Estimation and Conditional Inference in High-Dimensional Statistical Models

Arend L. Voorman

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

Doctor of Philosophy

University of Washington

2014

Reading Committee:

Ali Shojaie, Chair

Daniela Witten, Chair

Mathias Drton

Program Authorized to Offer Degree:
Biostatistics

University of Washington

Abstract

Estimation and Conditional Inference in High-Dimensional Statistical Models

Arend L. Voorman

Co-Chairs of the Supervisory Committee:
Dr. Ali Shojaie
Biostatistics

Dr. Daniela Witten
Biostatistics

In many areas of biology, recent advances in technology have facilitated the measurement of large numbers of features, while the number of observations in a data set may remain relatively modest. In this setting, lasso regression and related procedures have been extensively studied for prediction, while the problem of inference is relatively less studied. Most inference in high dimensions is based on simple marginal associations between variables. However, a richer characterization of the associations between variables can be obtained by examining conditional relationships, which account for the joint behavior of the variables. Inference on conditional relationships is more difficult, because it requires one to specify how features are related to one another, to estimate these relationships, and to characterize the uncertainty in the estimation procedure. In Chapters 2 and 3, we explore a few methods for testing hypotheses about conditional relationships in the high-dimensional setting. In Chapter 4, we note some strong distributional assumptions implicit in many treatments of high-dimensional graphical models, and propose a modification which treats this issue.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I would like to thank all of the people who have made this dissertation possible. My advisers Ali Shojaie and Daniela Witten have given me invaluable mentorship, space to explore ideas, and guidance in how to communicate them. I would also like to acknowledge my committee members Mathias Drton, Maryam Fazel, Benjamin Taskar, Elhanan Borenstein, and in particular Ken Rice, who has been influential throughout my research and coursework. I would also like to thank Thomas Lumley, Barbara McKnight, Bruce Psaty, and others I have worked with at Cardiovascular Health Research Unit.

NOTATION

Except where defined otherwise, in this dissertation vectors are denoted in lower case bold font, matrices in upper case bold font, while scalars are in lower case font. We use n to denote sample size, and d to denote the number of features, or the dimension. For instance, $\mathbf{X} \in \mathbb{R}^{n \times d}$ may denote a collection of d features measured on n individuals. We can write the features as $\mathbf{x}_j \in \mathbb{R}^n, j = 1, \ldots, d$, while x_{ij} may denote the value of feature j on the i^{th} individual.

Chapter 1

INTRODUCTION

In many areas of biology, recent advances in technology have facilitated the measurement of large numbers of features. However, the number of independent observations, or the sample size, in an analysis may be comparatively small. Examples of such technology include DNA sequencing, which measures millions of genetic variants, or gene expression arrays which measure the abundance of thousands of transcripts. These data are often termed 'high-dimensional', since they involve many more variables, or dimensions, than traditional data sets.

Using high-dimensional data for prediction has received much attention, especially in the machine learning community. On the other hand, methods for inference specific to the high-dimensional setting are relatively under-developed. A typical inferential analysis might model each feature separately, using classical techniques. Such analyses could be termed *marginal*, since they do not model the features in a multivariate framework. For instance, in genome-wide association studies, it is standard to calculate the correlation between individual genetic variants and a disease trait, and report those variants which produce p-values below a certain threshold [Pearson and Manolio, 2008]. However, correlation famously does not imply causation: a genetic variant may be correlated with disease, yet have no functional consequences. A richer characterization of association can be obtained through *conditioning* on, or adjusting for, other variables. For instance, under mild assumptions, a genetic variant which is correlated with disease after conditioning on all other genetic variants is actually causal of disease. Conditional associations cannot always be interpreted as causal, and are not always warranted in the high-dimensional setting. Nonetheless, the lack of tools

for conditional inference in high-dimensions is a conspicuous gap in the statistical literature.

Inference for conditional associations is more difficult than inference for marginal associations, since conditioning requires the analyst to (i) specify how features are related to the outcome via a statistical model, (ii) estimate these relationships, and (iii) quantify uncertainty in the estimation procedure. Specifying a correct model is difficult, even in the low-dimensional setting where we have few variables, and reasonable *a priori* understanding of how they relate to each other. This is even more challenging in high dimensions, where we may know little about the individual variables. Estimation is also difficult, since a model that accounts for all of the features necessarily involves many parameters, while the amount of information available to estimate these parameters (i.e. the sample size) is typically small. Estimation of high-dimensional parameters can often be improved by exploiting sparsity, or similar structure. However, it is notoriously challenging to express uncertainty in sparse parameter estimates [see e.g. Pötscher and Leeb, 2009]. Larry Wasserman characterizes the demand for methods that 'work' in high-dimensions, but do not rely too heavily on assumptions, as the key foundational issue in modern statistics [Wasserman, 2011].

The difficulty of conditioning in high dimensions is best exemplified in the setting of linear regression, where one seeks to model an outcome $\mathbf{y} \in \mathbb{R}^n$ using features $\{\mathbf{x}_j \in \mathbb{R}^n : j = 1, \ldots, d\}$. When d is much smaller than n, one can estimate how the $\mathbf{x}_j$'s are related to the outcome, in the conditional sense, and perform inference using classical tests, such as the Wald, likelihood ratio, or score test. Flexible modeling approaches, such as additive models [Hastie and Tibshirani, 1990], allow for non-parametric, and semi-parametric inference. Further, model-agnostic variance estimates permit valid inference even when the assumed model is violated [Huber, 1964, White, 1980]. However, when the dimension d is large, inference is more difficult. For each feature, we want to know whether it has an effect on the outcome that is distinct from the effects of the other features. When there are more features, it is increasingly

difficult to distinguish the effect of a single feature from the joint effect of all other features. When d is smaller than n, this difficulty is reflected in large standard errors for multiple linear regression parameter estimates, resulting in low power. When d is larger than n, linear regression estimates are undefined. Methods for variable selection, such as the lasso [Tibshirani, 1996], can be used to select which variables are conditionally related to the outcome [Zhao and Yu, 2006]. However, tools for formal inference, such as p-values or confidence intervals, are not available for the lasso or related procedures.

This dissertation is motivated by the unresolved challenges of estimation and conditional inference in high-dimensional statistics. In Chapter 2 we propose an inference procedure for high-dimensional linear models, termed the *penalized score test*. In order to use this method, we use penalized regression to account for the effects of all but a single feature, and then test the residuals from this penalized regression for correlation with the held-out feature. Interestingly, we show that this procedure is closely tied to high-dimensional variable selection techniques, such as the lasso, as well as classical mixed models, depending on the penalty used.

In Chapter 3, we extend the penalized score test framework to the context of M-estimation. This includes the important special cases of generalized linear models and graphical models, which we investigate in further detail. This extension also gives model-agnostic standard errors, which give valid inference even when the specified model is not correct.

In Chapter 4, we investigate the distributional assumptions implied by linear models in the graphical setting. In order to relax these assumptions, we propose a flexible, non-parametric alternative called 'Sparse Conditional Independence Graph Estimation with Joint Additive Models', or SpaCEJAM.

We end with a discussion in Chapter 5.

The methodologies described in this dissertation have been implemented as easy-to-use software, which makes them available to researchers. The R package `lassoscore`,

available at `http://cran.r-project.org/package=lassoscore`, implements the penalized score test in Chapters 2 and 3. The R package `spacejam`, available at `http://cran.r-project.org/package=spacejam`, implements the methods described in Chapter 4.

Chapter 2

INFERENCE IN HIGH DIMENSIONS WITH THE PENALIZED SCORE TEST

2.1 Introduction

Suppose we are interested in the association between an outcome variable $\mathbf{y} \in \mathbb{R}^n$ and a set of predictors $\mathbf{x}_j \in \mathbb{R}^n$, $j = 1, \ldots, d$. In order to assess this we might consider the model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2.1}$$

where $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_d] \in \mathbb{R}^{n \times d}$, $\boldsymbol{\beta} \in \mathbb{R}^d$ is vector of coefficients, and $\boldsymbol{\epsilon} \in \mathbb{R}^n$ is a vector of errors with mean zero and constant variance. If the number of variables d is much smaller than n, we could perform a formal statistical test for whether an element of $\boldsymbol{\beta}$ is zero using classical methods, such as the score, likelihood ratio, or Wald test. However, in the high-dimensional setting, when the number of variables d is large, these tests have low power, or are undefined.

In the case where d is large, penalized regression techniques, such as the lasso [Tibshirani, 1996], which takes the form

$$\hat{\boldsymbol{\beta}}_\lambda = \underset{\mathbf{b} \in \mathbb{R}^d}{\arg\min} \left\{ \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\mathbf{b}\|_2^2 + \lambda \|\mathbf{b}\|_1 \right\}, \tag{2.2}$$

can be used to provide a sparse estimate of $\boldsymbol{\beta}$. Under suitable conditions, explored by Zhao and Yu [2006], the sparsity pattern of the lasso coefficient vector $\hat{\boldsymbol{\beta}}_\lambda$ correctly identifies which elements of $\boldsymbol{\beta}$ are zero. However, the lasso and related procedures do not provide p-values or confidence intervals, and thus devising formal hypothesis tests for the parameter $\boldsymbol{\beta}$ remains an open problem.

In recent years there has been some work on formal inference in the high-dimensional setting: the bootstrap has been used to estimate the sampling distribution of lasso coefficients [Tibshirani, 1996, Bach, 2008, Chatterjee and Lahiri, 2011], Meinshausen and Bühlmann [2010] proposed *stability selection*, a sub-sampling technique which can control familywise error rates, and Fan and Li [2001] provided sandwich formulas for penalized coefficients. However, bootstrapping and sub-sampling are computationally expensive and their finite sample properties may not be desirable, especially for methods that involve thresholding, like the lasso. Available variance formulas also suffer shortcomings: they typically neglect the fact that the tuning parameter tends to be selected based on the data, and are often only available for the non-zero coefficients.

Recently, Lockhart et al. [2013] proposed the *covariance test*, which produces a sequence of p-values as λ decreases and features become non-zero in the lasso regression (2.2). Suppose the k^{th} feature to become non-zero does so when the tuning parameter in (2.2) is $\lambda_k - \eta$, for some arbitrarily small constant $\eta > 0$. Here, λ_k is the k^{th} knot in the lasso solution path. The covariance test statistic produces a p-value for each λ_k, which tests the null hypothesis that $\text{supp}(\hat{\boldsymbol{\beta}}_{\lambda_k}) \supseteq \text{supp}(\boldsymbol{\beta})$. Thus, the covariance test can be used to test whether all relevant variables are non-zero in the lasso coefficient vector. However, this does not give confidence intervals or p-values for any individual variable's coefficient.

Even more recently, Taylor et al. [2014] and Lee et al. [2013b] extended the covariance testing framework to test hypotheses about individual features, after conditioning on a model selected by the lasso. However, their framework permits inference only about features which have non-zero coefficients in a lasso regression; this set of features will vary across samples, making interpretation difficult. In contrast, our framework can be applied to all features in a data set, and can be used to understand why coefficients in a lasso regression are non-zero in the first place.

Alternatively, Zhang and Zhang [2011], van de Geer et al. [2013] and Javanmard

and Montanari [2013] proposed the *low-dimensional projection estimator* (LDPE) for inference in high dimensions based on inverting the stationary conditions for lasso regression. To do this, LDPE uses lasso regression among the covariates to estimate the inverse of $\mathbf{X}^T\mathbf{X}$. Under suitable assumptions, LDPE is asymptotically optimal, in that the variance of the estimator achieves the Gauss-Markov lower bound. However, unlike Lockhart et al. [2013], who give p-values associated with the knots in the lasso solution path, this method uses the lasso as a starting point for a different estimator. Decisions made using their confidence intervals need not correspond to variable selection using the lasso.

We applied the covariance test and LDPE to a diabetes data set, previously studied by Efron et al. [2004], and which we analyze in greater detail in Section 2.3.3. The data consist of a measure of diabetes disease progression for 442 patients, along with 10 variables. Table 2.1 lists the variables introduced or removed at each knot λ_k in the lasso solution path, along with their associated p-values from the covariance test and LDPE. For the sake of comparison, we also include the p-values produced by multiple linear regression with all variables, the p-values produced by simple linear regression of the outcome on each feature separately, and the p-values from our proposed lasso-penalized score test, described in Section 4.4. LDPE requires specification of a tuning parameter, which we chose using 10-fold cross-validation.

Using the covariance test to select λ, we might decrease λ and stop when some p-value greater than 0.05 is observed. Using this rule, we would stop upon reaching a p-value of 0.86 at knot 6, and report a model with 5 covariates. However, knot 7 produces a p-value of 0.04, which suggests that all variables were not in the model after knot 5. Should we use the model with 7 or 5 variables? In the model with 7 variables, how should we interpret the presence of glucose, which produced a p-value of 0.86? Further, using any reasonable stopping rule, we would include HDL in our model. However, HDL yields a p-value of 0.63 in multiple linear regression, suggesting there is little evidence for its association after accounting for trends in

Knot	Predictor	covTest	LDPE	Pen. score test	Multiple lin. reg.	Simple lin. reg.
1	BMI	3.7×10^{-9}	2.3×10^{-15}	5.1×10^{-22}	4.3×10^{-14}	3.4×10^{-42}
2	LTG	1.9×10^{-22}	2.9×10^{-10}	2.6×10^{-18}	1.6×10^{-5}	8.8×10^{-39}
3	MAP	0.005	1.4×10^{-6}	2.6×10^{-8}	9×10^{-8}	1.6×10^{-22}
4	HDL	0.003	0.93	3.6×10^{-15}	0.63	6.1×10^{-18}
5	Sex	0.008	1.7×10^{-4}	0.002	0.36	0.0012
6	Glu	0.86	0.31	0.057	0.079	7.6×10^{-17}
7	TC	0.04	0.07	0.14	0.058	6.9×10^{-6}
8	TCH	0.54	0.41	0.17	0.27	2.3×10^{-21}
9	LDL	0.87	0.30	0.21	0.16	2.3×10^{-4}
10	Age	0.98	0.87	0.87	0.87	7.1×10^{-5}
11	-HDL	-	-	-	-	
12	HDL	0.80	0.93	3.6×10^{-15}	0.63	6.1×10^{-18}

Table 2.1: The diabetes data set. Variables are ordered according to when their coefficients become non-zero in the lasso solution (2.2), as $\lambda \to 0$. 'covTest' refers to the method of Lockhart et al. [2013] and p-values for this method were produced by the covTest R package. 'LDPE' refers to the method of Zhang and Zhang [2011] and van de Geer et al. [2013], for which code was provided by the authors. Mutiple and simple linear regression refer to those p-values produced by the Wald test. 'Pen. score test' refers to the lasso-penalized score test, described in Section 4.4, with $\lambda = 4$. At the 11^{th} knot, HDL leaves the lasso solution, and no p-value is available for covTest. Of the p-values presented in this table, only those from covTest are ordered. For ease of display, here we present p-values for all methods in the ordering given by the covTest p-values.

all other features. How should we interpret this discrepancy? The answers to these questions are not clear.

Using LDPE, the results are broadly similar to those from multiple linear regression. Here, the sample size is sufficient so that the inverse of $\mathbf{X}^T\mathbf{X}$ can be accurately estimated. However, in higher dimensions, this may not be the case.

In this paper, we propose the *penalized score test*, which can be interpreted as a compromise between multiple linear regression on all features, and simple linear regression on each feature separately. We show that the sparsity pattern of the lasso results from a decision based on this test. Unlike the covariance test statistic, it gives p-values for the association of each individual feature with the outcome, and unlike LDPE, the resulting p-values are directly related to variable selection using the lasso.

The rest of the paper is organized as follows. In Section 4.4 we describe our proposed method, the penalized score test, for general penalties. In Section 2.3 we consider the special case of the lasso-penalized score test, and explore its asymptotic distribution, its relationship to consistent variable selection (Section 2.3.1), and the behavior of the test on simulated and real data (Sections 4.5.1, 2.3.3 and 2.3.4). In Section 2.4 we consider the special case of the ridge-penalized score test. In Section 2.5 we propose extensions to other sparsity-inducing penalties. We end with a discussion in Section 2.6.

2.2 The penalized score test

Throughout the paper, we assume that $\mathbf{y}^T\mathbf{1}_n = 0$, $\mathbf{x}_j^T\mathbf{x}_j = n$ and $\mathbf{x}_j^T\mathbf{1}_n = 0$ for $j = 1, \ldots, d$. Vectors are denoted in lowercase bold font, while matrices are in uppercase bold font. In order to simplify the notation, we will consider a single variable of interest $\mathbf{x} = \mathbf{x}_j$, and denote $\mathbf{Z} = [\,\mathbf{x}_k, k \neq j\,] \in \mathbb{R}^{n\times(d-1)}$ as the matrix containing all other features. Note that any procedure applied to $\mathbf{x}_j$ can be applied to all other

variables in turn. With this notation, we re-write the model (2.1) as

$$\mathbf{y} = \alpha\mathbf{x} + \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2.3}$$

where $\alpha \in \mathbb{R}$ and $\boldsymbol{\beta} \in \mathbb{R}^{d-1}$.

Our goal is to test $H_0 : \alpha = 0$. One way to do this is using the score test, based on the derivative of the log-likelihood of the model, also known as the score, evaluated under the null hypothesis. Using the model (2.3), and assuming normality of the errors, the (scaled) score statistic is

$$T = \mathbf{x}^T(\mathbf{y} - \mathbf{y}^0)/\sqrt{n}, \tag{2.4}$$

where $\mathbf{y}^0 = \mathbf{Z}\boldsymbol{\beta}$. We reject H_0 when $|T|$ is large, with respect to an appropriate reference distribution.

Typically, in applying the score test, the parameters are estimated under the constraints imposed by the null hypothesis. In the setting of (2.3), this corresponds to estimating $\boldsymbol{\beta}$ using multiple linear regression of $\mathbf{y}$ on $\mathbf{Z}$. However, when d is an appreciable fraction of n, multiple linear regression of $\mathbf{y}$ on $\mathbf{Z}$ yields highly variable coefficient estimates, resulting in low power, and when $d > n$ multiple linear regression of $\mathbf{y}$ on $\mathbf{Z}$ is undefined. As an alternative, we could use a small subset of the other features in estimating $\boldsymbol{\beta}$, with e.g. step-wise regression. However, selecting an appropriate model is challenging, and inference after model selection is notoriously difficult [Berk et al., 2013, Leeb and Pötscher, 2005].

Instead, we propose to estimate $\boldsymbol{\beta}$ in (2.3) using penalized regression. The proposed approach is as follows. We first calculate $\hat{\mathbf{b}}^0_\lambda$, which serves as an estimate of $\boldsymbol{\beta}$, using the penalized regression

$$\hat{\mathbf{b}}^0_\lambda = \arg\min_{\mathbf{b}} \left\{ \|\mathbf{y} - \mathbf{Z}\mathbf{b}\|_2^2/(2n) + \lambda J(\mathbf{b}) \right\}, \tag{2.5}$$

where $J(\mathbf{b})$ is a penalty function, such as the lasso ($J(\mathbf{b}) = \|\mathbf{b}\|_1$), ridge ($J(\mathbf{b}) = \|\mathbf{b}\|_2^2/2$), or subset selection ($J(\mathbf{b}) = \|\boldsymbol{\beta}\|_0$), and λ is a non-negative tuning parameter. We then set $\hat{\mathbf{y}}_\lambda^0 = \mathbf{Z}\hat{\mathbf{b}}_\lambda^0$, and form the test statistic

$$T_\lambda = \mathbf{x}^T(\mathbf{y} - \hat{\mathbf{y}}_\lambda^0)/\sqrt{n}, \tag{2.6}$$

a measure of association between $\mathbf{x}$ and $\mathbf{y} - \hat{\mathbf{y}}_\lambda^0$. We declare T_λ to be statistically significant, for a null hypothesis to be discussed in Section 2.2.1, when $|T_\lambda|$ is large, based on an appropriate reference distribution. Since T_λ looks superficially like the score statistic from linear regression (3.5), we refer to this procedure as the *penalized score test*.

Interestingly, if we choose $J(\mathbf{b}) = \|\mathbf{b}\|_1$, and declare T_λ to be significant when $|T_\lambda| > \sqrt{n}\lambda$, then T_λ is significant precisely when $\mathbf{x}$'s coefficient is non-zero in a lasso-penalized regression of $\mathbf{y}$ on $\mathbf{x}$ and $\mathbf{Z}$ together. That is, the sparsity pattern of the lasso solution for a given λ is the result of a decision based on the proposed test. We make this assertion precise in Proposition 1, the proof of which follows immediately from the Karush-Kuhn-Tucker conditions for lasso regression.

Proposition 1. *Let* $(\hat{a}_\lambda, \hat{\mathbf{b}}_\lambda)$ *be the lasso solution*

$$(\hat{a}_\lambda, \hat{\mathbf{b}}_\lambda) = \underset{(a,\mathbf{b})\in\mathbb{R}^d}{\arg\min} \left\{ \|\mathbf{y} - a\mathbf{x} - \mathbf{Z}\mathbf{b}\|_2^2/(2n) + \lambda\|(a,\mathbf{b})\|_1 \right\}, \tag{2.7}$$

and let T_λ *be as in* (2.5) *and* (2.6) *with* $J(\mathbf{b}) = \|\mathbf{b}\|_1$. *Then,* $\hat{a}_\lambda \neq 0$ *if and only if* $|T_\lambda| > \sqrt{n}\lambda$.

Further, if we choose $J(\mathbf{b}) = \|\mathbf{b}\|_2^2/2$, corresponding to ridge regression, then T_λ is precisely the score statistic from a mixed effects model, where the effects of $\mathbf{Z}$ are assumed to be random. We make this assertion precise in Proposition 2. This connection is further explored in Section 2.4.

Proposition 2. *Suppose that*

$$\begin{aligned} \mathbf{y} \mid \boldsymbol{\beta} &= \alpha\mathbf{x} + \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ \boldsymbol{\beta} &\sim N_{d-1}\left(0, \sigma_\epsilon^2 (n\lambda)^{-1}\mathbf{I}_{d-1}\right) \\ \boldsymbol{\epsilon} &\sim N_n(0, \sigma_\epsilon^2 \mathbf{I}_n), \end{aligned} \tag{2.8}$$

and let $l(\alpha)$ be the log-likelihood of $\mathbf{y}$, with score $\dot{l}(\alpha) = \frac{\partial}{\partial\alpha} l(\alpha)$. Let T_λ be as in (2.5) *and* (2.6) *with $J(\mathbf{b}) = \|\mathbf{b}\|_2^2/2$. Then $T_\lambda/\sigma_\epsilon^2 = \dot{l}(0)/\sqrt{n}$.*

2.2.1 What hypothesis is being tested?

When using penalized regression to estimate $\boldsymbol{\beta}$, some systematic bias is incurred: our estimate $\hat{\mathbf{b}}_\lambda^0$ is shrunken towards zero, relative to the unbiased multiple linear regression estimate. In this section, we describe how this bias affects inference, for a given tuning parameter λ. We will see that the hypothesis being tested using the penalized score test (2.6) depends on the tuning parameter λ.

For the moment, consider the case where $\sqrt{n}\lambda \to 0$ as $n \to \infty$ and the dimension d is fixed. As shown by Knight and Fu [2000], for bridge penalties ($J(\mathbf{b}) = \sum_j |b_j|^\gamma$ where $\gamma > 0$) we have that $\hat{\mathbf{b}}_\lambda^0 \to_p \hat{\mathbf{b}}_0^0 \equiv (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{y}$, where we recognize $\hat{\mathbf{b}}_0^0$ as the multiple linear regression estimate of $\boldsymbol{\beta}$ under $H_0 : \alpha = 0$. Thus, in this asymptotic setting, T_λ has the same limiting distribution as the classical score statistic, and one can interpret T_λ as a test of $H_0 : \alpha = 0$ vs. $H_1 : \alpha \neq 0$. However, this asymptotic treatment neglects the fact that in any finite sample, T_λ depends on the tuning parameter λ. For this reason, throughout this paper we predominantly consider an asymptotic scenario with λ fixed, a scenario also considered by Yu and Ruppert [2002] in the context of penalized spline estimation.

