat{\lambda}_i - \hat{\lambda}_i^*| &\leqslant \text{med}_j \left[\mathbb{V}(\mathbb{X}_{G_j}^T (\hat{\tilde{\boldsymbol{\gamma}}}_i - \boldsymbol{\gamma}_i)) \right] \\ &\leqslant \|\hat{\tilde{\boldsymbol{\gamma}}}_i - \boldsymbol{\gamma}_i\|^2 \text{med}_j \left[\text{Tr}(\mathbb{V}(\mathbb{X}_{G_j})) \right] \end{aligned}$$

using Cauchy-Schwarz inequality. Combining the facts $\|\hat{\tilde{\boldsymbol{\gamma}}}_i - \boldsymbol{\gamma}_i\| = O_P(1/\sqrt{n})$ and $\text{med}_j[\text{Tr}(\mathbb{V}(\mathbb{X}_{G_j}))] \xrightarrow{P} \text{Tr}(\mathbf{\Sigma})$ (Minsker, 2015) with (4.7.13), we get the needed.

□

Chapter 5

Nonconvex Penalized Regression using Depth-based Penalty

5.1 Introduction

Consider the multitask linear regression model:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}$$

where $\mathbf{Y} \in \mathbb{R}^{n\times q}$ is the matrix of responses, and $\mathbf{E}$ is $n \times q$ the noise matrix: each row of which is drawn from $\mathcal{N}_q(\mathbf{0}_q, \mathbf{\Sigma})$ for a $q \times q$ positive definite matrix $\mathbf{\Sigma}$. We are interested in sparse estimates of the coefficient matrix $\mathbf{B}$ through solving penalized regression problems of the form

$$\min_{\mathbf{B}} \mathrm{Tr}\{(\mathbf{Y} - \mathbf{XB})^T(\mathbf{Y} - \mathbf{XB})\} + P_\lambda(\mathbf{B}) \tag{5.1.1}$$

The frequently studied classical linear model may be realized as a special case of this for $q = 1$, where given a size-n sample of random responses $\mathbf{y} = (y_1, y_2, ..., y_n)^T$ and p-dimensional predictors $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n)^T$, the above model may now be written as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}; \quad \boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)^T \sim \mathcal{N}_n(\mathbf{0}_n, \sigma^2\mathbf{I}_p).$$

Here the typical objective is to estimate the parameter vector $\boldsymbol{\beta}$ by minimizing $\sum_{i=1}^{n} \rho(y_i - \mathbf{x}_i^T \boldsymbol{\beta})$, for some loss function $\rho(.)$. Selecting important variables in this setup is often significant from an inferential and predictive perspective it is generally achieved by obtaining an estimate of $\boldsymbol{\beta}$ that minimizes a linear combination of the loss function and a 'penalty' term $P(\boldsymbol{\beta}) = \sum_{j=1}^{p} p(|\beta_j|)$, instead of only the loss function:

$$\hat{\boldsymbol{\beta}}_n = \arg\min_{\boldsymbol{\beta}} \left[\sum_{i=1}^{n} \rho(y_i - \mathbf{x}_i^T \boldsymbol{\beta}) + \lambda_n P(\boldsymbol{\beta}) \right] \tag{5.1.2}$$

where λ_n is a tuning parameter depending on sample size. The penalty term is generally a measure of model complexity, providing a control against overfitting. Using a l_0 norm as penalty at this point, i.e. $p(z) = \mathbb{I}_{z \neq 0}$, gives rise to the information criterion-based paradigm of statistical model selection, which goes back to the Akaike Information Criterion (AIC: Akaike (1970)). Owing to the intractability of this problem due to an exponentially growing model space, researchers have been exploring the use of functions that are non-differentiable at the origin as $p(.)$. This dates back to the celebrated LASSO (Tibshirani, 1996) which uses l_1 norm, adaptive LASSO (Zou, 2006) that reweights the coordinate-wise LASSO penalties based on Ordinary Least Square (OLS) estimate of $\boldsymbol{\beta}$, and Fan and Li (2001); Zhang (2010) who used non-convex penalties to limit influence of large entries in the coefficient vector $\boldsymbol{\beta}$, resulting in improved estimation. Further, Zou and Li (2008) and Wang et al. (2013) provided efficient algorithms for computing solutions to the nonconvex penalized problems.

Two immediate extensions of the univariate-response penalized sparse regression paradigm are group-wise penalties and multivariate penalized regression. Applying penalties at variable group level instead of individual variables gives rise to Group LASSO (Bakin, 1999). From an application perspective, this utilizes additional relevant information on the natural grouping of predictors: for example multiple correlated genes, or blockwise wavelet shrinkage (Antoniadis and Fan, 2001). On the

other hand, for multitask regression, penalizing at the coefficient matrix-level results in better estimation and prediction performance compared to performing q separate LASSO regressions to recover its corresponding columns (Rothman et al., 2010).

Compared to sparse single-response regression where the penalty term can be broken down to elementwise penalties, in the multivariate response scenario we need to consider two levels of sparsity. The first level is recovering the set of predictors having non-zero effects on all the responses, as well as estimating their values. Assuming the coefficient matrix $\mathbf{B} \in \mathbb{R}^{p\times q}$ is made of rows $(\mathbf{b}_1, ..., \mathbf{b}_p)^T$, this means determining the set $\bigcup_k \mathcal{S}_k$, with $\mathcal{S}_k := \{k : b_{jk} \neq 0, j = 1, 2, ..., p\}$. This is called *support union recovery*, and is more effective in recovering non-zero elements of $\mathbf{B}$ compared to the naïve approach of performing q separate sparse regularized regressions and combining the results (Obozinski et al., 2011). The second level of sparsity is concerned with recovering non-zero elements *within* the non-zero rows obtained from the first step. Our method addresses both of these issues.

Specifically, we consider the case of performing support union recovery by considering the inverse depth functions introduced in Chapter 4 as row-level regularizers: $P_\lambda(\mathbf{B}) = \sum_{j=1}^p \lambda D^-(\mathbf{b}_j, F)$ where F is some probability distribution fixed beforehand. Section 5.2 motivates the use of a general depth-based regularization scheme in the multitask regression setup. From Section 5.3 onward we choose to concentrate on the scenario when $D^-(\mathbf{b}_j, F) = p_{,F}(\|\mathbf{b}_j\|_2)$, i.e. the row-level penalty is a potentially nonconvex scalar-valued function of the row-norm. This automatically tempers the effects of large regression coefficients in the case of general q-dimensional response: which is not the case for methods based on l_1-norm penalization, e.g. Lasso. We derive asymptotic results ensuring support union recovery, as well as provide an iterative algorithm for calculating the corresponding penalized estimator. We also show that a simple corrective thresholding on elements of the first level row-sparse estimator ensures sparse recovery of within-row elements as well. Additional theoretical

results in the orthogonal design case are discussed in Section 5.4, and simulation experiments are presented to compare our algorithm with other methods in Section 5.5. We present a data application of the algorithm in Section 5.6, followed by conclusions. Section 5.8 contains proofs of our theoretical results.

5.2 Depth-based regularization

We incorporate measures of data depth as a row-level penalty function in (5.1.1). Specifically, we estimate the coefficient matrix $\mathbf{B}$ by solving the following constrained optimization problem:

$$\hat{\mathbf{B}} = \arg\min_{\mathbf{B}} \left[\mathrm{Tr}\{(\mathbf{Y} - \mathbf{XB})^T(\mathbf{Y} - \mathbf{XB})\} + \lambda_n \sum_{j=1}^{p} P(\mathbf{b}_j, F) \right] \tag{5.2.1}$$

where $P(\mathbf{b}_j, F)$ is a function that measures the peripherality of the j-th row of $\mathbf{B}$ with respect to a fixed probability distribution F, as defined in Chapter 4. We refer to F as the *reference distribution*, and consider it fixed in the estimation process. In multitask learning, any additive penalty function of the form $P_\lambda(\mathbf{B}) = \sum_{j=1}^{p} \lambda p(\mathbf{b}_j)$ regularizes individual rows of the coefficient matrix by providing a control over their distance from the origin $\mathbf{0}_q$ through some norm (e.g. the l_1/l_q penalty: Neghaban and Wainwright (2011)), or a combination of norms (e.g. the Adaptive Multi-task Elastic-Net: Chen et al. (2012)). Through (5.2.1) we attempt to generalize this notion by proposing to regularize using the 'distance' from a *probability distribution* centered at the origin. Of course, any existing method of norm-based regularization arises as a special case by by using the norm (or combination of norms) as the peripherality function and taking as F the degenerate distribution centered at $\mathbf{0}_q$. While it is possible to use any peripherality function (or outlyingness functions, in the spirit of Zuo and Serfling (2000)) for this purpose, of special interest is the case of *inverse depth*

functions: $P(\mathbf{x}, F) = D^{-}(\mathbf{x}, F)$. Such functions essentially invert the funnel-shaped contour of the corresponding depth function (Figure 1.1). This immediately results in row-wise nonconvex penalties, where the penalty sharply increases for smaller entries inside the row but is bounded above for large values. This is easy to visualize for $p = 1$, which we show in panel a of Figure 5.1. This serves as our motivation of using data depth in regularized multitask regression.

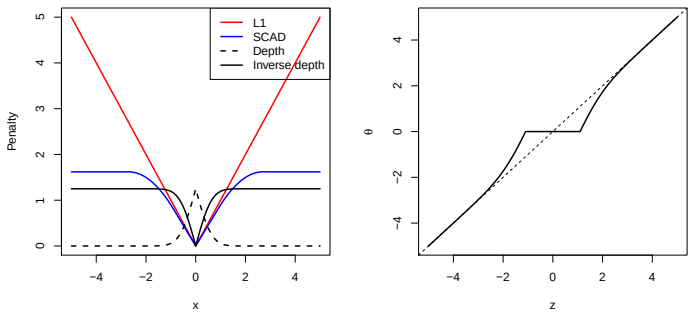
5.3 The LARN algorithm

5.3.1 Formulation

The reference distribution F is pivotal in the estimation problem in (5.2.1). While we believe that there is scope for a significant amount of theoretical analysis on the implications of different choices of F and its potential connections to Bayesian regularized support union recovery in multitask regression, here we shall work within a simplified setup. Specifically we assume that

(A1) The distribution F is spherically symmetric.

This is a fair assumption to make from a frequentist perspective, as we do not possess any extra information about the q responses being different from one another. Since F is spherically symmetric, depth at a point $\mathbf{b}$ becomes a function of $r = \|\mathbf{b}\|_2$ only, due to the affine invariance of $D(., F)$. In this situation, several depth functions have closed-form expressions: e.g. when D is projection depth and F is a p-variate standard normal distribution, $D(\mathbf{b}_j, F) = c/(c + \|\mathbf{b}_j\|); c = \Phi^{-1}(3/4)$ (Zuo, 2003), while for halfspace depth and any known F, $D(\mathbf{b}_j, F) = 1 - F_1(\|\mathbf{b}_j\|)$, F_1 being any univariate marginal of F (immediate from the definition of halfspace depth). Hence, the computational burden of calculating depths for rows of $\mathbf{B}$ becomes trivial.



(a) (b)

Figure 5.1: (a) Comparison of L1 and SCAD (Fan and Li, 2001) penalty functions with univariate halfspace depth: inverting the depth function helps obtain the nonconvex shape of the penalty function in the inverse depth; (b) Univariate thresholding rule for the LARN estimate assuming halfspace depth and max definition of inverse depth(see Section 5.4)

Because of the way we define inverse depth functions, the above holds for inverse depth functions $D^-(., F)$ as well. Thus we can write that $D^-(\mathbf{b}_j, F) = p_F(r_j)$ for some scalar-valued function $p_F(.)$. Any superscript or subscript in $\mathbf{B}$ or $\mathbf{b}_j$ will be passed accordingly to r_j. At this point we shall make the following technical assumption on $p_F(.)$:

(A2) The function $p_F(r)$ is concave in r, and continuously differentiable at every $r \neq 0$.

