UNIVERSITY OF MINNESOTA

This is to certify that I have examined
this copy of a doctoral dissertation by

Full Legal Name of Author

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Adviser Name Here
____
Name of Faculty Adviser

____
Signature of Faculty Adviser

____
Date

GRADUATE SCHOOL

Title of Your Fantastic Thesis

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

Full Legal Name of Author

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

Adviser Name Here, Adviser

Month 20XX

Acknowledgements

I'd like to thank ...

Dedication

This dissertation is dedicated to . . .

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 utility in achieving traditional parametric inferential goals is largely unexplored. In this proposal we develop different approaches to address this. In particular, firstly we define a multivariate rank vector using data depth and use them for robust principal component analysis. Second, we lay out a general model selection framework for supervised learning problems using a depth-based selection criterion and implement it in a sample setup using bootstrap. Finally we provide outlines and initial simulations for an iterated depth-weighted regression estimator.

Contents

List of Tables

List of Figures

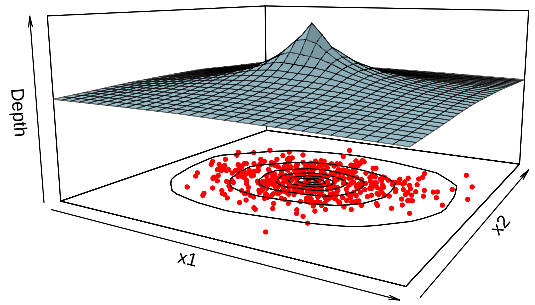
Chapter 1

Introduction

1.1 Background

The nonparametric concept of data-depth had first been proposed by ? 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 recently been popular for robust nonparametric classification (????). In parametric estimation, depth-weighted means (?) and covariance matrices (?) provide high-breakdown point as well as efficient estimators, although they do involve choice of a suitable weight function and tuning parameters. It is also possible to use statistical depth functions in hypothesis testing and an alternate notion of p-values (?). Approaching data depth as from the perspective of breakdown points, ? also introduced the concept of regression depth, which was later generalized by ?.



[]

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

1.2 Data depth and outlyingness

For any multivariate distribution $F = F_{\mathbf{X}}$, 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 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.

? outlines the desirable properties of a statistical depth function:

(D1) Affine invariance: $D(A\mathbf{x} + \mathbf{b}, F_{A\mathbf{X}+\mathbf{b}}) = D(\mathbf{x}, F_{\mathbf{X}})$;

(D2) Maximality at center: $D(\boldsymbol{\theta}, F_{\mathbf{X}}) = \sup_{\mathbf{x} \in \mathbb{R}^p} D(\mathbf{x}, F_{\mathbf{X}})$ for $F_{\mathbf{X}}$ having center of symmetry $\boldsymbol{\theta}$. This point is called the *deepest point* of the distribution.;

(D3) Monotonicity with respect to deepest point: $D(\mathbf{x}; F_{\mathbf{X}}) \leqslant D(\boldsymbol{\theta} + a(\mathbf{x} - \boldsymbol{\theta}), F_{\mathbf{X}})$, $\boldsymbol{\theta}$ being deepest point of $F_{\mathbf{X}}$.;

(D4) Vanishing at infinity: $D(\mathbf{x}; F_{\mathbf{X}}) \to 0$ as $\|\mathbf{x}\| \to \infty$.

In (D2) 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 (D2) and (D3) are actually restrictive towards the formulation of a reasonable depth function that captures the shape of the data cloud. In our derivations in chapter **??**, we replace these two by a slightly weaker condition: (D2*) The *existence* of a maximal point, i.e. $\sup_{\mathbf{x} \in \mathbb{R}^p} D(\mathbf{x}, F_{\mathbf{X}}) < \infty$. We denote this point by $M_D(F)$. Also the assumption (D4) does not need to hold literally as the limit 0 can be replaced by some quantity $c > 0$.

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, sensitive to multimodality, does not measure outlyingness or 'inlyingness' in general and the maximum likelihood point may not be a central point according to any definition of symmetry **?**.

A real-valued function measuring the *outlyingness* of a point with respect to the data cloud can be seen as the opposite of what data depth does. Indeed, such functions have been used to define several depth functions, for example simplicial volume depth, projection depth and L_p-depth. Let us now formally define such functions as a transformation on any depth function:

Definition 1.2.1. Given a random variable $\mathbf{X}$ following a probability distribution F, and a depth function $D(.,.)$, we define Htped of a point $\mathbf{x}$ as: $\tilde{D}(\mathbf{x}, F) = h(d_{\mathbf{x}})$ as any function of the data depth $D(\mathbf{x}, F) = d_{\mathbf{x}}$ so that $h(d_{\mathbf{x}})$ is bounded, monotonically decreasing in $d_{\mathbf{x}}$ and $\sup_{\mathbf{x}} \tilde{D}(\mathbf{x}, F) < \infty$.

For a fixed depth function, there are several choices of a corresponding htped. We develop our theory assuming a general htped function, but for the plots and simulations, fix our htped as $\tilde{D}(\mathbf{x}, F) = M_D(F) - D(\mathbf{x}, F)$, i.e. simply subtract the depth of a point from the maximum possible depth over all points in sample space.

Some measures of data depth are as follows:

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

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

  note here that $d_\Sigma(\mathbf{x}, \boldsymbol{\mu})$ can be seen as a valid htped function of $\mathbf{x}$ with respect to F.

- **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 $PD(\mathbf{x}, F) = 1/(1 + O(\mathbf{x}, F))$.

Computation-wise, MhD is easy to calculate since the sample mean and covariance matrix are generally used as estimates of μ and Σ, respectively. However this makes MhD less robust with respect to outliers. PD is generally approximated by taking maximum over a number of random projections. There have been several approaches for calculating HD. A recent unpublished paper (?) provides a general algorithm that computes exact HD in $O(n^{p-1}\log n)$ time. Here we shall use inbuilt functions in the R package `fda.usc` for calculating different depth functions.

1.3 Summary of work

Analyzing principal components for multivariate data from its spatial sign covariance matrix (SCM) has been proposed as a computationally simple and robust alternative to normal PCA (?), but it suffers from poor efficiency properties and is actually inadmissible with respect to the maximum likelihood estimator. In chapter ?? we use data depth-based spatial ranks in place of spatial signs to obtain the orthogonally equivariant Depth Covariance Matrix (DCM) and use its eigenvector estimates for PCA. We derive asymptotic properties of the sample DCM and influence functions of its eigenvectors. The shapes of these influence functions indicate robustness of estimated principal components, and good efficiency properties compared to the SCM. Finite sample simulation studies show that principal components of the sample DCM are robust with respect to deviations from normality, as well as are more efficient than the SCM and its affine equivariant version, Tyler's shape matrix. Through two real data examples, we also show the effectiveness of DCM-based PCA in analyzing high-dimensional data and outlier detection, and compare it with other methods of robust PCA.

In chapter ?? we introduce a one-step technique for general regression estimators to provide a solution to the problem of statistical model selection. Under very general assumptions, this method correctly identifies the set of non-zero values in the true coefficient (of length p) by comparing only $p+1$ models. We start by defining our selection criterion for a class of candidate models larger than considered before, and providing population-level results that differentiate between correct and wrong models within this class. After this we provide results for a general bootstrap scheme to estimate the criterion in a sample setup, and discuss its details for linear and linear mixed models. Simulations and a real data example demonstrate the efficacy of our method over existing model selection strategies in terms of detecting the correct set of predictors as well as accurate out-of-sample predictions. At the end we also discuss some immediate applications and possible extensions of this foundational methodology.

We provide an outline of a future project in chapter **??**. Here we propose an iterated reweighted least square algorithm for robust estimation of regression coefficients that uses depths of residuals as weights. Through a simulation study we demonstrate the commendable performance of the algorithm, and then provide a broad sketch of our plans on developing the concept.

Chapter 2

Multivariate ranking using data depth

2.1 Introduction

2.2 The robust location problem

Consider now $\mathcal{H}$ to be the p-dimensional Euclidean space $\mathbb{R}^p$, for some positive integer p. Following Fang et al. (1990a), elliptical distributions can be formally defined here using their characteristic function:

Definition 2.2.1. 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 $\Omega \equiv \Sigma^{-1} \in \mathbb{R}^{p \times p}$ 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 $\mathbf{t} \mapsto \phi(\mathbf{t}^T \Sigma \mathbf{t}), \mathbf{t} \in \mathbb{R}^p$.

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

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

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. For ease of notation, we shall denote such a distribution by $\mathcal{E}(\boldsymbol{\mu}, \Sigma, g)$.

Given the above formulation, we focus on the general situation of estimating or testing for the location parameter $\boldsymbol{\mu}$ using weighted sign vectors. 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 $\mathbf{Z} \equiv \Sigma^{-1/2}(\mathbf{X} - \boldsymbol{\mu})$. In other words, it is possible to write $w(\mathbf{X})$ as $f(r)$, with $r = \|\mathbf{Z}\|$. The simplest use of weighted signs here would be to construct an outlier-robust alternative to the Hotelling's T^2 test using their sample mean vector and covariance matrix. Formally, 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:

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

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 2.2.2. *Consider n random variables $Z = (\mathbf{Z}_1, ..., \mathbf{Z}_n)^T$ distributed independently and identically as $\mathcal{E}(\boldsymbol{\mu}, kI_p, g); k \in \mathbb{R}$, 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 but elliptic distributions as well.

2.2.1 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:

$$\boldsymbol{\mu}_w = \arg\min_{\boldsymbol{\mu}_0 \in \mathbb{R}^p} E(w(\mathbf{X})|\mathbf{X} - \boldsymbol{\mu}_0|) \tag{2.1}$$

This can be seen as a generalization of the Fermat-Weber location problem (which has the spatial median (Brown, 1983; Chaudhuri, 1996) as the solution) using data-dependent weights. Using affine equivariant weights in (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 2.2.3. *Let A_w, B_w be two matrices, dependent on the weight function w such that*

$$A_w = E\left[\frac{w(\boldsymbol{\epsilon})}{\|\boldsymbol{\epsilon}\|}\left(1 - \frac{\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T}{\|\boldsymbol{\epsilon}\|^2}\right)\right]; \quad B_w = E\left[\frac{(w(\boldsymbol{\epsilon}))^2\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T}{\|\boldsymbol{\epsilon}\|^2}\right]$$

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

$$\sqrt{n}(\hat{\boldsymbol{\mu}}_w - \boldsymbol{\mu}_w) \rightsquigarrow N_p(\mathbf{0}_p, A_w^{-1} B_w A_w^{-1}) \tag{2.2}$$

We provide a sketch of its proof in the supplementary material, which generalizes equivalent results for the spatial median (Oja, 2010). 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:

$$ARE(\boldsymbol{\mu}_w, \boldsymbol{\mu}_s) = \left[\frac{\det(A^{-1}BA^{-1})}{\det(A_w^{-1}B_w A_w^{-1})}\right]^{1/p}$$

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

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

Corollary 2.2.4. *For the spherical distribution $\mathcal{E}(\boldsymbol{\mu}, kI_p, g); k \in \mathbb{R}, \boldsymbol{\mu} \in \mathbb{R}^p$, we have*

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

Table 2.1 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 outperforms its unweighted counterpart for all data dimensions and distribution families. 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. Moreover, assuming an AR1 covariance structure, i.e. $\sigma_{ij} = \rho^{|i-j|}, \rho \in (0,1)$ results in largely similar ARE values as those obtained in Table 2.1 that assume $\Sigma = I_p$.

2.2.2 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 $n, p \rightarrow \infty$:

$$\frac{C_{n,w}}{\sqrt{\frac{n(n-1)}{2}\text{Tr}(B_w^2)}} \rightsquigarrow N(0,1) \tag{2.3}$$

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

$$\frac{C_{n,w} - \frac{n(n-1)}{2}\boldsymbol{\mu}_0^T A_w^2 \boldsymbol{\mu}_0 (1 + o(1))}{\sqrt{\frac{n(n-1)}{2}\text{Tr}(B_w^2)}} \rightsquigarrow N(0,1) \tag{2.4}$$

we provide the details behind deriving these two results in the supplementary material, which involve modified regularity conditions and sketches of proofs along the lines of Wang et al. (2015).

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

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

when $\Sigma = kI_p$, this again simplifies to $E^2(f(r)/r)/[Ef^2(r).E^2(1/r)]$. The ARE values will be exactly same as those in Table 2.1, 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 small 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

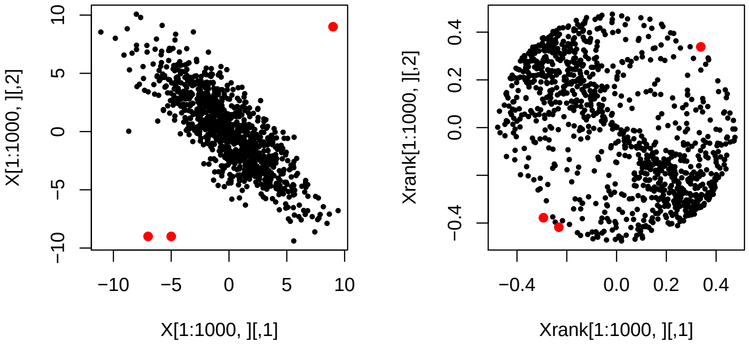
$\boldsymbol{\mu} = \text{rep}(.15, p)$				
p	n	CQ	WPL	$C_{n,w}$
500	20	0.051	0.376	0.418
500	50	0.060	0.832	0.866
1000	20	0.044	0.541	0.584
1000	50	0.039	0.973	0.987
$\boldsymbol{\mu} = \text{rep}(0, p)$				
p	n	CQ	WPL	$C_{n,w}$
500	20	0.049	0.061	0.063
500	50	0.039	0.061	0.064
1000	20	0.042	0.060	0.063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

Table 2.2: Table of empirical powers of level-0.05 tests for the Chen and Qin (CQ), WPL and $C_{n,w}$ statistics

utility. We do this in Table 2.2, 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 the null **better phrasing?**.

2.3 Depth-based rank covariance matrix

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, so for ease of notation while working with transformed random variables, from now on we shall be using $D_{\mathbf{X}}(\mathbf{x}) = D(\mathbf{x}, F)$ to denote the depth of a point $\mathbf{x}$. Now, given a depth function $D_{\mathbf{X}}(\mathbf{x})$ (equivalently, an htped function $\tilde{D}_{\mathbf{X}}(\mathbf{x}) = \tilde{D}(\mathbf{x}, F)$), transform the original random variable as: $\tilde{\mathbf{x}} = \tilde{D}_{\mathbf{X}}(\mathbf{x})\mathbf{S}(\mathbf{x} - \boldsymbol{\mu})$, $\mathbf{S}(.)$ being the spatial sign functional. The transformed random variable $\tilde{\mathbf{X}}$ 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



[t]

Figure 2.1: (Left) 1000 points randomly drawn from $\mathcal{N}_2\left((0,0)^T, \begin{pmatrix} 5 & -4 \\ -4 & 5 \end{pmatrix}\right)$ and (Right) their multivariate ranks based on halfspace depth

vector. Subsequent definitions of multivariate ranks were proposed by Möttönen and Oja (1995); Hallin and Paindaveine (2002) and Chernozhukov et al. (2014). Compared to these formulations, our definition of multivariate ranks works for any general depth function, and provides an intuitive extension to any spatial sign-based methodology.

Figure 2.1 gives an idea of how the multivariate 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 $\boldsymbol{\mu}$ that has radius $M_D(F)$ (which, for the case of halfspace depth, equals 0.5).

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

$$\begin{aligned}
\tilde{\mathbf{x}} &= \tilde{D}_{\mathbf{X}}(\mathbf{x})\mathbf{S}(\mathbf{x}-\boldsymbol{\mu}) \\
&= \tilde{D}_{\Gamma\Lambda^{1/2}\mathbf{Z}+\boldsymbol{\mu}}(\Gamma\Lambda^{1/2}\mathbf{z}+\boldsymbol{\mu}).\mathbf{S}(\Gamma\Lambda^{1/2}\mathbf{z}) \\
&= \Gamma\tilde{D}_{\mathbf{Z}}(\mathbf{z})\mathbf{S}(\Lambda^{1/2}\mathbf{z}) \\
&= \Gamma\Lambda^{1/2}\tilde{D}_{\mathbf{Z}}(\mathbf{z})\mathbf{S}(\mathbf{z})\frac{\|\mathbf{z}\|}{\|\Lambda^{1/2}\mathbf{z}\|} \qquad (2.5)
\end{aligned}$$

$\tilde{D}_{\mathbf{Z}}(\mathbf{z})$ is an even function in $\mathbf{z}$ because of affine invariance, as is $\|\mathbf{z}\|/\|\Lambda^{1/2}\mathbf{z}\|$. Since $\mathbf{S}(\mathbf{z})$ is odd in $\mathbf{z}$ for spherically symmetric $\mathbf{z}$, it follows that $E(\tilde{\mathbf{X}}) = \mathbf{0}$, and consequently we obtain an expression for the covariance matrix of $\tilde{\mathbf{X}}$:

Theorem 2.3.1. *Let the random variable $\mathbf{X} \in \mathbb{R}^p$ follow an elliptical distribution with center $\boldsymbol{\mu}$ and covariance matrix $\Sigma = \Gamma\Lambda\Gamma^T$, its spectral decomposition. Then, given a depth function $D_{\mathbf{X}}(.)$ the covariance matrix of the transformed random variable $\tilde{\mathbf{X}}$ is*

$$Cov(\tilde{\mathbf{X}}) = \Gamma\Lambda_{D,S}\Gamma^T, \quad with \quad \Lambda_{D,S} = \mathbb{E}_{\mathbf{Z}}\left[(\tilde{D}_{\mathbf{Z}}(\mathbf{z}))^2\frac{\Lambda^{1/2}\mathbf{z}\mathbf{z}^T\Lambda^{1/2}}{\mathbf{z}^T\Lambda\mathbf{z}}\right] \qquad (2.6)$$

where $\mathbf{Z} = (Z_1, ..., Z_p)^T \sim N(\mathbf{0}, I_p)$, so that $\Lambda_{D,S}$ a diagonal matrix with diagonal entries

$$\lambda_{D,S,i} = \mathbb{E}_{\mathbf{Z}}\left[\frac{(\tilde{D}_{\mathbf{Z}}(\mathbf{z}))^2\lambda_i z_i^2}{\sum_{j=1}^p \lambda_j z_j^2}\right] \qquad (2.7)$$

The matrix of eigenvectors of the covariance matrix, Γ, remains unchanged in the transformation $\mathbf{X} \to \tilde{\mathbf{X}}$. 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 $\Lambda_{D,S}$ do not change if a scale change is done on all entries of Λ, meaning the $\Lambda_{D,S}$ matrices corresponding to F and cF for some $c \neq 0$ will be same. This is the reason the DCM is not equivariant under affine transformations.

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

$$\Sigma_{Dw} = \frac{1}{Var(\tilde{Z}_1)} E \left[\frac{(\tilde{D}_{\mathbf{X}}(\mathbf{x}))^2 (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T}{(\mathbf{x} - \boldsymbol{\mu})^T \Sigma_{Dw}^{-1} (\mathbf{x} - \boldsymbol{\mu})} \right] \tag{2.8}$$

Its affine equivariance follows from the fact that the weights $(\tilde{D}_{\mathbf{X}}(\mathbf{x}))^2$ depend only on the standardized quantities $\mathbf{z}$ that come from the underlying spherical distribution G. We solve ((2.8)) iteratively by obtaining a sequence of positive definite matrices $\Sigma_{Dw}^{(k)}$ until convergence:

$$\Sigma_{Dw}^{(k+1)} = \frac{1}{Var(\tilde{Z}_1)} E \left[\frac{(\tilde{D}_{\mathbf{X}}(\mathbf{x}))^2 (\Sigma_{Dw}^{(k)})^{1/2} (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T (\Sigma_{Dw}^{(k)})^{1/2}}{(\mathbf{x} - \boldsymbol{\mu})^T (\Sigma_{Dw}^{(k)})^{-1} (\mathbf{x} - \boldsymbol{\mu})} \right]$$

To ensure existence and uniqueness of this estimator, let us consider the class of scatter estimators Σ_M that are obtained as solutions of the following equation:

$$E_{\mathbf{Z}_M} \left[u(\|\mathbf{z}_M\|) \frac{\mathbf{z}_M \mathbf{z}_M^T}{\|\mathbf{z}_M\|^2} - v(\|\mathbf{z}_M\|) I_p \right] = 0 \tag{2.9}$$

with $\mathbf{z}_M = \Sigma_M^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$. Under the following assumptions on the scalar valued functions u and v, the above equation produces a unique solution (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) $P(H) = E_{\mathbf{X}} 1_{\mathbf{x} \in H} < 1 -$

$pv(\infty)/u(\infty)$ and (ii) $P(H) \leqslant 1/p$.