In the fixed-λ regime, we define the population-level parameters

$$(a_\lambda, \mathbf{b}_\lambda) = \underset{a\in\mathbb{R},\mathbf{b}\in\mathbb{R}^{d-1}}{\arg\min} \left\{ \frac{1}{2n}\mathbb{E}\|\mathbf{y} - a\mathbf{x} - \mathbf{Z}\mathbf{b}\|_2^2 + \lambda J(\mathbf{b}) \right\}. \tag{2.9}$$

The stationary conditions of (2.9) imply that $a_\lambda = \mathbb{E}[\mathbf{x}^T(\mathbf{y} - \mathbf{Z}\mathbf{b}_\lambda)/n]$, which is a measure of linear association between $\mathbf{x}$ and $\mathbf{y} - \mathbf{Z}\mathbf{b}_\lambda$. That is, a_λ is a measure of correlation between the feature of interest $\mathbf{x}$, and the outcome $\mathbf{y}$ with the penalized effects of $\mathbf{Z}$ removed. Note that when $a_\lambda = 0$, we have that $\mathbb{E}[\,\mathbf{x}^T(\mathbf{y} - \mathbf{Z}\mathbf{b}_\lambda)/\sqrt{n}\,] = 0$, where we recognize $\mathbf{x}^T(\mathbf{y} - \mathbf{Z}\mathbf{b}_\lambda)/\sqrt{n}$ as the penalized score statistic with $\hat{\mathbf{b}}_\lambda^0$ replaced by $\mathbf{b}_\lambda$. Provided $\hat{\mathbf{b}}_\lambda^0$ converges quickly enough to $\mathbf{b}_\lambda$, then T_λ is centered around zero, asymptotically, when $a_\lambda = 0$. Thus, the statistic T_λ tests

$$H_{0,\lambda} : a_\lambda = 0 \quad \text{vs.} \quad H_{1,\lambda} : a_\lambda \neq 0.$$

We can write the parameter a_λ more simply as

$$a_\lambda = \alpha + \boldsymbol{\sigma}_{xz}^T(\boldsymbol{\beta} - \mathbf{b}_\lambda), \tag{2.10}$$

where $\boldsymbol{\sigma}_{xz} = \mathbf{Z}^T\mathbf{x}/n$. Recall that in multiple linear regression of $\mathbf{y}$ on $(\mathbf{x}, \mathbf{Z})$, the parameter associated with $\mathbf{x}$ is α, while in simple linear regression of $\mathbf{y}$ on $\mathbf{x}$ alone, the coefficient associated with $\mathbf{x}$ is $\alpha + \boldsymbol{\sigma}_{xz}^T\boldsymbol{\beta}$. Now, when $\lambda = 0$, then $a_\lambda = \alpha$. On the other hand, when λ is large, $\mathbf{b}_\lambda$ tends to zero and a_λ tends to $\alpha + \boldsymbol{\sigma}_{xz}^T\boldsymbol{\beta}$, provided $J(\mathbf{b}) > J(0)$ when $\mathbf{b} \neq \mathbf{0}$. For moderate values of λ, a_λ is thus a compromise between the multiple and simple linear regression parameters associated with $\mathbf{x}$.

To make this interpretation of the parameter a_λ more concrete, consider the case of $J(\mathbf{b}) = \|\mathbf{b}\|_0$, which corresponds to subset selection. This setting was studied by Berk et al. [2013], who discuss how the parameter associated with a feature $\mathbf{x}$ depends on which subset of the features $\mathbf{Z}$ are included in a regression model. Our framework, in which the parameter associated with $\mathbf{x}$ depends on the extent to which the effects

of $\mathbf{Z}$ are penalized, generalizes this concept.

In this fixed-λ regime, the distribution of T_λ depends on the choice of the penalty $J(\mathbf{b})$ and the value of λ in (2.5). In Section 2.3 we give asymptotic theory for the distribution of T_λ for the special case of lasso regression, $J(\mathbf{b}) = \|\mathbf{b}\|_1$. The distribution of the ridge-penalized score statistic, $J(\mathbf{b}) = \|\mathbf{b}\|_2^2/2$, comes from mixed-model theory, and is given in Section 2.4. We briefly discuss other penalty choices in Section 2.5.

2.3 The lasso-penalized score test

In this section, we examine in greater detail the penalized score test when the lasso is chosen as the penalty. That is, we first obtain the penalized coefficient vector

$$\hat{\mathbf{b}}_\lambda^0 = \underset{\mathbf{b}\in\mathbb{R}^{d-1}}{\arg\min} \left\{ \|\mathbf{y} - \mathbf{Z}\mathbf{b}\|_2^2/(2n) + \lambda\|\mathbf{b}\|_1 \right\}, \tag{2.11}$$

and then form the test statistic $T_\lambda = (\mathbf{y} - \mathbf{Z}\hat{\mathbf{b}}_\lambda^0)/\sqrt{n}$. Here we state Proposition 3, proven in the Appendix, which gives the asymptotic distribution of T_λ.

First, we require some notation. Let $\mathcal{A} = \text{supp}(\mathbf{b}_\lambda)$, where $|\mathcal{A}| = q$, and assume, without loss of generality, that $\mathbf{b}_\lambda$ is ordered so that $\mathbf{b}_\lambda = (b_{\lambda,1}, \ldots, b_{\lambda,q}, 0, \ldots, 0)^T$, and partition $\mathbf{Z}$ as $\mathbf{Z} = [\mathbf{Z}_\mathcal{A}, \mathbf{Z}_{\mathcal{A}^c}]$. Denote $\mathbf{P}_\mathcal{A} = \mathbf{Z}_\mathcal{A}(\mathbf{Z}_\mathcal{A}^T\mathbf{Z}_\mathcal{A})^{-1}\mathbf{Z}_\mathcal{A}^T$ as the projection onto the columns of $\mathbf{Z}_\mathcal{A}$, and let $\mathbf{\Sigma}_\mathcal{A} = \mathbf{Z}_\mathcal{A}^T\mathbf{Z}_\mathcal{A}/n$.

Note that the stationary conditions of (2.9) with $J(\mathbf{b}) = \|\mathbf{b}\|_1$ require that

$$\lambda\boldsymbol{\tau} = \mathbb{E}[\mathbf{Z}^T(\mathbf{y} - a_\lambda\mathbf{x} - \mathbf{Z}\mathbf{b}_\lambda)/n], \tag{2.12}$$

for some $\boldsymbol{\tau}$ satisfying $\boldsymbol{\tau}_\mathcal{A} = \text{sign}(\mathbf{b}_{\lambda\mathcal{A}})$ and $\|\boldsymbol{\tau}_{\mathcal{A}^c}\|_\infty \leq 1$.

We will require the following conditions. Note that some of these conditions depend on λ, and thus may hold for some values of λ, and not for others.

(**A1**) $\mathbf{y} = \alpha\mathbf{x} + \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{x}$ and $\mathbf{Z}$ are fixed, and $\boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_n]^T$ are independent and identically distributed with mean zero and variance σ_ϵ^2.

(**A2**) The covariates $[\mathbf{x}, \mathbf{Z}_{\mathcal{A}}]$ are such that $\lim_{n\to\infty} \|\mathbf{r}\|_\infty / \|\mathbf{r}\|_2 \to 0$, where $\mathbf{r} = (\mathbf{I}_n - \mathbf{P}_{\mathcal{A}})\mathbf{x} \in \mathbb{R}^n$.

Condition (**A2**) is needed in order to apply the Lindeberg-Feller Central Limit Theorem, and requires that no single element of $(\mathbf{I}_n - \mathbf{P}_{\mathcal{A}})\mathbf{x}$ is too large, relative to the other elements.

In order to allow d to grow more quickly than n, we require the following additional conditions.

(**A3**) λ, $(\alpha, \boldsymbol{\beta})$ and $(\mathbf{x}, \mathbf{Z})$ are such that $\|\boldsymbol{\tau}_{\mathcal{A}^c}\|_\infty \leq 1 - \delta$ for some $\delta > 0$.

(**A4**) The matrix $\mathbf{Z}$ is such that $\|\mathbf{Z}_{\mathcal{A}^c}^T \mathbf{Z}_{\mathcal{A}}/n\|_\infty = o(\sqrt{n/(q \log q)})$.

(**A5**) The errors $\boldsymbol{\epsilon}$ have sub-Gaussian tails. That is, there exists some constants $c, h > 0$ such that $\Pr(|\epsilon_i| > x) < 2\exp(-hx^2)$, $\forall x > c$. Furthermore, $\mathbf{Z} = [z_{ij}]$ has bounded entries, i.e. $|z_{ij}| < M$, $\forall i, j$.

(**A6**) The minimum eigenvalue of $\boldsymbol{\Sigma}_{\mathcal{A}}$ is bounded (i.e. $\Lambda_{\min}(\boldsymbol{\Sigma}_{\mathcal{A}}) \geq \eta > 0$), and the sample size n, the dimension d, the number of non-zero parameters q, and the minimum non-zero coefficient $b_{\min} \equiv \min\{\, |b_{\lambda,1}|, \ldots, |b_{\lambda,q}| \,\}$ are such that

$$\frac{\log(d)}{n} + \frac{q \log(q)}{n b_{\min}^2} \to 0.$$

Conditions (**A3**) and (**A4**) guarantee that the zero and non-zero elements of $\mathbf{b}_\lambda$ can be distinguished from one another. In particular, (**A3**) requires that the inactive variables $\mathbf{Z}_{\mathcal{A}^c}$ cannot be too correlated with the residuals $\mathbf{y} - a_\lambda \mathbf{x} - \mathbf{Z}\mathbf{b}_\lambda$. By the stationary conditions of (2.9), we know that $\|\boldsymbol{\tau}_{\mathcal{A}^c}\|_\infty \leq 1$; here (**A3**) ensures enough separation in this inequality as the dimension grows. Similarly, (**A4**) requires that correlation between the active and inactive features cannot grow too quickly. This is

related to the irrepresentable condition, given in Zhao and Yu [2006], which can be written as $\|\mathbf{Z}_{\mathcal{A}^{*c}}^T \mathbf{Z}_{\mathcal{A}^*}(\mathbf{Z}_{\mathcal{A}^*}^T \mathbf{Z}_{\mathcal{A}^*})^{-1}\text{sign}(\boldsymbol{\beta}_{\mathcal{A}^*})\|_\infty < 1$, where $\mathcal{A}^* = \text{supp}(\boldsymbol{\beta})$.

Conditions (**A5**) and (**A6**) are somewhat standard in high-dimensional statistics. The sub-Gaussian tails of $\boldsymbol{\epsilon}$ and boundedness of $\mathbf{Z}$ allow d to grow quickly, so long as $\log(d)/n \to 0$. We assume sub-Gaussian tails in (**A5**) for convenience; we could assume e.g. polynomial tails, at the cost of a slower rate of convergence. The condition on $q\log(q)/(nb_{\min}^2)$ ensures that the non-zero elements of $\mathbf{b}_\lambda$ are large enough to be detected in the estimate $\hat{\mathbf{b}}_\lambda^0$. The condition on $\Lambda_{\min}(\boldsymbol{\Sigma}_\mathcal{A})$ ensures that $\mathbf{b}_{\lambda\mathcal{A}}$ is identifiable.

Proposition 3. *Let* $\hat{\mathbf{b}}_\lambda^0$ *be as in* (2.11) *and define* $T_\lambda = \mathbf{x}^T(\mathbf{y} - \mathbf{Z}\hat{\mathbf{b}}_\lambda^0)/\sqrt{n}$. *Assume conditions (***A1***)-(***A6***) hold. Then under* $H_{0,\lambda} : a_\lambda = 0$,

$$\frac{T_\lambda}{\sigma_\epsilon \sqrt{\mathbf{x}^T(\mathbf{I}_n - \mathbf{P}_\mathcal{A})\mathbf{x}/n}} \to_d N(0,1). \tag{2.13}$$

Note that when λ is chosen large enough so that $\mathbf{b}_\lambda = \mathbf{0}$, then $\mathcal{A} = \emptyset$, and the variance of T_λ is approximately σ_ϵ^2, as in simple linear regression. On the other hand, if $\lambda = 0$, the variance of T_λ is approximately $\sigma_\epsilon^2 \mathbf{x}^T\left(\mathbf{I}_n - \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\right)\mathbf{x}/n$, the variance of the classical score statistic used to test $\alpha = 0$ in the model (2.3).

Unfortunately, the variance formula given in (2.13) depends on the support set $\mathcal{A} = \text{supp}(\mathbf{b}_\lambda)$ and the residual variance σ_ϵ^2, which are in general unknown. Estimating the residual variance σ_ϵ^2 is required in a number of other procedures, such as the covariance test [Lockhart et al., 2013], and there are a few options available [see e.g. Fan et al., 2012].

In order to estimate $\mathcal{A}$, we propose two options:

1. Use the observed support $\hat{\mathcal{A}} = \text{supp}(\hat{\mathbf{b}}_\lambda^0)$ as an estimate of $\mathcal{A}$:

$$\widehat{\text{var}}(T_\lambda) = \hat{\sigma}_\epsilon^2 \mathbf{x}^T(\mathbf{I}_n - \mathbf{P}_{\hat{\mathcal{A}}})\mathbf{x}/n. \tag{2.14}$$

We call this the *asymptotic variance estimate* since it relies on the property that $\hat{\mathcal{A}} = \mathcal{A}$ with high probability, asymptotically.

2. Replace $\mathbf{x}^T (\mathbf{I}_n - \mathbf{P}_{\mathcal{A}}) \mathbf{x}/n$ with the upper bound of 1:

$$\widehat{\text{var}}(T_\lambda) = \hat{\sigma}_\epsilon^2 \mathbf{x}^T \mathbf{x}/n = \hat{\sigma}_\epsilon^2. \tag{2.15}$$

We call this the *conservative variance estimate*.

In Section 4.5.1 we show that the asymptotic variance generally works well in practice. Using the conservative variance estimate has an appealing interpretation in light of Proposition 1. The penalized score test (2.6) tests the effect of a single feature $\mathbf{x} = \mathbf{x}_j$ in (2.3), adjusting for the other features $\mathbf{Z} = [\mathbf{x}_k, k \neq j]$ with lasso regression. When using the penalized score test for testing the effect of $\mathbf{x} = \mathbf{x}_j$ for each $j = 1 \ldots, d$ in turn, the conservative variance estimate will be the same for each $j = 1, \ldots, d$. Thus, the sparsity pattern of lasso regression, which results from comparing each T_λ to $\sqrt{n}\lambda$, is the same as the set of rejections that results from applying the penalized score test to each feature in turn, using the same (conservative) significance threshold.

2.3.1 Bias and the irrepresentable condition

As we saw in Section 2.2.1, when $\lambda > 0$ the penalized score test does not test the null hypothesis $H_0 : \alpha = 0$, as in multiple linear regression, but instead adopts the null hypothesis $H_{0,\lambda} : a_\lambda = 0$. Thus, if the penalized score test is used as a surrogate for classical tests of $H_0 : \alpha = 0$ versus $H_1 : \alpha \neq 0$, it may not be an unbiased test. In this section we investigate the relationship between the penalized score test and unbiased tests of $H_0 : \alpha = 0$, and show that, in the case of the lasso penalty, differences between the tests are closely related to the irrepresentable condition established by Zhao and Yu [2006].

As discussed in Section 2.2.1, our main interest in this paper is in the interpretation and behavior of the penalized score test at a particular tuning parameter λ. However, in this section, we consider the behavior of the lasso-penalized score test as $\lambda \to 0$ in order to establish a connection with variable selection consistency results. We note that when using the lasso-penalized score test in practice, variable selection consistency is not necessary in order to obtain a valid test of $H_{0,\lambda}$.

First, we show that the lasso-penalized score statistic (2.6) can be interpreted as a shifted version of a classical score statistic. Let $\hat{\mathcal{A}} = \text{supp}(\hat{\mathbf{b}}^0_\lambda)$ in (2.11), and consider testing for the effect of $\mathbf{x}$, adjusted for $\mathbf{Z}_{\hat{\mathcal{A}}}$, using the classical score test (3.5). In this case, the classical score statistic (3.5) takes the form

$$T_{\hat{\mathcal{A}}} = \mathbf{x}^T \left(\mathbf{I}_n - \mathbf{P}_{\hat{\mathcal{A}}}\right) \mathbf{y} / \sqrt{n}. \tag{2.16}$$

Suppose $\hat{\mathbf{b}}^0_\lambda$ is ordered such that $\hat{\mathbf{b}}^0_\lambda = (\hat{b}^0_{\lambda,1}, \ldots, \hat{b}^0_{\lambda,|\hat{\mathcal{A}}|}, 0, \ldots, 0)^T = (\hat{\mathbf{b}}^0_{\lambda\hat{\mathcal{A}}}, \mathbf{0}^T)^T$, where $\min\{|\hat{b}^0_{\lambda,1}|, \ldots, |\hat{b}^0_{\lambda,|\hat{\mathcal{A}}|}|\} > 0$. Let $\boldsymbol{\sigma}_{x\hat{\mathcal{A}}} = \mathbf{Z}^T_{\hat{\mathcal{A}}}\mathbf{x}/n$, $\boldsymbol{\Sigma}_{\hat{\mathcal{A}}} = \mathbf{Z}^T_{\hat{\mathcal{A}}}\mathbf{Z}_{\hat{\mathcal{A}}}/n$, and $\hat{\boldsymbol{\tau}}_{\hat{\mathcal{A}}} = \text{sign}(\hat{\mathbf{b}}^0_{\lambda\hat{\mathcal{A}}})$. Using the identity $\mathbf{Z}\hat{\mathbf{b}}^0_\lambda = \mathbf{P}_{\hat{\mathcal{A}}}\mathbf{y} - \lambda \mathbf{Z}_{\hat{\mathcal{A}}}\boldsymbol{\Sigma}^{-1}_{\hat{\mathcal{A}}}\hat{\boldsymbol{\tau}}_{\hat{\mathcal{A}}}$, given in Equation 21 of Tibshirani and Taylor [2012], we get that

$$T_\lambda = T_{\hat{\mathcal{A}}} + \sqrt{n}\lambda \boldsymbol{\sigma}^T_{x\hat{\mathcal{A}}}\boldsymbol{\Sigma}^{-1}_{\hat{\mathcal{A}}}\hat{\boldsymbol{\tau}}_{\hat{\mathcal{A}}}. \tag{2.17}$$

Thus, the penalized score statistic T_λ differs from the classical score statistic $T_{\hat{\mathcal{A}}}$ by $\sqrt{n}\lambda \boldsymbol{\sigma}^T_{x\hat{\mathcal{A}}}\boldsymbol{\Sigma}^{-1}_{\hat{\mathcal{A}}}\hat{\boldsymbol{\tau}}_{\hat{\mathcal{A}}}$.

Now, suppose that we wish to test $H_0 : \alpha = 0$. In Section 2.2.1, we saw that T_λ is centered around zero when $a_\lambda = 0$. Therefore, unless $\alpha = a_\lambda$, the penalized score test may not be unbiased as a test of $H_0 : \alpha = 0$. On the other hand, $T_{\hat{\mathcal{A}}}$ is centered around zero when $\hat{\mathcal{A}}$ contains all variables relevant to the outcome. A sufficient condition for $T_{\hat{\mathcal{A}}}$ to be centered around zero is then $\hat{\mathcal{A}} = \mathcal{A}^*$, where $\mathcal{A}^* = \text{supp}(\boldsymbol{\beta})$. In order to make use of (2.17) to compare the penalized score test to an unbiased

test of H_0, we thus consider the case where $(\mathbf{y}, \mathbf{Z}, \lambda)$ are such that $\Pr(\hat{\mathcal{A}} = \mathcal{A}^*) \to 1$ and $\Pr(\hat{\boldsymbol{\tau}}_{\hat{\mathcal{A}}} = \text{sign}(\boldsymbol{\beta}_{\mathcal{A}^*})) \to 1$ under $H_0 : \alpha = 0$. Zhao and Yu [2006] showed that this holds provided that $\lambda \to 0$ and $\sqrt{n}\lambda \to \infty$ as $n \to \infty$, in addition to some assumptions on $(\mathbf{y}, \mathbf{Z})$, which we omit here for ease of exposition. Note that we have not yet made any assumptions on the relationship between $\mathbf{x}$ and $\mathbf{Z}$. Under these assumptions, by (2.17) we have that

$$T_\lambda = T_{\mathcal{A}^*} + \sqrt{n}\lambda \boldsymbol{\sigma}_{x\mathcal{A}^*}^T \boldsymbol{\Sigma}_{\mathcal{A}^*}^{-1} \boldsymbol{\tau}_{\mathcal{A}^*} + o_p(1), \tag{2.18}$$

where $\boldsymbol{\tau}_{\mathcal{A}^*} = \text{sign}(\boldsymbol{\beta}_{\mathcal{A}^*})$. Here, we can consider $T_{\mathcal{A}^*}$ to be the 'oracle' test statistic: the score statistic for testing $H_0 : \alpha = 0$, which knows in advance the support of $\boldsymbol{\beta}$.

Now, in order for the lasso to recover the support of $(\alpha, \boldsymbol{\beta})$ in lasso regression of $\mathbf{y}$ on $(\mathbf{x}, \mathbf{Z})$, as in (2.7), Zhao and Yu [2006] showed that, in addition to conditions on $(\mathbf{y}, \mathbf{Z}, \lambda)$ required for (2.18) to hold, an *irrepresentable condition* must hold. This condition implies (among other things) that $|\boldsymbol{\sigma}_{x\mathcal{A}^*}^T \boldsymbol{\Sigma}_{\mathcal{A}^*}^{-1} \boldsymbol{\tau}_{\mathcal{A}^*}| < 1$. Examining Proposition 1, we can see the connection between the irrepresentable condition and recovery of the support of $(\alpha, \boldsymbol{\beta})$. In the lasso solution (2.7), $\hat{a}_\lambda = 0$ when $|T_\lambda| \leq \sqrt{n}\lambda$. Under $H_0 : \alpha = 0$, we have that $T_{\mathcal{A}^*} = O_p(1)$, and thus by (2.18), $\Pr_{H_0}(|T_\lambda| \leq \sqrt{n}\lambda) \to 1$ when $|\boldsymbol{\sigma}_{x\mathcal{A}^*}^T \boldsymbol{\Sigma}_{\mathcal{A}^*}^{-1} \boldsymbol{\tau}_{\mathcal{A}^*}| < 1$. That is, when the irrepresentable condition is satisfied, the penalized score statistic T_λ is close enough to the oracle statistic $T_{\mathcal{A}^*}$ for the decision rule used by the lasso (i.e. 'reject $H_{0,\lambda}$ when $|T_\lambda| > \sqrt{n}\lambda$') to correctly identify that $\alpha = 0$.

In the preceding discussion we assumed that λ was chosen in order to have $\hat{\mathcal{A}} = \mathcal{A}^*$. However, this is not necessary in order to for the penalized score test to yield meaningful results. How far the penalized score statistic T_λ deviates from zero under $H_0 : \alpha = 0$ depends more on the bias of $\hat{\mathbf{b}}_\lambda^0$ relative to $\boldsymbol{\beta}$, rather than the support set $\hat{\mathcal{A}}$ per se. Choosing a smaller λ will result in less bias in $\hat{\mathbf{b}}_\lambda^0$ relative to $\boldsymbol{\beta}$, at the expense of a larger number of degrees of freedom spent in the lasso regression

(2.5). We explore this issue numerically in Section 4.5.1, and discuss it further in Section 2.6.

2.3.2 Simulation study

In this section, we study the empirical behavior of the lasso-penalized score test. We show that the test serves as a useful proxy for tests of $H_0 : \alpha = 0$ provided λ is small enough, and that it behaves like tests of marginal correlation when λ is large.

First, we generated a matrix of correlated features $\mathbf{X} \in \mathbb{R}^{n \times d}$, where the rows were independently distributed $N_d(0, \mathbf{S})$ with $\mathbf{S}_{jk} = 0.5^{|j-k|}$. We then generated an outcome $\mathbf{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \mathbf{I}_n)$, where we set 10 elements of $\boldsymbol{\beta}$ at random to be 0.4, and the rest to be zero. We used sample sizes of $n = 50, 75, 100, 150, 200, 300$ and 400, and dimensions of $d = 100$ and $d = 300$. Results are averaged over $B = 500$ simulated data sets, where $\boldsymbol{\beta}$ was held constant over replications with the same dimension d.

We performed the lasso penalized score test on each feature in turn, for a sequence of values of λ. For the sake of comparison, we also performed simple linear regression of $\mathbf{y}$ on each feature, multiple linear regression of $\mathbf{y}$ on all features (for the cases where $d > n$), LDPE, and the 'oracle' score test, which tests for the association of feature $\mathbf{x}_j$, and knows the support of the other features $\{\beta_k : k \neq j\}$. In order to make results comparable, we used the same estimate of the residual variance $\sigma^2_{\boldsymbol{\epsilon}}$ in the penalized score test, multiple linear regression, and in simple linear regression, which we obtained using the refitted cross-validation method described by Fan et al. [2012]. Note that we do not compare to methods of Lee et al. [2013b] and Taylor et al. [2014], which describe inference regarding only those features with non-zero coefficients in lasso regression, or the covariance test of Lockhart et al. [2013], which does not provide inference for individual features.

We declared a test to be significant when the resulting p-value was less than $1/d$. Since 10 feature are truly associated with the outcome, we would expect $(d-10)/d \approx 1$ false positives per simulated data set for an unbiased test, which we will refer to as the

'nominal error rate'. For each test we calculated the expected false positives (EFP) and the power. For a particular test, if p_{jk} is the p-value for feature j on the k^{th} simulated data set, EFP and power are given by

$$\text{EFP} = \frac{1}{B}\sum_{k=1}^{B}\sum_{j:\beta_j=0} 1\{p_{jk} < 1/d\}$$
$$\text{power} = \frac{1}{B}\frac{1}{\|\boldsymbol{\beta}\|_0}\sum_{k=1}^{B}\sum_{j:\beta_j\neq 0} 1\{p_{jk} < 1/d\},$$

where $1\{\cdot\}$ is the indicator function and $\|\boldsymbol{\beta}\|_0 = 10$ is the cardinality of $\boldsymbol{\beta}$, where here we use the notation of Equation 2.1. We chose to use a significance threshold in order to control EFP, but in principle one could use any method of error control, such as a Ŝidák, Bonferroni, or FDR correction.