In general depth functions are assumed to have convex contours (Mosler, 2013), which implies quasi-concavity. Nevertheless, several depth functions adhere to concavity owing to their simplified closed forms for spherical distribution (e.g. halfspace depth and projection depth in the last paragraph). Continuous differentiability except at the origin, which is essential for admitting a sparse solution eventually, arises because of the same reason.

Keeping the above setup in mind, we now consider the first-order Taylor series

approximation of the overall penalty function:

$$\begin{aligned} P_{\lambda.F}(\mathbf{B}) &:= \lambda \sum_{j=1}^{p} p_F(r_j) \\ &\simeq \lambda \sum_{j=1}^{p} \left[p_F(r_j^*) + p_F'(r_j^*)(r_j - r_j^*) \right] \end{aligned} \tag{5.3.1}$$

for any $\mathbf{B}^*$ close to $\mathbf{B}$, and $r_j = \|\mathbf{b}_j\|_2, r_j^* = \|\mathbf{b}_j^*\|_2; j = 1, 2, ..., p$.

Thus, given a starting solution $\mathbf{B}^*$ close enough to the original coefficient matrix, $P_{\lambda.F}(\mathbf{B})$ is approximated by its conditional counterpart, say $P_{\lambda.F}(\mathbf{B}|\mathbf{B}^*)$. Following this a penalized maximum likelihood estimate for $\mathbf{B}$ can be obtained using the iterative algorithm below:

1. Take as starting value $\mathbf{B}^{(0)} = \hat{\mathbf{B}}_{\text{LS}} = (\mathbf{X}^T\mathbf{X})^-\mathbf{X}^T\mathbf{Y}$, i.e. the least square estimate of $\mathbf{B}$, set $k = 0$;

2. Calculate the next iterate by solving the penalized likelihood:

$$\mathbf{B}^{(k+1)} = \arg\min_{\mathbf{B}} \left[\text{Tr}\left\{ (\mathbf{Y} - \mathbf{X}\mathbf{B}^{(k)})^T (\mathbf{Y} - \mathbf{X}\mathbf{B}^{(k)}) \right\} + \lambda \sum_{j=1}^{p} p_F'(r_j^{(k)}) r_j \right] \tag{5.3.2}$$

3. Continue until convergence.

Taking $\hat{\mathbf{B}}_{\text{LS}}$ as a starting value ensures that $\|\hat{\mathbf{B}}_{\text{LS}} - \mathbf{B}\|_F = O(n^{-1/2})$ given the data, hence we get from (5.3.1) that

$$P_{\lambda,F}(\mathbf{B}) = P_{\lambda,F}(\mathbf{B}|\hat{\mathbf{B}}_{\text{LS}}) + \sum_{j=1}^{p} o(|r_j - \hat{r}_{j,\text{LS}}|) = P_{\lambda,F}(\mathbf{B}|\hat{\mathbf{B}}_{\text{LS}}) + \sum_{j=1}^{p} o(n^{-1/2})$$

for fixed p. This algorithm approximates contours of the nonconvex penalty function using gradient planes at successive iterates, and is a multivariate generalization of the local linear approximation algorithm of Zou and Li (2008). We call this the *Local*

Approximation by Row-wise Norm (LARN) algorithm.

LARN is a majorize-minimize (MM) algorithm where the actual objective function $Q(\mathbf{B})$ is being majorized by $R(\mathbf{B}|\mathbf{B}^{(k)})$, with

$$\begin{aligned} Q(\mathbf{B}) &= \mathrm{Tr}\left\{(\mathbf{Y}-\mathbf{XB})^T(\mathbf{Y}-\mathbf{XB})\right\} + P_{\lambda,F}(\mathbf{B}) \\ R(\mathbf{B}|\mathbf{B}^{(k)}) &= \mathrm{Tr}\left\{(\mathbf{Y}-\mathbf{XB})^T(\mathbf{Y}-\mathbf{XB})\right\} + P_{\lambda,F}(\mathbf{B}|\mathbf{B}^{(k)}) \end{aligned}$$

This is easy to see, because $Q(\mathbf{B})-R(\mathbf{B}|\mathbf{B}^{(k)}) = \lambda\sum_{j=1}^{p}\left[p_F(r_j) - p_F(r_j^*) - p_F'(r_j^*)(r_j - r_j^*)\right]$. And since $p_F(.)$ is concave in its argument, we have $p_F(r_j) \leqslant p_F(r_j^*) + p_F'(r_j^*)(r_j - r_j^*)$. Thus $Q(\mathbf{B}^{(k)}) \leqslant R(\mathbf{B}|\mathbf{B}^{(k)})$. Also by definition $Q(\mathbf{B}) = R(\mathbf{B}^{(k)}|\mathbf{B}^{(k)})$.

Now notice that $\mathbf{B}^{(k+1)} = \arg\min_{\mathbf{B}} R(\mathbf{B}|\mathbf{B}^{(k)})$. Thus $Q(\mathbf{B}^{(k+1)}) \leqslant R(\mathbf{B}^{(k+1)}|\mathbf{B}^{(k)}) \leqslant R(\mathbf{B}^{(k)}|\mathbf{B}^{(k)}) = Q(\mathbf{B}^{(k)})$, i.e. the value of the objective function decreases in each iteration. At this point, we make the following assumption to enforce convergence to a local solution:

(A3) $Q(\mathbf{B}) = Q(M(\mathbf{B}))$ only for stationary points of Q, where M is the mapping from $\mathbf{B}^{(k)}$ to $\mathbf{B}^{(k+1)}$ defined in (5.3.2).

Since the sequence of penalized losses i.e. $\{Q(\mathbf{B}^{(k)}\}$ is bounded below (by 0) and monotone, it has a limit point, say $\hat{\mathbf{B}}$. Also the mapping $M(.)$ is continuous as ∇p_F is continuous. Further, we have $Q(\mathbf{B}^{(k+1)}) = Q(M(\mathbf{B}^{(k)})) \leqslant Q(\mathbf{B}^{(k)})$ which implies $Q(M(\hat{\mathbf{B}})) = Q(\hat{\mathbf{B}})$. It follows that $\hat{\mathbf{B}}$ is a local minimizer following assumption (A3).

Remark. Although the LARN algorithm guarantees convergence to a stationary point, that point may not be a local solution. However, local linear approximation has been found to be effective in approximating nonconvex penalties and obtaining oracle solutions for single-response regression (Zou and Li, 2008) and support vector machines (Peng, 2016), and our method generalizes this concept for the multitask

situation. We plan to elaborate on the presence and influence of saddle points in our scenario, in a future extended version of this work.

5.3.2 The one-step estimate and its oracle properties

Due to the row-wise additive structure of our penalty function, supports of each of the iterates in the LARN algorithm have the same set of singular points as the solution to the original optimization problem, say $\hat{\mathbf{B}}$. Consequently each of these iterates $\hat{\mathbf{B}}^{(k)}$ are capable of producing sparse solutions. In fact, the first iterate itself possesses oracle properties desirable of row-sparse estimates, namely consistent recovery of the non-zero row support of $\mathbf{B}$, as well as of the elements in those rows. From our simulations there is little to differentiate between the first-step and multi-step estimates in terms of empirical efficiency. This is in line with the findings of Zou and Li (2008) and Fan and Chen (1999).

Given an initial solution $\mathbf{B}^*$, the first LARN iterate, say $\hat{\mathbf{B}}^{(1)}$, is a solution to the optimization problem:

$$\arg\min_{\mathbf{B}} R(\mathbf{B}|\mathbf{B}^*) \;=\; \arg\min_{\mathbf{B}} \left[\mathrm{Tr}\left\{(\mathbf{Y}-\mathbf{XB})^T(\mathbf{Y}-\mathbf{XB})\right\} + \lambda \sum_{j=1}^{p} p'_F(r_j^{(k)}) r_j \right] \quad (5.3.3)$$

At this point, without loss of generality we assume that the true coefficient matrix $\mathbf{B}$ has the following decomposition: $\mathbf{B} = (\mathbf{B}_1^T, \mathbf{0})^T, \mathbf{B}_1 \in \mathbb{R}^{p_1 \times q}$. Also denote the vectorized (i.e. stacked-column) version of a matrix $\mathbf{A}$ by $\mathrm{vec}(\mathbf{A})$. We are now in a position to to prove oracle properties of the one-step estimator in (5.3.3), in the sense that the estimator is able to consistently detect zero rows of $\mathbf{B}$ as well as estimate its non-zero rows for increasing sample size:

Theorem 5.3.1. *Assume that* $\mathbf{X}^T\mathbf{X}/n \to \mathbf{C}$ *for some positive definite matrix* $\mathbf{C}$*, and* $p'_F(r_j^*) = O((r_j^*)^{-s})$ *for* $1 \leqslant j \leqslant q, 0 < r_j^* < \delta$ *and some* $s > 0, \delta > 0$*. Consider now*

a sequence of tuning parameters λ_n *such that* $\lambda_n/\sqrt{n} \to 0$ *and* $\lambda_n n^{(s-1)/2} \to \infty$. *Then the following holds for the one-step estimate* $\hat{\mathbf{B}}^{(1)} = (\hat{\mathbf{B}}_1^T, \hat{\mathbf{B}}_0^T)^T$ *(with the component matrix having dimensions* $p_1 \times q$ *and* $p - p_1 \times q$, *respectively) as* $n \to \infty$:

- $\text{vec}(\hat{\mathbf{B}}_0) \to \mathbf{0}_{(p-p_1)q}$ *in probability;*
- $\sqrt{n}(\text{vec}(\hat{\mathbf{B}}_1) - \text{vec}(\mathbf{B}_1)) \rightsquigarrow \mathcal{N}_p(\mathbf{0}_{p_1 q}, \mathbf{\Sigma} \otimes \mathbf{C}_{11}^{-1})$

where $\mathbf{C}_{11}$ *is the first* $p_1 \times p_1$ *block in* $\mathbf{C}$.

The assumption on the covariate matrix $\mathbf{X}$ is standard, and ensures uniqueness of the asymptotic covariance matrix of our estimator. Note that the restricted eigenvalue condition, which has been used in the literature to establish finite sample error bounds of penalized estimators (Neghaban et al., 2009) is a stronger version of this. With respect to the general framework of nonconvex penalized M-estimation in Loh and Wainwright (2015), our penalty function $p_F(.)$ arising from assumptions (A1) and (A2) satisfies parts (i)-(iv) of Assumption 1 therein, and the conditions of theorem 5.3.1 adhere to part (v). Also note that the above oracle results depend on the assumption (A1), which simplifies depth as a function of the row-norm. We conjecture that similar oracle properties hold for weaker assumptions. From initial attempts into proving a broader result, we think it requires a more complex approach than the proof of Theorem 5.3.1, and plan to work on this in future.