In our case we take $v(r) = Var(\tilde{Z}_1)$, 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, $PD_{\mathbf{Z}}(\mathbf{z}) = (1 - G(\|\mathbf{z}\|); MhD_{\mathbf{Z}}(\mathbf{z}) = (1 + \|\mathbf{z}\|^2)^{-1}; HSD_{\mathbf{Z}}(\mathbf{z}) = (1 + \|\mathbf{z}\|)^{-1}$ etc., so that we can take as u square of the corresponding peripherality functions:

$$u_{PD}(r) = G^2(r); \quad u_{MhD}(r) = \frac{r^4}{(1+r^2)^2}; \quad u_{HSD}(r) = \frac{r^2}{(1 + r/G^{-1}(0.75))^2}$$

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

$$Var(\tilde{Z}_1) = Var\left(\tilde{D}_{\mathbf{Z}}(\mathbf{Z})\frac{Z_1}{\|\mathbf{Z}\|}\right) = Var(\tilde{D}_{\mathbf{Z}}(\mathbf{Z}))Var(S_1(\mathbf{Z})) = \frac{1}{p}Var(\tilde{D}_{\mathbf{Z}}(\mathbf{Z}))$$

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

2.3.1 Calculating the sample DCM and ADCM

Let us now consider n iid random draws from our elliptic distribution F, say $\mathbf{X}_1, ..., \mathbf{X}_n$. For ease of notation, denote $SS(\mathbf{x}; \boldsymbol{\mu}) = \mathbf{S}(\mathbf{x} - \boldsymbol{\mu})\mathbf{S}(\mathbf{x} - \boldsymbol{\mu})^T$. 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: $\sum_{i=1}^n (\tilde{D}_{\mathbf{X}}(\mathbf{x}_i))^2 SS(\mathbf{x}_i; \boldsymbol{\mu})/\sqrt{n}$ has an asymptotic multivariate normal distribution with mean $\sqrt{n}.vec(E[((\tilde{D}_{\mathbf{X}}(\mathbf{X}))^2 SS(\mathbf{x}; \boldsymbol{\mu})])$ and a certain covariance matrix by straightforward application of the central limit theorem (CLT). But in practice the population depth function $D_{\mathbf{X}}(\mathbf{x}) = D(\mathbf{x}, F)$ is estimated by the depth function based on the empirical distribution function, F_n. Denote this sample depth by $D^n_{\mathbf{X}}(\mathbf{x}) = D(\mathbf{x}, F_n)$. Here we make the following assumption regarding how it approximates $D_{\mathbf{X}}(\mathbf{x})$:

(D5) *Uniform convergence*: $\sup_{\mathbf{x} \in \mathbb{R}^p} |D^n_{\mathbf{X}}(\mathbf{x}) - D_{\mathbf{X}}(\mathbf{x})| \rightarrow 0$ as $n \rightarrow \infty$.

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 (Dümbgen, 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 $D^n_{\mathbf{X}}(.)$ 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 2.3.2. *Consider a random variable* $\mathbf{X} \in \mathbb{R}^p$ *having a continuous and symmetric distribution with location center* $\boldsymbol{\mu}$ *such that* $E\|\mathbf{x} - \boldsymbol{\mu}\|^{-3/2} < \infty$. *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*

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

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

Theorem 2.3.3. *Consider n iid samples from the distribution in Lemma 2.3.2. Then, given a depth function* $D_{\mathbf{X}}(.)$ *and an estimate of center* $\hat{\boldsymbol{\mu}}_n$ *so that* $\sqrt{n}(\hat{\boldsymbol{\mu}}_n - \boldsymbol{\mu}) =$

$O_P(1)$,

$$\sqrt{n}\left[vec\left\{\frac{1}{n}\sum_{i=1}^{n}(\tilde{D}^n_{\mathbf{X}}(\mathbf{x}_i))^2 SS(\mathbf{x}_i;\hat{\boldsymbol{\mu}}_n)\right\} - E\left[vec\left\{(\tilde{D}_{\mathbf{X}}(\mathbf{x}))^2 SS(\mathbf{x};\boldsymbol{\mu})\right\}\right]\right] \xrightarrow{D} N_{p^2}(\mathbf{0}, V_{D,S}(F))$$

$$with\ V_{D,S}(F) = Var\left[vec\left\{(\tilde{D}_{\mathbf{X}}(\mathbf{x}))^2 SS(\mathbf{x};\boldsymbol{\mu})\right\}\right]$$

In case F is elliptical, an elaborate form of the covariance matrix $V_{D,S}(F)$ explicitly specifying each of its elements (more directly those of its Γ^T-rotated version) can be obtained, which is given in Appendix Section 2.7. 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 $(\boldsymbol{\mu}_{Dw}, \Sigma_{Dw})$:

$$E\left[\frac{\Sigma_{Dw}^{-1/2}(\mathbf{x}-\boldsymbol{\mu}_{Dw})}{\|\Sigma_{Dw}^{-1/2}(\mathbf{x}-\boldsymbol{\mu}_{Dw})\|}\right] = \mathbf{0}_p \tag{2.10}$$

$$E\left[\frac{(\tilde{D}_{\mathbf{X}}(\mathbf{x}))^2\Sigma_{Dw}^{-1/2}(\mathbf{x}-\boldsymbol{\mu}_{Dw})(\mathbf{x}-\boldsymbol{\mu}_{Dw})^T\Sigma_{Dw}^{-1/2}}{(\mathbf{x}-\boldsymbol{\mu}_{Dw})^T\Sigma_{Dw}^{-1}(\mathbf{x}-\boldsymbol{\mu}_{Dw})}\right] = Var(\tilde{Z}_1)I_p \tag{2.11}$$

In the framework of ((2.8)), for any fixed Σ_M there exists a unique and fixed solution of the location problem $E_{\mathbf{Z}_M}(w(\|\mathbf{z}_M\|\mathbf{z}_M) = \mathbf{0}_p$ 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: $w(\|\mathbf{z}_M\|) = \tilde{D}_{\mathbf{Z}_M}(\mathbf{z}_M)/\|\mathbf{z}_M\|$. Uniqueness of simultaneous fixed point solutions of (2.10) and (2.11) is guaranteed when $\mathbf{X}$ has a symmetric distribution (Huber, 1981).

In practice it is difficult to calculate the scale multiple $Var(\tilde{Z}_1)$ analytically for known depth functions and an arbitrary F. Here we instead obtain its standardized version $\Sigma^*_{Dw} = \Sigma_{Dw}/Var(\tilde{Z}_1)$ (so that the determinant equals 1), alongwith $\boldsymbol{\mu}_{Dw}$

using the following iterative algorithm:

1. Start from some initial estimates $(\boldsymbol{\mu}_{Dw}^{(0)}, \Sigma_{Dw,(0)})$. Set $t = 0$;

2. Calculate the standardized observations $\mathbf{z}_i^{(t)} = \Sigma_{Dw,(t)}^{-1/2}(\mathbf{x}_i - \boldsymbol{\mu}_{Dw}^{(t)})$;

3. Update the location estimate:

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

4. Update the scatter estimate:

$$\Sigma_{Dw}^{*(t+1)} = \frac{1}{n}\sum_{i=1}^n \frac{(\tilde{D}_{\mathbf{X}}^n(\mathbf{x}_i))^2(\mathbf{x}_i - \boldsymbol{\mu}_{Dw}^{(t+1)})(\mathbf{x}_i - \boldsymbol{\mu}_{Dw}^{(t+1)})^T}{\|\mathbf{z}_i^{(t)}\|^2}; \quad \Sigma_{Dw}^{*(t+1)} \leftarrow \frac{\Sigma_{Dw}^{*(t+1)}}{\det(\Sigma_{Dw}^{*(t+1)})^{1/p}}$$

5. Continue until convergence.

2.3.2 Robust PCA using eigenvectors

Since we are mainly interested in using the DCM for robust principal components analysis, from now on we assume that the eigenvalues of Σ 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 with 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 the asymptotic distributions of their sample counterparts. 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

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

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 $T(F) = E_F g$ 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,

$$IF(\mathbf{x}_0; Cov(\tilde{\mathbf{X}}), F) = (\tilde{D}_{\mathbf{X}}(\mathbf{x}_0))^2 SS(\mathbf{x}_0; \boldsymbol{\mu}) - Cov(\tilde{\mathbf{X}})$$

Following Croux and Haesbroeck (2000), we now get the influence function of the i^{th} eigenvector of $Cov(\tilde{\mathbf{X}})$, say $\boldsymbol{\gamma}_D = (\boldsymbol{\gamma}_{D,1}, ..., \boldsymbol{\gamma}_{D,p}); i = 1, ..., p$:

$$\begin{aligned} IF(\mathbf{x}_0; \boldsymbol{\gamma}_{D,i}, F) &= \sum_{k=1; k\neq i}^{p} \frac{1}{\lambda_{D,S,i} - \lambda_{D,S,k}} \left\{ \boldsymbol{\gamma}_k^T IF(\mathbf{x}_0; Cov(\tilde{\mathbf{X}}), \boldsymbol{\gamma}_i) \right\} \boldsymbol{\gamma}_k \\ &= \sum_{k=1; k\neq i}^{p} \frac{1}{\lambda_{D,S,i} - \lambda_{D,S,k}} \left\{ \boldsymbol{\gamma}_k^T (\tilde{D}_{\mathbf{X}}(\mathbf{x}_0))^2 SS(\mathbf{x}_0; \boldsymbol{\mu}) \boldsymbol{\gamma}_i - \lambda_{D,S,i} \boldsymbol{\gamma}_k^T \boldsymbol{\gamma}_i \right\} \boldsymbol{\gamma}_k \\ &= \sum_{k=1; k\neq i}^{p} \frac{\sqrt{\lambda_i \lambda_k} z_{0i} z_{0k}}{\lambda_{D,S,i} - \lambda_{D,S,k}} . \frac{(\tilde{D}_{\mathbf{Z}}(\mathbf{z}_0))^2}{\mathbf{z}_0^T \Lambda \mathbf{z}_0} \boldsymbol{\gamma}_k \qquad (2.12) \end{aligned}$$

where $\Gamma^T \Lambda^{-1/2}(\mathbf{x}_0 - \boldsymbol{\mu}) = \mathbf{z}_0 = (z_{01}, ..., z_{0p})^T$. 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 can be expressed as

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

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

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

When $C = \Sigma_M$, i.e. the solution to (2.8), then Huber (1981) shows that

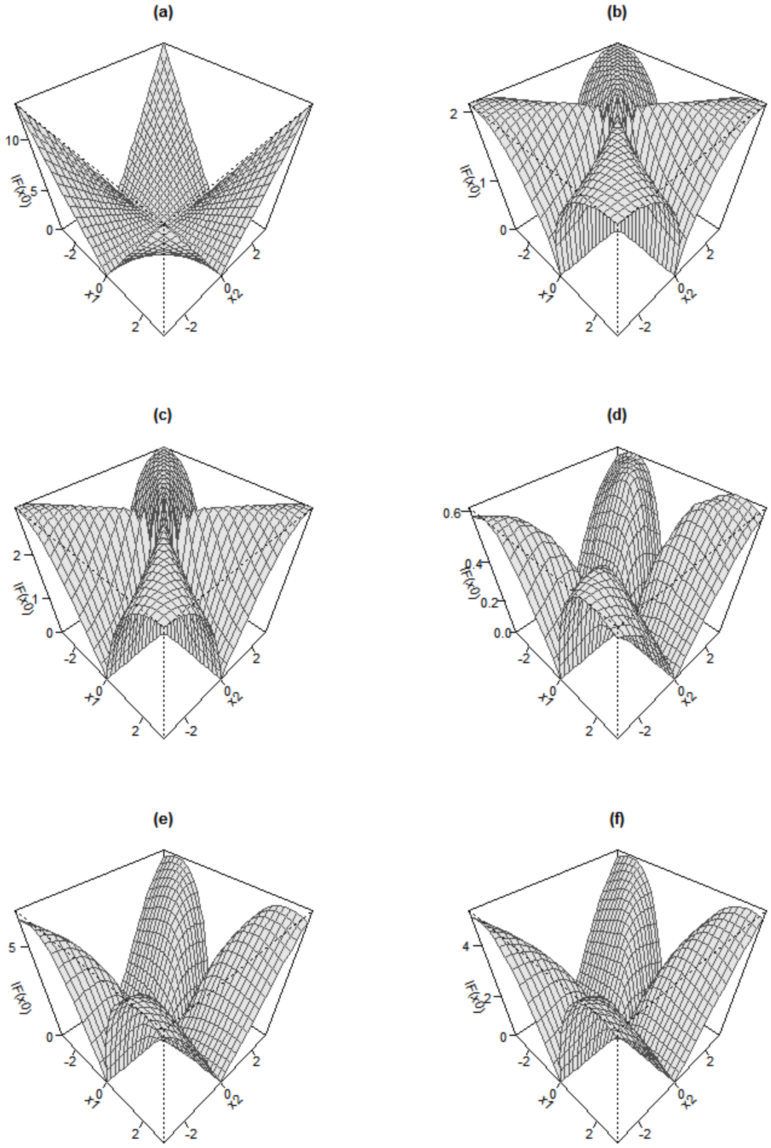
$$\alpha_C(\|\mathbf{z}_0\|) = \frac{p(p+2) u(\|\mathbf{z}_0\|)}{E_{F_0} \left[p u(\|\mathbf{y}\|) + u'(\|\mathbf{y}\|) \|\mathbf{y}\| \right]}$$

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

In Figure 2.2 we consider first eigenvectors of our scatter estimates, as well as teo well-known robust estimates of scatter: the Sign Covariance Matrix (SCM) and Tyler's shape matrix, 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:

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

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



[t]

Figure 2.2: 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 $\boldsymbol{\mu} = \mathbf{0}, \Sigma = \text{diag}(2, 1)$

$(p+2)$ 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 $\hat{C}$ is a $\sqrt{n}$-consistent estimator of a scatter functional C. Then the asymptotic variance of its eigenvectors are (Anderson, 2003)

$$AVar(\sqrt{n}\hat{\boldsymbol{\gamma}}_{c,i}) = \sum_{k=1;k\neq i}^{p} \frac{\lambda_i \lambda_k}{(\lambda_i - \lambda_k)^2} \boldsymbol{\gamma}_k \boldsymbol{\gamma}_k^T \tag{2.13}$$

The asymptotic relative efficiencies of eigenvectors from the sample DCM with respect to the sample covariance matrix can now be derived using (2.13) above and (2.8.2) from Corollary 2.8.2:

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

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 C is given by:

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

where $ASV(C_{12}, F_0)$ is the asymptotic variance of an off-diagonal element of C when

	PD-ACM				HSD-ACM			
Distribution	$p=2$	$p=5$	$p=10$	$p=20$	$p=2$	$p=5$	$p=10$	$p=20$
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 2.3: Table of AREs of the ADCM for different choices of p and data-generating distributions, and two choices of depth functions

the underlying distribution is F_0. Following Croux and Haesbroeck (2000) this equals

$$ASV(C_{12}, F_0) = E_{F_0}\left[\alpha_c(\|\mathbf{z}\|)^2 (S_1(\mathbf{z})S_2(\mathbf{z}))^2\right] = E_{F_0}\alpha_C(\|\mathbf{z}\|)^2 . E_{F_0}(S_1(\mathbf{z})S_2(\mathbf{z}))^2$$

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

$$ARE(\hat{\boldsymbol{\gamma}}_{\Sigma_M,i}, \hat{\boldsymbol{\gamma}}_{Cov,i}; F) = \frac{E_{F_0}\alpha_{Cov}(\|\mathbf{z}\|)^2}{E_{F_0}\alpha_C(\|\mathbf{z}\|)^2} = \frac{E_{F_0}\|\mathbf{z}\|^4 . \left[E_{F_0}(pu\|\mathbf{z}\| + u'(\|\mathbf{z}\|)\|\mathbf{z}\|)\right]^2}{E_{F_0}(u(\|\mathbf{z}\|))^2} \tag{2.14}$$

Table 2.3 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 (HSD-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 10000 samples each for sample sizes $n = 20, 50, 100, 300, 500$. All distributions are centered at $\mathbf{0}_p$, and have covariance matrix $\Sigma = \text{diag}(p, p-1, ...1)$. 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 $\hat{\boldsymbol{\gamma}}_1^E$ relative to that obtained from the sample covariance matrix, say $\hat{\boldsymbol{\gamma}}_1^{Cov}$ is obtained as:

$$FSE(\hat{\boldsymbol{\gamma}}_1^E, \hat{\boldsymbol{\gamma}}_1^{Cov}) = \frac{MSPA(\hat{\boldsymbol{\gamma}}_1^{Cov})}{MSPA(\hat{\boldsymbol{\gamma}}_1^E)}$$

Table 2.4, Table 2.5 and Table 2.6 give 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 (see Table 4 in Sirkiä et al. (2007), and quite close to the affine equivariant spatial sign covariance matrix (see 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 Σ

As we have seen in theorem 2.3.1, 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: $S = \hat\Gamma_D^T X$;

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 od S 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 Σ:

Theorem 2.3.4. *Consider the estimates $\hat\lambda_i$ obtained from the above algorithm, and the matrix of eigenvectors $\hat\Gamma_D$ estimated using the sample DCM. Define $\hat\Sigma = \hat\Gamma_D \hat\Lambda \hat\Gamma_D^T; \hat\Lambda = diag(\hat\lambda_1, ..., \hat\lambda_p)$. Then as $n \rightarrow \infty$,*

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

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

(put in 1 or 2 sentences?)

F = Bivariate t_5	SCM	Tyler	HSD-CM	MhD-CM	PD-CM	HSD-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
F = Bivariate t_6	SCM	Tyler	HSD-CM	MhD-CM	PD-CM	HSD-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
F = Bivariate t_{10}	SCM	Tyler	HSD-CM	MhD-CM	PD-CM	HSD-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
F = Bivariate t_{15}	SCM	Tyler	HSD-CM	MhD-CM	PD-CM	HSD-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
F = Bivariate t_{25}	SCM	Tyler	HSD-CM	MhD-CM	PD-CM	HSD-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
F = BVN	SCM	Tyler	HSD-CM	MhD-CM	PD-CM	HSD-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 2.4: Finite sample efficiencies of several scatter matrices: $p = 2$

3-variate t_5	SCM	Tyler	HSD-CM	MhD-CM	PD-CM	HSD-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
3-variate t_6	SCM	Tyler	HSD-CM	MhD-CM	PD-CM	HSD-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
3-variate t_{10}	SCM	Tyler	HSD-CM	MhD-CM	PD-CM	HSD-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
3-variate t_{15}	SCM	Tyler	HSD-CM	MhD-CM	PD-CM	HSD-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
3-variate t_{25}	SCM	Tyler	HSD-CM	MhD-CM	PD-CM	HSD-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
3-variate Normal	SCM	Tyler	HSD-CM	MhD-CM	PD-CM	HSD-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 2.5: Finite sample efficiencies of several scatter matrices: $p = 3$

4-variate t_5	SCM	Tyler	HSD-CM	MhD-CM	PD-CM	HSD-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
4-variate t_6	SCM	Tyler	HSD-CM	MhD-CM	PD-CM	HSD-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
4-variate t_{10}	SCM	Tyler	HSD-CM	MhD-CM	PD-CM	HSD-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
4-variate t_{15}	SCM	Tyler	HSD-CM	MhD-CM	PD-CM	HSD-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
4-variate t_{25}	SCM	Tyler	HSD-CM	MhD-CM	PD-CM	HSD-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
4-variate Normal	SCM	Tyler	HSD-CM	MhD-CM	PD-CM	HSD-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 2.6: Finite sample efficiencies of several scatter matrices: $p = 4$

2.4 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 $Cov(\mathbf{X})$ in modelling $E(Y|\mathbf{X})$. 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 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 $E(Y|\mathbf{X}) = E(Y|R(\mathbf{X}))$. 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:

$$\mathbf{X}_y = \bar{\boldsymbol{\mu}} + \Gamma \mathbf{v}_y + \boldsymbol{\epsilon} \tag{2.15}$$

where $\mathbf{X}_y = \mathbf{X}|Y = y, \bar{\boldsymbol{\mu}} = E\mathbf{X}$, Γ is a $p \times d$ semi-orthogonal basis for $\mathcal{S}_\Gamma$, the spanning subspace of $\{E\mathbf{X}_y - \bar{\boldsymbol{\mu}} | y \in S_Y\}$ (S_y is sample space of Y) and $\mathbf{v}_y = (\Gamma^T \Gamma)^{-1} \Gamma^T (E\mathbf{X}_y - \bar{\boldsymbol{\mu}}) \in \mathbb{R}^d$. The random error term $\boldsymbol{\epsilon}$ follows a multivariate normal distribution with mean $\mathbf{0}_p$ and covariance matrix Δ. 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 $R(\mathbf{X}) = \Gamma^T \Delta^{-1} \mathbf{X}$. The simplest case of this model is when $\Delta = \sigma^2 I_p$, for which the maximum likelihood estimator of $R(\mathbf{X})$ turns out to be the first d PCs of $Cov(\mathbf{X})$. Taking $\hat{E}\mathbf{X}_y = \bar{\mathbf{X}}_y$ and $\hat{\bar{\boldsymbol{\mu}}} = \bar{\mathbf{X}}$, one can now estimate σ^2 as: $\hat{\sigma}^2 = \sum_{i=1}^p s_{ii}/p$, where s_{ii} is the i^{th} diagonal element of $\hat{Cov}_Y(\mathbf{X}_Y - \bar{\mathbf{X}} - \hat{\Gamma}\hat{\mathbf{v}}_Y)$. Following this, predictions for a new observation

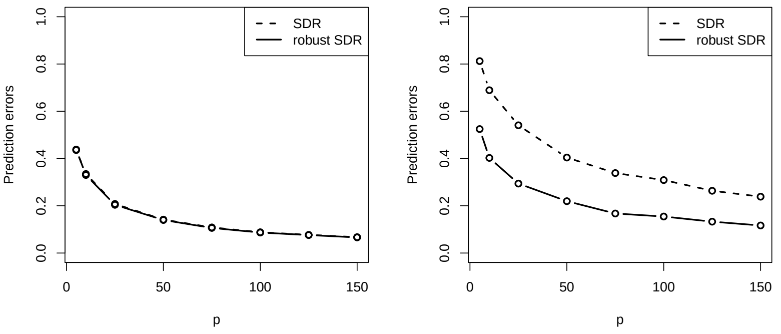


Figure 2.3: Plot of the norm of influence function for first eigenvector of (a) sample covariance matrix, (b) SCM, (c) Tyler's scatter matrix and DCMs for (d) Halfspace depth, (e) Mahalanobis depth, (f) Projection depth for a bivariate normal distribution with $\boldsymbol{\mu} = \mathbf{0}, \Sigma = \text{diag}(2, 1)$

$\mathbf{x}$ is obtained as a weighted sum of the responses:

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

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

- $\tilde{\Gamma}$ = first d eigenvectors of the sample DCM;
- $\tilde{\bar{\boldsymbol{\mu}}}$ = spatial median of the rows of X;
- $\tilde{\boldsymbol{\mu}}_y$ = spatial median of the rows of $(X|Y = y)$, for all $y \in S_Y$;
- $\tilde{\sigma}^2 = \sum_{i=1}^{p} [\widehat{\text{MAD}}_Y (X_{Y,i} - \tilde{\bar{\mu}}_i - \tilde{\boldsymbol{\gamma}}_i^T \tilde{\mathbf{v}}_Y)]^2 / p$, with $\tilde{\Gamma} = (\tilde{\boldsymbol{\gamma}}_1, ..., \tilde{\boldsymbol{\gamma}}_p)^T$.