In Figure 2.1 we examine the p-values produced by the lasso-penalized score test when $\lambda = 0.005$ and $\lambda = 0.6$, for a single simulated data set with $d = 100$ and $n = 200$, and compare them to the p-values produced by multiple and simple linear regression. When $\lambda = 0.005$, 88 of the 100 coefficients are non-zero in the lasso regression on all features. Consequently, we see that the penalized score test behaves much like multiple linear regression including all 100 features. On the other hand, when $\lambda = 0.6$, only 3 features are included in the lasso regression on all features, and the p-values are similar to those from simple linear regression performed on each feature separately. For the sake of comparison, we also plot the multiple linear regression p-values against those from simple linear regression, which demonstrates that the behavior of these two tests are quite different. Note that the penalized score test p-values are not identical to those from classical tests: inference with the penalized score test is with respect to the parameter a_λ, given in (2.10), which, as discussed in Section 2.2.1, measures different types of associations than those measured with simple or multiple linear regression.

Figure 2.2 summarizes the results of the experiment over the $B = 500$ replications.

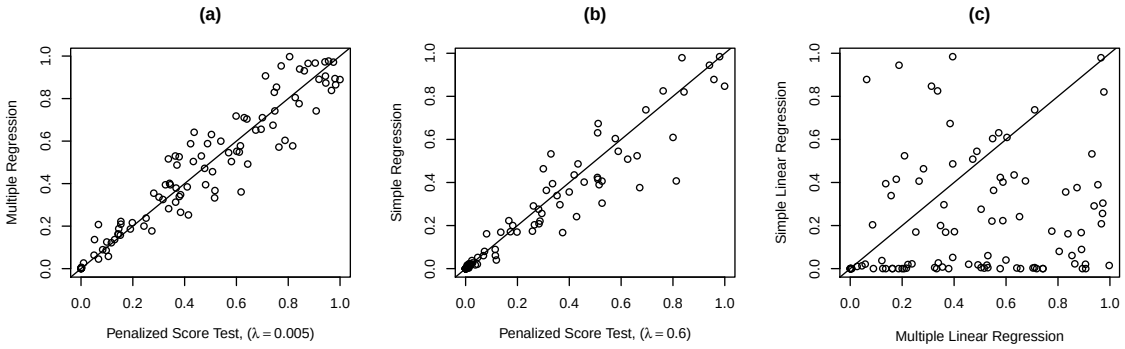

Figure 2.1: Comparison of lasso-penalized score test p-values for $H_{0,\lambda} : a_\lambda = 0$ to traditional p-values for $H_0 : \alpha = 0$, using the asymptotic variance formula (2.14). (a) The p-values from multiple linear regression plotted against those from the penalized score test with $\lambda = 0.005$. (b) The p-values from simple linear regression plotted against those from the penalized score test with $\lambda = 0.6$. (c) Multiple and simple linear regression p-values plotted against each other.

We see that simple linear regression provides high power, but also results in high EFP. Recall that simple linear regression will detect features which are marginally correlated with $\mathbf{y}$, whereas here we are interested in the conditional relationships. When λ is large, the penalized score test has nearly identical power and EFP to simple linear regression; this is not surprising, since in that setting the penalized score test is almost identical to simple linear regression (see e.g. Figure 2.1). As λ is decreased, the number of false positives converges to the nominal rate, for all sample sizes n and dimensions d considered, at the cost of lower power. This reduction in power should be expected: as λ is decreased, more features enter the lasso regression (2.11), making it increasingly difficult to distinguish the effect of the feature $\mathbf{x}$ from the effects of the other correlated features in the model. In the extreme case of $\lambda = 0$, or equivalently, using multiple linear regression, power is quite low. We see that the performance of our method using $\lambda = 0.05$ or $\lambda = 0.07$ typically yields comparable type-I error, and slightly higher power, than LDPE.

In the bottom panels of Figure 2.2, we also plot the number of non-zero coefficients

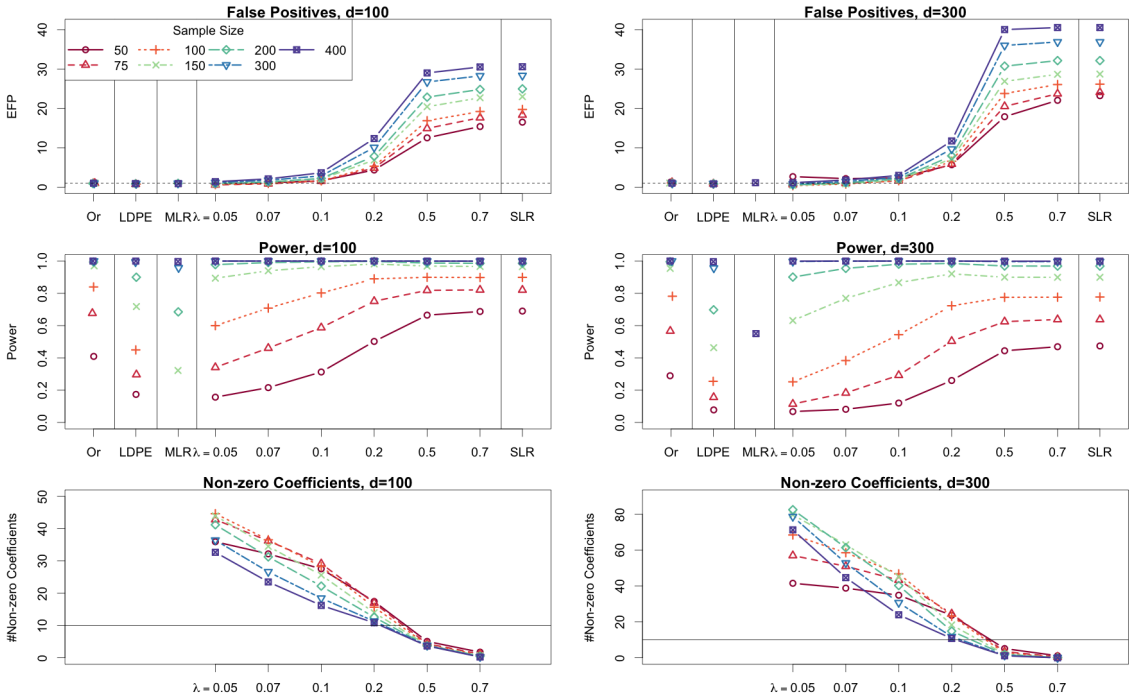

Figure 2.2: Simulation experiments. 'Or' indicates the oracle test, 'LDPE' indicates the method of Zhang and Zhang [2011] and van de Geer et al. [2013], 'MLR' indicates multiple linear regression, and 'SLR' indicates the results from simple linear regression. The six values of λ correspond to the lasso-penalized score test. In the top panels, the horizontal line indicates the nominal error rate $(d-10)/d \approx 1$. In the bottom panels, the horizontal line indicates the true number of non-zero coefficients $\|\boldsymbol{\beta}\|_0 = 10$. Results are averaged over 500 simulated data sets.

in lasso-penalized regression of $\mathbf{y}$ on $\mathbf{X}$. EFP is closest to the nominal rate when λ is smallest, and here we see that this results in many more non-zero lasso coefficients than the number of truly non-zero coefficients. That is, in order to strictly control the error rate, it seems beneficial to choose λ to be smaller than one would choose it if variable selection with the lasso were the goal. This assertion is supported by the theory in Section 2.3.1, where we showed that T_λ can diverge to $\pm\infty$ under $H_0 : \alpha = 0$, if λ is chosen so that the lasso selects the correct model (i.e. $\sqrt{n}\lambda \to \infty$ while $\lambda \to 0$). On the other hand, in Section 2.2.1 we showed that T_λ is centered around zero under $H_0 : \alpha = 0$ when $\sqrt{n}\lambda \to 0$.

2.3.3 Diabetes data

Here we re-examine the diabetes data set from Section 2.1. We apply the lasso-penalized score test, for a range of 300 values of λ between 0 and 50. We estimate σ^2_ϵ by the residual variance from multiple linear regression on all features.

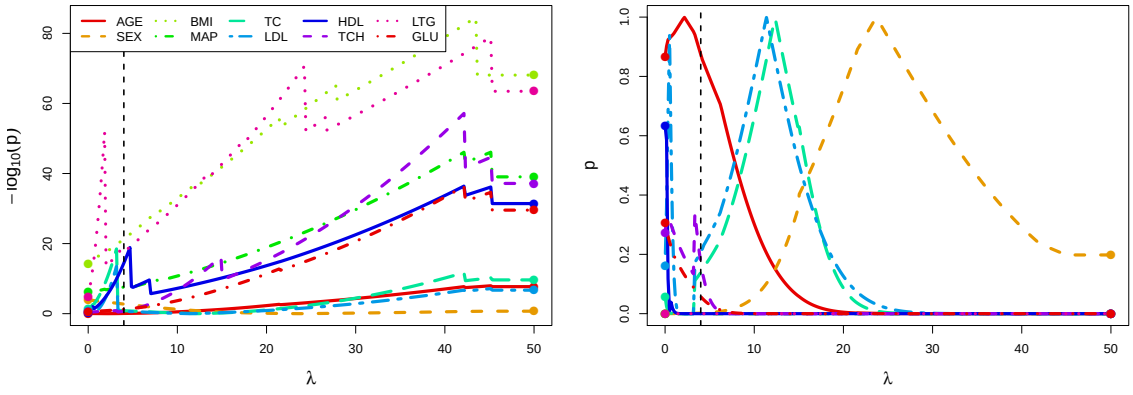

Figure 2.3: Diabetes data set. Lasso-penalized score test p-values were generated using the asymptotic variance formula (2.14). The vertical line at $\lambda = 4$ indicates the value chosen to produce p-values for the penalized score test given in Table 2.1. Dots at $\lambda = 0$ and $\lambda = 50$ indicate p-values from multiple linear regression on all features, and simple linear regression on each feature alone.

Figure 2.3 summarizes the results of this analysis, showing both $-\log_{10}$ and untransformed p-values for each of the λ values, using the asymptotic variance formula (2.14). For comparison, we also plot the multiple and simple linear regression p-values; these are displayed in Figure 2.3 at the far left side ($\lambda = 0$) and on the far right side ($\lambda = 50$) respectively. We see that the p-values from the penalized score test interpolate the multiple linear regression p-values and the simple linear regression p-values, and vary widely depending on the value of λ chosen. The p-values for each feature are piece-wise continuous, with jumps when the set $\text{supp}(\hat{\mathbf{b}}^0_\lambda) = \hat{\mathcal{A}}$ changes. The jumps typically result in smaller p-values immediately after an element of $\hat{\mathbf{b}}^0_\lambda$ becomes non-zero, since the variance of the test statistic, $\sigma^2_\epsilon \mathbf{x}^T \left(\mathbf{I}_n - \mathbf{P}_{\hat{\mathcal{A}}}\right) \mathbf{x}/n$, will be smaller when the size of the set $\hat{\mathcal{A}}$ is larger. This phenomenon is investigated in greater detail in Section 2.3.4.

The vertical line in Figure 2.3 indicates the value of $\lambda = 4$, chosen to produce the p-values in Table 2.1. With this value of λ, 4 of the 10 covariates (AGE, TC, LDL, and TCH) have zero coefficients in lasso regression on all features, while the rest are non-zero. Lasso regression on all features with this choice of λ yields an R^2 of 0.50, compared to the R^2 of 0.52 using multiple linear regression on all features. Both in Table 2.1 and in Figure 2.3, we see that this value of λ results in p-values which are qualitatively similar to those from multiple linear regression on all features. However, it is notable that HDL appears strongly associated in this lasso-penalized score test, but not in the multiple linear regression. The multiple linear regression results suggest that the effects of HDL can be explained by other features in the data set; using the penalized score test with $\lambda = 4$, the effects of other features are sufficiently shrunken towards zero so that HDL appears associated with the response.

In Figure 2.4 we once again show p-values corresponding to the lasso score test applied to the diabetes data. This time, however, we calculate p-values using the conservative variance estimate $\widehat{\text{var}}(T_\lambda) = \sigma^2_\epsilon$, in (2.15), for each feature. Since this variance formula does not depend on the support of $\hat{\mathbf{b}}^0_\lambda$, the p-values are continuous

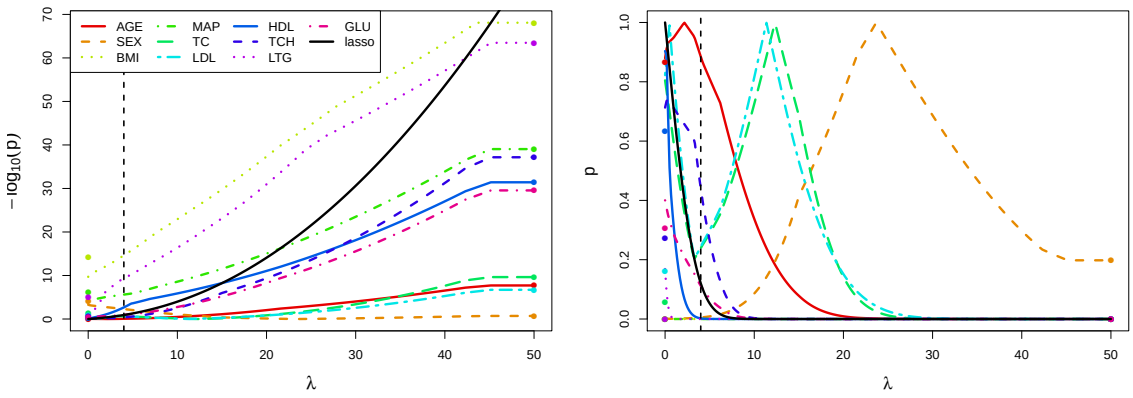

Figure 2.4: Diabetes data set. Lasso-penalized score test p-values generated using the conservative variance formula (2.15). The lasso decision rule, shown in black, corresponds to $\Phi(-2\sqrt{n}\lambda/\sigma_\epsilon)$, where $\Phi(\cdot)$ is the standard normal distribution function.

curves. Further, since the same reference distribution is used for each feature, the decision rule which yields the sparsity pattern of the lasso, 'reject $H_{0,\lambda}$ when $|T_\lambda| > \sqrt{n}\lambda$', corresponds to the same p-value threshold for each feature. This threshold is displayed as a thick black line in Figure 2.4; when the p-value for a feature crosses the line, its coefficient becomes non-zero in the lasso regression. For instance, at $\lambda = 4$, the p-value threshold is 0.12. The variables AGE, TC, LDL, and TCH have p-values above this threshold and thus have zero coefficients.

2.3.4 Assessing the impact of thresholding

It is well-known that the finite sample distributions of estimators that involve thresholding may be far from their large-sample limits. The canonical example of this phenomenon is Hodges' 'super-efficient' estimator [see e.g. Lehmann and Casella, 1998, page 589]; similar behavior has also been observed in lasso-type estimators [Knight and Fu, 2000, Pötscher and Leeb, 2009]. In this section, we explore how thresholding can impact the type-I error rate of the penalized score test, when the lasso is used

as the penalty. We follow the example of Leeb and Pötscher [2005], and consider the two-variable case, where the exact distribution of T_λ is simple to obtain. As we will see, our proposed test can be either conservative or anti-conservative, depending on the underlying parameters, and the nominal type-I error of the test.

Suppose we are interested in the effect of a variable $\mathbf{x} \in \mathbb{R}^n$, adjusted for an additional variable $\mathbf{z} \in \mathbb{R}^n$. Further suppose that $\mathbf{y} = \alpha\mathbf{x} + \beta\mathbf{z} + \boldsymbol{\epsilon}$, where $\mathbf{x}$ and $\mathbf{z}$ are fixed, $\mathbf{x}^T\mathbf{z}/n = \rho$, $\mathbf{x}^T\mathbf{x}/n = \mathbf{z}^T\mathbf{z}/n = 1$, and $\boldsymbol{\epsilon} \sim N_n(0, \mathbf{I}_n)$. As a reminder, we are testing the effect $a_\lambda = \alpha + \rho(\beta - b_\lambda)$, given in (2.10).

In this simple case, the lasso-penalized score test has two steps:

1. Regress $\mathbf{y}$ on $\mathbf{z}$ using the lasso. This corresponds to soft-thresholding the quantity $\mathbf{y}^T\mathbf{z}/n$. That is, we set $\hat{b}^0_\lambda = \text{sign}(\mathbf{y}^T\mathbf{z})(|\mathbf{y}^T\mathbf{z}/n| - \lambda)_+$, an estimate of b_λ under the null hypothesis $H_{0,\lambda} : a_\lambda = 0$.

2. Construct $T_\lambda = \mathbf{x}^T(\mathbf{y} - \hat{b}^0_\lambda\mathbf{z})/\sqrt{n}$, and compare to a normal reference distribution, with variance to be specified next.

Under $H_{0,\lambda} : a_\lambda = 0$, Proposition 3 shows that in large samples, T_λ should have variance $1 - \rho^2$ when $|\mathbb{E}\mathbf{y}^T\mathbf{z}/n| > \lambda$, and 1 otherwise. In finite samples, we can either use the asymptotic estimate (2.14) (i.e. use $\widehat{\text{var}}(T_\lambda) = 1 - \rho^2$ when $|\mathbf{y}^T\mathbf{z}/n| > \lambda$, and $\widehat{\text{var}}(T_\lambda) = 1$ otherwise), or the conservative estimate (2.15) (i.e. always use variance $\widehat{\text{var}}(T_\lambda) = 1$). We will investigate the behavior of the test for both estimators.

Here, T_λ can be written explicitly as

$$T_\lambda = \begin{cases} \sqrt{n}[(1-\rho^2)\alpha + \rho\lambda] + (\mathbf{x} - \rho\mathbf{z})^T\boldsymbol{\epsilon}/\sqrt{n} & \text{if} \quad \mathbf{z}^T\mathbf{y}/n \geq \lambda \\ \sqrt{n}(\alpha + \rho\beta) + \mathbf{x}^T\boldsymbol{\epsilon}/\sqrt{n} & \text{if} \quad |\mathbf{z}^T\mathbf{y}/n| < \lambda \\ \sqrt{n}[(1-\rho^2)\alpha - \rho\lambda] + (\mathbf{x} - \rho\mathbf{z})^T\boldsymbol{\epsilon}/\sqrt{n} & \text{if} \quad \mathbf{z}^T\mathbf{y}/n \leq -\lambda \end{cases} . \tag{2.19}$$

It is easy to see that, conditioned on $\mathbf{z}^T\mathbf{y} \sim N(n(\rho\alpha + \beta), n)$, T_λ is normally

distributed. To find the marginal distribution of T_λ, we simply calculate $\mathbb{E}_{\mathbf{z}^T\mathbf{y}}[\, T_\lambda \mid \mathbf{z}^T\mathbf{y}\,]$, by numerical integration.

Our goal is to determine the impact of thresholding on type-I error. In order to do this, we choose (α, β) such that (i) the null hypothesis $H_{0,\lambda} : a_\lambda = 0$ is true, and (ii) the probability $\Pr(\mathbf{z}^T\mathbf{y}/n \geq \lambda)$ is controlled to be γ. Since $\mathbf{z}^T\mathbf{y}/n \sim N(\rho\alpha + \beta, 1/n)$, we must have that $\beta = \Phi^{-1}(\gamma)/\sqrt{n} + \lambda - \rho\alpha$ in order to achieve $\Pr(\mathbf{z}^T\mathbf{y}/n \geq \lambda) = \gamma$, where $\Phi(\cdot)$ is the standard normal distribution function. In order to have $a_\lambda = 0$, we must have $\alpha + \rho(\beta - b_\lambda) = 0$, where $b_\lambda = \text{sign}(\mathbb{E}[\,\mathbf{y}^T\mathbf{z}\,])(|\mathbb{E}[\,\mathbf{y}^T\mathbf{z}/n\,]| - \lambda)_+ = \text{sign}(\alpha\rho + \beta)(\alpha\rho + \beta - \lambda)_+$. Thus, with restrictions (i) and (ii), we must have that

$$\alpha = \begin{cases} -\rho\lambda/(1-\rho^2) & \text{if} \quad \gamma > 0.5 \\ -\rho(\Phi^{-1}(\gamma)/\sqrt{n} + \lambda)/(1-\rho^2) & \text{if} \quad \Phi(-2\sqrt{n}\lambda) \leq \gamma \leq 0.5 \\ \rho\lambda/(1-\rho^2) & \text{if} \quad \gamma < \Phi(-2\sqrt{n}\lambda) \end{cases}$$

$$\beta = \Phi^{-1}(\gamma)/\sqrt{n} + \lambda - \rho\alpha.$$

The cases $\gamma > 0.5$, $\Phi(-2\sqrt{n}\lambda) \leq \gamma \leq 0.5$, and $\gamma < \Phi(-2\sqrt{n}\lambda)$ correspond to the cases $\mathbb{E}\mathbf{z}^T\mathbf{y}/n > \lambda$, $|\mathbb{E}\mathbf{z}^T\mathbf{y}/n| \leq \lambda$, and $\mathbb{E}\mathbf{z}^T\mathbf{y}/n < -\lambda$ respectively. Note that for fixed λ, $\Phi(-2\sqrt{n}\lambda) \approx 0$ for large n.

When $|\mathbb{E}\mathbf{z}^T\mathbf{y}/n| \leq \lambda$, or equivalently, when $\Phi(-2\sqrt{n}\lambda) \leq \gamma \leq 0.5$, then $b_\lambda = 0$ in truth. However, with probability γ, we will erroneously have $\hat{b}^0_\lambda > 0$. Likewise, when $\mathbb{E}\mathbf{z}^T\mathbf{y}/n > \lambda$, or equivalently, when $\gamma > 0.5$, we then have $b_\lambda > 0$, but with probability $1 - \gamma - \Phi[-2\sqrt{n}\lambda - \Phi^{-1}(\gamma)]$, which is approximately $1 - \gamma$ for $\sqrt{n}\lambda$ large enough, we erroneously set $\hat{b}^0_\lambda = 0$. Thus, by varying γ, we can examine the impact of erroneously including or excluding a feature in the lasso regression.

Figure 2.5 displays the relative type-I error of the test under $H_{0,\lambda} : a_\lambda = 0$, i.e. (observed type-I error)/(nominal type-I error), for a range of values of γ, using both the asymptotically derived variance estimate (2.14) and the conservative variance

estimate (2.15). We chose $\lambda = 0.2$, and $n = 500$. Note that by parametrizing the coefficients by γ, the results depend only very weakly on n; similar curves can be obtained for arbitrarily large sample sizes.

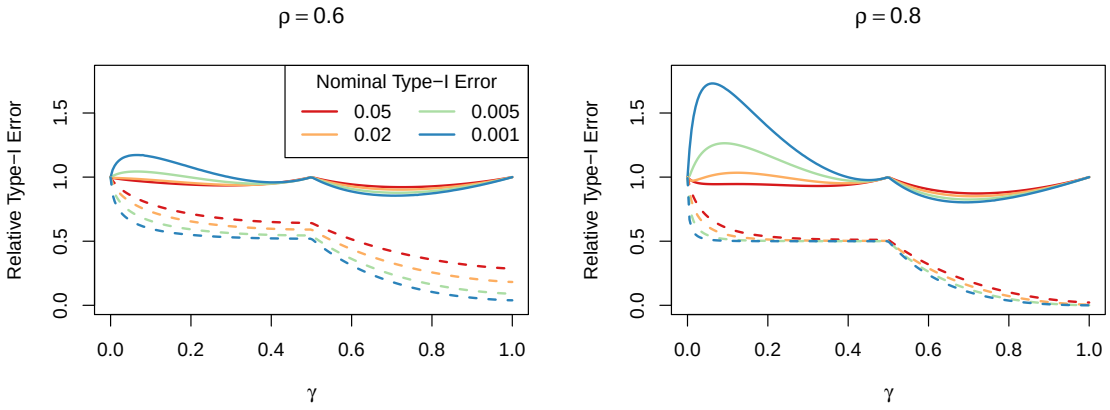

Figure 2.5: Relative type-I error rate, i.e. (observed type-I error)/(nominal type-I error), as a function of $\gamma = \Pr(\hat{b}^0_\lambda > 0)$. Solid lines indicate the error rates for the asymptotic variance formula (2.14), while dashed lines indicate the error rates when using the conservative variance (2.15). Note that the inflection point at $\gamma = 0.5$ is due to the fact that $b_\lambda = 0$ when $\Phi^{-1}(-2\sqrt{n}\lambda) < \gamma \leq 0.5$, while $b_\lambda > 0$ when $\gamma > 0.5$.

We see that when we use the conservative variance (dashed lines), the test is indeed conservative. When $\gamma = 0$ the observed error rate is identical to the nominal rate, while the test becomes increasingly conservative as γ increases.

On the other hand, when we use the asymptotic variance formula (solid lines), the test can be anti-conservative when both the nominal type-I error rate and γ are small, but is otherwise conservative. In general, the behavior of the test is worse for small type-I error rates, and when the correlation between the features is larger.