5.3.3 Recovering sparsity within a row

The set of variables with non-zero coefficients for each of the q univariate regressions may not be the same, and hence recovering the non-zero elements *within a row* is of interest as well. It turns out that consistent recovery at this level can be achieved by simply thresholding elements of the non-zero elements in the one-step estimate obtained in the preceding subsection. Obozinski et al. (2011) have shown that a

similar approach leads to consistent recovery of within-row supports in multivariate group lasso. The following result formalizes this in our scenario, provided that the non-zero signals in $\mathbf{B}$ are large enough:

Lemma 5.3.2. *Suppose the conditions of theorem 5.3.1 hold, and additionally all non-zero components of* $\mathbf{B}$ *have the following lower bound:*

$$|b_{jk}| \geqslant \sqrt{\frac{16 \log(qp_1)}{C_{min} n}}; \quad 1 \leqslant j \leqslant p_1, 1 \leqslant k \leqslant q$$

where $C_{\min} > 0$ *is a lower bound for eigenvalues of* $\mathbf{C}_1$*. Also define by* $\hat{\mathcal{S}}$ *the index set of non-zero rows estimated by the LARN algorithm. Then, for some constants* $c, c_0 > 0$*, the post-thresdolding estimator* $\mathbf{T}(\hat{\mathbf{B}}^{(1)})$ *defined by:*

$$t_{jk} = \begin{cases} 0 & \text{if } \hat{b}_{jk}^{(1)} \leqslant \sqrt{\frac{8 \log(q|\hat{\mathcal{S}}|)}{C_{min} n}} \\ \hat{b}_{jk}^{(1)} & \text{otherwise} \end{cases}; \quad j \in \hat{\mathcal{S}}, 1 \leqslant k \leqslant q$$

has the same set of non-zero supports within rows as $\mathbf{B}$ *with probability greater than* $1 - c_0 \exp(-cq \log p_1)$.

5.3.4 Computation

When the quantities $\mathbf{B}$ and $\mathbf{Y} - \mathbf{XB}$ are replaced with their corresponding vectorized versions, the optimization problem in (5.3.3) reduces to a weighted group lasso (Yang and Zou, 2015) setup, with group norms corresponding to l^2 norms of rows of $\mathbf{B}$ and inverse depths of corresponding rows of the initial estimate $\mathbf{B}^*$ acting as group weights. To solve this problem, we start from the following lemma, which gives necessary and sufficient conditions for the existence of a solution:

Lemma 5.3.3. *Given an initial value* $\mathbf{B}^*$*, a matrix* $\mathbf{B} \in \mathbb{R}^{p \times q}$ *is a solution to the*

optimization problem in (5.3.3) if and only if:

1. $2\mathbf{x}_j^T(\mathbf{Y}-\mathbf{XB}) + \lambda p'_F(r_j^*)\mathbf{b}_j/r_j = \mathbf{0}_q$ *if* $\mathbf{b}_j \neq \mathbf{0}_q$*;*

2. $\|\mathbf{x}_j^T(\mathbf{Y}-\mathbf{XB})\|_2 \leqslant \lambda/2$ *if* $\mathbf{b}_j = \mathbf{0}_q$.

This lemma is a modified version of lemma 4.2 in chapter 4 of Buhlmann and van de Geer (2011), and can be proved in a similar fashion. Following the lemma, we can now use a block coordinate descent algorithm (Li et al., 2015) to iteratively obtain $\hat{\mathbf{B}}^{(1)}$, given an appropriate starting value $\mathbf{B}^*$:

- Set $m = 1$ and $\hat{\mathbf{B}}^{(1,0)} = \mathbf{B}^*$;
- For $j = 1, 2, ..., p$ do:
  - If $\|\mathbf{x}_j^T(\mathbf{Y}-\mathbf{X}\hat{\mathbf{B}}^{(1,m-1)})\|_2 \leqslant (\lambda/2).p'_F(r_j^*)$, set $\hat{\mathbf{b}}_j^{(1,m)} = \mathbf{0}_q$;
  - Else update $\hat{\mathbf{b}}_j^{(1,m)}$ as

$$\hat{\mathbf{b}}_j^{(1,m)} = \frac{2\mathbf{s}_j^{(m-1)}}{2\|\mathbf{x}_j\|_2^2 + \lambda\frac{np'_F(r_j^*)}{\hat{r}_j^{(1,m-1)}}1_{\hat{r}_j^{(1,m-1)}>0}}$$

    where $\mathbf{s}_j^{(m-1)} = \mathbf{x}_j^T(\mathbf{Y}-\mathbf{X}\hat{\mathbf{B}}_{-j}^{(1,m-1)})$; $\hat{\mathbf{B}}_{-j}^{(1,m-1)}$ is the matrix obtained by replacing j-th row of $\hat{\mathbf{B}}^{(1,m-1)}$ by zeros.

- Set $m \leftarrow m+1$, check for convergence and continue until convergence.
- Apply the thresholding from lemma 5.3.2 to recover within-row supports.

The parameter λ controls row-sparsity in $\hat{\mathbf{B}}^{(1)}$: a larger or smaller λ corresponding to higher number of zero rows in $\hat{\mathbf{B}}^{(1)}$, or an estimate closer to the ordinary least square solution, respectively. Since we use block coordinate descent, rows can drop in or out of the solution path, i.e. zero rows can reappear to be nonzero for a smaller λ.

Given a fixed λ, an easy choice of $\mathbf{B}^*$ is $\hat{\mathbf{B}}_{\text{LS}}$, i.e. the least squares estimate. We use k-fold cross-validation to choose the optimal λ. Also notice that in a sample setup the quantity $C_{\min}$ in lemma 5.3.2 is unknown. For this reason, we choose a best threshold for within-row sparsity through the above cross-validation procedure as well. Even though this means that the cross-validation has to be done over a two-dimensional grid, the thresholding step is actually done *after* estimation. Thus for any fixed λ, only k models need to be calculated. Given a trained model for some value of λ we just cycle through the full range of thresholds to record their corresponding cross-validation errors. Also when optimizing over the range of tuning parameter values, say $\lambda_1 > ... > \lambda_m$, we use warm starts to speed up convergence. Denoting the solution corresponding to any tuning parameter λ as $\hat{\mathbf{B}}^{(1)}(\lambda)$, this means starting from the initial value $\mathbf{B}_0 = \hat{\mathbf{B}}^{(1)}(\lambda_{k-1})$ to obtain $\hat{\mathbf{B}}^{(1)}(\lambda_k)$, for $k = 2, ..., m$.

5.4 Orthogonal design and independent responses

We shed light on the workings of our penalty function by considering the simplified scenario when the predictor matrix $\mathbf{X}$ is orthogonal and all responses are independent. Independent responses make minimizing (5.2.1) equivalent to solving of q separate nonconvex penalized regression problems, while orthogonal predictors make the LARN estimate equivalent to a collection of coordinate-wise soft thresholding operators.

5.4.1 Thresholding rule

For the univariate thresholding rule, we are dealing with the simplified penalty function $p_F(|b_{jk}|) = D^-(b_{jk}, F)$, where D^- is a inverse depth function based on the univariate depth function D. In this case, depth calculation becomes simplified in exactly the same way as in Subection 5.3.1, only $|b_{jk}|$ replacing $\|\mathbf{b}_j\|$ therein, and $1 \leqslant k \leqslant q$.

Following Fan and Li (2001), a sufficient condition for the minimizer of the penalized least squares loss function

$$L(\theta; p_\lambda) = \frac{1}{2}(z-\theta)^2 + p_\lambda(|\theta|) \tag{5.4.1}$$

to be unbiased when the true parameter value is large is $p'_\lambda(|\theta|) = 0$ for large θ. In our formulation, this holds exactly when F has finite support, and approximately otherwise. A necessary condition for sparsity and continuity of the solution is $\min_{\theta\neq 0} |\theta| + p'_\lambda(|\theta|) > 0$. We ensure this by making a small assumption about the derivative of D^- (denoted by D_1^-):

(A4) $\lim_{\theta\to 0+} D_1^-(\theta, F) > 0$.

Subsequently we get the following thresholding rule as the solution to (5.4.1):

$$\begin{aligned} \hat{\theta}(F,\lambda) &= \text{sign}(z)\left[|z| - \lambda D_1^-(\theta, F)\right]_+ \\ &\simeq \text{sign}(z)\left[|z| - \lambda D_1^-(z, F)\right]_+ \end{aligned} \tag{5.4.2}$$

The approximation in the second step is due to Antoniadis and Fan (2001). A plot of the thresholding function in panel b of Figure 5.1 demonstrates the unbiasedness and continuity properties of this estimator.

We note here that thresholding rules due to previously proposed nonconvex penalty functions can be obtained as special case of our rule. For example, when we consider halfspace depth as our chosen depth function, and the max definition of inverse depth, i.e. $D^-(\mathbf{b}, F) = \max_{\mathbf{x}} D^-(\mathbf{x}, F) - D^-(\mathbf{b}, F)$, the MCP penalty (Zhang, 2010) corre-

sponds to $D_1^-(\theta, F) = |\theta|\mathbb{I}_{|\theta|<\lambda}$, while for the SCAD penalty (Fan and Li, 2001):

$$D_1^-(\theta, F) = \begin{cases} c\lambda & \text{if } |\theta| < 2\lambda \\ \frac{c}{a-2}(a\lambda - |\theta|) & \text{if } 2\lambda \leqslant |\theta| < a\lambda \\ 0 & \text{if } |\theta| > a\lambda \end{cases}$$

with $c = 1/(2\lambda^2(a+2))$.

5.4.2 Minimax optimal performance

In the context of estimating the mean parameters μ_i of independent and identically distributed observations with normal errors: $z_i = \theta_i + v_i, v_i \sim N(0,1)$, the minimax risk is $2\log n$ times the ideal risk $R(\text{ideal}) = \sum_{i=1}^n \min(\theta_i^2, 1)$ (Donoho and Johnstone, 1994). A major motivation of using lasso-type penalized estimators in linear regression is that they are able to approximately achieve this risk bound for large sample sizes (Donoho and Johnstone, 1994; Zou, 2006). We now show that our thresholding rule in (5.4.2) also, in fact, replicates this performance.

Theorem 5.4.1. *Suppose the inverse depth function $D^-(., F)$ is twice continuously differentiable, except at the origin, with first and second derivatives bounded above by c_1 and c_2 respectively. Then for $\lambda = (\sqrt{.5\log n} - 1)/c_1$, we have*

$$\begin{aligned} R(\hat{\theta}(F,\lambda)) &\leqslant (2\log n - 3) \\ &\quad \left[R(ideal) + \frac{c_1}{p_0(F)(\sqrt{.5\log n} - 1)} \right] \end{aligned} \tag{5.4.3}$$

where $p_0(F) := \lim_{\theta\to 0+} D_1^-(\theta, F)$.

Following the theorem, we easily see that for large n the minimax risk of $\hat{\theta}(F,\lambda)$ approximately achieves the $2\log n$ multiple bound.

The adaptive lasso proposed by Zou (2006) guarantees a similar minimax risk bound in the case of single-response regression. This is somewhat expected, given the similar weighted norm structure of the LARN penalty and the adaptive lasso penalty. However, this does *not* hold all weighted norm penalties: for example the SCAD and MCP penalties do not ensure near-minimax optimal performance because of their non-continuity in the second derivative. In this situation, using inverse depth functions that satisfy all the conditions in the theorem (both halfspace depth and projection depth do because of the simplification in Subection 5.3.1) allows us to go through with the result.

5.5 Simulation results

5.5.1 Methods and setup

We use the setup of Rothman et al. (2010) for our simulation study to compare the performance of LARN with other relevant methods. Specifically, we use performance metrics calculated after applying the following methods of predictor selection on simulated data for this purpose:

- *LARN*: We use halfspace depth as our chosen depth function, take $D^{-}(\mathbf{x}, F) = \max_{\mathbf{x}} D(\mathbf{x}, F) - D(\mathbf{x}, F)$, and consider the set of tuning parameters $\lambda \in 10^{\{100, 99.5, \ldots, 0.5, 0\}}$ and use 5-fold cross-validation to get the optimal solution;

- *Sparse Group Lasso (SGL: Simon et al. (2013))*: We adapt this method for single-response regression that uses group-level as well as element-level penalties on the coefficient vector in our scenario by taking $\text{vec}(\mathbf{Y})$ as the response vector, $\mathbf{X} \otimes \mathbf{I}_q$ as the matrix of predictors, and then transforming back the pq-length coefficient estimate into a $p \times q$ matrix. Default options in the R package `SGL` are used while fitting the model;

- *Group Lasso with thresholding (GL-t)*: This has been proposed by Obozinski et al. (2011), and performs element-wise thresholding on a row-level group lasso estimator to get final estimate of $\mathbf{B}$. It can also be realized as a special case of LARN, with weights of all row-norms set as 1.