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 X. For a fixed dimension p, we take $n = 200, d = 1$, generate the responses Y as independent standard normal, and the predictors as $\mathbf{X}_Y = \boldsymbol{\gamma}^* \upsilon_Y^* + \boldsymbol{\epsilon}$, with $\boldsymbol{\gamma}_{p \times 1}^* = (1, ..., 1)^T, \upsilon_Y =$

$Y + Y^2 + Y^3$ and $Var(\boldsymbol{\epsilon}) = 25 I_p$. We measure performance of both SDR models by their mean squared prediction error on another set of 200 observations $(Y^*, \mathbf{X}^*)$ 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 2.3 compares prediction errors using robust and maximum likelihood SDR estimates when X 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 observations in X, 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 our of sample observations for all values of p.

2.5 Robust inference with functional data

(Some technical notations)

We use the approach of Boente and Salibian-Barrera (2015) for performing robust PCA on functional data. Given data on n functions, say $f_1, f_2, ..., f_n \in L^2[0,1]$, each observed at a set of common design points $\{t_1, ..., t_m\}$, we model each function as a linear combination of p mutually orthogonal B-spline basis functions $\delta_1, ..., \delta_p$. Following this, we map data for each of the functions onto the coordinate system formed by the spline basis:

$$\tilde{x}_{ij} = \sum_{l=2}^{m} f_i(t_l)\delta_j(t_l)(t_l - t_{l-1}); \quad 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant p \tag{2.16}$$

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

$$\hat{\tilde{\mathbf{x}}}_i = \tilde{\boldsymbol{\mu}} + \tilde{P}\tilde{\mathbf{s}}_i$$

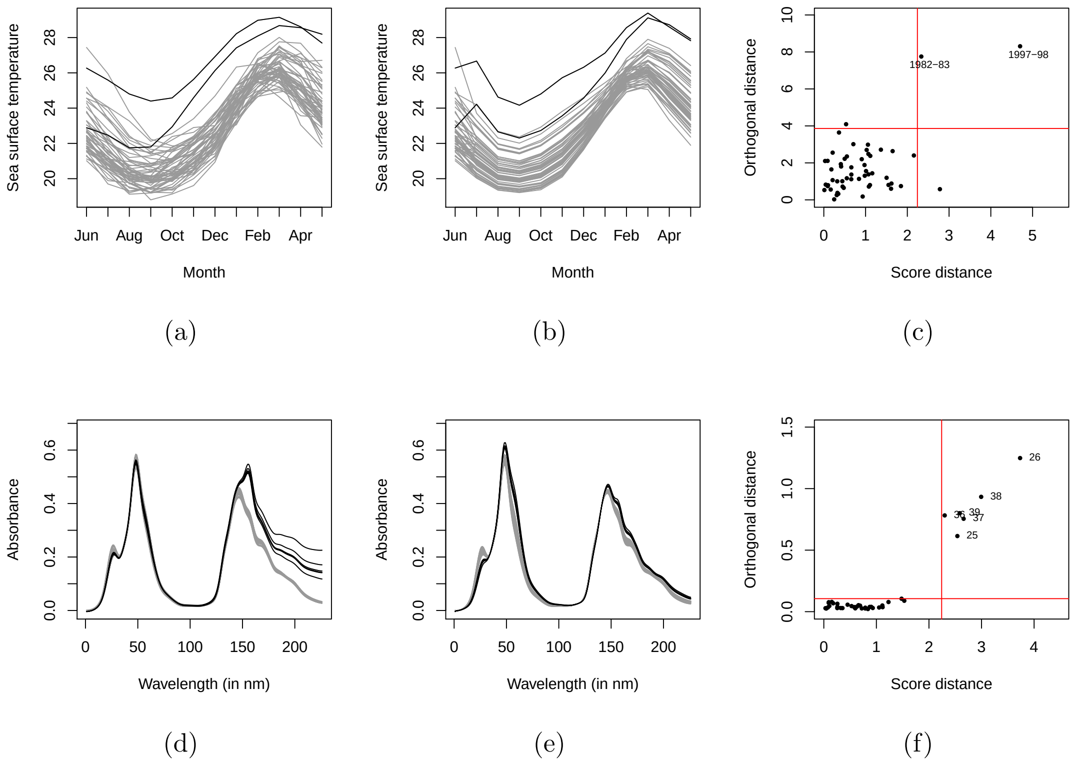
with $\tilde{\boldsymbol{\mu}}$ being the spatial median of $\tilde{X}$. Then we transform this approximation back to the original coordinates: $\hat{f}_i(t_l) = \sum_{j=1}^{p} \hat{\tilde{x}}_{ij}\delta_j(t_l)$.

Detection of anomalous observations is of importance in real-life problems involving functional data analysis. We now demonstrate the utility of our robust method for detecting functional outliers through two data examples.

(SD and OD definition, cutoffs... from previous manuscript)

We first look into 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/`. 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 Figure 2.4). Using a cubic spline basis with knots at alternate months starting in June gives a close approximation of the yearly time series data (panel b), and performing depth-based PCA with $q = 1$ results in two points having their SD and OD larger than cutoff (panel c). 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 b), 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



(a) (b) (c)

(d) (e) (f)

Figure 2.4: Actual sample curves, their spline approximations and diagnostic plots respectively for El-Niño (a-c) and Octane (d-f) datasets

outliers (panels d, e and f of Figure 2.4).

2.6 Conclusion

In the above sections we introduce a covariance matrix based on depth-based multivariate ranks that keeps the eigenvectors of the actual population unchanged for elliptical distributions. We provide asymptotic results for the sample DCM, its eigenvalues and eigenvectors. Bounded influence functions as well as simulation studies suggest that the eigenvector estimates obtained from the DCM are highly robust, yet do not lose much in terms of efficiency. Thus it provides a plausible alternative to existing approaches of robust PCA that are based on estimation of covariance matrices (for example SCM, Tyler's scatter matrix, Dümbgen's symmetrized shape matrix).

Appendix

2.7 Form of $V_{D,S}(F)$

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

$$V_{D,S}(F) = (\Gamma \otimes \Gamma) V_{D,S}(F_\Lambda)(\Gamma \otimes \Gamma)^T$$

where F_Λ has the same elliptic distribution as F, but with covariance matrix Λ. Now,

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

The matrix $vec(\Lambda_{D,S})vec^T(\Lambda_{D,S})$ 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 $E[\sqrt{\lambda_a\lambda_b\lambda_c\lambda_d} z_a z_b z_c z_d.(\tilde D_{\mathbf{Z}}(\mathbf{z}))^4/(\mathbf{z}^T\Lambda\mathbf{z})^2]$, with $1 \leqslant a,b,c,d \leqslant p$. Since $(\tilde D_{\mathbf{Z}}(\mathbf{z}))^4/(\mathbf{z}^T\Lambda\mathbf{z})^2$ 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:

$$(I_{p^2}+K_{p,p})\left\{\sum_{a=1}^p\sum_{b=1}^p \gamma_{ab}^D(\mathbf{e}_a\mathbf{e}_a^T\otimes\mathbf{e}_b\mathbf{e}_b^T) - \sum_{a=1}^p \gamma_{aa}^D(\mathbf{e}_a\mathbf{e}_a^T\otimes\mathbf{e}_a\mathbf{e}_a^T)\right\} + \sum_{a=1}^p\sum_{b=1}^p \gamma_{ab}^D(\mathbf{e}_a\mathbf{e}_b^T\otimes\mathbf{e}_a\mathbf{e}_b^T)$$

where $I_k = (\mathbf{e}_1, ..., \mathbf{e}_k), K_{m,n} = \sum_{i=1}^m\sum_{j=1}^n J_{ij}\otimes J_{ij}^T$ with J_{ij} the $m\times n$ matrix having 1 as $(i,j)^{\text{th}}$ element and 0 elsewhere, and $\gamma_{mn}^D = E[\lambda_m\lambda_n z_m^2 z_n^2.(\tilde D_{\mathbf{Z}}(\mathbf{z}))^4/(\mathbf{z}^T\Lambda\mathbf{z})^2]; 1 \leqslant m,n \leqslant p$.

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

- Variance of on-diagonal elements

$$AVar(\sqrt{n}\hat S_{aa}^D(F_\Lambda)) = E\left[\frac{(\tilde D_{\mathbf{Z}}(\mathbf{z}))^4\lambda_a^2 z_a^4}{(\mathbf{z}^T\Lambda\mathbf{z})^2}\right] - \lambda_{D,S,a}^2$$

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

$$AVar(\sqrt{n}\hat S_{ab}^D(F_\Lambda)) = E\left[\frac{(\tilde D_{\mathbf{Z}}(\mathbf{z}))^4\lambda_a\lambda_b z_a^2 z_b^2}{(\mathbf{z}^T\Lambda\mathbf{z})^2}\right]$$

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

$$ACov(\sqrt{n}\hat{S}^D_{aa}(F_\Lambda), \sqrt{n}\hat{S}^D_{bb}(F_\Lambda)) = E\left[\frac{(\tilde{D}_{\mathbf{Z}}(\mathbf{z}))^4 \lambda_a \lambda_b z_a^2 z_b^2}{(\mathbf{z}^T \Lambda \mathbf{z})^2}\right] - \lambda_{D,S,a}\lambda_{D,S,b}$$

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

$$ACov(\sqrt{n}\hat{S}^D_{ab}(F_\Lambda), \sqrt{n}\hat{S}^D_{cd}(F_\Lambda)) = 0$$

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

$$ACov(\sqrt{n}\hat{S}^D_{ab}(F_\Lambda), \sqrt{n}\hat{S}^D_{cc}(F_\Lambda)) = 0$$

2.8 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 2.8.1. *(Taskinen et al., 2012) Let F_Λ be an elliptical distribution with a diagonal covariance matrix Λ, and $\hat{C}$ be any positive definite symmetric $p \times p$ matrix such that at F_Λ the limiting distribution of $\sqrt{n}vec(\hat{C} - \Lambda)$ is a p^2-variate (singular) normal distribution with mean zero. Write the spectral decomposition of $\hat{C}$ as $\hat{C} = \hat{P}\hat{\Lambda}\hat{P}^T$. Then the limiting distributions of $\sqrt{n}vec(\hat{P} - I_p)$ and $\sqrt{n}vec(\hat{\Lambda} - \Lambda)$ are multivariate (singular) normal and*

$$\sqrt{n}vec(\hat{C}-\Lambda) = [(\Lambda \otimes I_p) - (I_p \otimes \Lambda)]\sqrt{n}vec(\hat{P}-I_p) + \sqrt{n}vec(\hat{\Lambda}-\Lambda) + o_P(1) \quad (2.8.1)$$

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. $V_{D,S}(F)$ to obtain limiting variance and covariances of eigenvalues and eigenvectors.

Corollary 2.8.2. *Consider the sample DCM $\hat{S}^D(F) = \sum_{i=1}^n (\tilde{D}^n_{\mathbf{X}}(\mathbf{x}_i))^2 SS(\mathbf{x}_i; \hat{\boldsymbol{\mu}}_{\mathbf{n}})/n$ and its spectral decomposition $\hat{S}^D(F) = \hat{\Gamma}_D \hat{\Lambda}_D \hat{\Gamma}_D^T$. Then the matrices $G = \sqrt{n}(\hat{\Gamma}_D - \Gamma)$ and $L = \sqrt{n}(\hat{\Lambda}_D - \Lambda_{D,S})$ have independent distributions. The random variable vec(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 $G = (\mathbf{g}_1, ..., \mathbf{g}_p)$ are as follows:*

$$AVar(\mathbf{g}_i) = \sum_{k=1; k \neq i}^{p} \frac{1}{(\lambda_{D,s,k} - \lambda_{D,S,i})^2} E\left[\frac{(\tilde{D}_{\mathbf{Z}}(\mathbf{z}))^4 \lambda_i \lambda_k z_i^2 z_k^2}{(\mathbf{z}^T \Lambda \mathbf{z})^2}\right] \boldsymbol{\gamma}_k \boldsymbol{\gamma}_k^T \tag{2.8.2}$$

$$ACov(\mathbf{g}_i, \mathbf{g}_j) = -\frac{1}{(\lambda_{D,s,i} - \lambda_{D,S,j})^2} E\left[\frac{(\tilde{D}_{\mathbf{Z}}(\mathbf{z}))^4 \lambda_i \lambda_j z_i^2 z_j^2}{(\mathbf{z}^T \Lambda \mathbf{z})^2}\right] \boldsymbol{\gamma}_j \boldsymbol{\gamma}_i^T; \quad i \neq j \tag{2.8.3}$$

where $\Gamma = (\boldsymbol{\gamma}_1, ..., \boldsymbol{\gamma}_p)$. The vector consisting of diagonal elements of L, say $\mathbf{l} = (l_1, ..., l_p)^T$ asymptotically has a p-variate normal distribution with mean $\mathbf{0}_p$ and variance-covariance elements:

$$AVar(l_i) = E\left[\frac{(\tilde{D}_{\mathbf{Z}}(\mathbf{z}))^4 \lambda_i^2 z_i^4}{(\mathbf{z}^T \Lambda \mathbf{z})^2}\right] - \lambda_{D,S,i}^2 \tag{2.8.4}$$

$$ACov(l_i, l_j) = E\left[\frac{(\tilde{D}_{\mathbf{Z}}(\mathbf{z}))^4 \lambda_i \lambda_j z_i^2 z_j^2}{(\mathbf{z}^T \Lambda \mathbf{z})^2}\right] - \lambda_{D,S,i} \lambda_{D,S,j}; \quad i \neq j \tag{2.8.5}$$

Proof of Corollary 2.8.2. In spirit, this corollary is similar to Theorem 13.5.1 in Anderson (2003). Due to the decomposition ((2.8.1)) we have, for the distribution F_Λ, the following relation between any off-diagonal element of $\hat{S}^D(F_\Lambda)$ and the corresponding element in the estimate of eigenvectors $\hat{\Gamma}_D(F_\Lambda)$:

$$\sqrt{n}\hat{\gamma}_{D,ij}(F_\Lambda) = \sqrt{n}\frac{\hat{S}^D_{ij}(F_\Lambda)}{\lambda_{D,S,i} - \lambda_{D,S,j}}; \quad i \neq j$$

So that for eigenvector estimates of the original F we have

$$\sqrt{n}(\hat{\boldsymbol{\gamma}}_{D,i} - \boldsymbol{\gamma}_i) = \sqrt{n}\Gamma(\hat{\boldsymbol{\gamma}}_{D,i}(F_\Lambda) - \mathbf{e}_i) = \sqrt{n}\left[\sum_{k=1;k\neq i}^{p} \hat{\gamma}_{D,ik}(F_\Lambda)\boldsymbol{\gamma}_k + (\hat{\gamma}_{D,ii}(F_\Lambda) - 1)\boldsymbol{\gamma}_i\right] \tag{2.8.6}$$

$\sqrt{n}(\hat{\gamma}_{D,ii}(F_\Lambda) - 1) = o_P(1)$ and $ACov(\sqrt{n}\hat{S}^D_{ik}(F_\Lambda), \sqrt{n}\hat{S}^D_{il}(F_\Lambda)) = 0$ for $k \neq l$, so the above equation implies

$$AVar(\mathbf{g}_i) = AVar(\sqrt{n}(\hat{\boldsymbol{\gamma}}_{D,i} - \boldsymbol{\gamma}_i)) = \sum_{k=1;k\neq i}^{p} \frac{AVar(\sqrt{n}\hat{S}^D_{ik}(F_\Lambda))}{(\lambda_{D,s,i} - \lambda_{D,S,k})^2}\boldsymbol{\gamma}_k\boldsymbol{\gamma}_k^T$$

For the covariance terms, from ((2.8.6)) we get, for $i \neq j$,

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

The exact forms given in the statement of the corollary now follows from the Form of $V_{D,S}$ in Appendix Section 2.7.

For the on-diagonal elements of $\hat{S}^D(F_\Lambda)$ Theorem 2.8.1 gives us $\sqrt{n}\hat{\lambda}_{D,s,i}(F_\Lambda) = \sqrt{n}\hat{S}^D_{ii}(F_\Lambda)$ for $i = 1, ..., p$. Hence

$$\begin{aligned} AVar(l_i) &= AVar(\sqrt{n}\hat\lambda_{D,s,i} - \sqrt{n}\lambda_{D,S,i}) \\ &= AVar(\sqrt{n}\hat\lambda_{D,s,i}(F_\Lambda) - \sqrt{n}\lambda_{D,S,i}(F_\Lambda)) \\ &= AVar(\sqrt{n}S^D_{ii}(F_\Lambda)) \end{aligned}$$

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

2.9 Proofs

Proof of Proposition 2.2.2. Under contiguous alternatives $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$, the weighted sign test statistic $T_{n,w}$ has mean $E(w(\mathbf{Z})\mathbf{S}(\mathbf{Z}))$. 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 $E(w(\mathbf{Z})\mathbf{S}(\mathbf{Z})) = Ew(\mathbf{Z}).E\mathbf{S}(\mathbf{Z})$. The same kind of decomposition holds for $Cov(w(\mathbf{Z})\mathbf{S}(\mathbf{Z}))$.

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

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

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

Sketch of proofs for equations (2.3) and (2.4). 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 2.9.1. *Under the conditions*
(C1) $Tr(\Sigma^4) = o(Tr^2(\Sigma^2))$,
(C2) $Tr^4(\Sigma)/Tr^2(\Sigma^2)\exp[-Tr^2(\Sigma)/128p\lambda^2_{\max}(\Sigma)] = o(1)$

when H_0 is true we have

$$E[(\boldsymbol{\epsilon}_{w1}^T\boldsymbol{\epsilon}_{w2})^4] = O(1)E^2[(\boldsymbol{\epsilon}_{w1}^T\boldsymbol{\epsilon}_{w2})^2] \tag{2.9.1}$$

$$E[(\boldsymbol{\epsilon}_{w1}^T B_w\boldsymbol{\epsilon}_{w1})^2] = O(1)E^2[(\boldsymbol{\epsilon}_{w1}^T B_w\boldsymbol{\epsilon}_{w1})^2] \tag{2.9.2}$$

$$E[(\boldsymbol{\epsilon}_{w1}^T B_w\boldsymbol{\epsilon}_{w2})^2] = o(1)E^2[(\boldsymbol{\epsilon}_{w1}^T B_w\boldsymbol{\epsilon}_{w1})^2] \tag{2.9.3}$$

with $\boldsymbol{\epsilon} \sim \mathcal{E}(\mathbf{0}_p, \Lambda, G)$ 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 2.9.2. *Given that condition (C3) holds, we have* $\lambda_{\max}(B_w) \leqslant 2\frac{\lambda_{\max}}{Tr(\Sigma)}(1+o(1))$.

Lemma 2.9.3. *Define* $D_w = E\left[\frac{w^2(\boldsymbol{\epsilon})}{\|\boldsymbol{\epsilon}\|^2}(I_p - \mathbf{S}(\boldsymbol{\epsilon})\mathbf{S}(\boldsymbol{\epsilon})^T)\right]$. *Then* $\lambda_{\max}(A_w) \leqslant E(w(\boldsymbol{\epsilon})/\|\boldsymbol{\epsilon}\|)$ *and* $\lambda_{\max}(D_w) \leqslant E(w(\boldsymbol{\epsilon})/\|\boldsymbol{\epsilon}\|)^2$. *Further, if (C3) and (C4) hold then* $\lambda_{\min}(A_w) \geqslant E(w(\boldsymbol{\epsilon})/\|\boldsymbol{\epsilon}\|)(1+o(1))/\sqrt{3}$.