A reviewer suggested that one could estimate the distribution of T_λ more accurately using the methods described by Andrews and Guggenberger [2009]. In order to implement this procedure, one would use a critical value for T_λ based on a hybrid of m-of-n bootstrap samples, and the critical values based on an asymptotic distribution

of the test. Alternately, the framework of Berk et al. [2013] could be used to obtain a conservative version of the test.

2.4 The ridge-penalized score test

In this section, we examine in greater detail the penalized score test where the ridge penalty is used in (2.5). That is, we first obtain the penalized coefficient vector

$$\hat{\mathbf{b}}_\lambda^0 = \underset{\mathbf{b}\in\mathbb{R}^{d-1}}{\arg\min} \left\{ \|\mathbf{y} - \mathbf{Z}\mathbf{b}\|_2^2/(2n) + \lambda\|\mathbf{b}\|_2^2/2 \right\}, \tag{2.20}$$

and then form the test statistic $T_\lambda = (\mathbf{y} - \mathbf{Z}\hat{\mathbf{b}}_\lambda^0)/\sqrt{n}$ in (2.6).

First, we prove Proposition 2. Let $\mathbf{H_Z} \equiv \mathbf{Z}(\lambda\mathbf{I}_{d-1} + \mathbf{Z}^T\mathbf{Z}/n)^{-1}\mathbf{Z}^T/n$ denote the $n \times n$ smoother matrix from ridge regression. Here we can write T_λ as

$$T_\lambda = \mathbf{x}^T(\mathbf{I}_n - \mathbf{H_Z})\mathbf{y}/\sqrt{n}. \tag{2.21}$$

We now show that (2.21) can be interpreted as the score statistic from a mixed model. If we assume (2.8), then the marginal distribution of $\mathbf{y}$ is

$$\mathbf{y} \sim N_n\left(\alpha\mathbf{x},\, \sigma_\epsilon^2[(n\lambda)^{-1}\mathbf{Z}\mathbf{Z}^T + \mathbf{I}_n]\right). \tag{2.22}$$

Writing $l(\alpha)$ as the log-likelihood of the data under (2.22), we can write the score, evaluated at $\alpha = 0$, as

$$\dot{l}(0) = \frac{1}{\sigma_\epsilon^2}\mathbf{x}^T\left[(n\lambda)^{-1}\mathbf{Z}\mathbf{Z}^T + \mathbf{I}_n\right]^{-1}\mathbf{y}.$$

Recognizing that $[(n\lambda)^{-1}\mathbf{Z}\mathbf{Z}^T + \mathbf{I}_n]^{-1} = (\mathbf{I}_n - \mathbf{H_Z})$, we see that the score for testing $\alpha = 0$ in (2.8) is

$$\dot{l}(0) = \frac{1}{\sigma_\epsilon^2}\mathbf{x}^T(\mathbf{I}_n - \mathbf{H_Z})\mathbf{y},$$

which is a scaled form of (2.21). That is, the ridge-penalized score statistic is equivalent to the score statistic for testing the effect of feature $\mathbf{x}$ in a mixed model, where all the other features have normally distributed random effects with variance $\sigma_\epsilon^2(n\lambda)^{-1}$.

The distribution of the ridge-penalized score statistic T_λ, or equivalently $\dot{l}(0)$, depends on whether we consider $\boldsymbol{\beta}$ to be fixed under the null hypothesis $H_{0,\lambda} : a_\lambda = 0$ in (2.9), or random under the null hypothesis $H_0 : \alpha = 0$ in (2.8). With $\boldsymbol{\beta}$ fixed, solving (2.9) when $a_\lambda = 0$ yields $\mathbf{Z}\mathbf{b}_\lambda = \mathbf{H}_\mathbf{Z}(\alpha\mathbf{x} + \mathbf{Z}\boldsymbol{\beta})$, and thus $\mathbb{E}[T_\lambda \mid \boldsymbol{\beta}] = \mathbf{x}^T(\mathbf{I}_n - \mathbf{H}_\mathbf{Z})(\alpha\mathbf{x} + \mathbf{Z}\beta)/\sqrt{n} = \sqrt{n}[\,\alpha + \boldsymbol{\sigma}_{xz}^T(\boldsymbol{\beta} - \mathbf{b}_\lambda)\,] = 0$. Thus, we can write the exact distribution of T_λ under $H_{0,\lambda} : a_\lambda = 0$ as

$$T_\lambda \mid \boldsymbol{\beta} \overset{H_{0,\lambda}}{\sim} N(0, \sigma_\epsilon^2 \mathbf{x}^T(\mathbf{I}_n - \mathbf{H}_\mathbf{Z})^2\mathbf{x}/n), \tag{2.23}$$

since T_λ is simply a linear function of a normal vector with mean zero. On the other hand, when $H_0 : \alpha = 0$ in the mixed model (2.8), we can use the marginal distribution of $\mathbf{y}$ in (2.22) to obtain

$$T_\lambda \overset{H_0}{\sim} N(0, \sigma_\epsilon^2 \mathbf{x}^T(\mathbf{I}_n - \mathbf{H}_\mathbf{Z})\mathbf{x}/n). \tag{2.24}$$

It is easy to show that $\mathbf{x}^T(\mathbf{I}_n - \mathbf{H}_\mathbf{Z})^2\mathbf{x} \leq \mathbf{x}^T(\mathbf{I}_n - \mathbf{H}_\mathbf{Z})\mathbf{x}$, and thus T_λ has a smaller variance if we assume that $\boldsymbol{\beta}$ is fixed.

In our usual interpretation of the penalized regression (2.5), we do not consider the effects of the features to be random draws from some population. Instead, we can motivate the use of ridge regression with the desire for an estimate of $\boldsymbol{\beta}$ with smaller variance than the multiple linear regression estimate [Draper and Van Nostrand, 1979]. Using a mixed-effects framework when $\boldsymbol{\beta}$ is non-random is also discussed by Hodges and Reich [2010] in the context of spatial statistics, and Greenland [2000] in the context of epidemiology.

2.5 Extension to other sparsity-inducing penalties

We have shown that the sparsity pattern of the lasso can be interpreted as resulting from a statistical test. In this section, we show that other sparsity-inducing penalties, such as *smoothly clipped absolute deviation* (SCAD) [Fan and Li, 2001] and the *elastic net* [Zou and Hastie, 2005], can also be interpreted in a similar framework.

Suppose we obtain sparse estimates $(\hat{\alpha}_\lambda, \hat{\mathbf{b}}_\lambda)$ of the linear regression parameters $(\alpha, \boldsymbol{\beta})$, by solving the optimization problem

$$(\hat{a}_\lambda, \hat{\mathbf{b}}_\lambda) = \underset{(a,\mathbf{b})\in\mathbb{R}^d}{\arg\min} \left\{ \frac{1}{2n}\|\mathbf{y} - a\mathbf{x} - \mathbf{Z}\mathbf{b}\|_2^2 + \lambda J(a, \mathbf{b}) \right\}. \tag{2.25}$$

A common characteristic of sparsity-inducing penalties is non-differentiability of $J(a, \mathbf{b})$ around zero. Suppose $J(a, \mathbf{b})$ is symmetric about zero, with subdifferential $\frac{\partial}{\partial a} J(a, \mathbf{b}) \mid_{a=0} = [-1, 1]$. The elastic net and SCAD are two such examples. Now, denote

$$\hat{\mathbf{b}}_\lambda^0 = \underset{\mathbf{b}\in\mathbb{R}^{d-1}}{\arg\min} \left\{ \frac{1}{2n}\|\mathbf{y} - \mathbf{Z}\mathbf{b}\|_2^2 + \lambda J(0, \mathbf{b}) \right\}. \tag{2.26}$$

A necessary condition for $\hat{a}_\lambda = 0$ in (2.25) is that $\mathbf{x}^T(\mathbf{y} - \mathbf{Z}\hat{\mathbf{b}}_\lambda^0)/n \in [-\lambda, \lambda]$. Using the notation from Section 4.4, we have that

$$\{\hat{a}_\lambda = 0\} \implies \{|T_\lambda| \leq \sqrt{n}\lambda\}.$$

In addition, if $J(a, \mathbf{b})$ is convex, as is the case for the elastic net, then

$$\{\hat{a}_\lambda = 0\} \iff \{|T_\lambda| \leq \sqrt{n}\lambda\}.$$

In other words, the sparsity pattern of coefficient vectors induced by any convex penalty with subdifferential $[-1, 1]$ at the origin can be interpreted as a decision made based on the penalized score statistic T_λ.

For non-convex penalties, such as SCAD, the penalized score test gives only a necessary condition for regression parameters to be zero. In practice, however, the penalized score test is in some sense both necessary and sufficient to determine the sparsity pattern produced by non-convex penalties. Solutions to non-convex problems like SCAD are often found using coordinate-descent procedures, which solve for a local optimum of (2.25) by iteratively minimizing with respect to each element of $(a, \mathbf{b})$ [Zou and Li, 2008, Breheny and Huang, 2011, Mazumder et al., 2011]. If we use $(0, \hat{\mathbf{b}}^0_\lambda)$ as initial values, and then solve (2.25) using coordinate descent, the algorithm will converge to $(0, \hat{\mathbf{b}}^0_\lambda)$ when $|T_\lambda| \leq \sqrt{n}\lambda$. That is, $\{|T_\lambda| \leq \sqrt{n}\lambda\}$ is also sufficient for $\{\hat{a}_\lambda = 0\}$, when using this algorithm.

In order to obtain p-values for these other sparsity-inducing penalties, we require an appropriate reference distribution, which we leave for future work.

2.6 Discussion

In this paper, we presented the penalized score test, in which the hypothesis being tested depends on the value of the tuning parameter λ. Therefore, λ should be chosen to yield a test which is scientifically meaningful. For instance, if the simple linear regression parameter $\alpha + \boldsymbol{\sigma}_{xz}^T \boldsymbol{\beta}$ is a scientifically meaningful target of inference, one should choose λ to be large (i.e. perform simple linear regression). On the other hand, λ should be chosen as small as possible when the multiple linear regression parameter α in (2.3) is of interest. Perhaps the simplest way to choose λ in this case is to specify how many degrees of freedom we are willing to invest in estimating the nuisance parameter $\boldsymbol{\beta}$. Whether this controls type-I error of the penalized score test at an acceptable level is highly context-specific, and in general difficult to ascertain. With any $\lambda > 0$, some bias in $\hat{\mathbf{b}}^0_\lambda$ relative to $\boldsymbol{\beta}$ will be incurred, in which case the penalized score test may be thought of as a pragmatic approximation to classical tests, when multiple linear regression is undefined, or produces coefficient estimates which are too variable to be useful.

In this manuscript, we focused on testing the hypothesis $H_{0,\lambda} : a_\lambda = 0$ using a score test. The test does estimate effect sizes or provide confidence intervals. However, estimates of effect size can be obtained by fitting the sample version of (2.9), i.e. $(\hat{a}_\lambda, \hat{\mathbf{b}}_\lambda) = \arg\min_{(a,\mathbf{b})} \{\|\mathbf{y} - a\mathbf{x} - \mathbf{Z}\boldsymbol{\beta}\|_2^2/(2n) + \lambda J(\mathbf{b})\}$. Confidence intervals for a_λ when using the ridge penalty ($J(\mathbf{b}) = \|\mathbf{b}\|_2^2/2$) are available from mixed-model theory. For the lasso ($J(\mathbf{b}) = \|\mathbf{b}\|_1$), slight modifications of our theory in Section 2.3 can be used to show that $\sqrt{n}(\hat{a}_\lambda - a_\lambda) \to_d N\left(0, \sigma_\epsilon^2 \left[\mathbf{x}^T(\mathbf{I}_n - \mathbf{P}_\mathcal{A})\mathbf{x}/n\right]^{-1}\right)$, under (**A1-6**). This result might be used for a penalized version of the Wald test.

Several possible extensions of the proposed method are outlined here. Instead of testing for the effect of a single feature $\mathbf{x} \in \mathbb{R}^n$, we can test for groups of k variables $\mathbf{X} \in \mathbb{R}^{n\times k}$, with the score statistic $T_\lambda = \mathbf{X}^T(\mathbf{y} - \mathbf{Z}\hat{\mathbf{b}}_\lambda^0)/\sqrt{n} \in \mathbb{R}^k$. The distribution of T_λ, under an appropriate null hypothesis, follows in a straightforward way from Proposition 2 for a ridge penalty, or from Proposition 3, for a lasso penalty. From Proposition 1, we know that in the lasso regression of $\mathbf{y}$ on $(\mathbf{X}, \mathbf{Z})$, one or more of the coefficients associated with $\mathbf{X}$ is non-zero if and only if $\|T_\lambda\|_\infty > \sqrt{n}\lambda$, where T_λ is constructed using the lasso penalty. Analogous to Proposition 1, the sparsity pattern of the group lasso [Yuan and Lin, 2007b] and the standardized group lasso [Simon and Tibshirani, 2012b] could also be understood in terms of restrictions on this score statistic T_λ, for an appropriate choice of penalty function $J(\mathbf{b})$.

The lasso-penalized score test is implemented in the `lassoscore` `R` package, available on the Comprehensive R Archive Network (CRAN).

Chapter 3

INFERENCE FOR ℓ_1-PENALIZED M-ESTIMATORS

3.1 Introduction

Suppose we observe n independent and identically distributed random variables $X_i \in \mathbb{R}^m, 1 \leq i \leq n, X_i \sim \mathcal{P}$, and are interested in some functional $\boldsymbol{\theta} \equiv \boldsymbol{\theta}(\mathcal{P}) \in \mathbb{R}^d$ of their distribution

$$\boldsymbol{\theta} = \underset{\boldsymbol{\omega} \in \mathbb{R}^d}{\arg\min}\, \mathbb{E}_{\mathcal{P}} l(\boldsymbol{\omega}; X_i), \tag{3.1}$$

where $l : \mathbb{R}^d \times \mathbb{R}^m \mapsto \mathbb{R}$ is a loss function. For instance, in linear regression, we might have $X_i = (Y_i, Z_i)$, where $Y_i \in \mathbb{R}$ is some outcome of interest, and $Z_i \in \mathbb{R}^d$ is a vector of covariates, and seek the minimizer of $\mathbb{E}_{\mathcal{P}} l(\boldsymbol{\omega}; Y_i, Z_i) = \mathbb{E}_{\mathcal{P}}(Y_i - \boldsymbol{\omega}^T Z_i)^2$. In this case $\boldsymbol{\theta}$ is the vector of linear regression coefficients. Alternately, $\boldsymbol{\theta}(\mathcal{P})$ may define the parameters of generalized linear model, or the covariance matrix in a multivariate normal distribution, cases considered in Sections 3.3 and 3.4 respectively.

Typically, when $d \ll n$ one uses the empirical version of (3.1) as an estimate of $\boldsymbol{\theta}$

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\omega} \in \mathbb{R}^d}{\arg\min} \left\{ l(\boldsymbol{\omega})/n \right\}, \tag{3.2}$$

where $l(\boldsymbol{\omega}) = \sum_{i=1}^n l(\boldsymbol{\omega}; X_i)$. Formally, $\hat{\boldsymbol{\theta}}$ is known as an M-estimator, and classical theory giving the asymptotic distribution of $\hat{\boldsymbol{\theta}}$ when $d \ll n$ can be found in e.g. van der Vaart [2000]. Using this theory, one can perform formal statistical hypothesis tests regarding the parameter $\boldsymbol{\theta}$.

However, when the number of parameters d is large relative to the sample size n, these classical tests may have low power, or may even be undefined. In this setting,

ℓ_1-penalized M-estimators such as

$$\hat{\boldsymbol{\theta}}_\lambda = \underset{\boldsymbol{\omega}\in\mathbb{R}^d}{\arg\min} \left\{ l(\boldsymbol{\omega})/n + \lambda\|\boldsymbol{\omega}\|_1 \right\}, \tag{3.3}$$

where $\lambda \geq 0$ is a tuning parameter, have become popular. These produce a sparse estimate $\hat{\boldsymbol{\theta}}_\lambda$, which, under appropriate conditions, can identify which elements of $\boldsymbol{\theta}$ are zero [Zhao and Yu, 2006, Bunea et al., 2008, Lee et al., 2013a]. Thus, we might use whether an element of $\hat{\boldsymbol{\theta}}_\lambda$ is non-zero as evidence for whether an element of $\boldsymbol{\theta}$ is non-zero. However, quantifying the strength of this evidence, with p-values or confidence intervals, is an open problem.

For the special case of ℓ_1-penalized linear regression,

$$\hat{\boldsymbol{\theta}}_\lambda = \underset{\boldsymbol{\omega}\in\mathbb{R}^d}{\arg\min} \left\{ \frac{1}{2n} \sum_{i=1}^{n} \left(Y_i - \boldsymbol{\omega}^T Z_i\right)^2 + \lambda\|\boldsymbol{\omega}\|_1 \right\}, \tag{3.4}$$

a few methods for inference have been proposed, based on inverting the stationary conditions for lasso regression [van de Geer et al., 2013, Zhang and Zhang, 2011, Javanmard and Montanari, 2013], conditioning on the selected model [Lee et al., 2013b, Taylor et al., 2014], or considering the change in model fit along the lasso solution path [Lockhart et al., 2013]. However, while these approaches are based on ℓ_1-penalized regression, they do not provide a connection between inference for the coefficient $\boldsymbol{\theta}$ and the sparsity pattern of $\hat{\boldsymbol{\theta}}_\lambda$. On the other hand, Meinshausen and Bühlmann [2010] proposed *stability selection*, a sub-sampling procedure which can control the family-wise error rate of *any* selection procedure. While stability selection can be applied to penalized M-estimators, it can be computationally expensive, and in some cases overly conservative.

The approach we follow in this paper builds on the recently proposed *penalized score test* for linear regression from Chapter 2, which can be used as an alternative to classical hypothesis tests in the high-dimensional setting. In order to use this test,

one performs penalized regression of the outcome on all but a single feature, and then computes the correlation of the residuals with the held-out feature. This procedure is then applied to all features in turn. Specifically, if $\mathbf{y} \in \mathbb{R}^n$ is the outcome, $\mathbf{x} \in \mathbb{R}^n$ is a feature of interest, and $\mathbf{Z} \in \mathbb{R}^{n \times (d-1)}$ is a matrix containing $d-1$ other features, then the penalized score statistic takes the form

$$T_\lambda = \frac{1}{\sqrt{n}} \mathbf{x}^T (\mathbf{y} - \mathbf{Z}\hat{\mathbf{b}}_\lambda^0),$$

where $\hat{\mathbf{b}}_\lambda^0 = \arg\min_{\mathbf{b}} \{\|\mathbf{y} - \mathbf{Z}\mathbf{b}\|_2^2/(2n) + \lambda J(\mathbf{b})\}$. The null hypothesis is rejected when $|T_\lambda|$ is large, with respect to an appropriate reference distribution. The null hypothesis being tested is of the form $H_{0,\lambda} : a_\lambda = 0$ vs $H_{1,\lambda} : a_\lambda \neq 0$, where a_λ can be viewed as a compromise between the effect of $\mathbf{x}$ in simple linear regression, and the effect of feature $\mathbf{x}$ in multiple linear regression of $\mathbf{y}$ on $\mathbf{x}$ and $\mathbf{Z}$ together. We showed that when $J(\mathbf{b}) = \|\mathbf{b}\|_1$, the sparsity pattern of lasso regression on all features results from a decision based on the penalized score test. Specifically, if $(\hat{a}_\lambda, \hat{\mathbf{b}}_\lambda) = \arg\min_{(a,\mathbf{b})} \{\|\mathbf{y} - a\mathbf{x} - \mathbf{Z}\mathbf{b}\|_2^2/(2n) + \lambda \|(a, \mathbf{b})\|_1\}$, then $\hat{a}_\lambda \neq 0$ if and only if $|T_\lambda| > \sqrt{n}\lambda$.

In this paper, we extend the framework of Chapter 2 to penalized M-estimation. We present a general approach for hypothesis testing in the high-dimensional setting. We will see that, just as in linear regression, model selection with ℓ_1-penalized M-estimators can be understood as a decision based on a statistical test.

The rest of the chapter is organized as follows. In Section 3.2 we develop general theory for testing in the context of high-dimensional M-estimators. We then investigate two special cases in more detail: in Section 3.3 we consider ℓ_1-penalized GLMs, in particular logistic regression, and in Section 3.4 we apply our framework to the Gaussian graphical model setting. We end with a discussion in Section 3.5.

3.2 Inference for ℓ_1-penalized M-estimators

As shown in Section 2.3 of Chapter 2, the sparsity pattern of ℓ_1-penalized linear regression, or the lasso, can be interpreted as a decision based on a modified score statistic. Here, we show how the same framework applies to M-estimators, and state Theorem 5, which gives the asymptotic distribution of the proposed test statistic.

Partition $\boldsymbol{\theta}$ as $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta})$ where $\alpha \in \mathbb{R}$ and $\boldsymbol{\beta} \in \mathbb{R}^{d-1}$, and suppose that we are interested in testing the hypothesis $H_0 : \alpha = 0$ vs. $H_1 : \alpha \neq 0$. Let $l(\alpha, \boldsymbol{\beta}) = l(\boldsymbol{\theta}) = \sum_{i=1}^{n} l(\boldsymbol{\theta}; X_i)$ denote the loss function, as in (3.2). When the dimension d is much smaller than n, one can test H_0 using the classical score statistic, based on the score, or derivative, of the loss function, evaluated under the restrictions imposed by the null hypothesis. The score statistic can be written as

$$T = \dot{l}_a(0, \hat{\mathbf{b}}^0)/\sqrt{n}, \tag{3.5}$$

where $\dot{l}_a(a, \mathbf{b}) = \frac{\partial}{\partial a} l(a, \mathbf{b})$, and $\hat{\mathbf{b}}^0 = \arg\min_{\mathbf{b}}\{l(0, \mathbf{b})/n\}$. The score statistic (3.5) measures how much the loss would be reduced if we were to relax the restrictions imposed by the null hypothesis. Under mild regularity conditions, given in van der Vaart [2000], T is asymptotically normally distributed with mean zero when $\alpha = 0$, and otherwise diverges to $\pm\infty$. Thus, one typically rejects H_0 when $|T|$ is large, with respect to the appropriate normal reference distribution.

Unfortunately, when the dimension d is large relative to n, the estimate $\hat{\mathbf{b}}^0$ of $\boldsymbol{\beta}$ may be highly variable, or even undefined. As an alternative, we can introduce bias in our estimate of $\boldsymbol{\beta}$ in order to reduce variance. One way of doing this is to use an ℓ_1-penalized estimate $\hat{\mathbf{b}}^0_\lambda = \arg\min_{\mathbf{b}}\{l(0, \mathbf{b})/n + \lambda\|\mathbf{b}\|_1\}$. The penalty $\|\mathbf{b}\|_1$ forces the estimate $\hat{\mathbf{b}}^0_\lambda$ to contain many zero entries when λ is large enough, and under certain assumptions, reasonably approximates $\boldsymbol{\beta}$ [van de Geer, 2008, Negahban et al., 2012].

This procedure suggests the penalized score statistic

$$T_\lambda = \dot{l}_a(0, \hat{\mathbf{b}}^0_\lambda)/\sqrt{n}. \tag{3.6}$$

Here, (3.6) can also be obtained by considering the score for the ℓ_1-penalized loss function, and so we call it a penalized score test. In Theorem 5 we give the asymptotic distribution of (3.6), under an appropriate null hypothesis, which can be used to calculate p-values.

The sparsity pattern of ℓ_1 penalized M-estimators corresponds precisely to a decision based on the penalized score statistic (3.6). We state this connection in Theorem 4, the proof of which follows immediately from the Karush-Kuhn-Tucker conditions.

Theorem 4. *Suppose $l(a, \mathbf{b}) + \lambda\|(a, \mathbf{b})\|_1$ is strictly convex, and let*

$$(\hat{a}_\lambda, \hat{\mathbf{b}}_\lambda) = \underset{(a,\mathbf{b})\in\mathbb{R}^d}{\arg\min}\{l(a, \mathbf{b})/n + \lambda\|(a, \mathbf{b})\|_1\}. \tag{3.7}$$

Then $\hat{a}_\lambda \neq 0$ if and only if $|T_\lambda| > \sqrt{n}\lambda$.

As discussed in Section 3.1, the estimate (3.7) has been suggested as a procedure to select which elements of $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta})$ are zero. Theorem 4 shows that the criteria on which the sparsity pattern of $\hat{\boldsymbol{\theta}}_\lambda = (\hat{a}_\lambda, \hat{\mathbf{b}}_\lambda)$ is based mirrors a hypothesis test using the classical score statistic (3.5).

3.2.1 The null hypothesis

In this section we describe how the use of the biased estimate $\hat{\mathbf{b}}_\lambda$ of $\boldsymbol{\beta}$ affects the null hypothesis being tested by the penalized score statistic (3.6), for a given tuning parameter λ. While $\hat{\mathbf{b}}_\lambda$ may be asymptotically unbiased when $\lambda \to 0$ [van de Geer, 2008, Negahban et al., 2012], some bias will always be incurred in finite samples if

$\lambda > 0$; we characterize the bias by considering the asymptotic scenario with $\lambda \geq 0$ fixed.