We generate rows of the model matrix $\mathbf{X}$ as $n = 50$ independent draws from $\mathcal{N}(\mathbf{0}_p, \mathbf{\Sigma}_X)$, where the positive $\mathbf{\Sigma}_X$ has an AR(1) covariance structure, with its $(i,j)^{\text{th}}$ element given by $0.7^{|i-j|}$. Rows of the random error matrix are generated as independent draws from $\mathcal{N}(\mathbf{0}_q, \mathbf{\Sigma})$: with $\mathbf{\Sigma}$ also having an AR(1) structure with correlation parameter $\rho \in \{0, 0.5, 0.7, 0.9\}$. Finally, to generate the coefficient matrix $\mathbf{B}$, we obtain the three $p \times q$ matrices: $\mathbf{W}$, whose elements are independent draws from $N(2,1)$; $\mathbf{K}$, which has elements as independent draws from Bernoulli(0.3); and $\mathbf{Q}$ whose rows are made all 0 or all 1 according to p independent draws of another Bernoulli random variable with success probability 0.125. Following this, we multiply individual elements of these matrices (denoted by $(*)$) to obtain a sparse $\mathbf{B}$:

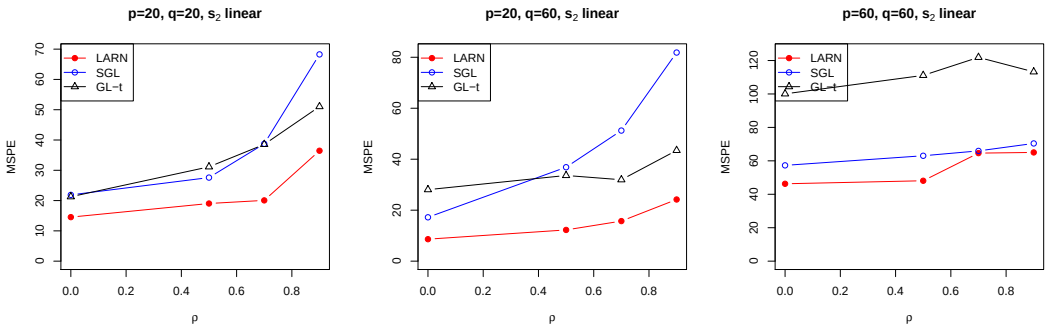
$$\mathbf{B} = \mathbf{W} * \mathbf{K} * \mathbf{Q}$$

Notice that the two levels of sparsity we consider: entire row and within-row, are imposed by the matrices $\mathbf{Q}$ and $\mathbf{K}$, respectively.

For a given value of ρ, we consider three settings of data dimensions for the simulations: (a) $p = 20, q = 20$, (b) $p = 20, q = 60$ and (c) $p = 60, q = 60$. Finally we replicate the full simulation 100 times for each set of (p, q, ρ).

5.5.2 Evaluation

To summarize the performance of an estimate matrix $\hat{\mathbf{B}}$ we use the following three performance metrics:



(a) (b) (c)

Figure 5.2: Mean squared testing errors for all three methods in different (p, q) settings

- *Mean Squared Testing Error (MSTE)*- Defined as

$$MSTE(\hat{\mathbf{B}}) = \frac{1}{pq}\text{Tr}\left[(\mathbf{Y}_{test} - \mathbf{X}_{test}\hat{\mathbf{B}})(\mathbf{Y}_{test} - \mathbf{X}_{test}\hat{\mathbf{B}})^T\right]$$

  with $(\mathbf{Y}_{test}, \mathbf{X}_{test})$ generated independently from $(\mathbf{Y}_{test}, \mathbf{X}_{test})$ using the simulation setup above, but using the same true $\mathbf{B}$;

- *True Positive Rate (TP)* - Defined as the proportion of non-zero entries in $\mathbf{B}$ detected as non-zero in $\hat{\mathbf{B}}$;

- *True Negative Rate (TN)* - Defined as the proportion of zero entries in $\mathbf{B}$ detected as zero in $\hat{\mathbf{B}}$.

A desirable estimate shall have low MSTE and high TP and TN proportions.

We summarize TP/TN rates of the three methods in Table 5.1, and MSTE performances in Figure 5.2. All across our method outperforms, GL-t, i.e. its unweighted version. Although its true negative detection is slightly worse than SGL, LARN makes up for that by a far superior signal detection ability (i.e. TP rate) for case (c), which has the highest feature and response space dimensions.

Replications were assigned randomly to any of the 8 threads of an Intel Core i7

ρ	GL-t	SGL	LARN
(a) $p = 20, q = 20$			
0.9	0.77/0.83	0.92/0.99	0.91/0.92
0.7	0.81/0.83	0.91/0.99	0.89/0.93
0.5	0.78/0.79	0.89/0.99	0.88/0.92
0.0	0.85/0.78	0.90/0.99	0.90/0.91
(b) $p = 20, q = 60$			
0.9	0.90/0.66	0.95/0.97	0.89/0.92
0.7	0.91/0.70	0.93/0.96	0.90/0.92
0.5	0.80/0.69	0.94/0.98	0.93/0.92
0.0	0.85/0.68	0.93/0.97	0.91/0.92
(c) $p = 60, q = 60$			
0.9	0.57/0.79	0.68/0.99	0.85/0.93
0.7	0.50/0.79	0.64/0.99	0.83/0.93
0.5	0.54/0.81	0.64/0.99	0.85/0.93
0.0	0.58/0.79	0.63/0.99	0.84/0.93

Table 5.1: Average true positive and true negative (TP/TN) rates for 3 methods, for $n = 50$ and AR1 covariance structure

Setting	GL-t	SGL	LARN
(a)	332	490	209
(b)	676	52	328
(c)	4994	39760	3883

Table 5.2: Total runtimes in seconds for SGL and LARN algorithms for the three simulation settings

3770 3.4 GHz processor-run machine with 8 GB of RAM and run in parallel for each set of values of (p, q, ρ). As seen in table Table 5.2, LARN is the most computationally efficient of the three methods. This advantage becomes widest for case (c). Although SGL uses accelerated generalized gradient descent to speed up computation from block coordinate descent, its advantage is no longer observed in our case since we apply it on vec($\mathbf{Y}$) and $\mathbf{X} \otimes \mathbf{I}_q$. Also note that GL-t is an unweighted version of LARN. In spite of that, LARN turns out to be faster than its unweighted counterpart: indicating faster convergence.

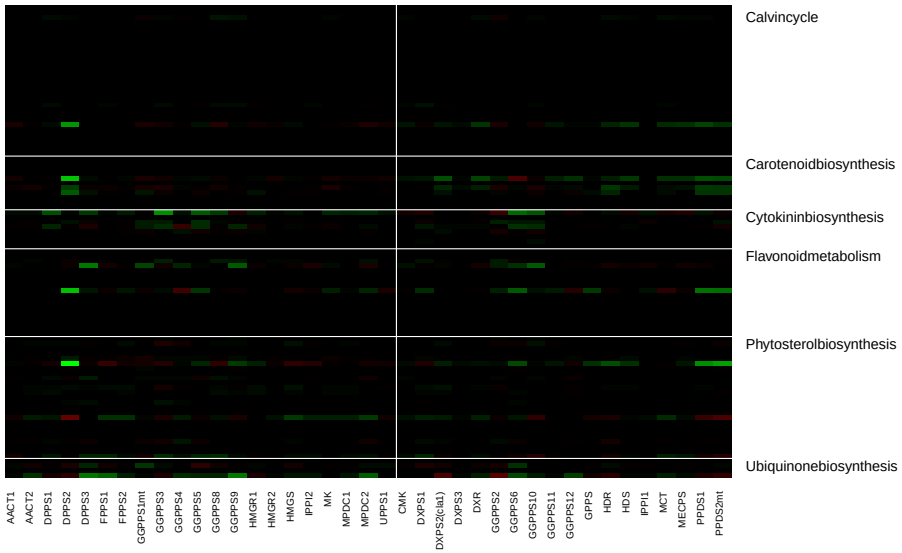


Figure 5.3: Estimated effects of different pathway genes on the activity of genes in Mevalonate and Non-mevalonate pathways (left and right of vertical line) in *A. thaliana*

5.6 Real data example

We apply the LARN algorithm on a microarray dataset containing expression of several genes in the flowering plant *Arabidopsis thaliana* (Wille et al, 2004). In this dataset, gene expressions are collected from $n = 118$ samples, which are plants grown under different experimental conditions. We take the expressions of $q = 40$ genes in two pathways for biosynthesis of isoprenoid compounds, which are key compounds affecting plant metabolism as our multiple responses. Expressions of 795 other genes corresponding to 56 other pathways are taken as predictors.

Our objective here is to find out the extent of crosstalk between isoprenoid pathway genes and those in the other pathways. We apply LARN, as well as the two methods mentioned before, on the data and evaluate them based on predictive accuracy of 100 random splits with 90 training samples. All three methods have similar mean squared prediction error (MSPE) (LARN and GL-t have MSPE 0.45 and SGL

Coeff	Gene	Pathway
0.18	DPPS2	Phytosterol biosynthesis
0.14	DPPS2	Carotenoid biosynthesis
0.14	DPPS2	Flavonoid metabolism
0.11	DPPS2	Calvin cycle
0.11	PPDS2mt	Phytosterol biosynthesis
0.10	GGPPS3	Cytokinin biosynthesis
0.10	PPDS1	Phytosterol biosynthesis
0.09	DPPS3	Flavonoid metabolism
0.09	DPPS3	Ubiquinone biosynthesis
0.09	GGPPS9	Ubiquinone biosynthesis

Table 5.3: Top 10 gene-pathway connections in *A. thaliana* data found by LARN

has 0.44), but LARN produces more sparse solutions on average: the mean proportion of non-zero elements in the coefficient matrix are 0.15, 0.21 and 0.29 for LARN, GL-t and SGL, respectively. Focusing on the coefficient matrix estimated by LARN, we summarize the 10 largest coefficients (in absolute values) in Table 5.3. We also visualize coefficients corresponding to genes in the 6 pathways in the table through a heatmap in Figure 5.3.

All of the four largest coefficients correspond to interactions of one gene, DPPS2, with four different pathways. Two of these pathways, Carotenoid and Phytosterol, directly use products from the isoprenoid pathways, and their connections with DPPS2 had been detected in Wille et al (2004). The large Calvin Cycle-DPPS2 coefficient reveals that compounds synthesized in Carotenoid and Phytosterol pathways get used in Calvin Cycle. In the heatmap, Carotenoid biosynthesis seems to be connected mostly to the non-mevalonate pathway genes (right of the vertical line), while the activities of genes in Cytokinin and Ubiquinone synthesis pathways seem to be connected with those in the mevalonate pathway. These are consistent with the findings of Wille et al (2004), Frebort et al. (2011) and Disch et al. (1998), respectively.

5.7 Conclusion

In this chapter we propose a class of nonconvex penalty functions, based on the idea of inverting data depths, for performing support union recovery in multitask linear regression. Although several nonconvex penalties exist in the literature, the strength of our penalization scheme lies in the significant scope of inference procedures that can rise from the choice of the reference distribution F. Here we consider a simplified reference distribution and provide asymptotic oracle results that ensure recovery of the non-zero row support in the coefficient matrix. We also show that a simple post-estimation thresholding recovers non-zero elements within non-zero rows of the estimated coefficient matrix with good accuracy. Although our method shares the weakness of all nonconvex penalties: small signals may go undetected or can be estimated in a biased fashion, the flexibility in choosing F provides enough motivation to fine tune similar penalization schemes. Our immediate plans for future studies include extending this specific setup to generalized linear models, dimensional asymptotics assuming the data dimension p to be a function of sample size n, as well as exploring the use of more efficient algorithms for calculating the sparse solutions, e.g. proximal gradient descent or Concave-Convex algorithms (Wang et al., 2013).