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 2.3.1. The proof follows directly from writing out the expression of $Cov(\tilde{\mathbf{X}})$:

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

□

Proof of Lemma 2.3.2. For two positive definite matrices A, B, we denote by $A > B$ that $A - B$ is positive definite. Also, denote

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

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

$$|(\tilde{D}_{\mathbf{X}}^{n_1}(\mathbf{x}_i))^2 - (\tilde{D}_{\mathbf{X}}(\mathbf{x}_i))^2| < \epsilon \tag{2.9.4}$$

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

$$\begin{aligned} S_{n_1} &< \epsilon \sqrt{n_1} \left[\frac{1}{n_1} \sum_{i=1}^{n_1} SS(\mathbf{x}_i; \hat{\boldsymbol{\mu}}_{n_1}) \right] \\ &= \epsilon \sqrt{n_1} \left[\frac{1}{n_1} \sum_{i=1}^{n_1} \left\{ SS(\mathbf{x}_i; \hat{\boldsymbol{\mu}}_{n_1}) - SS(\mathbf{x}_i; \boldsymbol{\mu}) \right\} + \frac{1}{n_1} \sum_{i=1}^{n_1} SS(\mathbf{x}_i; \boldsymbol{\mu}) \right] \end{aligned} \qquad (2.9.5)$$

We now construct a sequence of positive definite matrices $\{A_k(B_k + C_k) : k \in \mathbb{N}\}$ so that

$$A_k = \frac{1}{k}, \quad B_k = \sqrt{N_k} \left[\frac{1}{N_k} \sum_{i=1}^{N_k} \left\{ SS(\mathbf{x}_i; \hat{\boldsymbol{\mu}}_{N_k}) - SS(\mathbf{x}_i; \boldsymbol{\mu}) \right\} \right]$$

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

where $N_k \in \mathbb{N}$ gives the relation ((2.9.4)) in place of N when we take $\epsilon = 1/k$. Under conditions $E\|\mathbf{x} - \boldsymbol{\mu}\|^{-3/2} < \infty$ 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 $B_k = o_P(1)$, thus $A_k(B_k + C_k) \xrightarrow{P} 0$.

Now ((2.9.5)) implies that for any $\epsilon_1 > 0$, $S_{N_k} > \epsilon_1 \Rightarrow A_k(B_k + C_k) > \epsilon_1$, which means $P(S_{N_k} > \epsilon_1) < P(A_k(B_k + C_k) > \epsilon_1)$. Hence the subsequence $\{S_{N_k}\} \xrightarrow{P} 0$. Since the main sequence $\{S_k\}$ is bounded below by 0, this implies $\{S_k\} \xrightarrow{P} 0$. Finally, we have that

$$\begin{aligned} \sqrt{n} \left[\frac{1}{n} \sum_{i=1}^{n} (\tilde{D}_{\mathbf{X}}^n(\mathbf{x}_i))^2 SS(\mathbf{x}_i; \hat{\boldsymbol{\mu}}_n) - \frac{1}{n} \sum_{i=1}^{n} (\tilde{D}_{\mathbf{X}}(\mathbf{x}_i))^2 SS(\mathbf{x}_i; \boldsymbol{\mu}) \right] &\leqslant \\ S_n + \sqrt{n} \left[\frac{1}{n} \sum_{i=1}^{n} \{ SS(\mathbf{x}_i; \hat{\boldsymbol{\mu}}_n) - SS(\mathbf{x}_i; \boldsymbol{\mu}) \} \right] \end{aligned} \qquad (2.9.6)$$

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

Proof of Theorem 2.3.3. The quantity in the statement of the theorem can be broken down as:

$$\sqrt{n}\left[vec\left\{ \frac{1}{n}\sum_{i=1}^n (\tilde D^n_{\mathbf{X}}(\mathbf{x}_i))^2 SS(\mathbf{x}_i;\hat{\boldsymbol{\mu}}_n)\right\} - vec\left\{ \frac{1}{n}\sum_{i=1}^n (\tilde D_{\mathbf{X}}(\mathbf{x}_i))^2 SS(\mathbf{x}_i;\boldsymbol{\mu})\right\}\right] + \\ \sqrt{n}\left[vec\left\{ \frac{1}{n}\sum_{i=1}^n (\tilde D_{\mathbf{X}}(\mathbf{x}_i))^2 SS(\mathbf{x}_i;\boldsymbol{\mu})\right\} - E\left[vec\left\{ (\tilde D_{\mathbf{X}}(\mathbf{x}))^2 SS(\mathbf{x};\boldsymbol{\mu})\right\}\right]\right]$$

The first part goes to 0 in probability by Lemma 2.3.2, and applying Slutsky's theorem we get the required convergence. □

Proof of Theorem 2.3.4. We are going to prove the following:

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

as $n \to \infty$. For (1), we notice $\sqrt{n} vec(\hat\Gamma_D - \Gamma)$ asymptotically has a (singular) multivariate normal distribution following Corollary 2.8.2, so that $\|\hat\Gamma_D - \Gamma\|_F = O_P(1/\sqrt{n})$ 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 $\tilde\lambda_i$ as median-of-small-variances estimator of the *true* score vectors $\Gamma^T X$. For this we have

$$|\tilde\lambda_i - \lambda_i| \xrightarrow{P} 0 \qquad (2.9.7)$$

using Theorem 3.1 of Minsker (2015), with $\mu = \lambda_i$. Now $\hat\lambda_i = \text{med}_j(Var(X_{G_j}^T \hat{\boldsymbol{\gamma}}_{D,i}))$ and $\tilde\lambda_i = \text{med}_j(Var(X_{G_j}^T \boldsymbol{\gamma}_i))$, so that

$$\begin{aligned} |\hat\lambda_i - \tilde\lambda_i| &\leqslant \text{med}_j\left[Var(X_{G_j}^T(\hat{\boldsymbol{\gamma}}_{D,i} - \boldsymbol{\gamma}_i))\right] \\ &\leqslant \|\hat{\boldsymbol{\gamma}}_{D,i} - \boldsymbol{\gamma}_i\|^2 \text{med}_j\left[\text{Tr}(Cov(X_{G_j}))\right] \end{aligned}$$

using Cauchy-Schwarz inequality. Combining the facts $\|\hat{\boldsymbol{\gamma}}_{D,i} - \boldsymbol{\gamma}_i\| = O_P(1/\sqrt{n})$ and $\text{med}_j[\text{Tr}(Cov(X_{G_j}))] \xrightarrow{P} \text{Tr}(\Sigma)$ (Minsker, 2015) with (2.9.7), we get the needed.

□

Chapter 3

Generalized Model Discovery using Statistical Evaluation Maps

3.1 Introduction

In a typical statistical or data science exercise, both *data* and a *statistical model* is 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, which broadly constitute the realm of statistical models. In this paper, we interpret the term *statistical model* very broadly. We recognize 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 3.1.1 (Tree data)**.** Consider the data contained in `data(trees)` in the statistical software `R`. There are 31 observations, on tree volume, height, girth. The observed data is (X_{i1}, X_{i2}, Y_i) denoting the vector of tree girth, height and volume, for $i = 1, \ldots, n$. We denote $p = 2$ for the two explanatory variables *tree girth* and *height*, used to explain the properties of the response variable *volume*.

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

$$C(Y_i, \lambda) = \beta_0 + \beta\log(X_{i1}) + \beta\log(X_{i2}) + e_i$$

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 $\boldsymbol{\theta} = (\lambda, \beta_0, \beta_1, \beta_2, \sigma_1^2, \ldots, \sigma_n^2) \in \mathbb{R}^{p_{nmax}}$ where $p_{nmax} = n + 4$.

Evn more broad frameworks, for example involving nonparametric regression, may be considered. Even in this 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) the Gauss-Markov linear regression model with any other fixed, non-random value of λ, (iii) a model where λ is estimated form data but then a Gauss-Markov 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 **???**), (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 **??**, 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. This list is far from exhaustive, for example, in (iv) each alternative resampling scheme may be called a separate model.

The above list of possible models is far from exhaustive, 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 paper, we address the problem of elicitation of suitable models for analyzing data in a very general framework. We consider candiate 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). A technical definition of model-adequacy is given in Section Section 3.2. Each candidate model has its own set of unknown parameters, which are estimated using a model-specific optimization framework. Then all models are mapped to a common Euclidean reference frame for convenience, using user-defined functions. We then propose using an *evaluation map*, which compares each candidate model against a baseline model, that we call a *preferred model*. This may be the most complex candidate model, or a model in popular or current use, or a hypothesized model, or a model with known parsimony or computational advantage. Data-depth functions **?** are special cases of the kind of functions that may act as an evaluation map. An evaluation map typically compares the distribution of estimated characteristics of interest from any candidate model and the preferred model.

We then propose a suitably parallel computable, two-component resampling-based technique, to obtain a non-negative summary statistic from the evaluation map, that we call the *e-value* of any candidate model. The *e-value* of inadequate models asymptotically tend to zero, while the *e-values* for adequate models remain significantly higher than zero, or may even tend to infinity depending on the choice of the evaluation map. 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 that can be high-dimensional in nature. Thus, there is scope of evaluating models based on domain-knowledge preference, potential risks of various kinds in its usage, or standard statistical measures of skill.

In a typical scenario, some models will obtain higher *e-values* than the preferred model, while other models will obtain lower *e-values*, with models that are not competitive obtaining significantly lower scores. 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. Thus, our proposal 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.

Naturally, the traditional model selection problems of identifying necessary covariates in linear regression or choosing the lag-order in autoregression, are special cases of our framework. In such problems, there is a maximum number of parameters p_{nmax} to consider, and various candidate models consider subsets of a common set of p_{nmax} parameters. The candidate models can be arranged in a lattice, with the supremum being the *least parsimonious* or complete model that involves all p_{nmax} parameters. There are $2^{p_{nmax}}$ such models, and a full evaluation ought to consider all of these. However, owing to the computational challenge involved, especially in older generations and paradigms of computing, various algorithms to reduce computations by evaluating far fewer models have been proposed, which compromise optimality and other properties of the model selection procedure.

In this traditional model selection context, we propose a *fast and parallel model selection* (FPMS) algorithm, where only the least parsimonious model, which we consider as the preferred model in this case, is estimated from the data, and only a total of $p_{nmax} + 1$ models are evaluated. One of the evaluated models is the preferred model, while the other p_{nmax} models are those where only a single explanatory variable is dropped. These latter p_{nmax} model evaluations can be implemented by parallel computations extremely quickly. The final step of the FPMS algorithm is simple, if the model where the j-th parameter is dropped achieves an *e-value* that is lower than that of the preferred model, it is included in the finally selected model. Thus the finally selected model includes only those variables dropping which results in dropping of the model *e-value* below that of the preferred model. We establish that the model selected using this FPMS algorithm is the most parsimonious model that fits the data, termed as the "true model" in many studies, with probability tending to one. We do necessarily advocate selecting the most parsimonious model that fits the data irrespective of other considerations, and urge caution and evaluation of domain scientific principles and purpose before selecting any model.

One of the main advantages of the procedure we propose in this paper is that the distribution of parameter estimators is consistently approximated using the resampling techniques we propose. Thus, we have 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, 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 this paper, 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.

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 **?????** and several references therein. The general principle, and algorithms (excepting the *fast and parallel model selection* discussed in Section Section 3.5) proposed in this paper, advocate 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 paper, we have the ingredients at hand to conduct such studies and have already obtained some results on the conditions under which the proposed *e-value*-based procedures achieve multiple targets of optimal inference. However, we postpone discussion and presentation of such results to a future paper. This is primarily because that arm of study involves several other technical steps, which will greatly increase the length of this manuscript and defeat any attempt at clarity or conciseness.

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, 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" is not a goal of this paper, 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.

Our proposal 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 Section 3.2 we present details of the above steps. This includes notations, generic description of models including the preferred model, discussion on how unknown parameters are estimated in each model, and the broad resampling framework adopted in this paper. This is followed by a separate sub-section detailing the principle of mapping the different models into the common platform $\mathbb{R}^{d_n}$, along with the formal definition of *model inadequacy*. A separate sub-section describes the principle of using evaluation maps. We deliberately avoid presenting technical conditions in Section Section 3.2, so that a clear idea can be formed about the methodology proposed in this paper.

We then present in Section **??** the discussion on how the above four methodological elements are brought together to compute the *e-value* of any model, including the preferred model. Our model selection strategies are based on these *e-values*; broadly speaking, models with higher *e-values* are better. The traditional task for model selection is covariate elicitation. For instance, in Example 3.1.1 above models (*i*) an (*ii*), one may use model selection to discover whether the intercept term, tree girth or height are relevant for explaining tree volume data. In such contexts, in Section Section 3.5 we present the fast and parallel model selection (FPMS) algorithm.

This is followed by two sections on theoretical results. In Section **??** we establish consistency of the resampling procedure adopted in this paper, when one or more models are considered simultaneously. The results from this section establish the paradigm of *joint evaluation of models and inference within each model* advocated in this paper. In Section **??**, we establish that the *e-values* separate the adequate and inadequate models into two groups, and that the FPMS algorithm and variants of that organize the adequate models in increasing order of parsimony with probability tending to one.

Then, in Section Section 3.6 we present two illustrative examples on how our FPMS 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 Section 3.7 we discuss several caveats, future research plans and including some concluding comments.

3.2 The general framework

3.2.1 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 3.1.1 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 example 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 3.1.1 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 paper, 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 im not estimated in any model and im 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 3.1.1 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

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

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

(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 $\mathcal{S}_n$.

The sets $\mathcal{S}_n$ be finite, thus each model may include only a finite number of unknown real-valued constants.

The generic parameter vector corresponding to this model, say $\boldsymbol{\theta}(\mathcal{M}_n) \in \boldsymbol{\Theta}_{mn} \subseteq \boldsymbol{\Theta}_n = \bigtimes_j \boldsymbol{\Theta}_{n,j}$, will thus have the structure

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

Each $\theta_{nj} \subseteq \mathbb{R}$, 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 $\mathcal{C}_n$ are all zero.

We assume that at stage n, we have a *preferred model*, which we denote by $\mathcal{M}_{*n}$: identified with 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 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, or 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.

3.2.2 Transformation to a common platform

Suppose $\mathbf{G}_{mn} : \mathbf{\Theta}_n \to \mathbb{R}^{d_n}$ 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. 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 3.2.1 (Example 3.1.1 continued)**.** In Example 3.1.1 **fix example** we may be interested in *covariate selection*, where we consider which subset of regressors (X_i's) have an influence on the response (Y_i's in both example). Generally when there are p-regressors, there are 2^p possible models on covariate choices alone. In Example 3.1.1 where there are three regression parameters $(\beta_0, \beta_1, \beta_2) \in \mathbb{R}^3$, we have 8 possible models, even when we assume all other properties of the data as given, say in the Gauss-Markov structure. Suppose we consider the model where the entire regression coefficient vector $(\beta_0, \beta_1, \beta_2) \in \mathbb{R}^3$ is estimated as the preferred model. We may use $\mathbf{G}_{*n}(x) = x$ that takes $(\beta_0, \beta_1, \beta_2) \in \mathbb{R}^3$ to itself. For the submodel with no intercept term, we use $\mathbf{G}_{mn}\big((x_0, x_1, x_2)\big) = \big((0, x_1, x_2)\big) \in \mathbb{R}^3$, and so on.

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 $\mathbf{\Theta}(\mathcal{M}_n)$ 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 $\mathcal{G}_n$:

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

In this framework,

Definition 3.2.2. technical conditions for $\mathcal{G}'_n$? 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

$$\lim_{n \to \infty} d(\mathbf{g}_{1n}, \mathcal{G}_{2n}) = 0 \tag{3.2.1}$$

(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 $\mathcal{S}_{0n} = \emptyset$ and $\mathbf{c}_{0n} = \boldsymbol{\theta}_{0n}$, is nested within $\mathcal{M}_n$.

A model that is not adequate is an *inadequate* model. This notion of adequacy of a model depends on the choice of the preferred model, 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 $Y = \mathbf{X}_s \beta_s + \epsilon$, this concept of model adequacy captures standard notions of model 'correctness'. Given a full-rank covariate matrix $\mathbf{X}$, candidate models are fully specified by the index set $\mathcal{S}_n$, and for obvious choices of $\{\mathbf{G}_{mn}\}$, the condition for model adequacy reduces to $(\mathbb{E}Y - \mathbf{X}_s\beta_s) = 0$. The concept of the minimal adequate model merges with that of a 'true model' used in many studies. **mention edge cases**

3.2.3 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 $\mathcal{B} = \{B_{n1}, \ldots, B_{nk_n}\}$, 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 $\mathcal{B}$, 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 $\mathbf{\Theta}_{mn}$:

$$\Psi_{mn}(\boldsymbol{\theta}) = \mathbb{E} \sum_{i=1}^{k_n} \Psi_{mni}\big(\boldsymbol{\theta}, B_{ni}\big) \tag{3.2.2}$$

in any candidate model $\mathcal{M}_n$. Suppose this minimizer is $\boldsymbol{\theta}_{mn}$. We borrow the terminology *energy function* (**loss function?**) from optimization and other literature.

Such functions have also been called *contrast functions*, see **???**. The estimator $\hat{\boldsymbol{\theta}}_{mn}$ of $\boldsymbol{\theta}_{mn}$ is obtained as a minimizer of the sample analog of the above, i.e.

$$\hat{\boldsymbol{\theta}}_{mn} = \arg\min_{\boldsymbol{\theta}_{sn}} \sum_{i=1}^{k_n} \Psi_{mni}\big(\boldsymbol{\theta}_{sn}, B_{ni}\big) \tag{3.2.3}$$

Naturally, only the unknown elements of the generic model vector $\boldsymbol{\theta}(\mathcal{M}_n)$ and their sample equivalents are relevant for the above minimization problems.

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

$$\hat{\boldsymbol{\theta}}_{*n} = \arg\min_{\boldsymbol{\theta}_{*n}} \sum_{i=1}^{k_n} \Psi_{*ni}\big(\boldsymbol{\theta}_{*n}, B_{ni}\big) \tag{3.2.4}$$

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

We now assume the following very general technical 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 $\mathbb{R}^{d_n}$ in it as and when necessary, as the first d_n elements of ℓ_2. Denote by $[\boldsymbol{\theta}]$ the probability distribution of the random variable $\boldsymbol{\theta}$. Then for any candidate model there exists a tight sequence of probability measures $\mathbb{T}_{mn}$ on ℓ_2 with weak limit $\mathbb{T}_{m,\infty} \in \tilde{\ell}_2$, the set of probability measures on ℓ_2, and positive real numbers a_{mn} such that, for all n, $\left[a_{mn}\left(\hat{\boldsymbol{\theta}}_{mn} - \boldsymbol{\theta}_{mn}\right)\right]$ is the distribution of the marginal of $\mathbb{T}_{mn}$ under the first d_n coordinates.

We need only these assumptions to prove the population-level results in the next section. We shall eventually replace it by a few technical conditions in section Section 3.4 in order to prove consistency of the resampling scheme used **word this better**.

3.3 Statistical evaluation maps and e-values

3.3.1 A general evaluation map

We now introduce another function, the *statistical evaluation function* $E_n : \mathbb{R}^{d_n} \times \tilde{\mathcal{G}}_n \to [0, \infty)$, which takes as arguments a point from $\mathbb{R}^{d_n}$ and a probability measure from $\tilde{\mathcal{G}}_n$, the set of probability measures on $\mathcal{G}_n$, and maps that pair into non-negative real numbers. 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. In Section **??**, we discuss the nature of the evaluation map in detail. 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, the latter is a very rich collection of relevant functions: although properties are somewhat more restrictive than those our evaluation map satisfies. While we later use projection depth as our choice of evaluation map in numerical applications, for the 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 unnecessary in our context.

3.3.2 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:

$$e_n(\mathcal{M}_n) = \mathbb{E} E_n \left(\hat{\mathbf{G}}_{mn}, [\hat{\mathbf{G}}_{*n}] \right) \tag{3.3.1}$$

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 tail probability of a test statistic $\hat{T}_n$ with a null distribution, say $[T_{0n}]$. Here we can designate the model corresponding to the null hypothesis as the preferred model, and given the evaluation map $E_n(\hat{T}_n, [T_{0n}]) = \mathcal{I}(T_{0n} < \hat{T}_n)$ the e-value is calculated as $P(T_{0n} < \hat{T}_n)$.

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 the next sections. **mention bayesian interpretation?**

3.3.3 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. Furthermore, we are going to do this by fitting only a single model.