Let $(a_\lambda, \mathbf{b}_\lambda)$ be the population-level parameters:

$$(a_\lambda, \mathbf{b}_\lambda) = \underset{(a,\mathbf{b})\in\mathbb{R}^d}{\arg\min} \left\{ \mathbb{E}_{\mathcal{P}} l(a, \mathbf{b})/n + \lambda \|\mathbf{b}\|_1 \right\}. \tag{3.8}$$

When $a_\lambda = 0$, the stationary conditions of $\mathbb{E}_{\mathcal{P}} l(a, \mathbf{b})/n + \lambda\|\mathbf{b}\|_1$ imply that $\mathbb{E}_{\mathcal{P}} \dot{l}_a(0, \mathbf{b}_\lambda) = 0$, provided we can exchange the derivative and integral. This means that the penalized score statistic T_λ in (3.6) can be seen as the empirical version of $\dot{l}_a(0, \mathbf{b}_\lambda)$, with $\mathbf{b}_\lambda$ replaced by the estimate $\hat{\mathbf{b}}^0_\lambda$. Thus, provided $\hat{\mathbf{b}}^0_\lambda$ converges quickly enough to $\mathbf{b}_\lambda$, T_λ should be asymptotically centered around zero when $a_\lambda = 0$, and otherwise should diverge to $\pm\infty$. It follows that the penalized score statistic T_λ tests

$$H_{0,\lambda} : a_\lambda = 0 \qquad \text{vs} \qquad H_{1,\lambda} : a_\lambda \neq 0.$$

We now use Taylor's expansion in order to write the parameter a_λ in terms of $(\alpha, \boldsymbol{\beta})$. Let $\ddot{\mathbf{l}}_{a\mathbf{b}}(a, \mathbf{b}) = \frac{\partial^2}{\partial a \partial \mathbf{b}} l(a, \mathbf{b})$ and $\ddot{\mathbf{l}}_{aa}(a, \mathbf{b}) = \frac{\partial^2}{(\partial a)^2} l(a, \mathbf{b})$. From the definition of $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta})$, we know that $\mathbb{E}_{\mathcal{P}}[\dot{\mathbf{l}}(\alpha, \beta)] = \mathbf{0}$. Now, for some $(\tilde{a}, \tilde{\mathbf{b}})$ between $(a_\lambda, \mathbf{b}_\lambda)$ and $(\alpha, \boldsymbol{\beta})$, we can write

$$\begin{aligned} 0 &= \mathbb{E}_{\mathcal{P}} \dot{l}_a(\alpha, \boldsymbol{\beta}) \\ &= \mathbb{E}_{\mathcal{P}} \left[\dot{l}_a(a_\lambda, \mathbf{b}_\lambda) + \ddot{\mathbf{l}}_{aa}(\tilde{a}, \tilde{\mathbf{b}})(a_\lambda - \alpha) + \ddot{\mathbf{l}}^T_{a\mathbf{b}}(\tilde{a}, \tilde{\mathbf{b}})(\mathbf{b}_\lambda - \boldsymbol{\beta}) \right]. \end{aligned}$$

By the definition of a_λ, we know that $\mathbb{E}_{\mathcal{P}} \dot{l}_a(a_\lambda, \mathbf{b}_\lambda) = 0$, and so we can rearrange terms to write

$$a_\lambda = \alpha + \boldsymbol{\eta}^T(\boldsymbol{\beta} - \mathbf{b}_\lambda),$$

where $\boldsymbol{\eta} = \mathbb{E}_{\mathcal{P}} \ddot{\mathbf{l}}_{a\mathbf{b}}(\tilde{a}, \tilde{\mathbf{b}}) / \mathbb{E}_{\mathcal{P}} \ddot{l}_{aa}(\tilde{a}, \tilde{\mathbf{b}})$ is a measure of dependence between the parameters α and $\boldsymbol{\beta}$. That is, a_λ differs from α due to (i) the bias introduced by penalization

$\boldsymbol{\beta} - \mathbf{b}_\lambda$, and (ii) the dependence between the parameters $\boldsymbol{\eta}$.

In many models, such as those considered in Sections 3.3 and 3.4, α measures the conditional association between two features. Note that when $\lambda = 0$, we have $\mathbf{b}_\lambda = \boldsymbol{\beta}$, and thus $a_\lambda = \alpha$. On the other hand, when λ is large, a_λ can often be interpreted as a measure of marginal association between two features. Though the penalized score test is not guaranteed to be unbiased for the test $H_0 : \alpha = 0$ vs $H_1 : \alpha \neq 0$, it is still useful in practical high-dimensional settings, as we show in Sections 3.3.2 and 3.4.3.

3.2.2 The distribution of T_λ

Here we present Theorem 5, which gives the asymptotic distribution of T_λ under $H_{0,\lambda}$. First, we will introduce some notation and assumptions. Without loss of generality, suppose the first q elements of $\mathbf{b}_\lambda$ are non-zero, and the rest are zero, i.e. $\mathbf{b}_\lambda = (b_{\lambda,1}, \ldots, b_{\lambda,q}, 0, \ldots, 0)^T = (\mathbf{b}_{\lambda\mathcal{A}}, \mathbf{b}_{\lambda\mathcal{A}^c})$, where $b_{\min} = \min\{\, |b_{\lambda,1}|, \ldots, |b_{\lambda,q}| \,\} > 0$ and $\mathcal{A} = \text{supp}(\mathbf{b}_\lambda)$. Partition the vector of derivatives as $\dot{\mathbf{l}}(a, \mathbf{b}) = [\dot{l}_a(a, \mathbf{b}), \dot{\mathbf{l}}^T_{\mathcal{A}}(a, \mathbf{b}), \dot{\mathbf{l}}^T_{\mathcal{A}^c}(a, \mathbf{b})]^T$, and denote the matrix of second derivatives as

$$\ddot{\mathbf{l}}(a, \mathbf{b}) = \begin{pmatrix} \ddot{l}_{aa} & \ddot{\mathbf{l}}^T_{\mathcal{A}a} & \ddot{\mathbf{l}}^T_{\mathcal{A}^c a} \\ \ddot{\mathbf{l}}_{\mathcal{A}a} & \ddot{\mathbf{l}}_{\mathcal{A}\mathcal{A}} & \ddot{\mathbf{l}}^T_{\mathcal{A}^c a} \\ \ddot{\mathbf{l}}_{\mathcal{A}^c a} & \ddot{\mathbf{l}}_{\mathcal{A}^c \mathcal{A}} & \ddot{\mathbf{l}}^T_{\mathcal{A}^c \mathcal{A}^c} \end{pmatrix} (a, \mathbf{b}) \in \mathbb{R}^{d \times d}.$$

We will require the following regularity conditions on $l(a, \mathbf{b})$.

(**B1**) $l(a, \mathbf{b})$ is twice continuously differentiable in $(a, \mathbf{b})$, almost surely, in a neighborhood of $(a_\lambda, \mathbf{b}_\lambda)$. Further, the derivative and integral can be exchanged, so that (3.8) implies $\mathbb{E}_{\mathcal{P}}[\dot{\mathbf{l}}(a_\lambda, \mathbf{b}_\lambda)/n] = -\lambda\boldsymbol{\tau}$, where $\boldsymbol{\tau}_{\mathcal{A}} = \text{sign}(\mathbf{b}_{\lambda\mathcal{A}}) \in \mathbb{R}^q$, $\boldsymbol{\tau}_{\mathcal{A}^c} \in [-1, 1]^{d-q-1}$, and $\tau_a = 0$.

(**B2**) The information matrix

$$\mathbf{V} \equiv \mathrm{var}_{\mathcal{P}} \begin{bmatrix} \dot{l}_a(a_\lambda, \mathbf{b}_\lambda)/\sqrt{n} \\ \dot{\mathbf{l}}_{\mathcal{A}}(a_\lambda, \mathbf{b}_\lambda)/\sqrt{n} \end{bmatrix} \in \mathbb{R}^{(q+1)\times(q+1)}$$

exists and is positive definite.

(**B3**) The matrix

$$\mathbf{U} \equiv \mathbb{E}_{\mathcal{P}} \left[\frac{1}{n} \begin{pmatrix} \ddot{l}_{aa} & \ddot{\mathbf{l}}_{\mathcal{A}a}^T \\ \ddot{\mathbf{l}}_{\mathcal{A}a} & \ddot{\mathbf{l}}_{\mathcal{A}\mathcal{A}} \end{pmatrix} (a_\lambda, \mathbf{b}_\lambda) \right] \in \mathbb{R}^{(q+1)\times(q+1)}$$

exists and is positive definite.

Conditions (**B1-B3**) are somewhat standard regularity conditions for M-estimators [see e.g. Theorem 5.41 of van der Vaart, 2000]. However, our conditions on the derivatives of $l(a, \mathbf{b})$ differ from classical treatments in a few important ways. First, since we use a penalty, the scores will not, in general, have zero expectation: $\mathbb{E}_{\mathcal{P}}\left[\dot{\mathbf{l}}(a_\lambda, \mathbf{b}_\lambda)\right] = -\lambda\boldsymbol{\tau}$, whereas $\mathbb{E}_{\mathcal{P}}[\dot{\mathbf{l}}(\alpha, \boldsymbol{\beta})] = 0$. Second, our conditions on the derivatives of $l(a, \mathbf{b})$ depend on the support of $\mathbf{b}_\lambda$. Thus, we allow $\mathbb{E}_{\mathcal{P}}[\ddot{\mathbf{l}}(a_\lambda, \mathbf{b}_\lambda)]$ to be rank-deficient, so long as $\mathbf{U}$ is full-rank.

In order to allow d to grow more quickly than n, we require the following additional conditions.

(**B4**) The expectation of the score, $\mathbb{E}_{\mathcal{P}}[\,\dot{\mathbf{l}}(a_\lambda, \mathbf{b}_\lambda)/n\,] = -\lambda\boldsymbol{\tau}$, satisfies $\|\boldsymbol{\tau}_{\mathcal{A}^c}\|_\infty \leq 1 - \delta$ for some $\delta > 0$.

(**B5**) The elements of the score vectors $\dot{\mathbf{l}}(a_\lambda, \mathbf{b}_\lambda; X_i)$ have sub-Gaussian tails. That is, there exists a c such that for any $x > 0$ and each $\mathbf{e}_j = (0, \ldots, 0, 1, 0, \ldots, 0)^T \in \mathbb{R}^d$, $j = 1, \ldots, d$, we have that $\Pr(|\mathbf{e}_j^T\{\dot{\mathbf{l}}(a_\lambda, \mathbf{b}_\lambda; X_i) + \lambda\boldsymbol{\tau}\}| > x) < ge^{-hx^2}$ for all $x > c$ and for some $h, g > 0$ not depending on j, or x.

(**B6**) The minimum eigenvalue of $\mathbf{U}_{\mathcal{A}\mathcal{A}}$ is bounded, i.e. $\Lambda_{\min}(\mathbf{U}_{\mathcal{A}\mathcal{A}}) \geq \zeta > 0$. Further, the sample size n, the dimension d, the number of non-zero parameters q, and the minimum non-zero coefficient $b_{\min} \equiv \min\{\, |b_{\lambda,1}|, \ldots, |b_{\lambda,q}| \,\}$ satisfy

$$\frac{\log(d)}{n} + \frac{q \log q}{n b_{\min}^2} + \frac{q^2}{n} \to 0.$$

(**B7**) The matrix of second derivatives $\ddot{\mathbf{l}}(a, \mathbf{b})$ satisfies $\|\ddot{\mathbf{l}}_{\mathcal{A}\mathcal{A}^c}(a, \mathbf{b})/n\|_2 = o_p\left(\sqrt{n/(q \log q)}\right)$ in a neighborhood of $(a_\lambda, \mathbf{b}_\lambda)$.

(**B8**) The matrix of second derivatives $\ddot{\mathbf{l}}(a, \mathbf{b})$ is Lipschitz continuous in a neighborhood of $(a_\lambda, \mathbf{b}_\lambda)$. Further, $\ddot{\mathbf{l}}_{\mathcal{A}\mathcal{A}}(a_\lambda, \mathbf{b}_\lambda)/n$ converges sufficiently quickly to its expected value

$$\left\| \frac{1}{n} \begin{pmatrix} \ddot{l}_{aa} & \ddot{\mathbf{l}}_{\mathcal{A}a}^T \\ \ddot{\mathbf{l}}_{\mathcal{A}a} & \ddot{\mathbf{l}}_{\mathcal{A}\mathcal{A}} \end{pmatrix} (a_\lambda, b_\lambda) - \mathbf{U}_{\mathcal{A}\mathcal{A}} \right\|_2 = O_p(\sqrt{(q \log q)/n}).$$

Assumptions (**B4-7**) occur in similar form in Section 2.3 of Chapter 2, and are discussed in more detail there, in the context of linear regression. Here, we require the additional condition (**B8**) in order to show that $\ddot{\mathbf{l}}(\hat{a}_\lambda, \hat{\mathbf{b}}_\lambda)/n$ is sufficiently close to $\mathbb{E}_{\mathcal{P}}[\,\ddot{\mathbf{l}}(a_\lambda, \mathbf{b}_\lambda)/n\,]$. Lipschitz continuity in (**B8**) is also required by van de Geer [2008] in order to prove risk-consistency of ℓ_1-penalized M-estimators.

With these assumptions, we can now state our main theoretical result, which is proven in Section A.2 of the Appendix.

Theorem 5. *Suppose conditions (***B1-8***) hold, and* $H_{0,\lambda} : a_\lambda = 0$ *is true. Then, there exists a minimizer* $\hat{\mathbf{b}}_\lambda^0 = \arg\min_{\mathbf{b}}\{l(0, \mathbf{b})/n + \lambda\|\mathbf{b}\|_1\}$ *such that*

$$\frac{T_\lambda}{\sqrt{v}} \to_d N(0, 1),$$

where

$$v = v_{aa} + \mathbf{u}_{\mathcal{A}a}^T \mathbf{U}_{\mathcal{A}\mathcal{A}}^{-1}(\mathbf{V}_{\mathcal{A}\mathcal{A}}\mathbf{U}_{\mathcal{A}\mathcal{A}}^{-1}\mathbf{u}_{\mathcal{A}a} - 2\mathbf{v}_{\mathcal{A}a}). \tag{3.9}$$

Note that the term $\mathbf{U}_{\mathcal{A}\mathcal{A}}^{-1}\mathbf{V}_{\mathcal{A}\mathcal{A}}\mathbf{U}_{\mathcal{A}\mathcal{A}}^{-1}$ is a 'sandwich' variance formula, for misspecified models [Huber, 1967, White, 1980], which arises in other treatments of lasso-type estimators [e.g. Fan and Li, 2001]. Since in general $(a_\lambda, \mathbf{b}_\lambda) \neq (\alpha, \boldsymbol{\beta})$, we know that $l(a_\lambda, \mathbf{b}_\lambda)$ is not the negative log-likelihood of our data. Thus our model is in some sense misspecified, due to the penalty $\lambda\|\mathbf{b}\|_1$.

In order to apply Theorem 5 in practice, we need estimates of $\text{supp}(\mathbf{b}_\lambda) = \mathcal{A}$, $\mathbf{U}$, and $\mathbf{V}$. We suggest using $\hat{\mathcal{A}} = \text{supp}(\hat{\mathbf{b}}_\lambda^0)$ as an estimate of $\mathcal{A}$, and the 'plug-in' estimates of $\mathbf{U}$ and $\mathbf{V}$ given by

$$\hat{\mathbf{U}} = \frac{1}{n}\begin{bmatrix} \ddot{l}_{aa}(0, \hat{\mathbf{b}}_\lambda^0) & \ddot{\mathbf{l}}_{a\hat{\mathcal{A}}}(0, \hat{\mathbf{b}}_\lambda^0) \\ \ddot{\mathbf{l}}_{a\hat{\mathcal{A}}}^T(0, \hat{\mathbf{b}}_\lambda^0) & \ddot{\mathbf{l}}_{\hat{\mathcal{A}}\hat{\mathcal{A}}}(0, \hat{\mathbf{b}}_\lambda^0) \end{bmatrix} \tag{3.10}$$

$$\hat{\mathbf{V}} = \frac{1}{n}\sum_{i=1}^n (\mathbf{s}_i - \bar{\mathbf{s}})(\mathbf{s}_i - \bar{\mathbf{s}})^T, \tag{3.11}$$

where $\mathbf{s}_i = \begin{pmatrix} \dot{l}_a(0, \hat{\mathbf{b}}_\lambda^0; X_i) \\ \dot{\mathbf{l}}_{\hat{\mathcal{A}}}(0, \hat{\mathbf{b}}_\lambda^0; X_i) \end{pmatrix}$, and $\bar{\mathbf{s}} = \sum_{i=1}^n \mathbf{s}_i/n$. Using $\hat{\mathcal{A}}$ as an estimate of $\mathcal{A}$ is supported by Lemma 6 of the Appendix, where we show that $\Pr[\hat{\mathcal{A}} = \mathcal{A}] \to 1$. The estimate (3.11) is sometimes referred to as the Huber-White estimator, or the outer-product estimator [Huber, 1967, White, 1980]. In regression models, we find that the degrees-of-freedom adjustment $\hat{\mathbf{V}}' = n\hat{\mathbf{V}}/(n - |\hat{\mathcal{A}}|)$ can give better finite-sample performance than (3.11) [MacKinnon and White, 1985].

In Sections 3.3 and 3.4, we also give *model-based* standard errors for the special cases where $l(\cdot)$ is either the log-likelihood in a generalized linear model, or the log-likelihood of a multivariate Gaussian distribution.

3.3 Inference in high-dimensional generalized linear models

In this section we consider a special case of the theory developed in Section 3.2 for inference for generalized linear models (GLMs). Suppose we have an outcome $\mathbf{y} \in \mathbb{R}^n$, which we wish to predict using $\mathbf{X} \in \mathbb{R}^{n \times d}$. In a GLM, one relates the expectation of $\mathbf{y}$ to the covariates $\mathbf{X}$ through a link function g such that $g\left(\mathbb{E}_{\mathcal{P}}[\mathbf{y} \mid \mathbf{X}]\right) = g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\theta}$ [McCullagh and Nelder, 1989]. An important feature of GLMs is that the variance of the outcome $\mathbf{y}$ is allowed to depend on the mean $\boldsymbol{\mu}$ through the *variance function* $\text{var}(\mathbf{y}) = \phi V(\boldsymbol{\mu})$, where ϕ is the *dispersion parameter*. When the *canonical* link function $g(\cdot)$ is used, the loss function used by a GLM has the form

$$l(\boldsymbol{\theta}) = -\mathbf{y}^T \mathbf{X}\boldsymbol{\theta} + \mathbf{1}_n^T A(\mathbf{X}\boldsymbol{\theta}),$$

where $A(x)$ is a convex function satisfying $\partial A(x)/\partial x = g^{-1}(x)$, and $A(\mathbf{X}\boldsymbol{\theta}) \in \mathbb{R}^n$ denotes element-wise application of $A(x)$ to the vector $\mathbf{X}\boldsymbol{\theta}$. For simplicity, in this section we only consider canonical link functions. Table 3.1 gives some common examples of $g(x)$, $A(x)$, and $V(\mu)$, as well as the first and second derivatives of $l(\cdot)$, necessary to construct T_λ and the components of its variance.

Family	$g^{-1}(x)$	$A(x)$	$V(\mu)$	$\dot{\mathrm{l}}(\boldsymbol{\theta})$	$\ddot{\mathrm{l}}(\boldsymbol{\theta})$
Gaussian	x	$x^2/2$	1	$-\mathbf{X}^T(\mathbf{y} - \hat{\mathbf{y}})$	$\mathbf{X}^T \text{diag}\left(V(\hat{\mathbf{y}})\right)\mathbf{X}$
Binomial	expit(x)	$\log(1 + e^x)$	$\mu(1-\mu)$		
Poisson	e^x	e^x	μ		

Table 3.1: Cannonical link functions, variance functions, and derivatives of loss functions for common GLMs. Here $\hat{\mathbf{y}} = \hat{\mathbf{y}}(\boldsymbol{\theta}) = g^{-1}(\mathbf{X}\boldsymbol{\theta})$ denotes the fitted values.

As in Section 3.2, we partition the parameter $\boldsymbol{\theta}$ as $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta})$, where α corresponds to the regression coefficient of a single feature of interest $\mathbf{x} \in \mathbb{R}^n$, and $\boldsymbol{\beta}$ corresponds to the regression coefficients of all other features $\mathbf{Z} \in \mathbb{R}^{n \times (d-1)}$. In order to perform

the penalized score test in this setting, we first fit the restricted model

$$\hat{\mathbf{b}}_\lambda^0 = \underset{\mathbf{b}\in\mathbb{R}^{d-1}}{\arg\min} \left\{ \left[-\mathbf{y}^T\mathbf{Z}\mathbf{b} + \mathbf{1}_n^T A(\mathbf{Z}\mathbf{b}) \right] / n + \lambda \|\mathbf{b}\|_1 \right\}, \tag{3.12}$$

and then compare the penalized score statistic $T_\lambda = \dot{l}_a(0, \hat{\mathbf{b}}_\lambda^0)/\sqrt{n}$ to an appropriate reference distribution. From Table 3.1, we see that when using the canonical link function, the score statistic has a simple form: $T_\lambda = \mathbf{x}^T(\mathbf{y} - \hat{\mathbf{y}}_\lambda^0)/\sqrt{n}$, where $\hat{\mathbf{y}}_\lambda^0 = g^{-1}(\mathbf{Z}\hat{\mathbf{b}}_\lambda^0)$ are the fitted values from the regression (3.12). That is, the score statistic is simply a measure of correlation between the feature of interest $\mathbf{x}$ and the residuals $\mathbf{y} - \hat{\mathbf{y}}_\lambda^0$ from the penalized regression of $\mathbf{y}$ on $\mathbf{Z}$.

3.3.1 Variance estimation

We can apply Theorem 5 in order to obtain the asymptotic distribution of T_λ under $H_{0,\lambda}$. Applying this theorem requires an estimate of the variance of T_λ, which we consider next. In Section 3.2.2, we proposed using (3.10) and (3.11) to estimate the variance of T_λ. We will call that estimate the *sandwich variance*. However, for correctly-specified GLMs with a fixed design matrix the variance of the score has an explicit form:

$$\mathrm{var}_{\mathcal{P}} \left[\dot{\mathbf{l}}(0, \mathbf{b}_\lambda)/\sqrt{n} \right] = \mathbf{X}^T \mathrm{diag}(\mathrm{var}_{\mathcal{P}}(\mathbf{y}))\mathbf{X}/n = \phi \mathbf{X}^T \mathrm{diag}(V(\boldsymbol{\mu}))\mathbf{X}/n = \phi \ddot{\mathbf{l}}(\boldsymbol{\theta})/n.$$

Provided an estimate of $\boldsymbol{\theta}$ is available, this suggests an alternative to the empirical variance of the score in Equation 3.11. Though any number of estimators of $\boldsymbol{\theta}$ could be used, for simplicity we propose using $(0, \hat{\mathbf{b}}_\lambda^0)$, which leads to the estimate $\hat{\mathbf{V}} = \phi\hat{\mathbf{U}}$, where $\hat{\mathbf{U}}$ is as in (3.10). Plugging this into (3.9) we get

$$\widehat{\mathrm{var}}[\, T_\lambda \,] = \phi(\hat{u}_{aa} - \hat{\mathbf{u}}_{a\hat{\mathcal{A}}}^T \hat{\mathbf{U}}_{\hat{\mathcal{A}}\hat{\mathcal{A}}} \hat{\mathbf{u}}_{a\hat{\mathcal{A}}}) \tag{3.13}$$

where ϕ is the (known) dispersion parameter. This formulation for the variance of T_λ is precisely analogous to the 'asymptotic variance' discussed in Chapter 2. Here, we refer to (3.13) as the *model-based* variance.

In summary, we will consider two estimates of the variance of T_λ:

- *Sandwich variance*: Use (3.10) and (3.11) in the formulas in Theorem 5. This variance estimate does not require the GLM to be correctly specified, and allows for a stochastic design matrix.

- *Model-based variance*: Replace the empirical estimate of the score variance (3.11) with a model-based approximation, which yields (3.13). This formulation assumes a fixed design matrix, and that $l(\boldsymbol{\theta})$ is the log-likelihood of $\mathbf{y}$, up to constants.

For many forms of $l(\cdot)$, fitting the model (3.12) can be done efficiently using the `glmnet R` package [Friedman et al., 2010]. The penalized score statistic and its variance can be calculated using the `lassoscore R` package [Voorman, 2014], which is available on CRAN.

3.3.2 Simulation experiments

In this section, we illustrate the penalized score test for logistic regression. In Section 3.2.1 we argued that the penalized score test quantifies evidence against the null hypothesis $H_{0,\lambda} : a_\lambda = 0$, while in practice, we may be more interested in the null hypothesis $H_0 : \alpha = 0$, where in general $\alpha \neq a_\lambda$. Here, we will see that when $\alpha = 0$ the penalized score test behaves like a test of $H_0 : \alpha = 0$, provided λ is small enough.