5.8 Proofs

Proof of theorem 5.3.1. We shall prove a small lemma before going into the actual proof.

Lemma 5.8.1. *For matrices* $\mathbf{K} \in \mathbb{R}^{l\times k}, \mathbf{L} \in \mathbb{R}^{l\times m}, \mathbf{M} \in \mathbb{R}^{m\times k}$,

$$\mathrm{Tr}(\mathbf{K}^T\mathbf{L}\mathbf{M}) = \mathrm{vec}^T(\mathbf{K})(\mathbf{I}_k \otimes \mathbf{L})\,\mathrm{vec}(\mathbf{M})$$

Proof of lemma 5.8.1. From the property of Kronecker products, $(\mathbf{I}_k \otimes \mathbf{L})\,\mathrm{vec}(\mathbf{M}) =$

$\text{vec}(\mathbf{LM})$. The lemma follows since $\text{Tr}(\mathbf{K}^T\mathbf{LM}) = \text{vec}^T(\mathbf{K})\,\text{vec}(\mathbf{LM})$. □

Now, suppose $\mathbf{B} = \mathbf{B}_0 + \mathbf{U}/\sqrt{n}$, for some $\mathbf{U} \in \mathbb{R}^{p\times q}$, so that our objective function takes the form

$$\begin{aligned} T_n(\mathbf{U}) &= \text{Tr}\left[\left(\mathbf{Y} - \mathbf{X}\mathbf{B}_0 - \frac{1}{\sqrt{n}}\mathbf{X}\mathbf{U}\right)^T\left(\mathbf{Y} - \mathbf{X}\mathbf{B}_0 - \frac{1}{\sqrt{n}}\mathbf{X}\mathbf{U}\right)\right] \\ &\quad + \lambda_n \sum_{j=1}^{p} p'_F(r_j^*)\left\|\mathbf{b}_{0j} + \frac{\mathbf{u}_j}{\sqrt{n}}\right\|_2 \\ \Rightarrow T_n(\mathbf{U}) - T_n(\mathbf{0}_{p\times q}) &= \text{Tr}\left[\frac{1}{n}\mathbf{U}^T\mathbf{X}^T\mathbf{X}\mathbf{U} - \frac{2}{\sqrt{n}}\mathbf{E}^T\mathbf{X}\mathbf{U}\right] \\ &\quad + \frac{\lambda_n}{\sqrt{n}}\sum_{j=1}^{p} p'_F(r_j^*)\left(\|\sqrt{n}\mathbf{b}_{0j} + \mathbf{u}_j\|_2 - \|\sqrt{n}\mathbf{b}_{0j}\|_2\right) \\ &= \text{Tr}(\mathbf{V}_1 + \mathbf{V}_2) + V_3 \end{aligned} \tag{5.8.1}$$

Since $\mathbf{X}^T\mathbf{X}/n \to \mathbf{C}$ by assumption, we have $\text{Tr}(\mathbf{V}_1) \to \text{vec}^T(\mathbf{U})(\mathbf{I}_q \otimes \mathbf{C})\,\text{vec}(\mathbf{U})$ using lemma 5.8.1. Using the lemma we also get

$$\text{Tr}(\mathbf{V}_2) = \frac{2}{\sqrt{n}}\text{vec}^T(\mathbf{E})(\mathbf{I}_q \otimes \mathbf{X})\,\text{vec}(\mathbf{U})$$

Now $\text{vec}(\mathbf{E}) \sim \mathcal{N}_{nq}(\mathbf{0}_n, \mathbf{\Sigma}\otimes\mathbf{I}_q)$, so that $(\mathbf{I}_q\otimes\mathbf{X}^T)\,\text{vec}(\mathbf{E})/\sqrt{n} \rightsquigarrow \mathbf{W} \equiv \mathcal{N}_{pq}(\mathbf{0}_{pq}, \mathbf{\Sigma}\otimes\mathbf{C})$ using properties of Kronecker products and Slutsky's theorem.

Let us look at V_3 now. Denote by V_{3j} the j-th summand of V_3. Now there are two scenarios. Firstly, when $\mathbf{b}_{0j} \neq \mathbf{0}_q$, we have $p'_F(r_j^*) \xrightarrow{P} p'_F(r_{0j})$. Since $\lambda_n/\sqrt{n} \to 0$, this implies $V_{3j} \xrightarrow{P} 0$ for any fixed $\mathbf{u}_j$. Secondly, when $\mathbf{b}_{0j} = \mathbf{0}_q$, we have

$$V_{3j} = \lambda_n n^{(s-1)/2}.(\sqrt{n}r_j^*)^{-s}.\frac{p'_F(r_j^*)\|\mathbf{u}_j\|_2}{(r_j^*)^{-s}}$$

We now have $\mathbf{b}_j^* = O_p(1/\sqrt{n})$, and also each term of the gradient vector is $O((r_j^*)^{-s})$ by assumption. Thus $V_{3j} = O_P(\lambda_n n^{(s-1)/2}\|\mathbf{u}_j\|_2)$. By assumption, $\lambda_n n^{(s-1)/2} \to \infty$ as

$n \to \infty$, so $V_{3j} \xrightarrow{P} \infty$ unless $\mathbf{u}_j = \mathbf{0}_q$, in which case $V_{3j} = 0$.

Accumulating all the terms and putting them into (5.8.1) we see that

$$T_n(\mathbf{U}) - T_n(\mathbf{0}_{p\times q}) \rightsquigarrow \begin{cases} \mathrm{vec}^T(\mathbf{U}_1)\left[(\mathbf{I}_q \otimes \mathbf{C}_{11})\,\mathrm{vec}(\mathbf{U}_1) - 2\mathbf{W}_1\right] & \text{if } \mathbf{U}_0 = \mathbf{0}_{(p-p_1)q} \\ \infty & \text{otherwise} \end{cases} \tag{5.8.2}$$

where rows of $\mathbf{U}$ are partitioned into $\mathbf{U}_1$ and $\mathbf{U}_0$ according to the zero and non-zero rows of $\mathbf{B}_0$, respectively, and the random variable $\mathbf{W}$ is partitioned into $\mathbf{W}_1$ and $\mathbf{W}_0$ according to zero and non-zero *elements* of $\mathrm{vec}(\mathbf{B}_0)$. Applying epiconvergence results of Geyer (1994) and Knight and Fu (2000) we now have

$$\mathrm{vec}(\hat{\mathbf{U}}_{1n}) \rightsquigarrow (\mathbf{I}_q \otimes \mathbf{C}_{11}^{-1})\mathbf{W}_1 \tag{5.8.3}$$

$$\mathrm{vec}(\hat{\mathbf{U}}_{0n}) \rightsquigarrow \mathbf{0}_{(p-p_1)q} \tag{5.8.4}$$

where $\hat{\mathbf{U}}_n = (\hat{\mathbf{U}}_{1n}^T, \hat{\mathbf{U}}_{0n}^T)^T := \arg\min_{\mathbf{U}} T_n(\mathbf{U})$.

The second part of the theorem, i.e. asymptotic normality of $\sqrt{n}(\mathrm{vec}(\hat{\mathbf{B}}_{1n}) - \mathrm{vec}(\hat{\mathbf{B}}_{1n})) = \hat{\mathbf{U}}_{1n}$ follows directly from (5.8.3). It is now sufficient to show that $P(\hat{\mathbf{b}}_j^{(1)} \neq \mathbf{0}_q | \mathbf{b}_{0j} = \mathbf{0}_q) \to 0$ to prove the oracle consistency part. For this notice that KKT conditions of the optimization problem for the one-step estimate indicate

$$2\mathbf{x}_j^T(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}^{(1)}) = -\lambda_n p'_F(r_j^*)\frac{\mathbf{b}_j^{(1)}}{r_j^{(1)}} \quad \Rightarrow \quad \frac{2\mathbf{x}_j^T(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}^{(1)})}{\sqrt{n}} = -\frac{\lambda_n p'_F(r_j^*)}{\sqrt{n}} \cdot \frac{\mathbf{b}_j^{(1)}}{r_j^{(1)}} \tag{5.8.5}$$

for any $1 \leqslant j \leqslant p$ such that $\hat{\mathbf{b}}_j^{(1)} \neq \mathbf{0}_q$. Since $p'_F(r_j^*) = D^-((r_j^*)^{-s}) = O_P(\|(\mathbf{b}_{0j} + 1/\sqrt{n}\|^{-s})$ and $\lambda_n n^{(s-1)/2} \to \infty$, the right hand side goes to $-\infty$ in probability if

$\mathbf{b}_{0j} = \mathbf{0}_q$. As for the left-hand side, it can be written as

$$\frac{2\mathbf{x}_j^T(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}^{(1)})}{\sqrt{n}} = \frac{2\mathbf{x}_j^T\mathbf{X}.\sqrt{n}(\mathbf{B}_0 - \hat{\mathbf{B}}^{(1)})}{n} + \frac{2\mathbf{x}_j^T\mathbf{E}}{\sqrt{n}} = \frac{2\mathbf{x}_j^T\mathbf{X}\hat{\mathbf{U}}_n}{n} + \frac{2\mathbf{x}_j^T\mathbf{E}}{\sqrt{n}}$$

Our previous derivations show that vectorized versions of $\hat{\mathbf{U}}_n$ and $\mathbf{E}$ have asymptotic and exact multivariate normal distributions, respectively. Hence

$$\mathbb{P}\left[\hat{\mathbf{b}}_j^{(1)} \neq \mathbf{0}_q | \mathbf{b}_{0j} = \mathbf{0}_q\right] \leqslant P\left[2\mathbf{x}_j^T(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}^{(1)}) = -\lambda_n p'_F(r_j^*)\frac{\mathbf{b}_j^{(1)}}{r_j^{(1)}}\right] \to 0$$

□

Proof of theorem 5.3.2. See the proof of corollary 2 of Obozinski et al. (2011)in Appendix A therein. Our proof follows the same steps, only replacing $\mathbf{\Sigma}_{SS}$ with $\mathbf{\Sigma}\otimes\mathbf{C}_{11}$.