We consider the least parsimonious model with p_n parameters $\boldsymbol{\theta}_{*n} = (\theta_{*n1}, \ldots, \theta_{*np_n})$ as the preferred model, and first obtain a consistent estimator $\hat{\boldsymbol{\theta}}_n = (\hat{\theta}_{n1}, \ldots, \hat{\theta}_{np_n})$ for this. For a general model $\mathcal{M}_n$ specified by the set $\mathcal{S}_n = \{j_1, \ldots, j_{p_{sn}}\} \subseteq \{1, 2, \ldots\}$ and the vector of potentially non-zero constants $\mathbf{c}_n$, we define the parameter estimates to be

$$\hat{\boldsymbol{\theta}}_{mnj} = \begin{cases} \text{Unknown } \hat{\boldsymbol{\theta}}_{*nj} & \text{for } j \in \mathcal{S}_n; \\ \text{Known } c_{nj} & \text{for } j \notin \mathcal{S}_n. \end{cases} \tag{3.3.2}$$

Thus, we do not fit the model $\mathcal{M}_n$ separately, but simply plug-in the estimators from the preferred model at appropriate places, i.e. the indices in $\mathcal{S}_n$. The logic behind this is simple: for a candidate model $\mathcal{M}_n$, the joint distribution of the estimator of its parameters, i.e. $[\hat{\boldsymbol{\theta}}_{sn}]$, is an easily computable function of $[\hat{\boldsymbol{\theta}}_{*n}]$. 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 candiate model to the preferred model, whereby stronger results are obtained.

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 $\mathbb{R}^{p_{mn}}$ to $\mathbb{R}$, for $j = 1, \ldots, d_n$. We assume that such 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

$$G_j(\mathbf{x}) = G_j(\boldsymbol{\theta}) + \mathbf{G}_{1j}^T(\boldsymbol{\theta})\mathbf{t} + 2^{-1}\mathbf{t}^T\mathbf{R}_j(\boldsymbol{\theta} + c\mathbf{t})\mathbf{t} \tag{3.3.3}$$

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

$$\sup_{\mathbf{t}:\|t\|<\delta} \mathbf{R}_j(\boldsymbol{\theta} + c\mathbf{t}) < \mathbf{M}_j; \quad \lambda_{max}(\mathbf{M}_j) < \infty \tag{3.3.4}$$

We now state the technical conditions we assume on the sequence of evaluation maps.

(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 $BG \in \tilde{\mathcal{G}}_n$,

$$E_n(\mathbf{x}, \mathbb{G}) = E_n(a\mathbf{x} + \mathbf{b}, [a\mathbf{G} + \mathbf{b}]) \tag{3.3.5}$$

(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

$$|E_n(\mathbf{x}, \mathbb{G}) - E_n(\mathbf{y}, \mathbb{G})| < \|\mathbf{x} - \mathbf{y}\|^{\alpha_n} \tag{3.3.6}$$

(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

$$\lim_{n\to\infty} \mathbb{E}E_n(\mathbf{Y}_n, [\mathbf{Y}_n]) = \mathbb{E}E_\infty(\mathbf{y}, \mathbb{Y}_\infty) \tag{3.3.7}$$

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:

$$\lim_{n\to\infty} \mathbb{E}E_n(\mathbf{Z}_n, [\mathbf{Y}_n]) = 0 \tag{3.3.8}$$

Clearly, these properties are not mutually exclusive, and some may be derived from others, but we present these together for ease in verification. Additionally some properties like Lipschitz continuity are simply for technical convenience. Further, in all standard cases the limiting distributions are Gaussian with location parameter zero, and consequently $\boldsymbol{\mu}(\mathbb{G}_\infty) = \mathbf{0}$ is valid for all routine applications.

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

Theorem 3.3.1. *Consider a sequence of evaluation functions E_n satisfying properties (E1)-(E3). Then as n $\to \infty$:*

1. *When $\mathcal{M}_n$ is an adequate model, we have $|e_n(\mathcal{M}_n) - e_n(\mathcal{M}_{*n})| \to 0$;*
2. *When $\mathcal{M}_n$ is an inadequate model, $e_n(\mathcal{M}_n) \to 0$.*

Proof of theorem 3.3.1. Part 1. Assuming now that $\mathcal{M}_n$ is an adequate model, we again use the location invariance property of E_n:

$$E_n(\hat{\mathbf{G}}_{mn}, [\hat{\mathbf{G}}_n]) = E_n\left(\hat{\mathbf{G}}_{mn} - \mathbf{G}_{0n}, \left[\hat{\mathbf{G}}_{*n} - \mathbf{G}_{0n}\right]\right) \tag{3.3.9}$$

and decompose the first argument

$$\hat{\mathbf{G}}_{mn} - \mathbf{G}_{0n} = (\hat{\mathbf{G}}_{mn} - \hat{\mathbf{G}}_{*n}) + (\hat{\mathbf{G}}_{*n} - \mathbf{G}_{0n}) \tag{3.3.10}$$

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

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

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

$$\begin{aligned} G_j(\hat{\boldsymbol{\theta}}) &= G_j(\hat{\boldsymbol{\theta}}) + a_n^{-1}\mathbf{G}_{1j}^T(\hat{\boldsymbol{\theta}})\mathbf{T}_n + 2a_n^{-2}\mathbf{T}_n^T\mathbf{R}_j(\hat{\boldsymbol{\theta}}, \mathbf{T}_n)\mathbf{T}_n, \\ &= G_j(\hat{\boldsymbol{\theta}}) + a_n^{-1}\mathbf{G}_{1j}^T(\boldsymbol{\theta})\mathbf{T}_n + \mathbf{b}_n^{-1}(\hat{G}_{1j} - G_{1j})^T\mathbf{T}_n + 2a_n^{-2}\mathbf{T}_n^T\mathbf{R}_j(\hat{\boldsymbol{\theta}}, \mathbf{T}_n)\mathbf{T}_n, \\ &= G_j(\hat{\boldsymbol{\theta}}) + a_n^{-1}\mathbf{G}_{1j}^T(\boldsymbol{\theta})\mathbf{T}_n + a_n^{-1}\mathbf{R}_{nj1} + a_n^{-2}\mathbf{R}_{nj2} \end{aligned}$$

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

$$\mathbb{E}\left(\sum_{j=1}^{d_n} c_j\mathbf{R}_{nj1}\right)^2 = o(a_n^{-1}d_n); \quad \mathbb{E}\left(\sum_{j=1}^{d_n} c_j\mathbf{R}_{nj2}\right)^2 = O(a_n d_n),$$

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 ((3.3.10)) we get

$$\hat{\mathbf{G}}_{mn} - \hat{\mathbf{G}}_{*n} = \mathbf{G}_{mn} - \mathbf{G}_{*n} + O_P(\min\{a_{mn}, a_{0n}\}^{-1}) + o(1) \tag{3.3.11}$$

Since $\mathcal{M}_n$ is an adequate model, $\hat{\mathbf{G}}_{rmn} - \hat{\mathbf{G}}_{*n} = o(n)$. Thus, substituting the above right-hand side in ((3.3.10)) we get

$$\left|E_n\left(\hat{\mathbf{G}}_{mn} - \mathbf{G}_{0n}, \left[\hat{\mathbf{G}}_{*n} - \mathbf{G}_{0n}\right]\right) - E_n\left(\hat{\mathbf{G}}_{*n} - \mathbf{G}_{0n}, \left[\hat{\mathbf{G}}_{*n} - \mathbf{G}_{0n}\right]\right)\right| = o_P(\min\{a_{mn}, a_{0n}, n\}) \tag{3.3.12}$$

because of Lipschitz continuity ogf E_n. Adding back $\mathbf{G}_{0n}$ everywhere and applying (E1) again,

$$|E_n(\hat{\mathbf{G}}_{mn}, [\hat{\mathbf{G}}_{*n}]) - E_n(\hat{\mathbf{G}}_{*n}, [\hat{\mathbf{G}}_{*n}])| = o_P(\min\{a_{mn}, a_{0n}\}) \tag{3.3.13}$$

the proof of part 1 is immediate now.

Part 2. Since the evaluation map E_n is invariant under location and scale transformations, we have

$$E_n(\hat{\mathbf{G}}_{mn}, [\hat{\mathbf{G}}_n]) = E_n \left(a_{*n}(\hat{\mathbf{G}}_{mn} - \mathbf{G}_{0n}), \left[a_{*n}(\hat{\mathbf{G}}_{*n} - \mathbf{G}_{0n}) \right] \right) \tag{3.3.14}$$

Decomposing the first argument,

$$a_{*n}(\hat{\mathbf{G}}_{mn} - \mathbf{G}_{0n}) = a_{*n} \left\{ (\hat{\mathbf{G}}_{mn} - \mathbf{G}_{mn}) + (\mathbf{G}_{mn} - \mathbf{G}_{0n}) \right\} \tag{3.3.15}$$

Since $\mathcal{M}_n$ is inadequate, given $\delta > 0$ there exists a subsequence indexed by $\{k_n\}$ such that $\|\mathbf{G}_{mk_n} - \mathbf{G}_{0k_n}\| > \delta$. Because $a_{*n} \uparrow \infty$, this immediately implies $a_{*n}(\mathbf{G}_{mn} - \mathbf{G}_{0n}) \xrightarrow{P} \infty$. Finally $a_{*n}(\hat{\mathbf{G}}_{mn} - \mathbf{G}_{mn}) = O_P(1)$ using simular arguments as in proof of part 1 above, so that we get the needed by second part of assumption (E4). □

3.4 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 resulting procedure in estimating model e-values through imposing certain necessary conditions on the resampling scheme used.

Before proceeding further we state some necessary notations. For any function h of the parameters in any model, we will often simplify notations by using

$$\begin{aligned}
\mathbf{h} &\equiv \mathbf{h}_{mn} \equiv \mathbf{h}(\boldsymbol{\theta}_{mn}), \\
\hat{\mathbf{h}} &\equiv \hat{\mathbf{h}}_{mn} \equiv \mathbf{h}\left(\hat{\boldsymbol{\theta}}_{mn}\right), \\
\hat{\mathbf{h}}_r &\equiv \hat{\mathbf{h}}_{rmn} \equiv \mathbf{h}\left(\hat{\boldsymbol{\theta}}_{rmn}\right).
\end{aligned}$$

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 $\mathbf{R}_n$, $\mathbf{R}_{mn}$, $\mathbf{R}_{rmn}$ and so, are used as generic for remainder terms, which contribute asymptotically negligible terms in our results. **While we sometimes include algebraic details, often the tedious algebra behind moment calculations and probabilistic bound computations is omitted to contain this paper 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 paper, we have striven for simplicity and not on minimal requirements. Thus, the various assumptions made in this paper are often sufficient conditions, rather than necessary ones, for the theoretical results. (put in intro section maybe?)**

A special case of the family of resampling methods that we use in this paper 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, and all such cases are included in our framework. Since such cases are too numerous to list and review of resampling consistency im not central to this paper, we only demonstrate the properties of our resampling procedure in some interesting frameworks.

Recall that in ((3.2.3)) and ((3.2.4)) we obtain the estimator $\hat{\boldsymbol{\theta}}_n$ by minimizing the *energy functional* or *estimating functional* $\hat{\Psi}_{mn}(\boldsymbol{\theta}) = \sum_{i=1}^{k_n} \Psi_{mni}(\boldsymbol{\theta}, B_{ni})$. The parameter $\boldsymbol{\theta}_n$ is the unique minimizer of the expectation of the above. In this section, we occasionally drop the subscript $_s$ and $_*$ when there im 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 $\Psi_{ni}(\boldsymbol{\theta}) \equiv \Psi_{mni}(\boldsymbol{\theta}, B_{ni})$. Other notational simplifications in various contexts of this section, will be presented as related contexts arise.

We assume that the parameters $\boldsymbol{\theta}_n$ and their estimators have a bijective map to a subset of an Euclidean space, consequently we treat them as vectors of reals. This is a very natural assumption, since essentially all statistical estimation and inference is done on real Euclidean space. With little or no modifications, many parts of the developments below extend to general metric spaces or to infinite-dimensional vector spaces, though we do not explore such generalizations here. We assume in particular, that a distance metric exists on any parameter space under consideration, and we use the notation $d(x, y)$ to denote the distance between two elements x and y on such space.

3.4.1 Smooth estimating functional models

The first case is where the functions $\Psi_{mni}(\cdot, \cdot)$ is smooth in the first argument. This case covers a vast number of models routinely considered in statistics.

In a neighborhood of $\boldsymbol{\theta}_{mn}$, the functions Ψ_{mni} are thrice continuously differentiable in the first argument, with the successive derivatives being denoted by Ψ_{kmni}, $k = 0, 1, 2$. That is, there exists a $\delta > 0$ such that for any $\boldsymbol{\theta} = \boldsymbol{\theta}_{mn} + \mathbf{t}$ satisfying $d(\mathbf{0}, \mathbf{t}) < \delta$ we have

$$\frac{d}{d\boldsymbol{\theta}} \Psi_{mni}(\boldsymbol{\theta}) = \Psi_{0mni}(\boldsymbol{\theta}) \in \mathbb{R}^{p_{mn}}, \tag{3.4.1}$$

and for the a^{th} element of $\Psi_{0mni}(\boldsymbol{\theta}), a = 1, \dots p_{mn}$, denoted by $\Psi_{0mni(a)}(\boldsymbol{\theta})$, we have

$$\Psi_{0mni(a)}(\boldsymbol{\theta}) = \Psi_{0mni(a)}(\boldsymbol{\theta}_{mn}) + \Psi_{1mni(a)}(\boldsymbol{\theta}_{mn})t + 2^{-1}\mathbf{t}^T \Psi_{2mni(a)}(\boldsymbol{\theta}_{mn} + c\mathbf{t})\mathbf{t} \tag{3.4.2}$$

for some $c \in (0, 1)$ possibly depending on a.

We assume that for each $\mathcal{M}_n$ and n, there is a sequence of σ-fields $\mathcal{F}_{mn1} \subset \mathcal{F}_{mn2} \ldots \mathcal{F}_{mnk_n}$ such that $\{\sum_{i=1}^{j} \Psi_{0mni}(\boldsymbol{\theta}_{mn}), \mathcal{F}_{mnj}\}$ is a martingale.

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

$$\boldsymbol{\Gamma}_{0mn} = \mathbf{P}_{0mn}\boldsymbol{\Lambda}_{0mn}\mathbf{P}^T_{0mn}, \tag{3.4.3}$$

where $\mathbf{P}_{0mn} \in \mathbb{R}^{p_{mn}} \times \mathbb{R}^{p_{mn}}$ is an orthogonal matrix whose columns contain the eigenvectors, and $\boldsymbol{\Lambda}_{0mn}$ is a diagonal matrix contining the eigenvalues of $\boldsymbol{\Gamma}_{0mn}$. We assume that $\boldsymbol{\Gamma}_{0mn}$ is positive definite, that is, all the diagonal entries of $\boldsymbol{\Lambda}_{0mn}$ are positive numbers. We assume that there is a constant $\delta_{0s} > 0$ such that $\lambda_{min}(\boldsymbol{\Gamma}_{0mn}) > \delta_{0mn}$ for all sufficiently large n. The matrices $\boldsymbol{\Lambda}^c_{0mn}$ 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 $\boldsymbol{\Gamma}_{1mni}(\boldsymbol{\theta}_{mn})$ be the $p_{mn} \times p_{mn}$ matrix whose a-th row is $\mathbb{E}\Psi_{1mni(a)}(\boldsymbol{\theta}_{mn})$; we assume this expectation exists. Define

$$\boldsymbol{\Gamma}_{1mn}(\boldsymbol{\theta}_{mn}) = \sum_{i=1}^{k_n} \boldsymbol{\Gamma}_{1mni}(\boldsymbol{\theta}_{mn}). \tag{3.4.4}$$

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

$$\boldsymbol{\Gamma}_{1mn} = \mathbf{P}_{1mn}\boldsymbol{\Lambda}_{1mn}\mathbf{Q}^T_{1mn}, \tag{3.4.5}$$

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

We define the matrix

$$\mathbf{A}_{mn} := \mathbf{\Gamma}^{-1/2}_{0mn}\mathbf{\Gamma}_{1mn}. \tag{3.4.6}$$

We assume the following conditions:

(S2) The minimum eigenvalue of $\mathbf{A}^T_{mn}\mathbf{A}_{mn}$ tends to infinity. That is, there is a sequence $a_{mn} \uparrow \infty$ as $n \to \infty$ such that

$$\lambda_{min}\left(\mathbf{\Gamma}_{1mn}\mathbf{\Gamma}^{-1}_{0mn}\mathbf{\Gamma}^T_{1mn}\right) \asymp a^2_{mn}. \tag{3.4.7}$$

(S3)

$$\lambda_{max}\left(\mathbf{\Gamma}^{-1}_{1mn}\mathbf{\Gamma}^2_{0mn}\mathbf{\Gamma}^{-T}_{1mn}\right) = o(\tau^{-2}_{mn}) \tag{3.4.8}$$

as $n \to \infty$ for any $\mathcal{M}_n$ **define τ_{mn} first?**.

(S4)

$$\mathbb{E}\left\| a^{-1}_{mn}\left(\sum_{i=1}^{k_n}\Psi_{1mni} - \mathbf{\Gamma}_{1mn}\right) a^{-1}_{mn}\right\|^2_F = o(p_{mn}\tau^{-2}_{mn}) \tag{3.4.9}$$

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

(S5) For the symmetric matrix $\Psi_{2mni(a)}(\boldsymbol{\theta})$ and for some $\delta_0 > 0$, there exists a symmetric matrix $\mathbf{M}_{2mni(a)}$ such that

$$\sup_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_{mn}|<\delta_{\mathbf{0}}} \Psi_{2mni(a)}(\boldsymbol{\theta}) < \mathbf{M}_{2mni(a)}, \tag{3.4.10}$$

satisfying

$$\sum_{a=1}^{p_{mn}} \sum_{i=1}^{k_n} \mathbb{E}\lambda^2_{max} \left(\mathbf{M}_{2mni(a)}\right) = o\left(a^6_{mn} n^{-1} p_{mn} \tau^{-2}_{mn}\right) \tag{3.4.11}$$

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

$$\sum_{i=1}^{k_n} \mathbf{Z}^2_{mni} \xrightarrow{P} 1, \text{ and } \mathbb{E}\left[\max_i |\mathbf{Z}_{mni}|\right] \to 0. \tag{3.4.12}$$

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

We now state the conditions on the resampling weights $\mathbb{W}_{rmni}$, which for any n may be collected together in the vector $\mathcal{W}_{rmn} = (\mathbb{W}_{rmn1}, \ldots, \mathbb{W}_{rmnk_n})^T \in \mathbb{R}^{k_n}$. 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 $\mathcal{W}_{1mn}, \ldots, \mathcal{W}_{Rmn}$ for some sufficiently large integer R, and using them in a Monte Carlo procedure, where for any $r = 1, \ldots, R$, we minimize

$$\sum_{i=1}^{k_n} \mathbb{W}_{rmni} \Psi_{mni}\left(\theta, B_{ni}\right) \tag{3.4.13}$$

to obtain the resampling version of the estimator $\hat{\boldsymbol{\theta}}_{rmn} \in \mathbb{R}^{p_{mn}}$.

We assume that for each $i = 1, \ldots, k_n$, $\mathbb{E}\mathbb{W}_{rmni} = \mu_{mn}$ and $\mathbb{V}\mathbb{W}_{rmni} = \tau^2_{mn}$, consequently we write the centered and scaled resampling weights as

$$\mathbf{W}_{rmni} = \tau_{mn}^{-1} \left(\mathbb{W}_{rmni} - \mu_{mn} \right), \tag{3.4.14}$$

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

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

$$\mathbb{E}\mathbb{W}_{rmni} = \mu_{mn}, \tag{3.4.15}$$

$$\mathbb{V}\mathbb{W}_{rmni} = \tau_{mn}^2 \uparrow \infty, \tag{3.4.16}$$

$$\tau_{mn}^2 = o(a_{mn}^2), \tag{3.4.17}$$

$$\mathbb{E} W_{rmn1} W_{rmn2} = O(k_n^{-1}), \tag{3.4.18}$$

$$\mathbb{E} W_{rmn1}^2 W_{rmn2}^2 \to 1, \tag{3.4.19}$$

$$\mathbb{E} W_{rmn1}^4 < \infty. \tag{3.4.20}$$

Example 3.4.1 (The m-out-of-n (moon) bootstrap)**.** In our framework, the *moon*-bootstrap is identified with $\mathcal{W}_{rmn}$ 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 $\mathbb{E}\mathbb{W}_{rmni} = \mu_{mn} = (m^{-1}k_n)(m/k_n) = 1$, and $\mathbb{V}\mathbb{W}_{rmni} = \tau_{mn}^2 = (m^{-1}k_n)^2(mk_n^{-1}(1 - k_n^{-1}) = O(m^{-1}k_n)$. In typical applications of the *moon*-bootstrap, as in its application in this paper, we require that $m \to \infty$ and $m/k_n \to 0$ as $n \to \infty$. Thus we have $\tau_{mn}^2 \to \infty$ as $n \to \infty$, thus the *scale* factor of the resampling weights $\mathbb{W}_{rmni}$ 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 tend to infinity with n.