We generated the design matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ according to one of two schemes:

- *Correlated features*: Rows of $\mathbf{X}$ were independently distributed $N_d(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}_{jk} = 0.5^{|j-k|}$.

- *Independent features*: Rows of $\mathbf{X}$ were independently distributed $N_d(\mathbf{0}, \mathbf{I}_d)$.

We then generated the outcome $\mathbf{y}$ as a vector of independent Bernoulli random variables, with probability of success $\exp(\mathbf{X}\theta)/(1+\exp(\mathbf{X}\theta))$, where we randomly selected 10 elements of $\boldsymbol{\theta}$ to be 0.5, and the rest to be zero.

The scheme with correlated features serves to demonstrate that the penalized score test can behave like a test of $H_0 : \alpha = 0$, even when $a_\lambda \neq \alpha$. The setting with independent features serves to corroborate our theory, since, in this scheme $H_{0,\lambda} : a_\lambda = 0$ is true when $\alpha = \theta_j = 0$.

We let $d = 100$, and set n to be either 200 or 400. Though the sample size is larger than the number of features, multiple logistic regression is unstable. In fact, when $n = 200$, and we perform logistic regression of $\mathbf{y}$ on all 100 features, the Fisher scoring algorithm used by `R`'s `glm` function yields fitted probabilities near 0 or 1, and does not converge. Thus, if we wish to test which elements of $\boldsymbol{\theta}$ are zero, some alternative to classical testing procedures is required.

On each of the 100 features, we performed the penalized score test for several values of λ, over 2000 replications where both $\mathbf{X}$ and $\mathbf{y}$ were sampled from the distribution described above, with fixed coefficient vector $\boldsymbol{\theta}$. For the sake of comparison, we also performed multiple logistic regression of $\mathbf{y}$ on all features (for $n = 400$ only), simple logistic regression of each feature on the outcome, and an *oracle test* in which we calculated the classical score test for each feature, adjusting only for other features with non-zero coefficients in $\boldsymbol{\theta}$. That is, the oracle test reflects the unattainable scenario where we test $H_0 : \theta_j = 0$ for $j = 1, \ldots, 100$, and know the support of $\boldsymbol{\theta}_{-j} = \{\theta_k : k \neq j\}$.

We declared a test to be significant when its p-value was less than 0.01. Thus, for an unbiased test, we would expect on average $0.01 \times 90 = 0.9$ false positives per simulated data set, which we will refer to as the nominal, or advertised error rate. We calculated the average number of false positives per simulated data set, or Expected

False Positives (EFP), and the power for each test. Specifically, for a given test, if p_{jm} is the p-value resulting from the j^{th} feature on the m^{th} simulation run, then the EFP and power are given by

$$\text{EFP} = \frac{1}{B}\sum_{m=1}^{B}\sum_{j:\theta_j=0} 1\{p_{jm} < 0.01\}$$

$$\text{power} = \frac{1}{B}\sum_{m=1}^{B}\frac{1}{\|\boldsymbol{\theta}\|_0}\sum_{j:\theta_j\neq 0} 1\{p_{jm} < 0.01\}$$

where $1\{\cdot\}$ is the indicator function, $B = 2000$ is the number of simulations, and $\|\boldsymbol{\theta}\|_0 = 10$ is the number of non-zero parameters.

Figure 3.1 summarizes the results of the experiment. When the features are correlated, $\mathbf{y}$ is marginally dependent on all features, while it is independent of $\{\mathbf{x}_j : \theta_j = 0\}$ after conditioning on $\{\mathbf{x}_j : \theta_j \neq 0\}$. Thus, simple logistic regression, which tests for marginal associations, has higher EFP than the nominal rate. Further, EFP increases with the sample size, as the power of simple logistic regression to detect these marginal associations increases. Similarly, when testing for the effect of features $\mathbf{x}_k : \theta_k = 0$, the penalized score test does not fully account for the effects of $\{\mathbf{x}_j : \theta_j \neq 0\}$, and thus tends to have higher EFP than the advertised rate. For $\lambda = 0.1$ the results of the penalized score test are qualitatively similar to simple logistic regression. On the other hand, we see that the number of false positives produced by the penalized score test is closer to the nominal level when λ is smaller, since less bias is incurred in the estimate $\hat{\mathbf{b}}^0_\lambda$ relative to $\boldsymbol{\beta}$.

When the features are independent, all tests, except multiple logistic regression, have EFP around the nominal rate of 0.9. This might be expected, since, in this case, the distribution of $\mathbf{y}$ depends on a feature $\mathbf{x}_j$, either marginally or conditionally, only when $\theta_j \neq 0$. Thus, estimating the effects of $\{\mathbf{x}_j : \theta_j \neq 0\}$ is not necessary in order

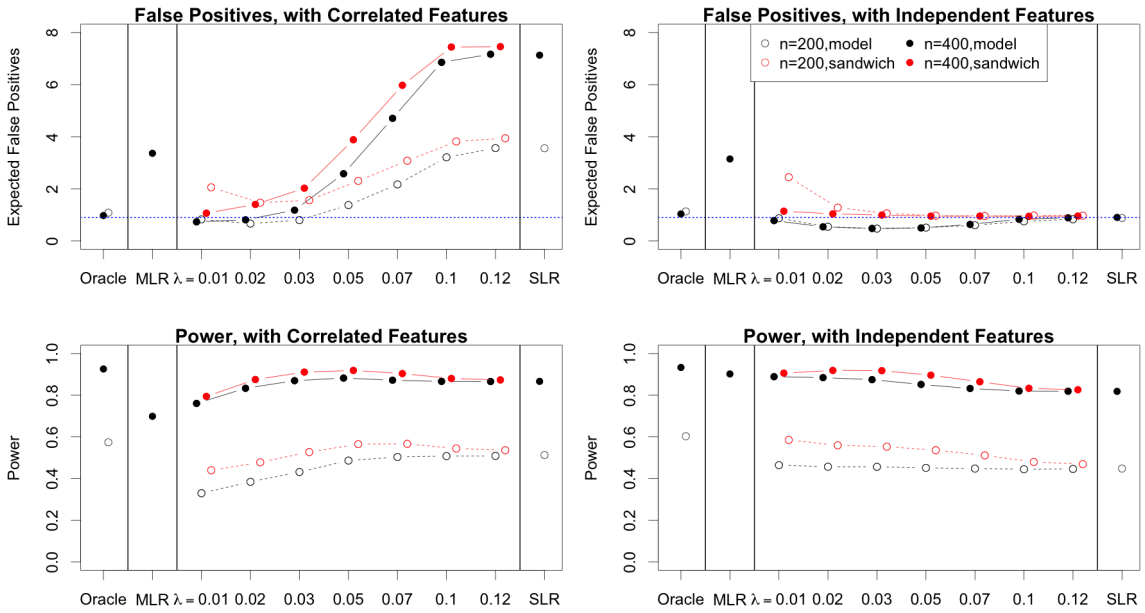

Figure 3.1: Expected false positives (EFP) and power. 'MLR' indicates multiple logistic regression, 'SLR' indicates simple logistic regression, and the seven values of λ indicate results for the penalized score test. Note that multiple linear regression is only available for $n = 400$, as the results with $n = 200$ do not converge in R. In the top panels, the horizontal line indicates the nominal error of 0.9 expected false positives per simulated data set.

to control the error rate.

It is notable that multiple logistic regression gives a larger number of false positives than the advertised rate both when features are correlated, and when features are independent. This may be due to the fact that the number of parameters $d = 100$ is an appreciable fraction of the sample size $n = 400$, whereas classical theory requires that $d/n \to 0$. Indeed, repeating the experiment with $n = 600$ yielded an average of 1.8 false positives per simulated data set, with both correlated and independent features, which is much closer to the nominal level.

In general, the model-based variance formula gives lower EFP than the sandwich variance formula. Recall that in the model-based variance estimate, we use $(0, \hat{\mathbf{b}}_\lambda)$ as an ad-hoc estimate of $\boldsymbol{\theta}$. This results in fitted values $\hat{\mathbf{y}}^0_\lambda$, used in (3.13), which are shrunk towards the mean of $\mathbf{y}$. For logistic regression the variance $V(\cdot)$ is largest when the fitted probabilities are near 0.5, and thus this shrinkage tends to increase the variance $V(\hat{\mathbf{y}}^0_\lambda)$, resulting in a smaller number of false positives than if $\boldsymbol{\theta}$ were known.

The sandwich formula yields a relatively high number of false positives when $n = 200$ and $\lambda = 0.1$, under both schemes of generating $\mathbf{X}$. This may be due to the large number of non-zero coefficients in the associated ℓ_1-penalized regression when λ is small, whereas our theory requires the number of non-zero coefficients in (3.8) to be small, relative to n (see Assumption (**B6**)). This is supported by the fact that when $\lambda = 0.1$ and n is increased to 400, EFP is much closer to the nominal level.

We also note that decreasing λ gives higher power when the features are independent, while it gives lower power when the features are correlated. Decreasing λ reduces variance of the residuals $\mathbf{y} - \hat{\mathbf{y}}^0_\lambda$. When the features are independent, this reduction of variance simply makes identifying associations easier, since associated features explain a larger proportion of the residual variance. On the other hand, reducing λ also lets more features enter the model, which increases co-linearity and reduces power, in the presence of correlated features.

3.4 Inference in high-dimensional Gaussian graphical models

Suppose we observe a matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_m] \in \mathbb{R}^{n \times m}$, and are interested in the pattern of association between the features $\{\mathbf{x}_j : j = 1, \ldots, m\}$. Here, we assume that the features are centered and scaled so that $\mathbf{x}_j^T \mathbf{x}_j / n = 1$ and $\mathbf{x}_j^T \mathbf{1}_n = 0$ for $j = 1, \ldots, m$. These associations are sometimes summarized with a graphical model, in which the features correspond to vertices, and the edges describe relationships between features. In particular, suppose we are interested in the conditional dependence graph, where presence of an edge indicates dependence between two features, after conditioning on all other features. When the rows of $\mathbf{X}$ are independently distributed $N_m(\mathbf{0}, \mathbf{\Theta}^{-1})$, features j and k are conditionally dependent if and only if $\mathbf{\Theta}_{jk} \neq 0$. Thus, in this setting, an estimate of the conditional independence graph can be obtained by inferring which elements of $\mathbf{\Theta}$ are non-zero. Note that m denotes the number of features, while there are $d \equiv \binom{m}{2}$ parameters which define the conditional independence graph.

In this section, we will investigate two ways to perform approximate inference on $\mathbf{\Theta}$ in the high-dimensional setting, using the framework introduced in Section 3.2. The first method, discussed in Section 3.4.1, is based on the graphical lasso [Yuan and Lin, 2007a, Friedman et al., 2008], which involves the joint distribution of $\mathbf{X}$. In Section 3.4.2 we consider an alternate method based on neighborhood selection [Meinshausen and Buhlmann, 2006], which involves the conditional distributions $\{\mathbf{x}_j \mid \mathbf{x}_k : k \neq j\}$. We compare the results obtained using the two procedures in Section 3.4.3.

3.4.1 Inference with the graphical lasso

When $m \ll n$, $\mathbf{\Theta}$ can be estimated by minimizing the negative log-likelihood of the multivariate normal distribution

$$\hat{\mathbf{\Theta}} = \underset{\mathbf{W} \in \mathcal{S}_m}{\arg\min} \left\{ \operatorname{tr}(\mathbf{WC}) - \log\det(\mathbf{W}) \right\},$$

where $\mathcal{S}_m$ is the space of $m \times m$ symmetric positive definite matrices, and $\mathbf{C} = \mathbf{X}^T\mathbf{X}/n$ is the sample covariance matrix of $\mathbf{X}$. In this setting, a number of methods can be applied to test whether individual elements of $\boldsymbol{\Theta}$ are zero [Drton and Perlman, 2004, 2007, Dempster, 1972]. In particular, Drton and Perlman [2004] provide simultaneous confidence intervals for each element of $\hat{\boldsymbol{\Theta}}$, which can be used to perform formal hypothesis tests.

When the dimension m is large, we can instead consider the *graphical lasso* [Yuan and Lin, 2007a, Friedman et al., 2008]:

$$\hat{\boldsymbol{\Theta}}_\lambda = \underset{\mathbf{W} \in \mathcal{S}_m}{\arg\min} \left\{ \operatorname{tr}(\mathbf{WC}) - \log\det(\mathbf{W}) + \lambda \sum_{q \neq r} |w_{qr}| \right\}, \tag{3.14}$$

where w_{qr} denotes the qr^{th} entry of $\mathbf{W}$[1]. For sufficiently large λ, the resulting estimate $\hat{\boldsymbol{\Theta}}_\lambda$ is sparse, and noting which elements of $\hat{\boldsymbol{\Theta}}_\lambda$ are zero in a sample provides a guess for the sparsity pattern of $\boldsymbol{\Theta}$ in the population. Notably, however, the method does not provide measures of uncertainty, such as p-values, for the selection procedure.

Here, we show how Theorems 4 and 5 can be applied to the graphical lasso (4.4). The score of the objective function takes the form

$$\dot{\mathrm{l}}(\boldsymbol{\Theta})/n = \mathbf{C} - \boldsymbol{\Theta}^{-1} \tag{3.15}$$

Given a particular pair of variables $(\mathbf{x}_j, \mathbf{x}_k)$ with $j, k \in \{1, \ldots, m\}$ and $j \neq k$, we can parametrize $l(\mathbf{W})$ as

$$l(a, \mathbf{b})/n \equiv l(\mathbf{W})/n = \operatorname{tr}(\mathbf{WC}) - \log\det(\mathbf{W}),$$

where $a = w_{jk}$ and $\mathbf{b} = \{w_{qr} : (q, r) \neq (j, k)\}$, where $\mathbf{b}$ is a vector of $d - 1$ elements.

---

[1] In Friedman et al. [2008]'s formulation, the diagonal of $\mathbf{W}$ is penalized as well

For this choice of $l(a, \mathbf{b})$, the penalized score statistic takes the form

$$T_\lambda = \sqrt{n}\left(c_{jk} - \left[\hat{\mathbf{\Theta}}_{0,\lambda}^{-1}\right]_{jk}\right), \tag{3.16}$$

where c_{jk} is the jk^{th} entry of $\mathbf{C}$, and

$$\hat{\mathbf{\Theta}}_{0,\lambda} = \underset{\mathbf{W}\in\mathcal{S}_m : w_{jk}=0}{\arg\min}\left\{\text{tr}(\mathbf{WC}) - \log\det(\mathbf{W}) + \lambda\sum_{q\neq r}|w_{qr}|\right\}. \tag{3.17}$$

Applying Theorem 4, we find that $[\hat{\mathbf{\Theta}}_\lambda]_{jk} \neq 0$ if and only if $|T_\lambda| > \sqrt{n}\lambda$. Note that when λ is large, then $\hat{\mathbf{\Theta}}_{0,\lambda} = \mathbf{I}_m$, and hence $T_\lambda = \sqrt{n}c_{jk}$. Therefore, when λ is large, T_λ is a measure of marginal, or Pearson, correlation between $\mathbf{x}_j$ and $\mathbf{x}_k$.

Theorem 5 gives the distribution of T_λ, under $H_{0,\lambda} : a_\lambda = 0$. In order to estimate the variance of T_λ in (3.16), we also require the second derivatives of $l(\mathbf{\Theta})$, which are given by

$$\frac{\partial^2 l(\mathbf{\Theta})/n}{\partial\mathbf{\Theta}_{jk}\partial\mathbf{\Theta}_{qr}} = \left[\mathbf{\Theta}^{-1}\right]_{jq}\left[\mathbf{\Theta}^{-1}\right]_{kr} + \left[\mathbf{\Theta}^{-1}\right]_{jr}\left[\mathbf{\Theta}^{-1}\right]_{kq}. \tag{3.18}$$

We can then plug (3.15) and (3.18), evaluated at $\hat{\mathbf{\Theta}}_{0,\lambda}$, into (3.10) and (3.11) in order to compute the *sandwich variance*, as in Section 3.3.1.

As with GLMs in Section 3.3.1, we can also derive a model-based variance for T_λ in (3.16). When the rows of $\mathbf{X}$ are independently distributed $N_m(\mathbf{0}, \mathbf{\Theta}^{-1})$, then, for any matrix $\mathbf{W}$, we have that $\text{var}[\,\dot{\mathbf{l}}(\mathbf{W})/\sqrt{n}\,] = \text{var}(\mathbf{C}/\sqrt{n}) = \text{var}[\,\dot{\mathbf{l}}(\mathbf{\Theta})/\sqrt{n}\,] = \ddot{\mathbf{l}}(\mathbf{\Theta})/n$, using the fact that the information matrix ($\text{var}[\,\dot{\mathbf{l}}(\mathbf{\Theta})/\sqrt{n}\,]$) is also the second derivative matrix when the model is correctly specified[2]. Thus, if we have an estimate of $\mathbf{\Theta}$, we can plug it into $\ddot{\mathbf{l}}(\mathbf{\Theta})$ as an alternative to the empirical information matrix $\hat{\mathbf{V}}$ in (3.11). For simplicity, we propose using $\hat{\mathbf{\Theta}}_{0,\lambda}$ as an estimate of $\mathbf{\Theta}$, which gives

[2]This fact can also be seen by noting that $n\mathbf{C}$ has a Wishart($\mathbf{\Theta}^{-1}, n$) distribution.

$\hat{\mathbf{V}} = \hat{\mathbf{U}}$, and leads to the variance estimate

$$\widehat{\text{var}}(T_\lambda) = \hat{u}_{aa} - \hat{\mathbf{u}}_{\hat{\mathcal{A}}a}^T \hat{\mathbf{U}}_{\hat{\mathcal{A}}\hat{\mathcal{A}}}^{-1} \hat{\mathbf{u}}_{\hat{\mathcal{A}}a}, \tag{3.19}$$

where $\hat{\mathbf{U}}$ is as in (3.10). We refer to (3.19) as the *model-based* variance estimate. This is precisely analogous to the model-based estimate (3.13) discussed in Section 3.3.1.

In order to perform the penalized score test, we use the `glasso` package to fit (3.17) [Friedman et al., 2011]. Both the model-based and sandwich variances are calculated in the R package `lassoscore`.

3.4.2 Inference with neighborhood selection

As an alternative to working with the joint distribution of $\mathbf{X}$, we could instead consider the conditional distributions $\mathbf{x}_j \mid \{\mathbf{x}_k : k \neq j\}$. When the rows of $\mathbf{X}$ are distributed $N_m(\mathbf{0}, \mathbf{\Theta}^{-1})$, the conditional distributions are

$$\mathbf{x}_j \mid \{\mathbf{x}_l : l \neq j\} = \sum_{l \neq j} \gamma_{jl} \mathbf{x}_l + \boldsymbol{\epsilon}_j, \tag{3.20}$$

where $\gamma_{jl} = \mathbf{\Theta}_{jl}/\mathbf{\Theta}_{ll}$ and $\boldsymbol{\epsilon}_j \sim N(0, 1/\mathbf{\Theta}_{jj})$. Thus, determining whether the jl^{th} element of $\mathbf{\Theta}$ is non-zero is equivalent to determining whether γ_{jl} is non-zero in (3.20). In the low-dimensional setting, we can perform formal inference regarding linear regression parameters using ordinary least squares. In the high-dimensional setting, we can use node-wise lasso regression, sometimes called *neighborhood selection* [Meinshausen and Buhlmann, 2006], where, for each $j = 1, \ldots, m$, we perform lasso regression of $\mathbf{x}_j$ on all other features

$$\{\hat{\gamma}_{jl,\lambda} : 1 \leq l \leq m, l \neq j\} = \underset{w_{jl}: l \neq j}{\arg\min} \left\{ \|\mathbf{x}_j - \sum_{l \neq j} w_{jl} \mathbf{x}_l\|_2^2/(2n) + \lambda \sum_{l \neq j} |w_{jl}| \right\}. \tag{3.21}$$

Just as with the graphical lasso, the sparsity pattern of the estimates (3.21) is used to provide an estimate for which elements of $\boldsymbol{\Theta}$ are non-zero. Lasso regression is a special case of the GLMs considered in Section 3.3, and thus we can use the penalized score test to frame the sparsity pattern of (3.21) as a statistical test. For instance, the penalized score statistic for the effect of $\mathbf{x}_k$ using $\mathbf{x}_j$ as the outcome is

$$T_\lambda = \frac{1}{\sqrt{n}}\mathbf{x}_k^T\left(\mathbf{x}_j - \sum_{l\neq j}\hat{\gamma}^0_{jl,\lambda}\mathbf{x}_l\right), \tag{3.22}$$

where

$$\left\{\hat{\gamma}^0_{jl,\lambda} : 1 \leq l \leq m, l \neq j\right\} = \underset{w_{jl}:l\neq j,\, w_{jk}=0}{\arg\min}\left\{\|\mathbf{x}_j - \sum_{l\neq j} w_{jl}\mathbf{x}_l\|_2^2/(2n) + \lambda\sum_{l\neq j}|w_{jl}|\right\}. \tag{3.23}$$

By Theorem 4, we know that $\hat{\theta}_{jk,\lambda} \neq 0$ if and only if $|T_\lambda| > \sqrt{n}\lambda$.

Theorem 5 gives the asymptotic distribution of T_λ under $H_{0,\lambda} : a_\lambda = 0$. As in Section 3.3.1, we can use either model-based or sandwich variance estimates to obtain p-values.

Note that when we apply this procedure to each pair of features (j, k), we can obtain a p-value from the lasso regression using $\mathbf{x}_j$ as the outcome, and also when using $\mathbf{x}_k$ as the outcome, which need not be the same. We treat this issue by simply averaging the two test statistics produced for each (j, k) in order to obtain a single test statistic, and compare to a $N(0, 1)$ reference distribution. This procedure is conservative when the null hypotheses are true.

3.4.3 Simulation experiments

In this section, we investigate the behavior of the penalized score test applied to the graphical lasso and neighborhood selection, as described in Sections 3.4.1 and 3.4.2, respectively. As with the simulations in Section 3.3.2, we will seek to demonstrate

that, even though the penalized score test measures evidence against $H_{0,\lambda} : a_\lambda = 0$, it serves as a useful approximation to classical tests of $H_{0,\lambda} : \alpha = 0$ when λ is small enough.

We generated a matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$ of $m = 50$ features according one of two schemes:

- *Correlated features*: Rows of $\mathbf{X}$ were independently distributed $N_m(\mathbf{0}, \mathbf{\Theta}^{-1})$, where $\mathbf{\Theta}_{kk} = 1$ for $k = 1, \ldots, m$, and for 50 randomly chosen pairs $(j, k) : j < k$ we set $\mathbf{\Theta}_{jk} = \mathbf{\Theta}_{kj} = 0.3$. Figure 3.2 shows the resulting conditional dependence graph.

- *Independent features*: Rows of $\mathbf{X}$ were independently distributed $N_m(\mathbf{0}, \mathbf{I}_m)$.

The scheme with correlated features demonstrates the ability of the penalized score test to behave like a test of $H_0 : \alpha = 0$, while the scheme with uncorrelated features validates the theory for the case where $H_{0,\lambda} : a_\lambda = 0$ is true.

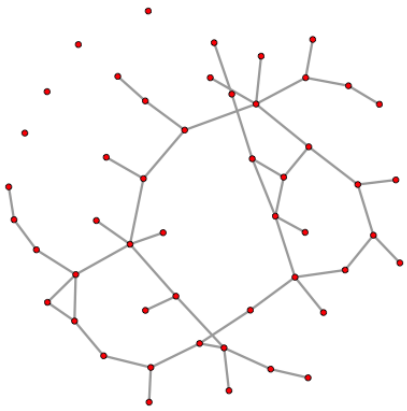

Figure 3.2: Conditional dependence graph used in the "correlated features" scheme. Each vertex corresponds to a features, and edges indicate conditional dependence between features.

We performed the penalized score test based on the graphical lasso and neighbor-

hood selection for several values of λ, over $B = 500$ simulated data sets, with $n = 100$ or $n = 200$. For the sake of comparison, we tested whether pairwise Pearson correlations were non-zero, as well as whether the partial correlations were zero [Drton and Perlman, 2004, 2007].

To account for multiplicity in testing the $d \equiv \binom{m}{2}$ edges in the conditional dependence graph, we declared a test to be significant if the resulting p-value was smaller than $1/d = 0.0008$. We calculated the EFP and power for each test, using the same formulas as in Section 3.3.2. In the scenario with correlated features, since 50 edges are present, an unbiased test should have an EFP of $0.96 = (1/d) \times (d - 50)$, while when the features are independent, an unbiased test should have an EFP of 1. Note that when the features are independent, power is undefined.