□

Proof of Lemma 5.4.1. We broadly proceed in a similar fashion as the proof of Theorem 3 in Zou (2006). As a first step, we decompose the mean squared error:

$$\begin{aligned} E[\hat{\theta}(F,\lambda) - \theta]^2 &= E[\hat{\theta}(F,\lambda) - z]^2 + E(z-\theta)^2 + 2E[\hat{\theta}(F,\lambda)(z-\theta)] - 2E[z(z-\theta)] \\ &= E[\hat{\theta}(F,\lambda) - z]^2 + E\left[\frac{d\hat{\theta}(F,\lambda)}{dz}\right] - 1 \end{aligned}$$

by applying Stein's lemma (Stein, 1981). We now use Theorem 1 of Antoniadis and Fan (2001) to approximate $\hat{\theta}(F,\lambda)$ in terms of y only. By part 2 of the theorem,

$$\hat{\theta}(F,\lambda) = \begin{cases} 0 & \text{if } |z| \leqslant \lambda p_0(F) \\ z - \text{sign}(z).\lambda D_1^-(\hat{\theta}(F,\lambda), F) & \text{if } |z| > \lambda p_0(F) \end{cases} \tag{5.8.6}$$

Moreover, applying part 5 of the theorem,

$$\hat{\theta}(F,\lambda) = z - \text{sign}(z).\lambda D_1^-(z,F) + o(D_1^-(z,F)) \tag{5.8.7}$$

for $|z| > \lambda p_0(F)$. Thus we get

$$[\hat{\theta}(F,\lambda) - z]^2 = \begin{cases} z^2 & \text{if } |z| \leqslant \lambda p_0(F) \\ \lambda^2 D_1^-(z,F)^2 + k_1(|z|) & \text{if } |z| > \lambda p_0(F) \end{cases} \tag{5.8.8}$$

and

$$\frac{d\hat{\theta}(F,\lambda)}{dz} = \begin{cases} 0 & \text{if } |z| \leqslant \lambda p_0(F) \\ 1 + \lambda D_2^-(z,F) + k_1'(|z|) & \text{if } |z| > \lambda p_0(F) \end{cases} \tag{5.8.9}$$

where $k_1(|z|) = o(|z|)$, and $D_2^-(z,F) = d^2D^-(z,F)/dz^2$. Thus

$$\begin{aligned} E[\hat{\theta}(F,\lambda) - \theta]^2 &= E[z^2 \mathbb{I}_{|z| \leqslant \lambda p_0(F)}] + E\left[\left(\lambda^2 D_1^-(|z|,F)^2 + 2\lambda D_2^-(|z|,F) + 2 + \right.\right. \\ &\qquad \left.\left. k_1(|z|) + k_1'(|z|)\right) \mathbb{I}_{|z| > \lambda p_0(F)}\right] - 1 \end{aligned} \tag{5.8.10}$$

Now

$$\begin{aligned} k_1(|z|) &= \lambda^2 \left[D_1^-(z,F)^2 - D_1^-(\hat{\theta}(F,\lambda),F)^2\right] \quad \leqslant \quad \lambda^2 c_1^2, \text{ and} \\ |k_1'(|z|)| &= \lambda \left| D_2^-(z,F) - \frac{dD_1^-(\hat{\theta}(F,\lambda),F)}{dz} \right| \quad \leqslant \quad 2\lambda c_2 \end{aligned}$$

Substituting these in (5.8.10) above we get

$$\begin{aligned} E[\hat{\theta}(F,\lambda)-\theta]^2 &\leqslant \lambda^2 p_0(F)^2 P[|z| \leqslant \lambda p_0(F)] + E\left[\left(\lambda^2 f^2(|z|) + 2\lambda D_2^-(z,F)\right) BI_{|z|>\lambda p_0(F)}\right] \\ &\quad +\lambda^2 c_1^2 + 2\lambda c_2 + 1 \\ &\leqslant 2\lambda^2 c_1^2 + 4\lambda c_2 + 1 \\ &\leqslant 4\lambda^2 c_1^2 + 8\lambda c_2 + 1 \end{aligned} \tag{5.8.11}$$

Adding and subtracting $z^2 \mathbb{I}_{|z|>\lambda p_0(F)}$ to the first and second summands of (5.8.10) above, we also have

$$\begin{aligned} E[\hat{\theta}(F,\lambda)-\theta]^2 &= Ez^2 + E\left[\left(\lambda^2 D_1^-(z,F)^2 + 2\lambda D_2^-(z,F) + 2 - y^2 + \lambda^2 c_1^2\right.\right. \\ &\quad \left.\left. +2\lambda c_2\right)\mathbb{I}_{|z|>\lambda p_0(F)}\right] - 1 \\ &\leqslant (2\lambda^2 c_1^2 + 4\lambda c_2) P[|z| > \lambda p_0(F)] + \theta^2 \end{aligned} \tag{5.8.12}$$

Following Zou (2006), $P[|z| > \lambda p_0(F)] \leqslant 2q(\lambda p_0(F)) + 2\theta^2$, with $q(x) = \exp[-x^2/2]/(\sqrt{2\pi}x)$. Thus

$$\begin{aligned} E[\hat{\theta}(F,\lambda)-\theta]^2 &\leqslant 2(2\lambda^2 c_1^2 + 4\lambda c_2)[q(\lambda p_0(F)) + \theta^2] + \theta^2 \\ &\leqslant (4\lambda^2 c_1^2 + 8\lambda c_2 + 1)[q(\lambda p_0(F)) + \theta^2] \end{aligned} \tag{5.8.13}$$

Combining this with (5.8.11) we get

$$E[\hat{\theta}(F,\lambda)-\theta]^2 \leqslant [4(\lambda c_1 + 1)^2 - 3][q(\lambda p_0(F)) + \min(\theta^2, 1)] \tag{5.8.14}$$

assuming without loss of generality that $c_1 \geqslant c_2$. Since $R(\text{ideal}) = \min(\theta^2, 1)$ and $q(x) \leqslant (\sqrt{2\pi}x)^{-1} < 1/x$, we have the needed. □

Chapter 6

Future Work

6.1 Characterization of depth in general normed spaces

As mentioned in the beginning of this thesis, a normed space and a probability measure in it are the only requirements to motivate the definition of a depth-like quantity. For example consider the Wasserstein metric on a Radon space (M, d), i.e. a metric space for which every probability measure defined on its borel subsets is inner regular. For any two measures μ and ν that values in this space and have finite p-th moment for some $p \in \mathbb{N}$, the p-th Wasserstein distance is defined as:

$$W_p(\mu, \nu) = \left[\inf_{G \in \mathbb{G}(\mu,\nu)} \int_{M \times M} d(x, y)^p dG(x, y) \right]^{1/p} \tag{6.1.1}$$

where $\mathbb{G}(\mu, \nu)$ is the collection of all 'couplings' of the measures μ and ν: a coupling being a probability measure in $M \times M$ with μ and ν at its corresponding marginals.

Now just consider $\mu = \delta_{x_0}$ i.e. the degenerate distribution at some $x_0 \in M$. In that case the infimum in (6.1.1) shall be taken over all possible *conditional couplings* of ν and x_0, i.e. random vectors in $M \times M$ such that x_0 in the first part can come from any distribution, but is fixed at x_0, while the marginal in second part is ν. It is

easy to show that

$$W_p(\delta_{x_0}, \nu) = \left[\int d(x_0, x)^p d\nu(x) \right]^{1/p}$$

Here we can look at the quantity $W_p(\delta_{x_0}, \nu)$ as a generalized outlyingness function.

Similar formulations are possible for other distributional distance measures as well. We plan to investigate this idea in future. Some relevant theoretical machinery for this is possibly available in Leskelä and Vihola (2015) and Dedecker and Michel (2011).

6.2 Future of e-values

In Chapter 2 we only discuss a very specialized implementation of the e-values, in statistical model selection. As seen in Section 3.3 of Chapter 3, using other functionals of the evaluation map distribution can give more refined inference that can be tweaked to suit the need for the data-analytic task in hand and inference objectives of the practitioner. We plan to investigate this in future. We know that the concept of using tail probabilities that yields favorable results in Section 3.3 works in practice, but need to formalize this concept, as well as study in detail the contrasting tail behaviors of adequate and inadequate models. Specific implementations of the variable selection technique are also of interest, for example in robust regression through the usage of robust bootstrap (Salibian-Barrera and Van Aelst, 2008; Salibian-Barrera and Zamar, 2002), or even as a measure of variable importance in machine learning methods like random forest or boosting.

6.3 Others

Reiterating from the conclusions of Chapter 4, the scope of application for the inverse depth-based rank transformation needs to be investigated. A low-hanging fruit in this

respect can be its application in functional data. The generic regularization structure presented in Chapter 5, i.e. (5.2.1) therein, is very interesting. The choice of the reference distribution F represents an initial belief on the correlation structure of responses, and can easily be interpreted as a prior distribution. A study of the methods and modelling algorithms that can possibly arise from such a formulation of bayesian multitask penalized regression is something we want to pursue in future.

References

Adragni, K. P. and Cook, R. D. (2009). Sufficient dimension reduction and prediction in regression. *Phil. Trans. R. Soc. A*, 367:4385–4405.

Akaike, H. (1970). Statistical predictor identification. *Ann. Inst. Statist. Math.*, 22:203–217.

Anderson, T. (3rd ed. 2003). *An Introduction to Multivariate Statistical Analysis*. Wiley, Hoboken, NJ.

Antoniadis, A. and Fan, J. (2001). The Adaptive Lasso and Its Oracle Properties. *J. Amer. Statist. Assoc.*, 96:939–967.

Aulchenko, Y. S., Koning, D. J. D., and Haley, C. (2007). Genome-wide rapid association using mixed model and regression: a fast and simple method for genome-wide pedigree-based quantitative trait loci association analysis. *Nat. Genet.*, 177:577–585.

Bakin, S. (1999). *Adaptive regression and model selection in data mining problems*. PhD thesis, Australian National University, Canberra.

Boente, G. and Salibian-Barrera, M. (2015). S-Estimators for Functional Principal Component Analysis. *J. Amer. Statist. Assoc.*, 110:1100–1111.

Bogdan, M., Chakrabarti, A., Frommelet, F., and Ghosh, J. K. (2011). Asymptotic

Bayes-optimality under sparsity of some multiple testing procedures. *Ann. Statist.*, 39:1551–1579.

Bondell, H. D., Krishna, A., and Ghosh, S. K. (2010). Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics*, 66:1069–1077.

Bose, A. and Chatterjee, S. (2003). Generalized bootstrap for estimators of minimizers of convex functions. *J. Statist. Plan. Inf.*, 117:225–239.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations. *J. R. Statist. Soc. B*, 26:211–252.

Brown, B. (1983). Statistical Use of the Spatial Median. *J. R. Statist. Soc. B*, 45:25–30.

Buhlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data. Methods, Theory and Applications*. Springer.

Calhoun, V. D., Adali, T., and Mcginty, V. B. (2010). fMRI activation in a visual-perception task: network of areas detected using the general linear model and independent components analysis. *Neuroimage*, 14:1080–1088.

Chakraborty, B. and Chaudhuri, P. (1996). On a Transformation and Re-Transformation Technique for Constructing an Affine Equivariant Multivariate Median. *Proc. Amer. Math. Soc.*, 124:2539–2547. doi: 10.1017/S1461145713001296.

Chakraborty, B., Chaudhuri, P., and Oja, H. (1998). Operating Transformation Retransformation Spatial Median and Angle Test. *Stat. Sinica*, 8:767–784. doi: 10.1017/S1461145713001296.

Chang, C.-H., Huang, H.-C., and Ing, C.-K. (2014). Asymptotic theory of generalized information criterion for geostatistical regression model selection. *Ann. Statist.*, 42:2441–2468.

Chatterjee, S. and Bose, A. (2005). Generalized bootstrap for estimating equations. *Ann. Statist.*, 33:414–436.

Chaudhuri, P. (1996). On a geometric notion of quantiles for multivariate data. *J. Amer. Statist. Assoc.*, 91:862–872.

Chen, S. X. and Qin, Y. L. (2010). A Two-sample Test for High-dimensional Data with Application to Gene-Set Testing. *Ann. Statist.*, 38:808–835.

Chen, W. M. and Abecasis, G. (2007). Family-based association tests for genome-wide association scans. *Am. J. Hum. Genet.*, 81:913–926.

Chen, X., He, J., Lawrence, R., and Carbonell, J. G. (2012). Adaptive Multi-task Sparse Learning with an Application to fMRI Study. In *Proceedings of the 2012 SIAM International Conference on Data Mining*, volume 12. DOI: http://dx.doi.org/10.1137/1.9781611972825.19.

Chernozhukov, V., Galichon, A., Hallin, M., and Henry, M. (2017). MongeKantorovich depth, quantiles, ranks and signs. *Ann. Statist.*, 45:223–256.

Claeskens, G. and Hjort, N. L. (2008). *Model Selection and Model Averaging*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, first edition.