Example 3.4.2 (The scale-enhanced Bayesian bootstrap)**.** A version of Bayesian bootstrap may be constructed by choosing $\mathbb{W}_{rmni}$ to be independent and identically distributed Gamma random variables, with mean $\mu_{mn} = 1$ and variance $\tau_{mn}^2 \to \infty$ 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 3.4.3. *Assume conditions (S2)-(S5) and that* $p_{mn}^2 k_n^{-1} \to 0$ *as* $n \to \infty$*. Additionally, assume that the resampling weights* $\mathbb{W}_{rmni}$ *are exchangeable random variables satisfying the conditions* ((3.4.15))-((3.4.20))*. Define* $\hat{\mathbf{B}}_{mn} := \tau_{mn}^{-1}\hat{\mathbf{\Gamma}}_{0mn}^{1/2}\hat{\mathbf{\Gamma}}_{1mn}^{-1}$*, where* $\hat{\mathbf{\Gamma}}_{0mn}$ *and* $\hat{\mathbf{\Gamma}}_{1mn}$ *are sample equivalents of* $\mathbf{\Gamma}_{0mn}$ *and* $\mathbf{\Gamma}_{1mn}$*, respectively. Then* $\mathbf{A}_{mn}(\hat{\boldsymbol{\theta}}_{sn} - \boldsymbol{\theta}_{sn})$ *converges weakly to the standard Normal distribution in* p_{sn}*-dimension, and conditional on the data,* $\hat{\mathbf{B}}_{mn}(\hat{\boldsymbol{\theta}}_{rsn} - \hat{\boldsymbol{\theta}}_{sn})$ *also converges weakly to the same distribution in probability.*

We note that we only consider the case where $\tau_{mn}^2 \to \infty$ as $n \to \infty$ here, since that is essential for the rest of the paper.

3.4.2 Bootstrap estimation of e-values

state simplifying assumption that $\mathbf{A}_{mn}$ is a diagonal matrix, define b_{mn}

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 propose using two independent bootstrap samples to approximate $[\hat{\mathbf{G}}_{mn}]$ and $[\hat{\mathbf{G}}_{*n}]$, indexed by $_{r.}$ and $_{r_1.}$ respectively, for technical simplicity, and define the sample e-value to be

$$\hat{e}(\mathcal{M}_n) := \mathbb{E}_r E_n(\hat{\mathbf{G}}_{rmn}, [\hat{\mathbf{G}}_{r_1 * n}]) \tag{3.4.21}$$

The expectation above is taken on the first set of bootstrap samples. Following this we have

Theorem 3.4.4. *Consider a resampling scheme satisfying technical conditions in the previous subsection, and an evaluation map E_n satisfying the assumptions (E1)-(E4). Then as $n \to \infty$,*

1. *For $b_{*n} = O(\min\{a_{*n}, a_{mn}, b_{mn}\})$ and any adequate model $\mathcal{M}_n$ we have $|\hat{e}_n(\mathcal{M}_n) - e_n(\mathcal{M}_n)| \xrightarrow{P} 0$;*

2. *For $b_{*n} = o(\min\{a_{*n}, a_{mn}, b_{mn}\})$ and any inadequate model $\mathcal{M}_n$ we have $\hat{e}_n(\mathcal{M}_n) \xrightarrow{P} 0$.*

Proof of theorem 3.4.4. Part 1. Taking a similar approach as in the proof of theorem 3.3.1, we get

$$b_{*n}^{-1}(\hat{\mathbf{G}}_r - \hat{\mathbf{G}}) = \mathbf{G}_1^T \mathbf{T}_{rn} + \mathbf{R}_n \tag{3.4.22}$$

with $\mathbb{E}_r \|\mathbf{R}_n\|^2 = o_P(1)$.

Now following assumption (E1),

$$E_n(\hat{\mathbf{G}}_{rmn}, [\hat{\mathbf{G}}_{r_1 * n}]) = E_n((\hat{\mathbf{G}}_{rmn} - \hat{\mathbf{G}}_{*n}), [\hat{\mathbf{G}}_{r_1 * n} - \hat{\mathbf{G}}_{*n}]) \tag{3.4.23}$$

Expanding first argument of the right-hand side

$$\hat{\mathbf{G}}_{rmn} - \hat{\mathbf{G}}_{*n} = (\hat{\mathbf{G}}_{rmn} - \hat{\mathbf{G}}_{r_1 * n}) + (\hat{\mathbf{G}}_{r_1 * n} - \hat{\mathbf{G}}_{*n}) \tag{3.4.24}$$

Again borrowing from the proof of theorem 3.3.1, this time using an equivalent decomposition as ((3.3.11)) and proceeding further we get

$$\left| E_n(\hat{\mathbf{G}}_{rmn}, [\hat{\mathbf{G}}_{r_1 * n}]) - E_n(\hat{\mathbf{G}}_{r_1 * n}, [\hat{\mathbf{G}}_{r_1 * n}]) \right| = o_{P_n}(b_{*n}^{-1}) \tag{3.4.25}$$

where $s_n = o_{P_n}(t_n)$ meaning $s_n / t_n \to 0$ in probability conditional on the data. **make it clearer?**

Part 2. Following location and scale invariance of E_n, rewrite $E_n(\hat{\mathbf{G}}_{rmn}, [\hat{\mathbf{G}}_{r_1 * n}])$ as:

$$E_n(\hat{\mathbf{G}}_{rmn}, [\hat{\mathbf{G}}_{r_1 * n}]) = E_n(b_{*n}(\hat{\mathbf{G}}_{rmn} - \hat{\mathbf{G}}_{*n}), [b_{*n}(\hat{\mathbf{G}}_{r_1 * n} - \hat{\mathbf{G}}_{*n})]) \tag{3.4.26}$$

and then

$$b_{*n}(\hat{\mathbf{G}}_{rmn} - \hat{\mathbf{G}}_{*n}) = \frac{b_{*n}}{b_{mn}}.b_{mn}(\hat{\mathbf{G}}_{rmn} - \hat{\mathbf{G}}_{mn}) + b_{*n}(\hat{\mathbf{G}}_{mn} - \hat{\mathbf{G}}_{*n}) \tag{3.4.27}$$

From theorem 3.4.4 $b_{mn}(\hat{\mathbf{G}}_{rmn} - \hat{\mathbf{G}}_{mn})$ and $a_{mn}(\hat{\mathbf{G}}_{mn} - \mathbf{G}_{mn})$ converge to the same limiting distribution for almost every data sequence; and choice of b_{*n} means that, conditioned on the data, our first summand in the right side above goes to $\mathbf{0}$ in probability. For the second summand, notice that the approximation ((3.3.11)) holds for any candidate model. Since $b_{*n} \uparrow \infty$, this implies $\|b_{*n}(\hat{\mathbf{G}}_{mn} - \hat{\mathbf{G}}_{*n})\| \xrightarrow{P} \infty$. The proof is now complete through application of the second part of (E4). □

Proving the above theorem largely requires the same arguments used in the proof of its population counterpart, i.e. theorem 3.3.1. Note that the convergence occurs for a broader range of the bootstrap rate parameter b_{*n}, while a slower rate is required to separate e-value estimates of inadequate models from those of the adequate models. In practice when dealing with $\sqrt{n}$-consistent estimators (i.e. $a_{mn} = a_{*n} = \sqrt{n}$ for all $\mathcal{M}_n$, this would mean choosing the variance parameter τ_n^2 of the resampling weight distribution $\mathbb{W}_{mn} \equiv \mathbb{W}$ such that $\tau_n^2 \to \infty$ and $\tau_n^2/n \to 0$ as $n \to \infty$. Interestingly, the bootstrap model selection criterion by Shao Shao (1996) uses 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). **better language / mention in short here and elaborate in discussion?**

3.5 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 methods **??**, the fence method **?**.

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. Now, all candidate models are sub-models of the least parsimonious model, which we shall refer as the 'full model' from now on, 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 at most 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) **?**, a full-scale analysis would require analyzing all 2^p possible candidate models. This is an NP-Hard problem **?**, 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.

3.5.1 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 $\mathbf{G}_{mn}$ 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 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 $E : \mathbb{R}^d \times \tilde{\mathbb{R}}^d \rightarrow [0, \infty)$ 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 $\mathbb{R}^p$,

$$E(\mathbf{x}, \mathbb{G}) = E(\mathbf{A}\mathbf{x} + \mathbf{b}, [\mathbf{A}\mathbf{G} + \mathbf{b}]) \tag{3.5.1}$$

(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 $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$,

$$|E(\mathbf{x}, \mathbb{G}) - E(\mathbf{y}, \mathbb{G})| < \|\mathbf{x} - \mathbf{y}\|^{\alpha} \tag{3.5.2}$$

(D3) For any $\mathbb{G} \in \tilde{\mathbb{R}}^p$, $\lim_{\|\mathbf{x}\| \rightarrow \infty} E(\mathbf{x}, \mathbb{G}) = 0$.

(D4) 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 $E(\mathbf{y}, \mathbb{Y})$.

(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 $\mathbf{x} \in \mathbb{R}^p$

$$E(\mathbf{x}, \mathbb{G}) < E(\boldsymbol{\mu}(\mathbb{G}) + t(\mathbf{x} - \boldsymbol{\mu}(\mathbb{G})), \mathbb{G}) < E(\boldsymbol{\mu}(\mathbb{G}), \mathbb{G}) = \sup_{\mathbf{x} \in \mathbb{R}^p} E(\mathbf{x}, \mathbb{G}) < \infty \tag{3.5.3}$$

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 first property is a stronger version of (E1), and the second a restatement of (E2), both assuming a common evaluation map for all n. The third and fourth properties together imply (E3), while (D5) will be essential in proving the theoretical results that follow. Also note that (D1), (D3) and (D5) have traditionally been used for data depth functions , and the other two arise implicitly for many implementations of data depth . 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 specify a p-dimensional elliptical distribution is specified using the mean vector $\boldsymbol{\mu}$, positive definite covariance matrix $\boldsymbol{\Sigma}$ and a scalar valued density generator function g Fang et al. (1990b); with the density

$$f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, g) = |\boldsymbol{\Sigma}|^{-1/2} g\left((\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} \mathbf{x} - \boldsymbol{\mu})\right) \tag{3.5.4}$$

We shall assume elliptical asymptotic distributions for full model estimators $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_*$. Specifically, we denote elliptical distributions as $\mathcal{E}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, g)$ and have

(S1a) The limiting distribution $\mathbb{T}$ of the full model estimate, i.e. $a_n(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0), (a_{*n} \equiv a_n)$ 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 $\text{plim}_{n \rightarrow \infty} \mathbf{V}_n = \mathbf{V}$.

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

3.5.2 Derivation of the algorithm

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

Theorem 3.5.1. *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.

Proof of theorem 3.5.1. 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 $a_n \mathbf{V}_n^{-1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \Rightarrow \mathbb{T}_0$. 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. $e_n(\mathcal{M}_1) > e_n(\mathcal{M}_2)$.

□

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 now elicit the following result:

Corollary 3.5.2. *Consider the subcollection of candidate models* $\mathbb{M}_0 = \{\mathcal{M} : c_j = 0 \forall j \notin \mathcal{S}\}$*. 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* $n_1 > N$*,*

$$\mathcal{M}_0 = \underset{\mathcal{M} \in \mathbb{M}_0}{\arg\max} \left[e_{n_1}(\mathcal{M}) \right] \tag{3.5.5}$$

Proof of corollary 3.5.2. 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 3.5.1 implies $e_n(\mathcal{M}_0) > e_n(\mathcal{M}^c)$ for any adequate model $\mathcal{M}^c \in \mathbb{M}_0$ and a 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 2 of theorem 3.3.1 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 $\mathbb{M}_0$. □

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. In our setup, we designate this true parameter vector as $\boldsymbol{\theta}_0$, the preferred vector of parameters. 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 $\mathcal{S}$. 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, for example, 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 3.5.3. *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 3.5.2 we shall have*

$$\mathcal{S}_0 = \{j : e_{n_1}(\mathcal{S}_{-j}) < e_{n_1}(\mathcal{S}_*)\} \tag{3.5.6}$$

for any positive integer $n_1 > N$.

Proof of corollary 3.5.3. 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 3.5.1: leading to a contradiction and thus completing the proof. □

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 the following simple and generic steps:

1. Start from the full model and compute its e-value;
2. Take the marginal models by dropping each covariate, compute the corresponding e-values;
3. Collect covariates that cause a decrease in e-value compared to the full model.

A safer version of this recipe can be to keep on dropping covariates until no submodel achieves a lower e-value. In numeric studies we conducted, a sample of which is reported in Section Section 3.6, we did not find substantial difference between selecting covariates directly based on whether $e_n(\mathcal{S}_{-j}) < e_n(\mathcal{S}_*)$, and this stepwise-deletion method. Also in an empirical data-analytic setup, the performance of this 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.

3.5.3 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

$$\hat{e}_n(\mathcal{S}_{-j}) = \mathbb{E}_r D(\hat{\boldsymbol{\theta}}_{r,-j}, [\hat{\boldsymbol{\theta}}_{r_1}]) \tag{3.5.7}$$

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 $\hat{\mathcal{S}}_0 = \{j : \hat{e}_{n_1}(\mathcal{S}_{-j}) < \hat{e}_{n_1}(\mathcal{S}_*)\}$

To ensure that the sample e-values appropriately mimic the population level process, the bootstrap method used must adhere to the guidelines in section Section 3.4.

Example 3.5.4 (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 3.4.1) or the scale-enhanced Gamma bootstrap (example 3.4.2). 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 $\tau_{mn}^2 \equiv \tau_n^2$ is required as per theorem 3.4.4. 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 $m_n \rightarrow \infty, m_n/n \rightarrow 0$ as $n \rightarrow \infty$.

Example 3.5.5 (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) and obtain new sets of observations as

$$(\mathbf{y}_{ri}, \mathbf{X}_{rij}) : \quad \mathbf{y}_{rij} = \mathbf{X}_{ij}\hat{\boldsymbol{\beta}} + \tau_n w_{ri}(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}), \quad \mathbf{X}_{ri} = \mathbf{X}_i$$

for $i = 1, \ldots, m$, and $\hat{\boldsymbol{\beta}}$ the maximum likelihood estimate of $\boldsymbol{\beta}$. We take equal resampling weights $w_{ri} \sim N(0,1)$ inside a group, while τ_n satisfies similar conditions as last example. In this case, a simple relationship exists between $\hat{\boldsymbol{\beta}}$ and its bootstrap counterpart $\hat{\boldsymbol{\beta}}_r$:

$$\hat{\boldsymbol{\beta}}_r = \hat{\boldsymbol{\beta}} + \frac{\sqrt{n}}{\tau_n}(\mathbf{X}^T\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{W}_r\mathbf{X}^T\hat{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) + \mathbf{R}_{rn} \tag{3.5.8}$$

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 ((3.5.8)) above. Although a similar approach can certainly be used for GLMs as well, although computationally they are much more effective here. **is this okay?**

3.6 Simulation studies

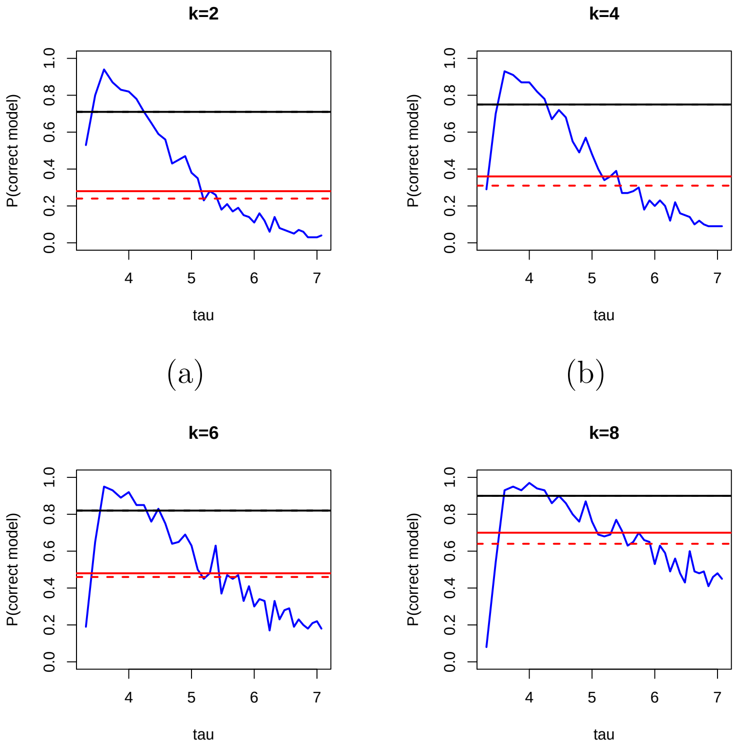
We now present the results of two simulation studies to compare the performance of our proposed fast variable selection method using model e-values, with the model selection procedures obtained from backward deletion and all subset regression versions that aim to minimize the Akaike Information Criterion (AIC: **?**) or the Bayesian Information Criterion (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 $\mathbb{E}Y$.

3.6.1 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 τ_n, the standard deviation of the resampling weights, 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 ((3.5.7)), 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".

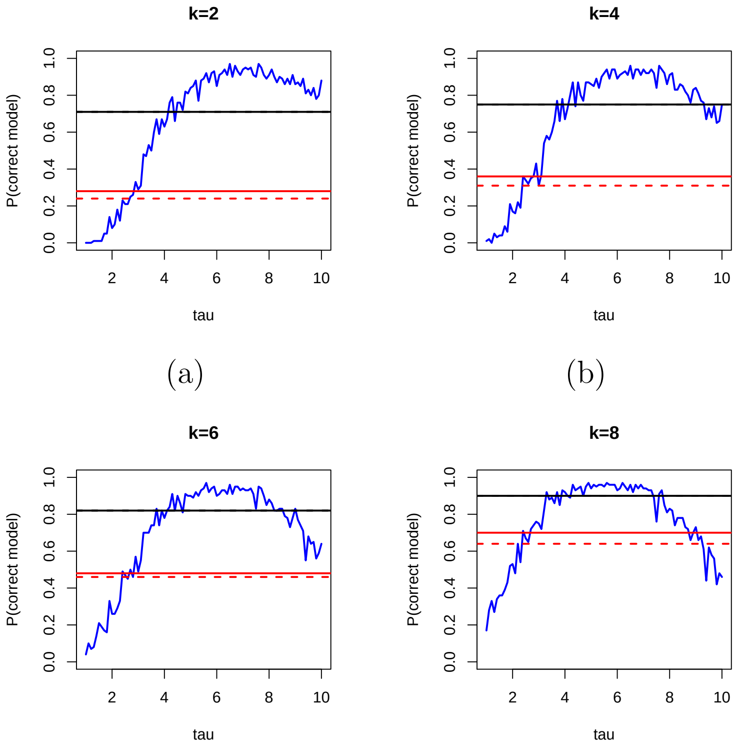
We use the backward deletion and all-subset regression search strategies while using AIC and BIC as the model selection criterion. We use the leaps-and-bound algorithm, implemented in the `R` package `leaps`, for all-subset search. We display the results of this study in Figure Figure 5.2 for the gamma bootstrap. 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. We



(a) (b)

(c) (d)

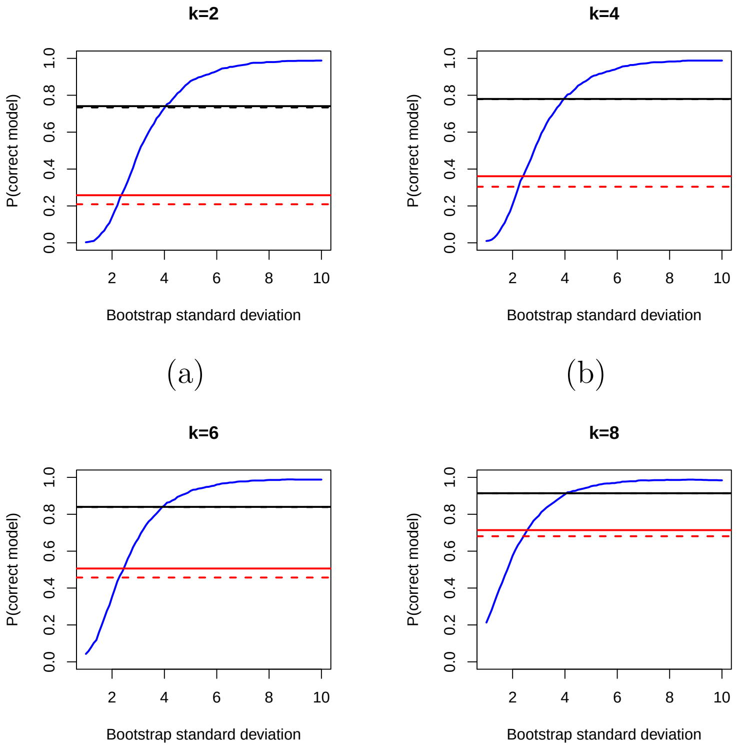
Figure 3.1: Empirical probabilities of selecting the correct model through moon bootstrap for several levels of sparsity: The e-values method- blue solid, AIC backward deletion- red dotted, AIC all subset- red solid, BIC backward deletion- black dotted, BIC all subset- black solid



(a) (b)

(c) (d)

Figure 3.2: 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



(a) (b)

(c) (d)

Figure 3.3: 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

experimented with other choices of n, p, R_1, R_2, and it seems considering $\tau_n \in (4, 8)$ 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.