Figures 3.3 and 3.4 summarize the results of the experiment for correlated and independent features, respectively. In addition, Table 3.2 gives the average number of non-zero coefficients in the associated ℓ_1-penalized M-estimators (4.4) and (3.21). In Figure 3.3, we see that the penalized score test, using either neighborhood selection or the graphical lasso, can control EFP at the nominal level when λ is small, while giving similar EFP to marginal correlation when λ is large. EFP is closest to the nominal level when $\lambda \leq 0.1$ in which case we see from Table 3.2 there are a few hundred non-zero coefficient estimates in both the graphical lasso and neighborhood selection, many more than the 50 truly non-zero parameters. That is, in order to control false positives, it seems beneficial to choose λ to be smaller than one would if variable selection were the goal. This was also observed in Section 4.5.1 of Chapter 2 in the context of linear regression.

It is notable, however, that the sandwich variance for the graphical lasso gives a high number of false positives when $n = 100$, regardless of the choice of λ. When λ is large, the high EFP may be due to discrepancies between a_λ and α, whereas when λ is small, the high EFP is likely due to instability in the sandwich variance estimate $\hat{\mathbf{V}}$. Indeed, from Table 3.2, we see that the dimension of $\hat{\mathbf{V}}$ is on the order of 645×645

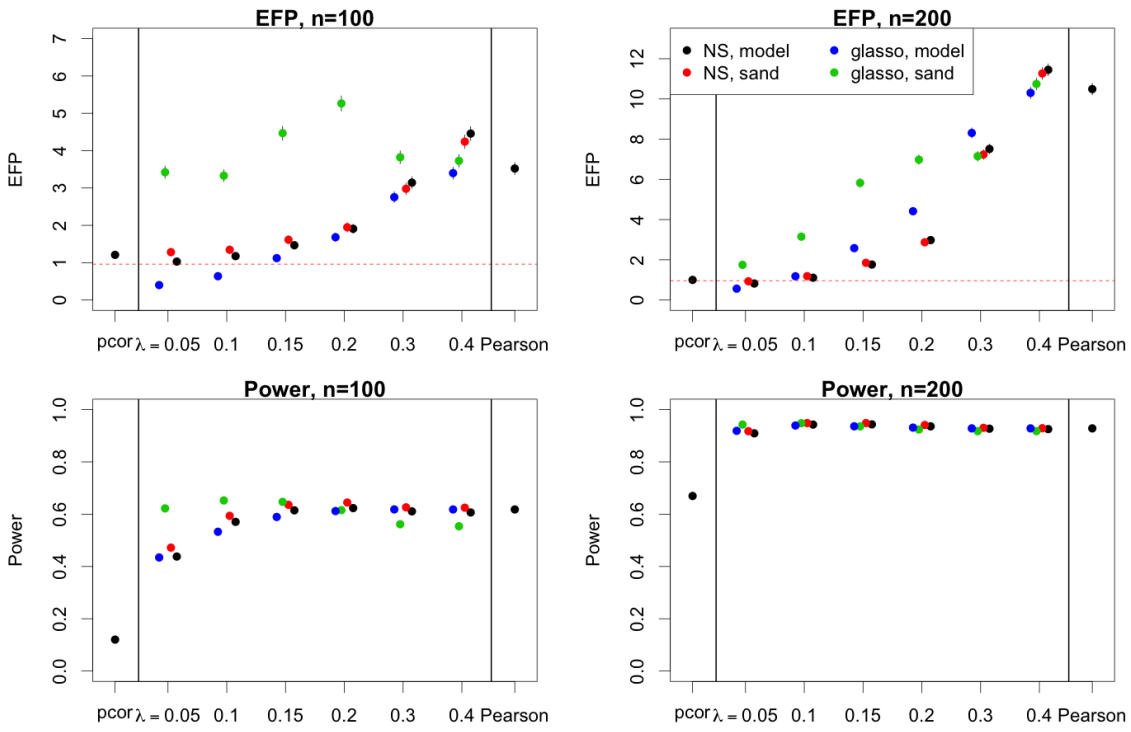

Figure 3.3: Results, for graphical model simulations with correlated features. Vertical bars indicate 95% confidence intervals, from Monte-Carlo error. 'pcor' indicates the test for partial correlations, 'glasso' indicates the penalized score test using the graphical lasso, 'NS' indicates the penalized score test for neighborhood selection, and 'Pearson' indicates the test for marginal correlation.

when $\lambda = 0.05$ and $n = 100$; consequently, $\hat{\mathbf{V}}$ may be a poor estimate of $\mathbf{V}$. This explanation for high EFP when λ is small is corroborated by the simulations with independent features, displayed in Figure 3.4, where we also observe EFP higher than the nominal rate for the graphical lasso when λ is small, using the sandwich variance estimate. Note that we do not observe such bad performance using neighborhood selection. This may be due to the fact that variance estimates in neighborhood selection depend only on the non-zero coefficients in a single lasso regression, while the graphical lasso variance estimates are based on all non-zero coefficients in the matrix $\hat{\mathbf{\Theta}}_{0,\lambda}$. A similar phenomenon was observed in Section 3.3.2, when using the sandwich variance for GLMs when both λ and the sample size are small.

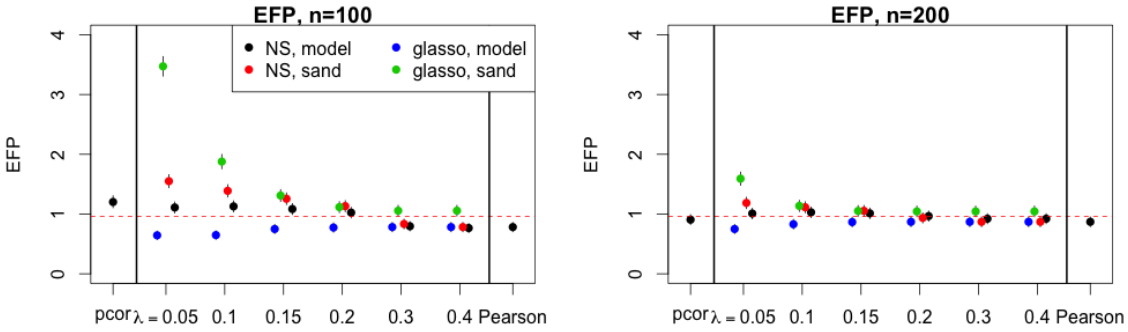

Figure 3.4: Results for graphical model simulations, with independent features. Vertical bars indicate 95% confidence intervals, from Monte-Carlo error. 'pcor' indicates the test for partial correlations, 'glasso' indicates the penalized score test using the graphical lasso, 'NS' indicates the penalized score test using neighborhood selection, and 'Pearson' indicates test for marginal correlation.

We see that the power of the penalized score test decreases with λ, for both the graphical lasso and neighborhood selection, and that the trend is more pronounced when $n = 100$. However, the penalized score test still enjoys substantially higher power than the classical test of partial correlation. The low power of tests for partial correlation may be due to the large number of degrees-of-freedom, relative to the sample size.

	Correlated Features				**Independent Features**			
	$n = 100$		$n = 200$		$n = 100$		$n = 200$	
λ	NS	glasso	NS	glasso	NS	glasso	NS	glasso
0.4	15	17	14	15	0.04	0.04	0	0
0.3	35	41	36	40	2	2	0.008	0.006
0.2	78	105	52	61	48	51	5	5
0.15	142	192	74	99	138	154	39	39
0.1	284	351	167	217	322	360	177	188
0.05	583	645	467	517	655	695	541	562

Table 3.2: Average number of non-zero parameters in neighborhood selection (NS) and the graphical lasso (glasso), at each value of λ. For neighborhood selection, the number reported is half the number of non-zero coefficient estimates from summed over all of the node-wise lasso regressions, in order to make results comparable with the graphical lasso.

The fact that the penalized score test controls false positives, provided λ is small enough, demonstrates that the penalized score test serves as a useful proxy for classical tests of conditional independence, such as partial correlation, which have low power in these simulations.

3.5 Discussion

In this chapter, we proposed an extension of the penalized score test for linear models to the setting of ℓ_1-penalized M-estimators, and explored the special cases of ℓ_1-penalized generalized linear models and Gaussian graphical models. The test we proposed serves as an alternative to classical tests for a single parameter α, when there are a large number of nuisance parameters $\boldsymbol{\beta}$. The test is most accurately viewed as a test of $H_{0,\lambda} : a_\lambda = 0$, where a_λ differs from α due to bias from the penalized estimate of $\boldsymbol{\beta}$. However, through simulation, we showed that this test serves as a useful proxy for tests of $H_0 : \alpha = 0$, provided that λ is small enough.

In Theorem 4 we showed that the penalized score test is closely tied to variable selection with ℓ_1-penalized M-estimators. However, the simulations in Sections 3.3.2

and 3.4.3 demonstrated that in order to control type-I error at the nominal rate, λ must be chosen smaller than if variable selection with ℓ_1-penalized M-estimators were the goal. That is, the test performs best when the number of non-zero parameter estimates $\|\hat{\mathbf{b}}_\lambda\|_0$ is larger than the number of truly non-zero parameters $\|\boldsymbol{\beta}\|_0$.

In practice, choosing λ in order to achieve a powerful test that also controls type-I error may not be straightforward. Bias in the parameter estimate $\mathbf{b}_\lambda$ decreases with λ, while our theory requires $\|\mathbf{b}_\lambda\|_0$ to be small relative to n. Thus, the most pragmatic option may be to choose λ such that $\|\hat{\mathbf{b}}_\lambda\|_0$ is a small fraction of the sample size.

An interesting extension of this work would be effect size estimation for a single parameter of interest using penalization to estimate nuisance parameters. Theory for confidence intervals follows in a straightforward way from Theorem 5. We leave implementation details to future work.

Chapter 4

GRAPH ESTIMATION WITH JOINT ADDITIVE MODELS

4.1 Introduction

In recent years, there has been considerable interest in developing methods to estimate the joint pattern of association among a set of random variables. The relationships between d random variables can be summarized with an undirected graph $\Gamma = (V, E)$ in which the random variables are represented by the vertices $V = \{1, \ldots, d\}$ and the conditional dependencies between pairs of variables are represented by edges $E \subset V \times V$. That is, for each $j \in V$, we want to determine a minimal set of variables on which the conditional densities $p_j(x_j \mid \{x_k, k \neq j\})$ depend,

$$p_j(x_j \mid \{x_k, k \neq j\}) = p_j(x_j \mid \{x_k : (k, j) \in E\}).$$

Recently there has also been considerable work in estimating marginal associations between a set of random variables [see e.g. Basso et al., 2005, Meyer et al., 2008, Liang and Wang, 2008, Hausser and Strimmer, 2009, Chen et al., 2010]; however, in this chapter we focus on conditional dependencies, which provide richer information about the relationships among the variables.

Estimating the conditional independence graph Γ based on a set of n observations is an old problem [Dempster, 1972]. In the case of high-dimensional continuous data, most prior work has assumed either (a) multivariate Gaussianity [see e.g. Friedman et al., 2008, Rothman et al., 2008, Yuan and Lin, 2007c, Banerjee et al., 2008] or (b) linear conditional means [see e.g. Meinshausen and Bühlmann, 2006, Peng et al.,

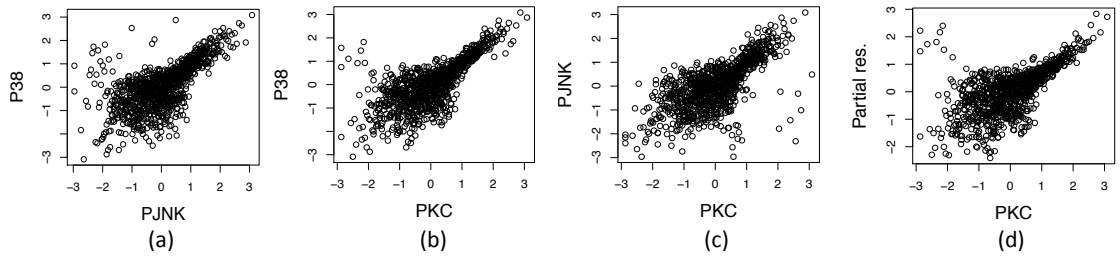

Figure 4.1: Cell signaling data from Sachs et al. [2005]. (a)-(c) Pairwise scatterplots for PKC, P38 and PJNK. (d) Partial residuals from the linear regression of P38 on PKC and PJNK. The data are standardized to have normal marginal distributions, but are clearly not multivariate normal.

2009] for the features. However, as we will see, these two assumptions are essentially equivalent. As an illustration, consider the cell signaling data set from Sachs et al. [2005], which consists of protein concentrations measured under a set of perturbations. We analyze the data set in more detail in Section 4.5.3. Pairwise scatterplots of three of the variables are given in Figure 4.1 (a)-(c) for one of 14 perturbations. Here, the data have been transformed to be marginally normal, as suggested by Liu et al. [2009]. The transformed data clearly are not multivariate normal, given the non-constant variance in the bivariate scatterplots, and as confirmed by a Shapiro-Wilk test ($p < 2 \times 10^{-16}$).

Can the data in Figure 4.1 be well-represented by linear relationships? In Figure 4.1 (d), we see strong evidence that the conditional mean of the protein P38 given PKC and PJNK is nonlinear. This is corroborated by the fact that the p-value for including quadratic terms in the linear regression of P38 onto PKC and PJNK is small ($p < 2 \times 10^{-16}$). Therefore in this data set, the features are not multivariate Gaussian, and marginal transformations do not remedy the problem.

In order to flexibly model conditional mean relationships, we could specify a more flexible joint distribution. However, joint distributions are difficult to construct and computationally challenging to fit, and the resulting conditional models need not be

easy to obtain or interpret. Alternatively we can specify the conditional distributions directly. This has the advantage of simpler interpretation and greater computational tractability. In this chapter, we will model the conditional means of non-Gaussian random variables with generalized additive models [Hastie and Tibshirani, 1990], and will use these in order to construct conditional independence graphs.

Throughout this chapter, we will assume that we are given n independent and identically distributed observations from a d-dimensional random vector $x = (x_1, \ldots, x_d) \sim \mathcal{P}$. Our observed data can be written as $\mathbf{X} = [\mathbf{x}_1, \ldots \mathbf{x}_d] \in \mathbb{R}^{n \times d}$. Note that lower-case x is a vector, in contrast to our usual notation, to distinguish the random variables x from the observed data $\mathbf{X}$.

The rest of the chapter is organized as follows. In Sections 4.2 and 4.3 we review methods for modeling conditional dependence relationships among a set of variables, and discuss their limitations. In Section 4.4 we propose our method (SpaCE JAM) and an algorithm for its computation. We illustrate our method on real and simulated data in Section 4.5, and compare with available methods. In Section 4.6 we extend the method to the estimation of directed acyclic graphs with known causal ordering. In Section 4.7 we prove consistency of our algorithm, and in Section 4.8 we propose a screening rule for estimation in high dimensions.

4.2 Modeling conditional dependence relationships

Suppose we are interested in estimating the conditional independence graph Γ for a random vector $x \in \mathbb{R}^d$. If the joint distribution is known up to some finite dimensional parameter $\boldsymbol{\theta}$, then to estimate Γ it suffices to estimate $\boldsymbol{\theta}$ via e.g. maximum likelihood. One practical difficulty that arises in estimating Γ is specification of a plausible joint distribution. Specifying a conditional distribution, such as in a regression model, is typically much less daunting. We therefore consider pseudo-likelihoods [Besag, 1974,

1975] of the form

$$\log(p_{PL}(x;\boldsymbol{\theta})) = \sum_{j=1}^{d} \log\left(p_j(x_j \mid \{x_k : (j,k) \in E\};\boldsymbol{\theta})\right).$$

For a set of arbitrary conditional distributions, there need not be a compatible joint distribution [Wang and Ip, 2008]. However, the conditionally specified graphical model has an appealing theoretical justification, in that it minimizes the Kullback-Leibler distances to the conditional distributions [Varin and Vidoni, 2005]. Furthermore, in estimating conditional independence graphs, our scientific interest is in the conditional independence relationships rather than in the joint distribution. So in a sense, modeling the conditional distribution rather than the joint distribution amounts to a more direct approach to graph estimation. We therefore advocate for an approach for non-Gaussian graphical modeling based on conditionally specified models [Varin et al., 2011].

4.3 Previous work

4.3.1 Estimating graphs with Gaussian data

Suppose for now that x has a joint Gaussian distribution with mean 0 and precision matrix $\boldsymbol{\Theta}$. One can write the negative log-likelihood of the joint distribution, up to constants, as

$$-\log\det(\boldsymbol{\Theta}) + \operatorname{tr}\left(xx^T\boldsymbol{\Theta}\right). \tag{4.1}$$

In this case, the conditional relationships are linear,

$$x_j \mid \{x_k,\, k \neq j\} = \sum_{k\neq j} \beta_{jk} x_k + \epsilon_j, \quad j = 1,\ldots,d, \tag{4.2}$$

where $\beta_{jk} = -\boldsymbol{\Theta}_{jk}/\boldsymbol{\Theta}_{kk}$ and $\epsilon_j \sim N_1(0, 1/\boldsymbol{\Theta}_{jj})$. To estimate the graph Γ, we must determine which β_{jk} are zero in (4.2), or equivalently which $\boldsymbol{\Theta}_{jk}$ are 0 in (4.1). This

is simple when $n \gg d$.

In the high-dimensional setting, when the maximum likelihood estimate is unstable or undefined, a number of approaches have been proposed to estimate the conditional independence graph Γ, which we review here. Meinshausen and Bühlmann [2006] proposed fitting (4.2) using an ℓ_1-penalized regression. This is referred to as neighborhood selection:

$$\left\{\hat{\beta}_{jk} : 1 \leq j, k \leq d\right\} = \underset{\beta_{jk}:1\leq j,k\leq d}{\arg\min} \left\{ \frac{1}{2} \sum_{j=1}^{d} \|\mathbf{x}_j - \sum_{k \neq j} \mathbf{x}_k \beta_{jk}\|^2 + \lambda \sum_{j=1}^{d} \sum_{k \neq j} |\beta_{jk}| \right\}. \tag{4.3}$$

Here λ is a nonnegative tuning parameter that encourages sparsity in the coefficient estimates. Peng et al. [2009] improved upon the neighborhood selection approach by applying ℓ_1 penalties to the partial correlations; this is known as sparse partial correlation estimation.

As an alternative to (4.3), many authors have considered estimating $\mathbf{\Theta}$ under the multivariate normality assumption by maximizing an ℓ_1-penalized joint log likelihood [see e.g. Yuan and Lin, 2007c, Banerjee et al., 2008, Friedman et al., 2008]. This amounts to the optimization problem

$$\hat{\mathbf{\Theta}} = \underset{\mathbf{\Theta} \succ 0}{\arg\min} \left\{ - \log\det(\mathbf{\Theta}) + \operatorname{tr}\left(\mathbf{X}^T \mathbf{X} \mathbf{\Theta}\right) / n + \lambda \|\mathbf{\Theta}\|_1 \right\}, \tag{4.4}$$

known as the graphical lasso. The solution $\hat{\mathbf{\Theta}}$ to (4.4) serves as an estimate for $\mathbf{\Theta}$, and hence the sparsity pattern of $\hat{\mathbf{\Theta}}$ (induced by the ℓ_1 penalty) provides an estimate of Γ.

At first glance, neighborhood selection and sparse partial correlation may seem semi-parametric: a linear model may hold in the absence of multivariate normality. However, while (4.2) can accurately model each conditional dependence relationship semi-parametrically, the accumulation of these specifications is very restrictive in terms of the joint distribution. In fact, Khatri and Rao [1976] proved that if (4.2)

holds, along with some other mild assumptions, then the joint distribution must be multivariate normal. Notably, this is true regardless of the distribution of the errors $\epsilon_1, \ldots, \epsilon_d$ in (4.2). In other words, even though (4.3) does not explicitly involve the multivariate normal likelihood, normality is implicitly assumed. This means that if we wish to model non-normal continuous data, then non-linear conditional models are necessary.

4.3.2 Estimating graphs with non-Gaussian data

We now briefly review three existing methods for modeling conditional independence graphs with non-Gaussian data. The normal copula or nonparanormal model (Liu et al. 2009, Liu et al. 2012, Xue and Zou 2012, studied in the Bayesian context by Dobra and Lenkoski 2011) assumes that x has a nonparanormal distribution: that is, $(h_1(x_1), \ldots, h_d(x_d)) \sim N_d(0, \mathbf{\Theta})$ for functions $h_1(\cdot), \ldots, h_d(\cdot)$. After $h_1(\cdot), \ldots, h_d(\cdot)$ are estimated, one can apply any of the methods mentioned in Section 4.3.1 to the transformed data. The conditional model implicit in this approach is

$$h_j(x_j) \mid \{x_k,\, k \neq j\} = \sum_{k \neq j} \beta_{jk} h_k(x_k) + \epsilon_j, \quad j = 1, \ldots, d. \tag{4.5}$$

This is itself a restrictive assumption, which may not hold, as seen in Figure 4.1.

Forest density estimation [Liu et al., 2011] replaces the need for distributional assumptions with graphical assumptions: the underlying graph is assumed to be a forest. Then bivariate densities are estimated non-parametrically. Unfortunately, the restriction to acyclic graphs may be inappropriate in applications, and maximizing over all possible forests is infeasible.

The graphical random forests [Fellinghauer et al., 2013] approach uses random forests to flexibly model conditional means, and allows for interaction terms. But this does not correspond to a well-defined statistical model, and guarantees on feature selection consistency are unavailable.

4.4 Method

4.4.1 Jointly additive models

In order to estimate a conditional independence graph using a pseudolikelihood approach, we must estimate the variables on which the conditional distributions $p_j(\cdot)$ depend. However, since density estimation is generally a challenging task, especially in high dimensions, we focus on the simpler problem of estimating the conditional mean $\mathbb{E}[\, x_j \mid \{x_k : (j,k) \in E\}\,]$, under the assumption that the conditional distribution and the conditional mean depend on the same set of variables. Thus, we seek to estimate the conditional mean $f_j(\cdot)$ in the regression model

$$x_j|\{x_k, k \neq j\} = f_j\left(x_k : k \neq j\right) + \epsilon_j,$$

where ϵ_j is a mean-zero error term. Since estimating arbitrary functions $f_j(\cdot)$ is infeasible in high dimensions, we restrict ourselves to additive models of the form

$$x_j|\{x_k,\, k \neq j\} = \sum_{k \neq j} f_{jk}(x_k) + \epsilon_j, \tag{4.6}$$

where $f_{jk}(\cdot) \in \mathcal{F}$ for some space of functions $\mathcal{F}$. This amounts to modeling each variable using a generalized additive model [Hastie and Tibshirani, 1990]. Unlike Fellinghauer et al. [2013], we do not assume that the errors ϵ_j are independent of the additive components $f_{jk}(\cdot)$, but merely that the conditional independence structure can be recovered from the additive components $f_{jk}(\cdot)$.

4.4.2 Estimation with SpaCE JAM

Since we believe that the conditional independence graph is sparse, we fit (4.6) using a penalty that performs simultaneous estimation and selection of the $f_{jk}(\cdot)$. Specifically, we link together d sparse additive models [Ravikumar et al., 2009] using a penalty

that groups the parameters corresponding to a single edge in the graph. This results in the problem

$$\underset{f_{jk}\in\mathcal{F},1\leq j,k\leq d}{\text{minimize}}\left\{\frac{1}{2n}\sum_{j=1}^{d}\|\mathbf{x}_j-\sum_{k\neq j}f_{jk}(\mathbf{x}_k)\|_2^2+\lambda\sum_{k>j}\left(\|f_{jk}(\mathbf{x}_k)\|_2^2+\|f_{kj}(\mathbf{x}_j)\|_2^2\right)^{1/2}\right\}. \tag{4.7}$$

We consider $f_{jk}(\mathbf{x}_k)=\boldsymbol{\Psi}_{jk}\boldsymbol{\beta}_{jk}$, where $\boldsymbol{\Psi}_{jk}$ is a $n\times r$ matrix whose columns are basis functions used to model the additive components f_{jk}, and $\boldsymbol{\beta}_{jk}$ is an r-vector containing the associated coefficients. For instance, if we use a linear basis function, i.e. $\boldsymbol{\Psi}_{jk}=\mathbf{x}_k$, then $r=1$ and we are modeling only linear conditional means, as in Meinshausen and Bühlmann [2006]. Higher-order terms allow us to model more complex dependencies. The standardized group lasso penalty [Simon and Tibshirani, 2012a] encourages sparsity and ensures that the estimates of $f_{jk}(\cdot)$ and $f_{kj}(\cdot)$ will be simultaneously zero or non-zero. Problem (4.7) is the natural extension of sparse additive modeling [Ravikumar et al., 2009] to graphs, and generalizes neighborhood selection [Meinshausen and Bühlmann, 2006] and sparse partial correlation [Peng et al., 2009] to allow for flexible conditional means. We call the solution to (4.7) SpaCE JAM (for SPArse Conditional Estimation with Joint Additive Models), to reflect its ties with the aforementioned techniques.

Algorithm 1 uses block coordinate descent to solve (4.7). Since (4.7) is convex, the algorithm converges to the global minimum [Simon and Tibshirani, 2012a]. Performing Step 2 requires an $r\times r$ matrix inversion, where r is the number of basis functions; this must be performed only twice per pair of variables. Estimating 30 conditional independence graphs with $r=3$ on a simulated data set with $n=50$ and $d=100$ takes 1.1 seconds on a 2.8 GHz Intel Core i7 Macbook Pro.