Cook, R. D., Li, B., and Chiaromonte, F. (2010). Envelope models for parsimonious and efficient multivariate linear regression. *Biometrika*, 20:927–1010.

Coombes, B. J. (2016). *Tests for detection of rare variants and gene-environment interaction in cohort and twin family studies*. PhD thesis, University of Minnesota.

Croux, C. and Haesbroeck, G. (2000). Principal component analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies. *Biometrika*, 87:603–618.

Cui, W. Y., Seneviratne, C., Gu, J., and Li, M. D. (2012). Genetics of GABAergic signaling in nicotine and alcohol dependence. *Hum. Genet.*, 131:843–855. doi: 10.1007/s00439-011-1108-4.

De Neve, J.-E., Mikhaylov, S., Dawes, C. T., et al. (2013). Born to Lead? A Twin Design and Genetic Association Study of Leadership Role Occupancy. *Leadersh Q*, 24:45–60.

Dedecker, J. and Michel, B. (2011). Deconvolution for the Wasserstein metric and geometric inference. *Electron. J. Statist.*, 5:1394–1423.

Dietz, L. R. and Chatterjee, S. (2014). Logit-normal mixed model for Indian monsoon precipitation. *Nonlin. Processes Geophys.*, 21:939–953.

Dietz, L. R. and Chatterjee, S. (2015). Investigation of Precipitation Thresholds in the Indian Monsoon Using Logit-Normal Mixed Models. In Lakshmanan, V., Gilleland, E., McGovern, A., and Tingley, M., editors, *Machine Learning and Data Mining Approaches to Climate Science*, pages 239–246. Springer.

Disch, A., Hemmerlin, A., Bach, T. J., and Rohmer, M. (1998). Mevalonate-derived isopentenyl diphosphate is the biosynthetic precursor of ubiquinone prenyl side chain in tobacco BY-2 cells. *J. Exp. Bot.*, 331:615–621.

Donoho, D. and Johnstone, I. (1994). Ideal Spatial Adaptation via Wavelet Shrinkages. *Biometrika*, 81:425–455.

Dumbgen, L. (1992). Limit theorems for the simplicial depth. *Statist. Probab. Lett.*, 14:119–128.

Dürre, A., Vogel, D., and Tyler, D. (2014). The spatial sign covariance matrix with unknown location. *J. Multivariate Anal.*, 130:107–117.

Dutta, S. and Ghosh, A. (2012). On robust classification using projection depth. *Ann. Inst. Stat. Math.*, 64-3:657–676.

Dyckerhoff, R. and Mozharovskyi, P. (2016). Exact computation of the halfspace depth. *Comput. Statist. Data Anal.*, 98:19–30.

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *Ann. Statist.*, 1:1–26.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Springer Science, first edition.

El Karoui, N. (2009). Concentration of Measure and Spectra of Random Matrices: with Applications to Correlation Matrices, Elliptical Distributions and Beyond. *Ann. Applied Probab.*, 19:2362–2405.

Eloyan, A., Li, S., Muschelli, J., et al. (2014). Analytic programming with FMRI data: a quick-start guide for statisticians using R. *PLoS One*, 9:e89470. doi: 10.1371/journal.pone.0089470.

Esbensen, K. H., Schönkopf, S., and Midtgaard, T. (1994). *Multivariate Analysis in Practice*. CAMO, Trondheim, Germany.

Fan, J. and Chen, J. (1999). One-Step Local Quasi-Likelihood Estimation. *J. R. Statist. Soc. B*, 61:927–943.

Fan, J. and Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *J. Amer. Statist. Assoc.*, 96:1348–1360.

Fan, Y. and Li, R. (2012). Variable selection in linear mixed effect models. *Ann. Statist.*, 40(4):2043–2068.

Fang, K. T., .Kotz, S., and Ng, K. W. (1990). *Symmetric multivariate and related distributions*. Monographs on Statistics and Applied Probability 36. Chapman and Hall Ltd., London, United Kingdom.

Frebort, I., Kowalska, M., Huska, T., Frebortova, J., and Galuszka, P. (2011). Evolution of cytokinin biosynthesis and degradation. *J. Exp. Bot.*, 62:2431–2452.

Frommelet, F., Ruhaltinger, F., Twaróg, P., and Bogdan, M. (2012). Modified versions of Bayesian Information Criterion for genome-wide association studies. *Comput. Stat. Data Anal.*, 56:1038–1051.

Gelernter, J., Kranzler, H. R., Sherva, R., Almasy, L., et al. (2014). Genome-wide association study of alcohol dependence: significant findings in African- and European-Americans including novel risk loci. *Mol. Psychiatry*, 19:41–49. doi:10.1038/mp.2013.145.

Geyer, C. (1994). On the Asymptotics of Constrained M-Estimation. *Ann. Statist.*, 22:1993–2010.

Ghosh, A. and Chaudhuri, P. (2005). On Maximum Depth and Related Classifiers. *Scand. J. Statist.*, 32:327–350.

Ghosh, S., Vittal, H., Sharma, T., et al. (2016). Indian Summer Monsoon Rainfall: Implications of Contrasting Trends in the Spatial Variability of Means and Extremes. *PLoS ONE*, 11:e0158670. https://doi.org/10.1371/journal.pone.0158670.

Gosswami, B. N. (2005). *The Global Monsoon System: Research and Forecast*, chapter The Asian Monsoon: Interdecadal Variability. World Scientific.

Gross, L. (1967). Potential theory on Hilbert space‘. *J. Funct. Analysis*, 1:123–181.

Haldane, J. (1948). Note on the Median of a Multivariate Distribution. *Biometrika*, 35:414–415.

Hallin, M. and Paindaveine, D. (2002). Optimal tests for multivariate location based on interdirections and pseudo-Mahalanobis ranks. *Ann. Statist.*, 30:1103–1133.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Staehl, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. Wiley.

Hicks, B. M., Schalet, B. D., Malone, S., Iacono, W. G., and McGue, M. (2011). Psychometric and Genetic Architecture of Substance Use Disorder and Behavioral Disinhibition Measures for Gene Association Studies. *Behav Genet.*, 41:459–475. doi:10.1007/s10519-010-9417-2.

Huber, P. J. (1981). *Robust Statistics*. Wiley series in probability and mathematical statistics. Wiley.

Hubert, M., Rousseeuw, P. J., and Branden, K. V. (2005). ROBPCA: A New Approach to Robust Principal Component Analysis. *Technometrics*, 47-1:64–79.

Iacono, W. G., Carlson, S. R., Taylor, J., Elkins, I. J., and McGue, M. (1999). Behavioral disinhibition and the development of substance use disorders: Findings from the Minnesota Twin Family Study. *Dev. Psychopathol.*, 11:869–900.

Irons, D. E. (2012). *Characterizing specific genetic and environmental influences on alcohol use*. PhD thesis, University of Minnesota.

Jiang, J., Rao, S., Gu, Z., and Nguyen, T. (2008). Fence methods for mixed model selection. *Ann. Statist.*, 36:1669–1692.

Jornsten, R. (2004). Clustering and classification based on the l_1 depth. *J. Multivariate Anal.*, 90-1:67–89.

Karpyak, V. M., Biernacka, J. M., Weg, M. W., et al. (2010). Interaction of SLC6A4 and DRD2 polymorphisms is associated with a history of delirium tremens. *Addict. Biol.*, 15:23–34. doi: 10.1111/j.1369-1600.2009.00183.x.

Ke, X. (2012). Presence of multiple independent effects in risk loci of common complex human diseases. *Am. J. Hum. Genet.*, 91:185–192.

Keyes, M. A., Malone, S. M., Elkins, I. J., Legrand, L., McGue, M., and Iacono, W. G. (2009). The Enrichment Study of the Minnesota Twin Family Study: Increasing the yield of twin families at high risk for externalizing psychopathology. *Twin Res. Hum. Genet.*, 12:489–501.

Kirschen, M. P., Chen, S. H., and Desmond, J. E. (2010). Modality specific cerebro-cerebellar activations in verbal working memory: an fMRI study. *Behav. Neurol.*, 23:51–63.

Knight, K. and Fu, W. (2000). Asymptotics for Lasso-Type Estimators. *Ann. Statist.*, 28:1356–1378.

Knutti, R., Furrer, R., Tebaldi, C., et al. (2010). Challenges in Combining Projections from Multiple Climate Models. *J. Clim.*, 23:2739–2758.

Konishi, S. and Kitagawa, G. (1996). Generalised information criteria in model selection. *Biometrika*, 83:875–890.

Krishnamurthy, C. K. B., Lall, U., and Kwon, H.-H. (2009). Changing Frequency and Intensity of Rainfall Extremes over India from 1951 to 2003. *J. Clim.*, 22:4737–4746.

Krishnamurthy, V. and Kinter III, J. L. (2003). The Indian Monsoon and its Relation to Global Climate Variability. In Rodo, X. and Comin, F. A., editors, *Global Climate: Current Research and Uncertainties in the Climate System*. Springer.

Lamparter, D., Marbach, D., Rueedi, R., et al. (2016). Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics. *PLoS Comput. Biol.*, 12:e1004714. doi:10.1371/journal.pcbi.1004714.

Leeb, H. and Potscher, B. M. (2005). Model Selection and Inference: Facts and Fiction. *Econometric Theory*, 21:21–59.

Leskelä, L. and Vihola, M. (2015). Conditional convex orders and measurable martingale couplings. arXiv:1404.0999.

Li, R., Zhong, W., and Zhu, L. (2012). Feature Screening via Distance Correlation Learning. *J. Amer. Statist. Assoc.*, 107:1129–1139.

Li, X., Basu, S., Miller, M. B., Iacono, W. G., and McGue, M. (2011). A Rapid Generalized Least Squares Model for a Genome-Wide Quantitative Trait Association Analysis in Families. *Hum. Hered.*, 71:67–82. doi:10.1159/000324839.

Li, Y., Nan, B., and Zhu, J. (2015). Multivariate Sparse Group Lasso for the Multivariate Multiple Linear Regression with an Arbitrary Group Structure. *Biometrics*, 71:354–363.

Lind, P. A., Macgregor, S., Heath, A. C., and Madden, P. A. F. (2012). Association between *in vivo* alcohol metabolism and genetic variation in pathways that metabolize the carbon skeleton of ethanol and NADH reoxidation in the Alcohol Challenge Twin Study. *Alcohol Clin. Exp. Res.*, 36:2074–2085. doi:10.1111/j.1530-0277.2012.01829.x.

Liu, R. (1990). On a notion of data depth based on random simplices. *Ann. Statist.*, 18:405–414.

Liu, R., Parelius, J., and Singh, K. (1999). Multivariate analysis by data depth: Descriptive statistics, graphics and inference (with discussion). *Ann. Statist.*, 27:783–858.

Liu, R. Y. and Singh, K. (1997). Notions of Limiting P Values Based on Data Depth and Bootstrap. *J. Amer. Statist. Assoc.*, 92(437):266–277.

Locantore, N., Marron, J., Simpson, D., Tripoli, N., Zhang, J., and Cohen, K. (1999). Robust principal components of functional data. *TEST*, 8:1–73.

Loh, P.-L. and Wainwright, M. J. (2015). Regularized M-estimators with Nonconvexity: Statistical and Algorithmic Theory for Local Optima. *J. Mach. Learn. Res.*, 16:559–616.

Magyar, A. and Tyler, D. (2014). The asymptotic inadmissibility of the spatial sign covariance matrix for elliptically symmetric distributions. *Biometrika*, 101:673–688.

Mammen, E. (1993). Bootstrap and Wild Bootstrap for High Dimensional Linear Models. *Ann. Statist.*, 21(1):255–285.

Manolio, T. A., Collins, F. S., Cox, N. J., et al. (2009). Finding the missing heritability of complex diseases. *Nature*, 461:747–753.