3.6.2 Model selection in the presence of random effects

Here we use the repeated measures simulation setup from **?**, 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:

$$D = \begin{pmatrix} 9 & & & \\ 4.8 & 4 & & \\ 0.6 & 1 & 1 & \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{3.6.1}$$

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 $\tau_n \in \{1, \ldots, 8\}$ here. 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 **?**, the tuning parameter can be selected using several

Method	Tuning	FPR%	FNR%	Model size	FPR%	FNR%	Model size
		$n_i = 5, m = 30$			$n_i = 10, m = 60$		
e-value based	$\tau = 1$	60.1	0.0	5.35	56.7	0.0	4.96
	$\tau = 2$	30.8	0.0	3.21	29.4	0.0	3.09
	$\tau = 3$	11.1	0.0	2.37	9.6	0.0	2.32
	$\tau = 4$	2.4	0.0	2.14	1.8	0.0	2.01
	$\tau = 5$	1	0.0	2.03	0.0	0.0	2.00
	$\tau = 6$	0.2	0.0	2.01	0.0	0.0	2.00
	$\tau = 7$	0.0	0.0	2.00	0.0	0.0	2.00
	$\tau = 8$	0.0	0.0	2.00	0.0	0.0	2.00
?	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 3.1: Comparison between our method and that proposed by **?** through average false positive percentage, false negative percentage and model size

Method		Setting 1	Setting 2
e-value based	$\tau = 1$	1	1.5
	$\tau = 2$	29.5	29
	$\tau = 3$	70	73.5
	$\tau = 4$	93	94.5
	$\tau = 5$	97	100
	$\tau = 6$	99.5	100
	$\tau = 7$	100	100
	$\tau = 8$	100	100
?		73	83
?		49	86
?		90	100

Table 3.2: Comparison of our method and three sparsity-based methods of mixed effect model selection through accuracy of selecting correct fixed effects

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 Table 3.1. It can be seem the *e-value*-based method handsomely outperforms the method proposed by **?**, especially in smaller sample sizes, as long as $\tau_n \geqslant 4$.

We also compare the percentages of times the correct model was identified, and these results are presented in Table Table 5.1, along with the corresponding results from two other papers. The proposed e-value based procedure performs best here for $\tau_n \geqslant 4$ for the smaller sample setting, and for $\tau_n \geqslant 5$ for larger sample setting.

3.7 Discussion and conclusion

We present above an expansive framework and principle, where the definition of a statistical model is very broad, estimation procedures very general, resampling algorithms broad and 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, termed FPMS algorithm, obtains consistent true model selection with probability tending to one, essentially using a two step algorithm, the second of the steps involving parallel computations and both steps requiring resampling. Simulation results show that the FPMS performs better than traditional methods in two illustrative examples, and a case study on Indian summer precipitation identifies several important physical drivers of monsoon precipitation. Theoretical consistency results of multiple kinds are provided.

While the above framework is extremely open-ended, multiple details require cautious approach and more detailed studies. The choice of the resampling algorithm, the tuning parameter τ_n associated with it, should be subject to further scrutiny. Our results suggest excellent *asymptotic* properties and 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 more deep 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 study. A further, perhaps philosophic, issue 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.

Chapter 4

Applications of the Evaluation Maps Framework

4.1 Simultaneous Selection of Multiple Genetic and Environmental Factors in Twin Studies

Sketch:

- Describe the problem.
- How we apply *e*-values to select variables here.
- New innovation: Bonferroni correction on *e*-values.
- Simulation: setup and results
- Real data analysis: report some relevant genes.
- Describe the integration of gene selection and SNP selection: by applying *e*-values at the two levels.

4.2 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 **????**. For example, **?** discovered an upward trend in frequency and magnitude of extreme rain events, using daily central Indian rainfall data on a 10° × 12° grid, but a similar study on a 1° × 1° gridded data by **?** suggested that there are both increasing and decreasing trends of extreme rainfall events, depending on the location. Additionally, **?** 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. There are significant spatial and temporal variabilities at various scales discovered by **?** and **?**.

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.

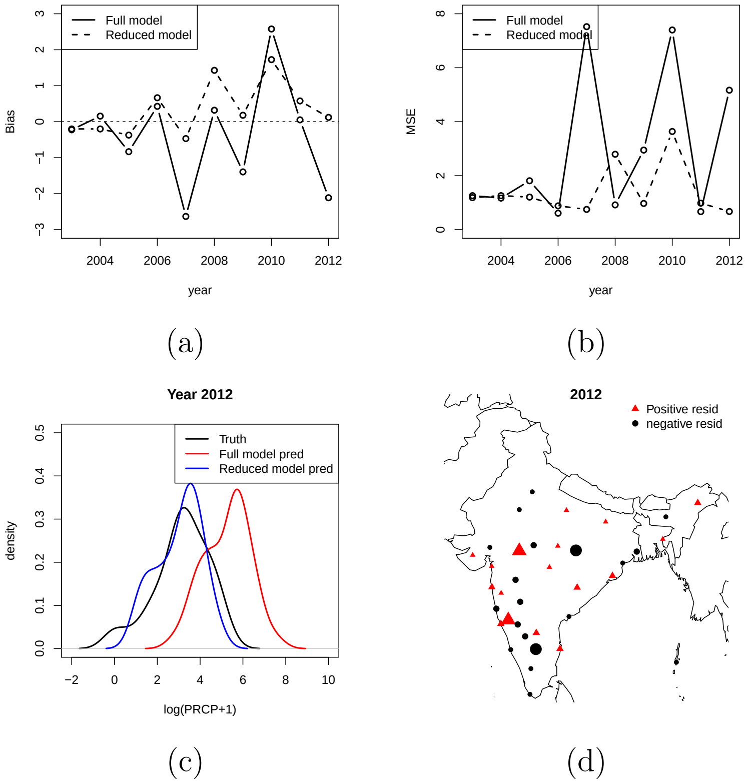
As a modeling step, we consider the annual medians of all the above covariates as fixed effects, the log yearly rainfall at a weather station as response variable, and include year-specific random intercepts. Table Table 4.1 lists the estimated $\hat{e}(\mathcal{S}_{-j})$ values in increasing order for the full model as well as all 35 models where a single variable is dropped. We use resample Monte Carlo sizes $R = R_1 = 1000$. The variables listed above *none* appears in Table Table 4.1 are considered relevant by our e-value criterion.

All the variables selected by our procedure have documented effects on Indian monsoon **??**. 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

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 4.1: 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



(a) (b)

(c) (d)

Figure 4.1: 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

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 Figure 4.1 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).

4.3 Spatio-temporal Dependence Analysis in fMRI data

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 **?**. 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 $t \in \{1, 2, ..., T\}$.

4.3.1 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 **?**.

In this analysis we consider $K = 5$ and $q = 2$, i.e. an AR(5) model with quadratic drift.

4.3.2 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 Figure 4.2 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, whose role in visual perception is well-documented **?**.

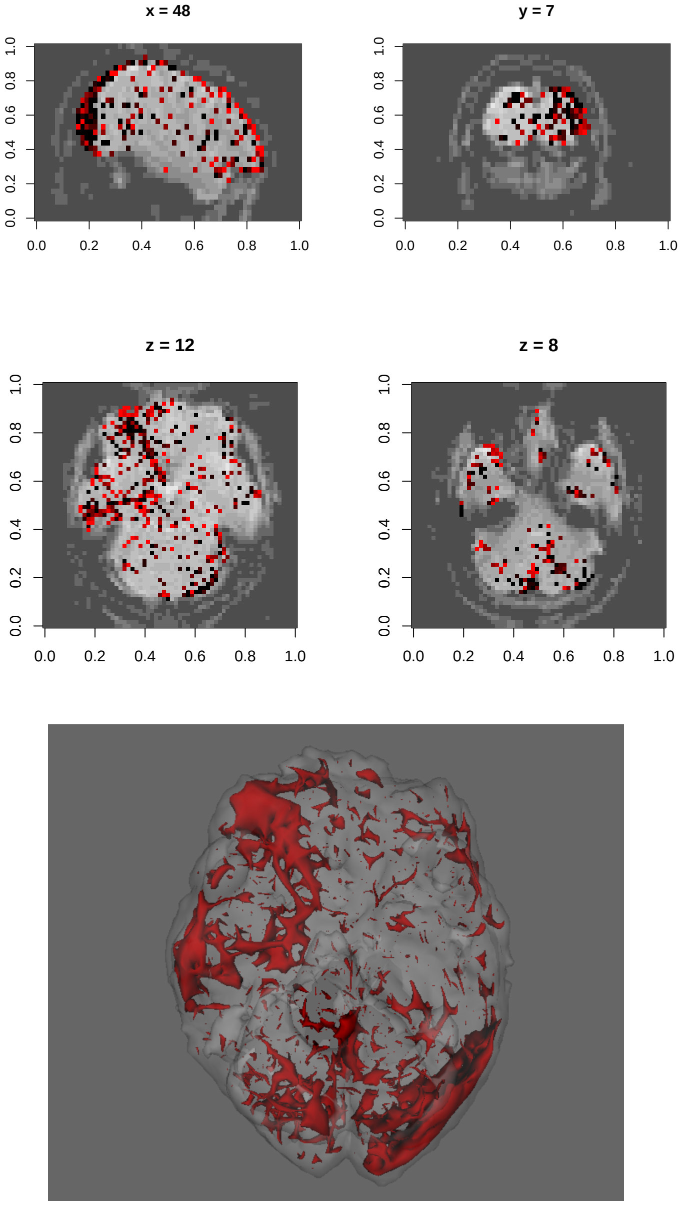


Figure 4.2: (Top) Plot of significant p-values at 95% confidence level at the specified cross-sections; (bottom) a smoothed surface obtained from the p-values clearly shows high spatial dependence in right optic nerve, auditory nerves, auditory cortex and left visual cortex areas

Chapter 5

Nonconvex Penalized Regression using Depth-based Penalty

5.1 Introduction

Consider the multitask linear regression model:

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E}$$

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

$$\min_{\mathbf{B}} \mathrm{Tr}\{(\mathbf{Y} - \mathbf{X}\mathbf{B})^T(\mathbf{Y} - \mathbf{X}\mathbf{B})\} + P_\lambda(\mathbf{B}). \tag{5.1.1}$$

The frequently studied classical linear model may be realized as a special 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. $p(z) = 1(z \neq 0)$, 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 $\bigcup_k S_k$, with $S_k := \{k : b_{jk} \neq 0, j = 1, 2, ..., p\}$. 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. The LARN algorithm addresses both of these issues.

Specifically, we perform support union recovery by considering penalties with row-wise decomposition: $P_\lambda(\mathbf{B}) = \sum_{j=1}^{p} p_\lambda(\|\mathbf{b}_j\|_2)$. In this paper, we shall concentrate on the scenario when $p_\lambda(\|\mathbf{b}_j\|_2)$ is a potentially nonconvex function of the row-norm. This automatically tempers the effects of large regression coefficients in the case of general q-dimensional response: this is not the case for methods based on l_1-norm penalization, e.g. Lasso. 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.

Our work provides a detailed treatment of using nonconvex penalties in the context of multivariate responses. We define the regularizing function in terms of *data depth* functions, which quantify the center-outward ranking of multivariate data (Zuo and Serfling, 2000). Inverse depth functions, or *peripherality functions* can be defined as some reverse ranking based on data depth, and we use such peripherality functions as regularizers in this paper. In section **??** we discuss data-depth and illustrate some instances of peripherality functions, followed by detailed presentation of the LARN algorithm in section Section 5.2. Additional theoretical results in the orthogonal design case are discussed in section Section 5.3, and some simulation experiments are presented to compare the LARN algorithm with other methods in section Section 5.4. We present a data application of the LARN algorithm in section Section 5.5, followed by conclusions. The appendix contains proofs of the theoretical results.

5.2 The LARN algorithm

5.2.1 Formulation

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:

$$\begin{aligned} \hat{\mathbf{B}} &= \arg\min_{\mathbf{B}} \Big[\mathrm{Tr}\{(\mathbf{Y} - \mathbf{XB})^T (\mathbf{Y} - \mathbf{XB})\} + \\ & \lambda_n \sum_{j=1}^{p} D^-(\mathbf{b}_j, F) \Big] \end{aligned} \tag{5.2.1}$$

where $D^-(\mathbf{x}, F)$ is an *inverse depth* function that measures the peripherality or outlyingness of the point $\mathbf{x}$ with respect to a fixed probability distribution F. Given a measure of data depth, any nonnegative-valued monotonically decreasing transformation on that depth function can be taken as a inverse depth function. Some examples include but are not limited to $D^-(\mathbf{x}, F) := \max_{\mathbf{x}} D(\mathbf{x}, F) - D(\mathbf{x}, F)$ and $D^-(\mathbf{x}, F) := \exp(-D(\mathbf{x}, F))$. This helps us obtain the nonconvex shape for our row-wise penalty function, where the penalty sharply increases for smaller entries inside the row but is bounded above for large values (see figure Figure 5.1 panel b).

We first focus on the support union recovery problem in ((5.2.1)), starting with the first-order Taylor series approximation of $D^{-}(\mathbf{b}_j, F)$. At this point, we make the following assumptions:

(A1) The function $D^{-}(\mathbf{b}, F)$ is concave in $\mathbf{b}$, and continuously differentiable at every $\mathbf{b} \neq \mathbf{0}_q$ with bounded derivatives;

(A2) The distribution F is spherically symmetric.

Since F is spherically symmetric, due to affine invariance of $D(., F)$ hence $D^{-}(., F)$, depth at a point $\mathbf{b}$ becomes a function of $r = \|\mathbf{b}\|_2$ only. 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, $D(\mathbf{b}_j, F) = c/(c + \|\mathbf{b}_j\|); c = \Phi^{-1}(3/4)$, while for halfspace depth and any F, $D(\mathbf{b}_j, F) = 1 - F_1(\|\mathbf{b}_j\|)$, F_1 being any univariate marginal of F. Hence, the computational burden of calculating depths becomes trivial.

Keeping the above in mind, we can write $D^{-}(\mathbf{b}_j, F) = p_F(r_j)$, and thus

$$\begin{aligned} P_{\lambda.F}(\mathbf{B}) &:= \lambda \sum_{j=1}^{p} p_F(r_j) \\ &\simeq \lambda \sum_{j=1}^{p} \left[p_F(r_j^*) + p_F'(r_j^*)(r_j - r_j^*) \right] \end{aligned} \tag{5.2.2}$$

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 $\mathbf{B}^{(0)} = \mathbf{B}^* = (\mathbf{X}^T\mathbf{X})^{-}\mathbf{X}^T\mathbf{Y}$, i.e. the least square estimate of $\mathbf{B}$, set $k = 0$;

2. Calculate the next iterate by solving the penalized likelihood:

$$\begin{aligned}\mathbf{B}^{(k+1)} &= \arg\min_{\mathbf{B}} \Big[\mathrm{Tr}\left\{ (\mathbf{Y} - \mathbf{X}\mathbf{B}^{(k)})^T (\mathbf{Y} - \mathbf{X}\mathbf{B}^{(k)}) \right\} + \\ &\quad \lambda \sum_{j=1}^{p} p'_F(r_j^{(k)}) r_j \Big] \end{aligned} \tag{5.2.3}$$

3. Continue until convergence.

We take the least square estimate as a starting value since it is within $O(1/\sqrt{n})$ distance of $\mathbf{B}$, and the upper bound on p'_F ensures that $P_{\lambda,F}(\mathbf{B}) = P_{\lambda,F}(\mathbf{B}|\mathbf{B}^*) + O(1/\sqrt{n})$ 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.

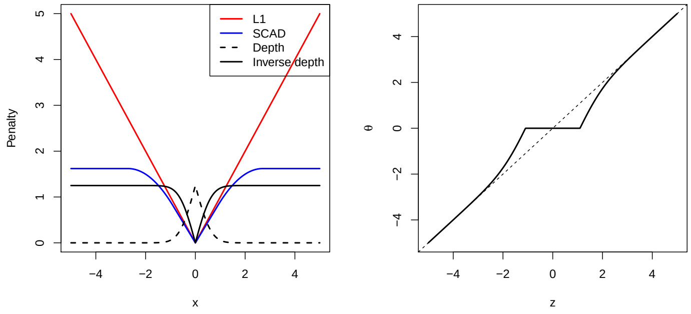
LARN is a majorize-minimize (MM) algorithm where the actual objective function $Q(\mathbf{B})$ is being majorized by $R(\mathbf{B}|\mathbf{B}^{(k)})$, with

$$\begin{aligned} Q(\mathbf{B}) &= \mathrm{Tr}\left\{ (\mathbf{Y} - \mathbf{X}\mathbf{B})^T (\mathbf{Y} - \mathbf{X}\mathbf{B}) \right\} + P_{\lambda,F}(\mathbf{B}) \\ R(\mathbf{B}|\mathbf{B}^{(k)}) &= \mathrm{Tr}\left\{ (\mathbf{Y} - \mathbf{X}\mathbf{B})^T (\mathbf{Y} - \mathbf{X}\mathbf{B}) \right\} + P_{\lambda,F}(\mathbf{B}|\mathbf{B}^{(k)}) \end{aligned}$$

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 $Q(\mathbf{B}) \leqslant R(\mathbf{B}|\mathbf{B}^{(k)})$. 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



(a) (b)

Figure 5.1: (a) Comparison of L1 and SCAD (Fan and Li, 2001) penalty functions with depth at a scalar point: 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 (see section Section 5.3)

from $\mathbf{B}^{(k)}$ to $\mathbf{B}^{(k+1)}$ defined in ((5.2.3)).

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 (**?**), 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.

5.2.2 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 support union $\bigcup_k S_k$ as well as the corresponding rows in $\mathbf{B}$. 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:

$$\arg\min_{\mathbf{B}} R(\mathbf{B}|\mathbf{B}^*) = \arg\min_{\mathbf{B}} \left[\text{Tr}\left\{ (\mathbf{Y} - \mathbf{XB})^T (\mathbf{Y} - \mathbf{XB}) \right\} + \lambda \sum_{j=1}^{p} p'_F(r_j^{(k)}) r_j \right] (5.2.4)$$

At this point, without loss of generality we assume that the true coefficient matrix $\mathbf{B}_0$ has the following decomposition: $\mathbf{B}_0 = (\mathbf{B}_{01}^T, \mathbf{0})^T, \mathbf{B}_1 \in \mathbb{R}^{p_1 \times q}$. Also denote the vectorized (i.e. stacked-column) version of a matrix $\mathbf{A}$ by vec($\mathbf{A}$). We are now in a position to to prove oracle properties of the one-step estimator in (**??**), 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 5.2.1. *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 $1 \leqslant j \leqslant q$ and some $s > 0$. 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 $\hat{\mathbf{B}}^{(1)} = (\hat{\mathbf{B}}_{01}^T, \hat{\mathbf{B}}_{00}^T)^T$ (with the component matrix having dimensions $p_1 \times q$ and $p - p_1 \times q$, respectively) as $n \to \infty$:*

- *vec($\hat{\mathbf{B}}_{00}$) $\to \mathbf{0}_{(p-p_1)q}$ in probability;*
- $\sqrt{n}(\mathit{vec}(\hat{\mathbf{B}}_{01}) - \mathit{vec}(\mathbf{B}_{01})) \rightsquigarrow \mathcal{N}_p(\mathbf{0}_{p_1 q}, \mathbf{\Sigma} \otimes \mathbf{C}_{11}^{-1})$

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 **?** is a stronger version of this. With respect to the general framework of nonconvex penalized M-estimation in **?**, our modified form of p_F arising from assumption (A2) satisfies parts (i)-(iv) of Assumption 1 therein, and the conditions of theorem 5.2.1 adhere to part (v). Also note that the above oracle results depend on the assumption (A2), 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 5.2.1, and plan to work on this in the extended version of the work.

5.2.3 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 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 5.2.2. *Suppose the conditions of theorem 5.2.1 hold, and additionally all non-zero components of* $\mathbf{B}$ *have the following lower bound:*

$$|b_{jk}| \geqslant \sqrt{\frac{16 \log(qs)}{C_{min} n}}; \quad j \in S, 1 \leqslant k \leqslant q$$

where $C_{\min} > 0$ is a lower bound for eigenvalues of $\mathbf{C}_1$. Then, for some constants $c, c_0 > 0$, the post-thresdolding estimator $\mathbf{T}(\hat{\mathbf{B}}^{(1)})$ defined by:

$$t_{jk} = \begin{cases} 0 & \textit{if } \hat{b}^{(1)}_{jk} \leqslant \sqrt{\frac{8\log(q|\hat{S}|)}{C_{min}n}} \\ \hat{b}^{(1)}_{jk} & \textit{otherwise} \end{cases} \quad ; \quad j \in \hat{S}, 1 \leqslant k \leqslant q$$

has the same set of non-zero supports within rows as $\mathbf{B}$ with probability greater than $1 - c_0 \exp(-cq \log s)$.

5.2.4 Computation

When the quantities $\mathbf{B}$ and $\mathbf{Y} - \mathbf{XB}$ are replaced with their corresponding vectorized versions, the optimization problem in (**??**) 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 5.2.3. *Given an initial value $\mathbf{B}^*$, a matrix $\mathbf{B} \in \mathbb{R}^{p \times q}$ is a solution to the optimization problem in (**??**) 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 $\mathbf{B}_0$:

- Start with the OLS estimate $\mathbf{B}^*$, set $m = 1$ and $\hat{\mathbf{B}}^{(1,0)} = \mathbf{B}_0$;

- 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 5.2.2 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 $\mathbf{B}_0$ is $\mathbf{B}^*$. We use k-fold cross-validation to choose the optimal λ. Also notice that in a sample setup the quantity $C_{\min}$ in lemma 5.2.2 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 $k = 2, ..., m$.

5.3 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.

5.3.1 Thresholding rule

For the univariate thresholding rule, we are dealing with the simplified penalty function $p(|b_{jk}|) = D^{-}(b_{jk}, F)$, 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 subsection Section 5.2.1, 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

$$L(\theta; p_\lambda) = \frac{1}{2}(z - \theta)^2 + p_\lambda(|\theta|) \tag{5.3.1}$$

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 ((5.3.1)):

$$\begin{aligned} \hat{\theta}(F, \lambda) &= \operatorname{sign}(z)\left[|z| - \lambda D_1^-(\theta, F)\right]_+ \\ &\simeq \operatorname{sign}(z)\left[|z| - \lambda D_1^-(z, F)\right]_+ \end{aligned} \tag{5.3.2}$$

the approximation in the second step being due to Antoniadis and Fan (2001). A plot of the thresholding function in panel c of figure 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, the MCP penalty (Zhang, 2010) corresponds to $D_1^-(\theta, F) = |\theta| I(|\theta| < \lambda)$, 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 $c = 1/(2\lambda^2(a+2))$.

5.3.2 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 ((5.3.2)) also, in fact, replicates this performance.

Theorem 5.3.1. *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*

$$\begin{aligned} R(\hat{\theta}(F, \lambda)) \leqslant & \ (2 \log n - 3) \\ & \left[R(ideal) + \frac{c_1}{p_0(F)(\sqrt{.5 \log n} - 1)} \right] \end{aligned} \tag{5.3.3}$$

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.

5.4 Simulation results

5.4.1 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 projection 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 Graphical 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;

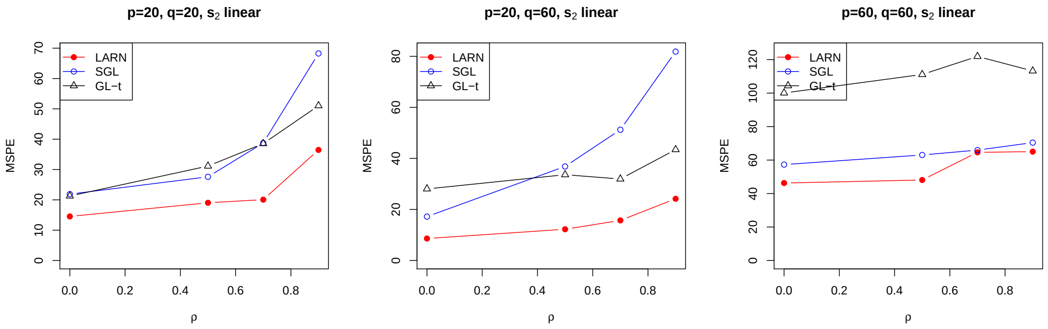
Group Lasso with thresholding (GL-t): This has been proposed by Obozinski et al. (2011), and performs element-wise thresholding on a row-level group lasso estimator to get final estimate of $\mathbf{B}$. It can also be realized as a special case of LARN, with weights of all row-norms set as 1.

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).



(a) (b) (c)

Figure 5.2: Mean squared testing errors for all three methods in different (p, q) settings

5.4.2 Evaluation

To summarize the performance of an estimate matrix $\hat{\mathbf{B}}$ we use the following three performance metrics:

- *Mean Squared Testing Error (MSTE)*- Defined as $(1/pq)\text{Tr}\left[(\mathbf{Y}_{test} - \mathbf{X}_{test})(\mathbf{Y}_{test} - \mathbf{X}_{test})^T\right]$, with $(\mathbf{Y}_{test}, \mathbf{X}_{test})$ generated from the same simulation setup 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 Table 5.1, and MSTE performances in figure 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.

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

All simulations were run in parallel on 8 threads of an Intel Core i7 3770 3.4 GHz processor-run machine with 8 GB of RAM. 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.

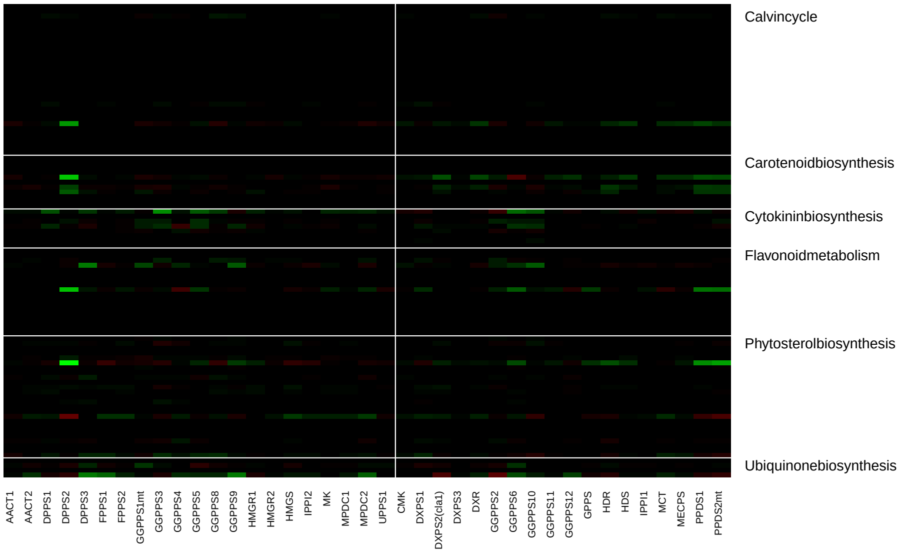


Figure 5.3: Estimated effects of different pathway genes on the activity of genes in Mevalonate and Non-mevalonate pathways (left and right of vertical line) in *A. thaliana*

5.5 Real data example

Table 5.3: Top 10 gene-pathway connections in *A. thaliana* data found by LARN

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

We apply the LARN algorithm on a microarray dataset containing expression of several genes in the flowering plant *Arabidopsis thaliana* (Wille et al, 2004). 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 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 **??**. We also visualize coefficients corresponding to genes in the 6 pathways in the table through a heatmap in figure 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.

5.6 Conclusion

In the above sections we propose general nonconvex penalty functions based on data depth 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 its general nature and the instant extension to multitask regression. For the multitask case, we further show that a simple post-estimation thresholding recovers non-zero elements within rows in the coefficient matrix with good accuracy. It shares the weakness of all nonconvex penalties: small signals may go undetected or can be estimated in a biased fashion. Future studies in this direction include extending the setup to include generalized linear models, as well as exploring the use of more efficient algorithms for calculating the sparse solutions.

5.7 Proofs

Proof of theorem 5.2.1. We shall prove a small lemma before going into the actual proof.

Lemma 5.7.1. *For matrices* $\mathbf{K} \in \mathbb{R}^{l \times k}, \mathbf{L} \in \mathbb{R}^{l \times m}, \mathbf{M} \in \mathbb{R}^{m \times k}$,

$$\mathrm{Tr}(\mathbf{K}^T \mathbf{L} \mathbf{M}) = vec^T(\mathbf{K})(\mathbf{I}_k \otimes \mathbf{L})\, vec(\mathbf{M})$$

Proof of lemma 5.7.1. From the property of Kronecker products, $(\mathbf{I}_k \otimes \mathbf{L})\,\mathrm{vec}(\mathbf{M}) = \mathrm{vec}(\mathbf{L}\mathbf{M})$. The lemma follows since $\mathrm{Tr}(\mathbf{K}^T \mathbf{L}\mathbf{M}) = \mathrm{vec}^T(\mathbf{K})\,\mathrm{vec}(\mathbf{L}\mathbf{M})$. □

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

$$\begin{aligned}
T_n(\mathbf{U}) &= \operatorname{Tr}\left[\left(\mathbf{Y}-\mathbf{X}\mathbf{B}_0-\frac{1}{\sqrt{n}}\mathbf{X}\mathbf{U}\right)^T\left(\mathbf{Y}-\mathbf{X}\mathbf{B}_0-\frac{1}{\sqrt{n}}\mathbf{X}\mathbf{U}\right)\right]\\
&\quad+\lambda_n\sum_{j=1}^{p}p'_F(r_j^*)\left\|\mathbf{b}_{0j}+\frac{\mathbf{u}_j}{\sqrt{n}}\right\|_2\\
\Rightarrow T_n(\mathbf{U})-T_n(\mathbf{0}_{p\times q}) &= \operatorname{Tr}\left[\frac{1}{n}\mathbf{U}^T\mathbf{X}^T\mathbf{X}\mathbf{U}-\frac{2}{\sqrt{n}}\mathbf{E}^T\mathbf{X}\mathbf{U}\right]\\
&\quad+\frac{\lambda_n}{\sqrt{n}}\sum_{j=1}^{p}p'_F(r_j^*)\left(\|\sqrt{n}\mathbf{b}_{0j}+\mathbf{u}_j\|_2-\|\sqrt{n}\mathbf{b}_{0j}\|_2\right)\\
&= \operatorname{Tr}(\mathbf{V}_1+\mathbf{V}_2)+V_3 \qquad (5.7.1)
\end{aligned}$$

Since $\mathbf{X}^T\mathbf{X}/n \to \mathbf{C}$ by assumption, we have $\operatorname{Tr}(\mathbf{V}_1) \to \operatorname{vec}^T(\mathbf{U})(\mathbf{I}_q\otimes\mathbf{C})\operatorname{vec}(\mathbf{U})$ using lemma 5.7.1. Using the lemma we also get

$$\operatorname{Tr}(\mathbf{V}_2)=\frac{2}{\sqrt{n}}\operatorname{vec}^T(\mathbf{E})(\mathbf{I}_q\otimes\mathbf{X})\operatorname{vec}(\mathbf{U})$$

Now $\operatorname{vec}(\mathbf{E})\sim\mathcal{N}_{nq}(\mathbf{0}_n,\mathbf{\Sigma}\otimes\mathbf{I}_q)$, so that $(\mathbf{I}_q\otimes\mathbf{X}^T)\operatorname{vec}(\mathbf{E})/\sqrt{n}\rightsquigarrow\mathbf{W}\equiv\mathcal{N}_{pq}(\mathbf{0}_{pq},\mathbf{\Sigma}\otimes\mathbf{C})$ 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 $D^-((r_j^*)^{-s})$ 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 ((5.7.1)) we see that

$$T_n(\mathbf{U}) - T_n(\mathbf{0}_{p\times q}) \rightsquigarrow \begin{cases} \text{vec}^T(\mathbf{U}_1)[(\mathbf{I}_q \otimes \mathbf{C}_{11})\,\text{vec}(\mathbf{U}_1) - 2\mathbf{W}_1] & \text{if } \mathbf{U}_0 = \mathbf{0}_{(p-p_1)q} \\ \infty & \text{otherwise} \end{cases} \tag{5.7.2}$$

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 vec$(\mathbf{B}_0)$. Applying epiconvergence results of **?** and **?** we now have

$$\text{vec}(\hat{\mathbf{U}}_{1n}) \rightsquigarrow (\mathbf{I}_q \otimes \mathbf{C}_{11}^{-1})\mathbf{W}_1 \tag{5.7.3}$$

$$\text{vec}(\hat{\mathbf{U}}_{0n}) \rightsquigarrow \mathbf{0}_{(p-p_1)q} \tag{5.7.4}$$

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 $\sqrt{n}(\text{vec}(\hat{\mathbf{B}}_{1n}) - \text{vec}(\hat{\mathbf{B}}_{1n})) = \hat{\mathbf{U}}_{1n}$ follows directly from ((5.7.3)). It is now sufficient to show that $P(\hat{\mathbf{b}}_j^{(1)} \neq \mathbf{0}_q | \mathbf{b}_{0j} = \mathbf{0}_q) \rightarrow 0$ to prove the oracle consistency part. For this notice that KKT conditions of the optimization problem for the one-step estimate indicate

$$2\mathbf{x}_j^T(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}^{(1)}) = -\lambda_n p'_F(r_j^*)\frac{\mathbf{b}_j^{(1)}}{r_j^{(1)}} \quad \Rightarrow \quad \frac{2\mathbf{x}_j^T(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}^{(1)})}{\sqrt{n}} = -\frac{\lambda_n p'_F(r_j^*)}{\sqrt{n}}.\frac{\mathbf{b}_j^{(1)}}{r_j^{(1)}} \tag{5.7.5}$$

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 $\lambda_n n^{(s-1)/2} \rightarrow \infty$, 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

$$P\left[\hat{\mathbf{b}}_j^{(1)} \neq \mathbf{0}_q | \mathbf{b}_{0j} = \mathbf{0}_q\right] \leqslant P\left[2\mathbf{x}_j^T(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}^{(1)}) = -\lambda_n p_F'(r_j^*)\frac{\mathbf{b}_j^{(1)}}{r_j^{(1)}}\right] \to 0$$

□

Proof of theorem 5.2.2. 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 5.3.1. 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 (**?**). 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,

$$\hat{\theta}(F,\lambda) = \begin{cases} 0 & \text{if } |z| \leqslant \lambda p_0(F) \\ z - \text{sign}(z).\lambda D_1^-(\hat{\theta}(F,\lambda), F) & \text{if } |z| > \lambda p_0(F) \end{cases} \tag{5.7.6}$$

Moreover, applying part 5 of the theorem,

$$\hat{\theta}(F,\lambda) = z - \operatorname{sign}(z).\lambda D_1^-(z,F) + o(D_1^-(z,F)) \tag{5.7.7}$$

for $|z| > \lambda p_0(F)$. Thus we get

$$[\hat{\theta}(F,\lambda) - z]^2 = \begin{cases} z^2 & \text{if } |z| \leqslant \lambda p_0(F) \\ \lambda^2 D_1^-(z,F)^2 + k_1(|z|) & \text{if } |z| > \lambda p_0(F) \end{cases} \tag{5.7.8}$$

and

$$\frac{d\hat{\theta}(F,\lambda)}{dz} = \begin{cases} 0 & \text{if } |z| \leqslant \lambda p_0(F) \\ 1 + \lambda D_2^-(z,F) + k_1'(|z|) & \text{if } |z| > \lambda p_0(F) \end{cases} \tag{5.7.9}$$

where $k_1(|z|) = o(|z|)$, and $D_2^-(z,F) = d^2 D^-(z,F)/dz^2$. Thus

$$\begin{aligned} E[\hat{\theta}(F,\lambda) - \theta]^2 &= E[z^2 1_{|z|\leqslant\lambda p_0(F)}] + E\left[\left(\lambda^2 D_1^-(|z|,F)^2 + 2\lambda D_2^-(|z|,F) + 2 + \right.\right. \\ & \quad \left.\left. k_1(|z|) + k_1'(|z|)\right) 1_{|z|>\lambda p_0(F)}\right] - 1 \end{aligned} \tag{5.7.10}$$

Now

$$\begin{aligned} k_1(|z|) &= \lambda^2 \left[D_1^-(z,F)^2 - D_1^-(\hat{\theta}(F,\lambda),F)^2\right] &\leqslant& \lambda^2 c_1^2, \text{ and} \\ |k_1'(|z|)| &= \lambda \left| D_2^-(z,F) - \frac{dD_1^-(\hat{\theta}(F,\lambda),F)}{dz} \right| &\leqslant& 2\lambda c_2 \end{aligned}$$

Substituting these in ((5.7.10)) above we get

$$\begin{aligned} E[\hat{\theta}(F,\lambda)-\theta]^2 &\leqslant \lambda^2 p_0(F)^2 P[|z| \leqslant \lambda p_0(F)] + E\left[\left(\lambda^2 f^2(|z|) + 2\lambda D_2^-(z,F)\right) 1_{|z|>\lambda p_0(F)}\right] \\ &\quad +\lambda^2 c_1^2 + 2\lambda c_2 + 1 \\ &\leqslant 2\lambda^2 c_1^2 + 4\lambda c_2 + 1 \\ &\leqslant 4\lambda^2 c_1^2 + 8\lambda c_2 + 1 \qquad (5.7.11) \end{aligned}$$

Adding and subtracting $z^2 1_{|z|>\lambda p_0(F)}$ to the first and second summands of ((5.7.10)) above, we also have

$$\begin{aligned} E[\hat{\theta}(F,\lambda)-\theta]^2 &= Ez^2 + E\left[\left(\lambda^2 D_1^-(z,F)^2 + 2\lambda D_2^-(z,F) + 2 - y^2 + \lambda^2 c_1^2 + 2\lambda c_2\right) 1_{|z|>\lambda p_0(F)}\right] - 1 \\ &\leqslant (2\lambda^2 c_1^2 + 4\lambda c_2) P[|z| > \lambda p_0(F)] + \theta^2 \qquad (5.7.12) \end{aligned}$$

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

$$\begin{aligned} E[\hat{\theta}(F,\lambda)-\theta]^2 &\leqslant 2(2\lambda^2 c_1^2 + 4\lambda c_2)[q(\lambda p_0(F)) + \theta^2] + \theta^2 \\ &\leqslant (4\lambda^2 c_1^2 + 8\lambda c_2 + 1)[q(\lambda p_0(F)) + \theta^2] \qquad (5.7.13) \end{aligned}$$

Combining this with ((5.7.11)) we get

$$E[\hat{\theta}(F,\lambda)-\theta]^2 \leqslant [4(\lambda c_1 + 1)^2 - 3][q(\lambda p_0(F)) + \min(\theta^2, 1)] \qquad (5.7.14)$$

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

Conclusion and future work

Donec gravida posuere arcu. Nulla facilisi. Phasellus imperdiet. Vestibulum at metus. Integer euismod. Nullam placerat rhoncus sapien. Ut euismod. Praesent libero. Morbi pellentesque libero sit amet ante. Maecenas tellus. Maecenas erat. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas.

6.1 More Work

Cras dictum. Maecenas ut turpis. In vitae erat ac orci dignissim eleifend. Nunc quis justo. Sed vel ipsum in purus tincidunt pharetra. Sed pulvinar, felis id consectetuer malesuada, enim nisl mattis elit, a facilisis tortor nibh quis leo. Sed augue lacus, pretium vitae, molestie eget, rhoncus quis, elit. Donec in augue. Fusce orci wisi, ornare id, mollis vel, lacinia vel, massa. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas.

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.

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.

Brown, B. (1983). Statistical Use of the Spatial Median. *J. Royal Statist. Soc. B*, 45:25–30.

Buhlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data. Methods, Theory and Applications.* Springer.

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.

Chernozhukov, V., Galichon, A., Hallin, M., and Henry, M. (2014). Monge-Kantorovich Depths, Quantiles, Ranks and Signs. `http://arxiv.org/abs/1412.8434`.

Cook, R. D., Li, B., and Chiaromonte, F. (2010). Envelope models for parsimonious and efficient multivariate linear regression. *Biometrika*, 20:927–1010.

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.

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.

Dümbgen, 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. Mult. Anal.*, 130:107–117.

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.

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.

Fang, K. T., .Kotz, S., and Ng, K. W. (1990a). *Symmetric multivariate and related distributions.* Monographs on Statistics and Applied Probability 36. Chapman and Hall Ltd., London, United Kingdom.

Fang, K. T., Kotz, S., and Ng, K. W. (1990b). *Symmetric multivariate and related distributions. Monographs on Statistics and Applied Probability*, volume 36. Chapman and Hall Ltd., London.

Frebort, I., Kowalska, M., Huska, T., Frebortova, J., and Galuszka, P. (2011). Evolution of cytokinin biosynthesis and degradation. *J. Exp. Bot.*, 62:2431–2452.

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 loca-tion 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.

Huber, P. J. (1981). *Robust Statistics*. Wiley series in probability and mathematical statistics. Wiley.

Konishi, S. and Kitagawa, G. (1996). Generalised information criteria in model selection. *Biometrika*, 83:875–890.

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.

Magyar, A. and Tyler, D. (2014). The asymptotic inadmissibility of the spatial sign covariance matrix for elliptically symmetric distributions. *Biometrika*, 101:673–688.

Miao, J. and Ben-Israel, A. (1992). On principal angles between subspaces in $\mathbb{R}^n$. *Lin. Algeb. Applic.*, 171:81–98.

Minsker, S. (2015). Geometric median and robust estimation in Banach spaces. *Bernoulli*, 21:2308–2335.

Möttönen, J. and Oja, H. (1995). Multivariate spatial sign and rank methods. *J. Nonparametric Stat.*

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. Mult. Anal.*, 87:328–355.

Puri, M. L. and Sen, P. K. (1971). *Nonparametric Methods in Multivariate Analysis*. Wiley, New York, NY.

Rothman, A. J., Levina, E., and Zhu, J. (2010). Sparse Multivariate Regression With Covariance Estimation. *J. Comp. Graph. Stat.*, 19:947–962.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, 6:461–464.

Serfling, R. (2006). Depth Functions in Nonparametric Multivariate Inference. In *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, volume 72, pages 1–16.

Shao, J. (1996). Bootstrap model selection. *J. Amer. Statist. Assoc.*, 91:655–665.

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. Mult. Anal.*, 98:1611–1629.

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).

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.

Wille et al, A. (2004). Sparse graphical Gaussian modeling of the isoprenoid gene network in *Arabidopsis thaliana*. *Genome Biol.*, 5:R92.

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.

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 Serfling, R. (2000). General notions of statistical depth functions. *Ann. Statist.*, 28-2:461–482.