McGue, M., Keyes, M., Sharma, A., Elkins, I. J., Legrand, L. N., Johnson, W., and Iacono, W. G. (2007). The environments of adopted and non-adopted youth: Evidence on range restriction from the Sibling Interaction and Behavior Study (SIBS). *Behav. Genet.*, 37:449–462.

McGue, M., Zhang, Y., Miller, M. B., et al. (2013). A Genome-Wide Association Study of Behavioral Disinhibition. *Behav Genet.*, 43. doi:10.1007/s10519-013-9606-x.

Miao, J. and Ben-Israel, A. (1992). On principal angles between subspaces in $\mathbb{R}^n$. *Lin. Algeb. Applic.*, 171:81–98.

Michel, R. and Pfanzagl, J. (1971). The accuracy of the normal approximation for minimum contrast estimates. *Z. Wahrsch Verw. Gebiete*, 18:73–84.

Miller, M. B., Basu, S., Cunningham, J., et al. (2012). The Minnesota Center for Twin and Family Research Genome-Wide Association Study. *Twin Res Hum Genet.*, 15:767–774. doi:10.1017/thg.2012.62.

Minsker, S. (2015). Geometric median and robust estimation in Banach spaces. *Bernoulli*, 21:2308–2335.

Mizera, I. (2002). On Depth and Deep Points: A Calculus. *Ann. Statist.*, 30:1681–1736.

Moon, J.-Y., Wang, B., and Ha, K. J. (2012). Teleconnections associated with Northern Hemisphere summer monsoon intraseasonal oscillation. *Clim. Dyn.*, 40(11):2761–2774.

Mosler, K. (2013). Depth statistics. In Becker, C., Fried, R., and Kuhnt, S., editors, *Robustness and Complex Data Structures*, pages 17–34. Springer Berlin Heidelberg.

Möttönen, J. and Oja, H. (1995). Multivariate spatial sign and rank methods. *J. Nonparametric Stat.*

Narisetty, N. N. and He, X. (2014). Bayesian variable selection with shrinking and diffusing priors. *Ann. Statist.*, 42:789–817.

Narisetty, N. N. and Nair, V. N. (2016). Extremal Depth for Functional Data and Applications. *J. Amer. Statist. Assoc.*, 111:1705–1714.

Natarajan, B. K. (1995). Sparse Approximate Solutions to Linear Systems. *Siam. J. Comput.*, 24:227–234.

Neghaban, S. and Wainwright, M. J. (2011). Simultaneous support recovery in high dimensions: Benefits and perils of block l_1/l_∞-regularization. *IEEE Trans. Inf. Theory*, 57:3841–3863. doi: 10.1017/S1461145713001296.

Neghaban, S. N., Yu, B., Wainwright, M. J., and Ravikumar, P. (2009). A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. In *Advances in Neural Information Processing Systems*, pages 1348–1356.

Obozinski, G., Wainwright, M. J., and Jordan, M. I. (2011). Support Union Recovery in High-dimensional Multivariate Regression. *Ann. Statist.*, 39:1–47.

Oja, H. (1983). Descriptive Statistics for Multivariate Distributions. *Statist. and Prob. Lett.*, 1:327–332.

Oja, H. (2010). *Multivariate Nonparametric Methods with R: An Approach Based on Spatial Signs and Ranks*. Lecture Notes in Statistics. Springer.

Ollilia, E., Oja, H., and Croux, C. (2003). The affine equivariant sign covariance matrix: asymptotic behavior and efficiencies. *J. Multivariate Anal.*, 87:328–355.

Paindaveine, D. and bever, G. V. (2013). From Depth to Local Depth: A Focus on Centrality. *J. Amer. Statist. Assoc.*, 108:1105–1119.

Peng, B. (2016). *Methodologies and Algorithms on Some Non-convex Penalized Models for Ultra High Dimensional Data*. PhD thesis, University of Minnesota Twin Cities.

Peng, H. and Lu, Y. (2012). Model selection in linear mixed effect models. *J. Multivariate Anal.*, 109:109–129.

Pfanzagl, J. (1969). On the measurability and consistency of minimum contrast estimates. *Metrika*, 14:249–272.

Puri, M. L. and Sen, P. K. (1971). *Nonparametric Methods in Multivariate Analysis*. Wiley, New York, NY.

Rosenbloom, K. R., Armstrong, J., Barber, G. P., et al. (2015). The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res.*, 43:D670–81. doi:10.1007/s10519-010-9417-2.

Rothman, A. J., Levina, E., and Zhu, J. (2010). Sparse Multivariate Regression With Covariance Estimation. *J. Comp. Graph. Stat.*, 19:947–962.

Ročkovà, V. and George, E. I. (2016). The Spike-and-Slab LASSO. *J. Amer. Statist. Assoc.*, 0:0–0. http://dx.doi.org/10.1080/01621459.2016.1260469.

Rousseeuw, P. and Hubert, M. (1999). Regression Depth. *J. Amer. Statist. Assoc*, 94:388–402.

Salibian-Barrera, M. and Van Aelst, S. (2008). Robust model selection using fast and robust bootstrap. *Comp. Statist. Data Anal.*, 52:5121–5135.

Salibian-Barrera, M. and Zamar, R. H. (2002). Bootstrapping robust estimates of regression. *Ann. Statist.*, 20(2):556–582.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, 6:461–464.

Segal, I. E. (1958). Distributions in Hilbert space and canonical systems of operators. *Trans. Amer. Math. Soc.*, 88:12–41.

Serfling, R. (2006). Depth Functions in Nonparametric Multivariate Inference. In *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, volume 72, pages 1–16.

Sguera, C., Galeano, P., and Lillo, R. (2014). Spatial depth-based classification for functional data. *TEST*, 23-4:725–750.

Sguera, C., Galeano, P., and Lillo, R. E. (2016). Functional outlier detection by a local depth with application to NOx levels. *Stoch. Envir. Res. and Risk Assess.*, 30:1115–1130.

Shao, J. (1993). Linear model selection by cross validation. *J. Amer. Statist. Assoc.*, 88:484–494.

Shao, J. (1996). Bootstrap model selection. *J. Amer. Statist. Assoc.*, 91:655–665.

Shao, J. and Tu, D. (1995). *The Jackknife and Bootstrap*. Springer.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A Sparse-Group Lasso. *J. Comp. Graph. Stat.*, 22:231–245.

Sirkiä, S., Taskinen, S., and Oja, H. (2007). Symmetrised M-estimators of scatter. *J. Multivariate Anal.*, 98:1611–1629.

Stein, C. (1981). Estimation of the Mean of a Multivariate Normal Distribution. *Ann. Statist.*, 9:1135–1151.

Taskinen, S., Koch, I., and Oja, H. (2012). Robustifying principal component analysis with spatial sign vectors. *Statist. and Prob. Lett.*, 82:765–774.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc. B*, 58(267–288).

Tibshirani, R. J., Rinaldo, A., Tibshirani, R., and Wasserman, L. (2015). Uniform asymptotic inference and the bootstrap after model selection. Preprint. Available at `arXiv:1506.06266`.

Tibshirani, R. J., Taylor, J., Lockhart, R., and Tibshirani, R. (2016). Exact Post-Selection Inference for Sequential Regression Procedures. *J. Amer. Statist. Assoc.*, 111:600–620.

Trenberth, K. E. (2011). Changes in precipitation with climate change. *Clim. Res.*, 47:123–138.

Trenberth, K. E., Dai, A., Rasmussen, R. M., and Parsons, D. B. (2003). The changing character of precipitation. *Bull. Am. Meteorol. Soc.,*, 84:1205–1217. http://dx.doi.org/10.1175/BAMS-84-9-1205.

Tukey, J. (1975). Mathematics and picturing data. In James, R., editor, *Proceedings of the International Congress on Mathematics*, volume 2, pages 523–531.

Tyler, D. (1987). A distribution-free M-estimator of multivariate scatter. *Ann. Statist.*, 15:234–251.

Vardi, Y. and Zhang, C.-H. (2000). The multivariate l_1-median and associated data depth. *Proc. Natl. Acad. Sci.*, 97-4:1423–1426.

Visscher, P. M., Brown, M. A., McCarthy, M. I., and Yang, J. (2012). Five Years of GWAS Discovery. *Amer. J. Hum. Genet.*, 90:7–24. doi: 10.1016/j.ajhg.2011.11.029.

Voisey, J., Swagell, C. D., Hughes, I. P., et al. (2011). A novel SNP in COMT is associated with alcohol dependence but not opiate or nicotine dependence: a case control study. *Behav. Brain Funct.*, 7(51). doi: 10.1186/1744-9081-7-51.

Wakeman, D. G. and Henson, R. N. (2015). A multi-subject, multi-modal human neuroimaging dataset. *Scientif. Data*, 2:article 15001. DOI: 10.1038/sdata.2015.1.

Wang, B., Ding, Q., Fu, X., et al. (2005). Fundamental challenge in simulation and prediction of summermonsoon rainfall. *Geophys. Res. Lett.*, 32:L15711. doi:10.1029/2005GL022734.

Wang, L., Kim, Y., and Li, R. (2013). Calibrating Nonconvex Penalized Regression in Ultra-high Dimension. *Ann. Statist.*, 41:2505–2536.

Wang, L., Peng, B., and Li, R. (2015). A High-Dimensional Nonparametric Multivariate Test for Mean Vector. *J. Amer. Statist. Assoc.*, 110:1658–1669.

Wang, T. Y., Lee, S. Y., Chen, S. L., et al. (2014). Gender-specific association of the SLC6A4 and DRD2 gene variants in bipolar disorder. *Int. J. Neuropsychopharmacol.*, 17:211–222. doi: 10.1017/S1461145713001296.

Westfall, P. H. and Young, S. S. (1993). *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. Wiley; New York.

Wille et al, A. (2004). Sparse graphical Gaussian modeling of the isoprenoid gene network in *Arabidopsis thaliana*. *Genome Biol.*, 5:R92.

Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *Ann. Statist.*, 14(4):1261–1295.

Xu, K., Kranzler, H. R., Sherva, R., Sartor, C. E., et al. (2015). Genomewide Association Study for Maximum Number of Alcoholic Drinks in European Americans and African Americans. *Alcohol Clin. Exp. Res.*, 39:1137–1147. doi: 10.1111/acer.12751.

Yang, J., Ferreira, T., Morris, A. P., et al. (2012). Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet.*, 44:369–375 S361S363.

Yang, Y. (2005). Can The Strengths of AIC and BIC Be Shared? *Biometrika*, 92:937–950.

Yang, Y. and Zou, H. (2015). A fast unified algorithm for solving group-lasso penalize learning problems. *Statist. and Comput.*, 25:1129–1141.

Zhang, C. H. (2010). Nearly Unbiased Variable Selection under Minimax Concave Penalty. *Ann. Statist.*, 38:894–942.

Zhang, H., Shi, J., Liang, F., et al. (2014). A fast multilocus test with adaptive SNP selection for large-scale genetic-association studies. *Eur. J. Hum. Genet.*, 22:696–701.

Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *J. Amer. Statist. Assoc.*, 101:1418–1429.

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.*, 36:1509–1533.

Zuo, Y. (2003). Projection-based depth functions and associated medians. *Ann. Statist.*, 31:1460–1490.

Zuo, Y. and Cui, M. (2005). Depth weighted scatter estimators. *Ann. Statist.*, 33-1:381–413.

Zuo, Y., Cui, M., and He, X. (2004). On the Staehl-Donoho estimator and depth-weighted means of multivariate data. *Ann. Statist.*, 32-1:167–188.

Zuo, Y. and Serfling, R. (2000). General notions of statistical depth functions. *Ann. Statist.*, 28-2:461–482.