Initialize $\hat{\boldsymbol{\beta}}$'s
Repeat until convergence:
For $(j,k) \in V \times V$:
1: Calculate the vector of residuals for the jth and kth variables:
$\mathbf{r_{jk}} \leftarrow \mathbf{x_j} - \sum_{\mathbf{i} \neq \mathbf{j},\mathbf{k}} \boldsymbol{\Psi}_{\mathbf{ji}} \hat{\boldsymbol{\beta}}_{\mathbf{ji}}$
$\mathbf{r_{kj}} \leftarrow \mathbf{x_k} - \sum_{\mathbf{i} \neq \mathbf{j},\mathbf{k}} \boldsymbol{\Psi}_{\mathbf{ki}} \hat{\boldsymbol{\beta}}_{\mathbf{ki}}$
2: Regress the residuals on the specified basis functions:
$\hat{\boldsymbol{\beta}}_{jk} \leftarrow \left(\boldsymbol{\Psi}_{jk}^T \boldsymbol{\Psi}_{jk}\right)^{-1} \boldsymbol{\Psi}_{jk}^T \mathbf{r_{jk}}$
$\hat{\boldsymbol{\beta}}_{kj} \leftarrow \left(\boldsymbol{\Psi}_{kj}^T \boldsymbol{\Psi}_{kj}\right)^{-1} \boldsymbol{\Psi}_{kj}^T \mathbf{r_{kj}}$
3: Threshold:

$$\hat{\boldsymbol{\beta}}_{jk} \leftarrow \left(1 - n\lambda \left(\|\boldsymbol{\Psi}_{jk}\hat{\boldsymbol{\beta}}_{jk}\|_2^2 + \|\boldsymbol{\Psi}_{kj}\hat{\boldsymbol{\beta}}_{kj}\|_2^2\right)^{-1/2}\right)_+ \hat{\boldsymbol{\beta}}_{jk}$$

$$\hat{\boldsymbol{\beta}}_{kj} \leftarrow \left(1 - n\lambda \left(\|\boldsymbol{\Psi}_{jk}\hat{\boldsymbol{\beta}}_{jk}\|_2^2 + \|\boldsymbol{\Psi}_{kj}\hat{\boldsymbol{\beta}}_{kj}\|_2^2\right)^{-1/2}\right)_+ \hat{\boldsymbol{\beta}}_{kj}$$

Algorithm 1: SpaCE JAM algorithm

4.4.3 Tuning

A number of options for tuning parameter selection are available, such as generalized cross-validation [Tibshirani, 1996], the Bayesian information criterion [Zou et al., 2007], and stability selection [Meinshausen and Bühlmann, 2010]. We take an approach motivated by the Bayesian information criterion, as in Peng et al. [2009]. For the jth variable, the criterion is

$$\text{BIC}_j(\lambda) = n \log(\text{RSS}_j(\lambda)) + \log(n)\text{DF}_j(\lambda), \tag{4.8}$$

where $\text{RSS}_j(\lambda) = \|\mathbf{x}_j - \sum_{k \neq j} \boldsymbol{\Psi}_{jk} \hat{\boldsymbol{\beta}}_{jk}^{(\lambda)}\|_2^2$ is the residual sum of squares from minimizing (4.7) with tuning parameter λ, and $\text{DF}(\lambda)_j$ is the degrees of freedom used in this regression. We seek the value of λ that minimizes $\sum_{j=1}^d \text{BIC}_j(\lambda)$. When a single basis function is used, we can approximate the degrees of freedom by the number of nonzero parameters in the regression [Zou et al., 2007, Peng et al., 2009]. But when $r > 1$

basis functions are used, we use

$$\text{DF}_j(\lambda) = |S_j^{(\lambda)}| + (r-1)\sum_k \frac{\|\boldsymbol{\Psi}_{jk}\hat{\boldsymbol{\beta}}_{jk}^{(\lambda)}\|_2^2}{\|\boldsymbol{\Psi}_{jk}\hat{\boldsymbol{\beta}}_{jk}^{(\lambda)}\|_2^2 + \lambda}, \tag{4.9}$$

where $S_j^{(\lambda)} = \{k : \|\hat{\boldsymbol{\beta}}_{jk}^{(\lambda)}\| \neq 0\}$. Though (4.9) was derived under the assumption of an orthogonal design matrix, it is a good approximation for the non-orthogonal case [Yuan and Lin, 2006].

In order to perform SpaCE JAM, we must select a set of basis functions. In the absence of domain knowledge, we use cubic polynomials, which can approximate a wide range of functions.

4.5 Numerical experiments

4.5.1 Simulation setup

As discussed in Section 4.2, it can be difficult to specify flexible non-Gaussian distributions for continuous variables. However, construction of multivariate distributions via conditional distributions is straightforward when the variables can be represented with a directed acyclic graph. The joint probability distribution of variables in a directed acyclic graph can be decomposed as $p(x_1, \ldots, x_d) = \prod_{j=1}^d p_j(x_j|\{x_k : (k,j) \in E_D\})$, where E_D denotes the directed edge set of the graph. This is a valid joint distribution regardless of the choice of conditional distributions $p_j(x_j|\{x_k : (k,j) \in E_D\})$ [Pearl, 2000, Chapter 1.4]. We chose structural equations of the form

$$x_j|\{x_k : (k,j) \in E_D\} = \sum_{(k,j)\in E_D} f_{jk}(x_k) + \epsilon_j, \tag{4.10}$$

with $\epsilon_j \sim N(0,1)$. If the f_{jk} are chosen to be linear, then the data are multivariate normal, and if the f_{jk} are non-linear, then the data will typically not correspond to a well-known multivariate distribution. We moralized the directed graph in order to

obtain the conditional independence graph [Cowell et al., 2007, Chapter 3.2]. Note that here we have used directed acyclic graphs simply as a tool to generate non-Gaussian data, and that the full conditional distributions of the random variables created using this approach are not necessarily additive.

We first generated a directed acyclic graph with $d = 100$ nodes and 80 edges chosen at random from the $\binom{100}{2}$ possible edges. We used two schemes to construct a distribution on this graph. In the first setting, we chose $f_{jk}(x_k) = b_{jk1}x_k + b_{jk2}x_k^2 + b_{jk3}x_k^3$, where the b_{jk1}, b_{jk2}, and b_{jk3} are independent and normally distributed with mean zero and variance 1, 0.5, and 0.5, respectively. In the second case, we chose $f_{jk}(\mathbf{x}_k) = \mathbf{x}_k$, resulting in multivariate normal data. In both cases we scaled the $f_{jk}(\mathbf{x}_k)$ to have unit variance.

We generated $n = 50$ observations, and compared SpaCE JAM to sparse partial correlation [Peng et al., 2009, `R` package `space`], graphical lasso [Yuan and Lin, 2007c, `R` package `glasso`], neighborhood selection [Meinshausen and Bühlmann, 2006, `R` package `glasso`], nonparanormal [Liu et al., 2012, Xue and Zou, 2012, `R` package `glasso`], forest density estimation [Liu et al., 2011, code provided by authors], the method of Basso et al. [2005, `R` package `minet`], and graphical random forests [Fellinghauer et al., 2013, code provided by authors]. In performing neighborhood selection, we declared an edge between the jth and kth variables if $\hat{\beta}_{jk} \neq 0$ or $\hat{\beta}_{kj} \neq 0$. We performed SpaCE JAM using three sets of basis functions: $\mathbf{\Psi}_{jk} = [\,\mathbf{x}_k, \mathbf{x}_k^2\,]$, $\mathbf{\Psi}_{jk} = [\,\mathbf{x}_k, \mathbf{x}_k^3\,]$, and $\mathbf{\Psi}_{jk} = [\,\mathbf{x}_k, \mathbf{x}_k^2, \mathbf{x}_k^3\,]$.

4.5.2 Simulation results

Figure 4.2 summarizes the results of our simulations. For each method, the numbers of correctly and incorrectly estimated edges were averaged over 100 simulated data sets for a range of 100 tuning parameter values. When the $f_{jk}(\cdot)$ are non-linear, SpaCE JAM with the basis $\mathbf{\Psi}_{jk} = [\,\mathbf{x}_k, \mathbf{x}_k^2, \mathbf{x}_k^3\,]$ dominates SpaCE JAM with the basis sets $\mathbf{\Psi}_{jk} = [\,\mathbf{x}_k, \mathbf{x}_k^2\,]$ or $[\,\mathbf{x}_k, \mathbf{x}_k^3\,]$, which in turn tend to enjoy superior performance

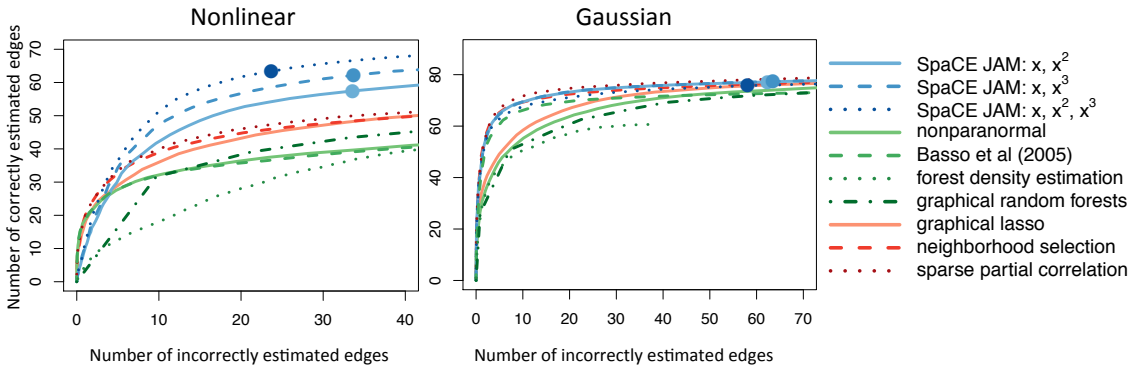

Figure 4.2: Simulation study. The number of correctly estimated edges is displayed as a function of incorrectly estimated edges, for a range of tuning parameter values, in the non-linear (left) and Gaussian (right) set-ups, averaged over 100 simulated data sets. Dots indicate the average model size chosen using the BIC criterion. In the order of appearance in the legend, the competing methods are those of Liu et al. [2012], Basso et al. [2005], Liu et al. [2011], Fellinghauer et al. [2013], Yuan and Lin [2007c], Meinshausen and Bühlmann [2006], Peng et al. [2009].

relative to all other methods (left panel of Figure 4.2). Furthermore, even though the basis sets $\mathbf{\Psi}_{jk} = [\,\mathbf{x}_k, \mathbf{x}_k^2\,]$ and $[\,\mathbf{x}_k, \mathbf{x}_k^3\,]$ do not entirely capture the functional forms of the data-generating mechanism, they still outperform methods that assume linearity, as well as competitors intended to model non-linear relationships.

When the conditional means are linear and the number of estimated edges is small, all methods perform roughly equally (right panel of Figure 4.2). As the number of estimated edges is increased, sparse partial correlation performs best, while the graphical lasso, the nonparanormal and the forest-based methods perform worse. This agrees with the observations of Peng et al. [2009] that sparse partial correlation and neighborhood selection tend to outperform the graphical lasso. In this setting, since non-linear terms are not needed to model the conditional dependence relationships, sparse partial correlation outperforms SpaCE JAM with two basis functions, which performs better than SpaCE JAM with three basis functions. Nonetheless, the loss in accuracy due to the inclusion of non-linear basis functions is not dramatic, and

SpaCE JAM still tends to outperform other methods for non-Gaussian data, as well as the graphical lasso.

4.5.3 Application to cell signaling data

We apply SpaCE JAM to a data set consisting of measurements for 11 proteins involved in cell signaling, under 14 different perturbations [Sachs et al., 2005]. To begin, we consider data from one of the 14 perturbations ($n = 911$), and compare SpaCE JAM using cubic polynomials to neighborhood selection, the nonparanormal skeptic, and graphical random forests with stability selection. Minimizing the BIC for SpaCE JAM yielded a graph with 16 total edges. We compared SpaCE JAM to competing methods, selecting tuning parameters such that each resulting estimated graph contained 16 edges, as well as 10 and 20 edges for the sake of comparison. Figure 4.3 displays the estimated graphs, along with the directed graph presented in Sachs et al. [2005].

The graphs estimated using different methods are qualitatively different. If we treat the directed graph from Sachs et al. [2005] as the ground truth, then SpaCE JAM with 16 edges correctly identifies 12 of the edges, compared to 11, 9, and 8 using sparse partial correlation, the nonparanormal skeptic, and random forests, respectively.

Next we examined the other 13 perturbations, and found that for graphs with 16 edges, SpaCE JAM chooses on average 0.93, 0.64 and 0.2 more correct edges than sparse partial correlation, nonparanormal skeptic, and graphical random forests, respectively ($p = 0.001$, 0.19 and 0.68 using the paired t-test). Since graphical random forests does not permit arbitrary specification of graph size, when graphs with 16 edges could not be obtained, we used the next largest graph.

In Section 4.1, we showed that these data are not well-represented by linear models even after the nonparanormal transformation. The superior performance of SpaCE JAM in this section confirms this observation. The differences between the SpaCE JAM and graphical random forests results indicate that the approach taken for mod-

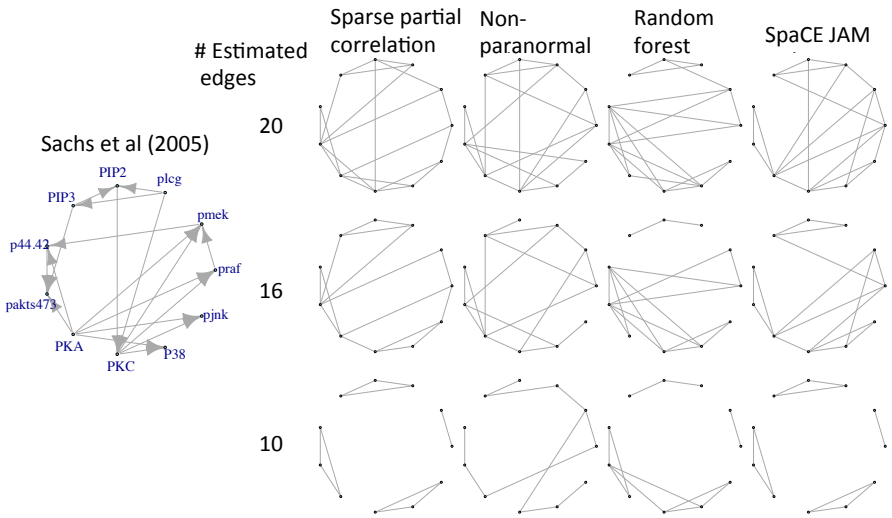

Figure 4.3: Cell signaling data set; graph reported in Sachs et al. [2005] is shown on the left. On the right, graphs were estimated using data from one perturbation of the data set. From top to bottom, panels contain graphs with 20, 16 and 10 edges. From left to right, comparisons are to Peng et al. [2009], Liu et al. [2012], Fellinghauer et al. [2013]. We cannot specify an arbitrary graph size using graphical random forests, so graph sizes for that approach do not match exactly.

eling non-linearity does affect the results obtained.

4.6 Extension to directed graphs

In certain applications, it can be of interest to estimate the causal relationships underlying a set of features, typically represented as a directed acyclic graph. Though directed acyclic graph estimation is in general NP-hard, it is computationally tractable when the causal ordering is known. In fact, in this case, a modification of neighborhood selection is equivalent to the graphical lasso [Shojaie and Michailidis, 2010]. We extend the penalized likelihood framework of Shojaie and Michailidis [2010] to

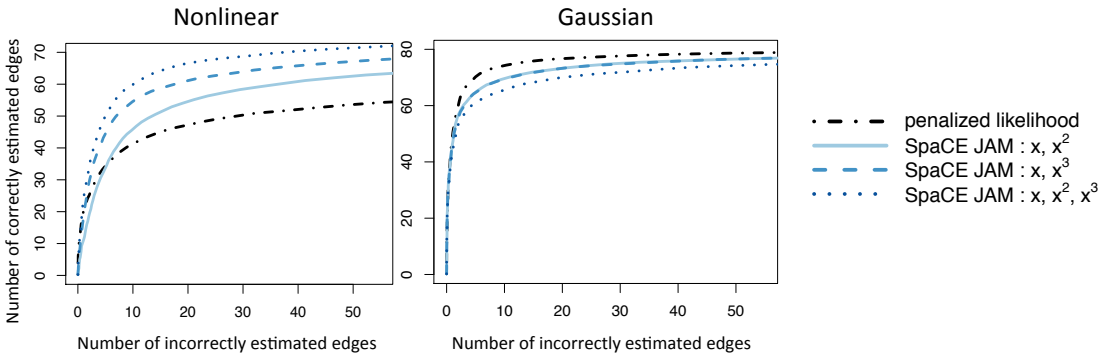

Figure 4.4: Simulation example with directed acyclic graphs. The simulation is exactly as in Section 4.5.1 and Figure 4.2. For each method, the number of correctly and incorrectly estimated edges are averaged over 100 simulated data sets, for a range of 100 tuning parameter values. The competing method is that of Shojaie and Michailidis [2010].

non-linear additive models by solving

$$\underset{\boldsymbol{\beta}_{jk}, 2 \le j \le p, k \prec j}{\text{minimize}} \left\{ \frac{1}{2n} \|\mathbf{x}_j - \sum_{k \prec j} \boldsymbol{\Psi}_{jk} \boldsymbol{\beta}_{jk}\|_2^2 + \lambda \sum_{k \prec j} \|\boldsymbol{\Psi}_{jk} \boldsymbol{\beta}_{jk}\|_2 \right\},$$

where $k \prec j$ indicates that k precedes j in the causal ordering. When $\boldsymbol{\Psi}_{jk} = \mathbf{x}_k$, the model is exactly the penalized Gaussian likelihood approach of Shojaie and Michailidis [2010].

Figure 4.4 displays the same simulation scenario as Section 4.5.1, but with the directed graph estimated using the (known) causal ordering. Results are compared to the penalized Gaussian likelihood approach of Shojaie and Michailidis [2010]. SpaCE JAM performs best when the true relationships are non-linear, and performs competitively when the relationships are linear.

4.7 Theoretical Results

In this section, we provide theory for consistency of the SpaCE JAM graph estimate. Here, we focus on theory for undirected graphs. Similar results also hold for directed graphs, but we omit them due to space considerations. The theoretical development follows that of sparsistency results for sparse additive models with orthogonal series smoothers [Ravikumar et al., 2009].

First, we must define the graph for which SpaCE JAM is consistent. Recall that we have the random vector $x = (x_1, \ldots, x_d) \sim \mathcal{P}$, and $X = [\mathbf{x}_1, \ldots, \mathbf{x}_d] \in \mathbb{R}^{n \times d}$ is a matrix where each row is an independent draw from $\mathcal{P}$. For each $(j, k) \in V \times V$ consider the orthogonal set of basis functions $\psi_{jkt}(\cdot),\ t \in \mathbb{N}$. Define the population level parameters $\boldsymbol{\beta}^*_{jk} \in \mathbb{R}^\infty$ as

$$\left\{\boldsymbol{\beta}^*_{jk}, k = 1, \ldots, d\right\} \equiv \underset{\boldsymbol{\beta}_{jk}\,:\,k=1,\ldots,d}{\arg\min} \left\{ \mathbb{E}|x_j - \sum_{k \neq j} \sum_{t=1}^{\infty} \psi_{jkt}(x_k)\beta_{jkt}|^2 \right\}, \quad j = 1, \ldots, d.$$

Let $S_j = \{k : \|\boldsymbol{\beta}^*_{jk}\| \neq 0\}$ and $s_j = |S_j|$. Let $f_{jk}(x_k) = \sum_{t=1}^{\infty} \psi_{jkt}(x_k)\beta^*_{jkt} \in \mathcal{F}$. Then

$$x_j = \sum_{k \in S_j} f_{jk}(x_k) + \epsilon_j, \quad j = 1, \ldots, d,$$

where $\epsilon_1, \ldots, \epsilon_d$ are residuals, and $\sum_{k \in S_j} f_{jk}(x_k)$ is the best additive approximation to $\mathbb{E}[x_j \mid \{x_k : k \neq j\}]$, in the least-squares sense. We wish to determine which of the $f_{jk}(\cdot)$ are zero.

On observed data, we use a finite set of basis functions to model the $f_{jk}(\cdot)$. Denote the set of r orthogonal basis functions used in the regression of $\mathbf{x}_j$ on $\mathbf{x}_k$ as $\boldsymbol{\Psi}_{jk} = [\psi_{jk1}(\mathbf{x}_k), \ldots, \psi_{jkr}(\mathbf{x}_k)]$, a matrix of dimension $n \times r$ such that $\boldsymbol{\Psi}^T_{jk}\boldsymbol{\Psi}_{jk}/n = \mathbf{I}_r$. Let $\boldsymbol{\beta}^{*(r)}_{jk} = [\boldsymbol{\beta}^*_{jk1}, \ldots, \beta^*_{jkr}]^T$ denote the first r components of $\boldsymbol{\beta}^*_{jk}$. Further, let $\boldsymbol{\Psi}_{S_j}$ be the concatenated basis functions in $\{\boldsymbol{\Psi}_{jk} : k \in S_j\}$, thus $\boldsymbol{\Psi}_{S_j}$ is a matrix of dimension $n \times s_j r$. Also let $\boldsymbol{\Sigma}_{S_j,S_j} = n^{-1}\boldsymbol{\Psi}^T_{S_j}\boldsymbol{\Psi}_{S_j}$ and $\boldsymbol{\Sigma}_{jk,S_j} = n^{-1}\boldsymbol{\Psi}^T_{jk}\boldsymbol{\Psi}_{S_j}$. Define the sub-

gradient of the penalty in (4.7) with respect to $\boldsymbol{\beta}_{jk}$ as $\mathbf{g}_{jk}(\beta)$. On the set S_j, we write the concatenated sub-gradients as $\mathbf{g}_{S_j}$, a vector of length $s_j r$.

Let $\hat{\boldsymbol{\beta}}$ be the parameter estimates from solving (4.7), let $\hat{E}_n = \{(j,k) : \|\hat{\boldsymbol{\beta}}_{jk}\|_2^2 + \|\hat{\boldsymbol{\beta}}_{kj}\|_2^2 \neq 0\}$ be the corresponding estimated edge set, and let $E^* = \{(j,k) : k \in S_j \text{ or } j \in S_k\}$ be the graph obtained from the population level parameters. In Theorem 6, we give precise conditions under which $\text{pr}(\hat{E}_n = E^*) \to 1$ as $n \to \infty$.

Theorem 6. *Let the functions f_{jk} be sufficiently smooth, in the sense that if $f_{jk}^{(r)} = \sum_{t=1}^{r} \psi_{jkt}(x_k)\beta_{jkt}^*$, then $|f_{jk}^{(r)}(x_k) - f_{jk}(x_k)| = O_p(1/r^m)$ uniformly in $(j,k) \in V \times V$ for some $m \in \mathbb{N}$. For $j = 1, \ldots, d$, assume the basis functions satisfy $\Lambda_{min}(\boldsymbol{\Sigma}_{S_j,S_j}) \geq C_{\min} > 0$ with probability tending to 1. Assume the irrepresentability condition,*

$$\|\boldsymbol{\Sigma}_{jk,S_j}\boldsymbol{\Sigma}_{S_j,S_j}^{-1}\hat{\mathbf{g}}_{S_j}\|_2^2 + \|\boldsymbol{\Sigma}_{kj,S_k}\boldsymbol{\Sigma}_{S_k,S_k}^{-1}\hat{\mathbf{g}}_{S_k}\|_2^2 \leq 1 - \delta, \tag{4.11}$$

holds for $(j,k) \notin E^*$ and some $\delta > 0$ with probability tending to 1, where $\hat{\mathbf{g}}_{S_j} = \mathbf{g}_{S_j}(\hat{\beta})$. Assume the following conditions on the number of edges $|E^*|$, the neighborhood size s_j, the regularization parameter λ, and the truncation dimension r:

$$\frac{r\log(r|E^{*c}|)}{\lambda^2 n} \to 0, \quad \max_j \frac{r s_j \log(r|E^*|)}{\lambda^2 n} \to 0, \quad \max_j \frac{s_j}{r^m \lambda} \to 0, \quad \text{and}$$

$$\frac{1}{\rho^*}\max_j \left[\left(\frac{s_j r\log(r|E^*|)}{n}\right)^{1/2} + \frac{s_j}{r^m} + \lambda(r s_j)^{1/2}\right] \to 0$$

where $\rho^* = \min_j \min_{k \in S_j} \|\beta_{jk}^*\|_\infty$. Further, assume the variables

$$\xi_{jkt} \equiv \psi_{jkt}(x_k)\epsilon_j \quad \text{for } j,k \in V, \text{ and } j = 1, \ldots, d$$

have exponential tails, that is $\text{pr}[\,|\xi_{jkt}| > z] \leq ae^{-bz^2}$ for some $a, b > 0$.

Then, the SpaCE JAM graph estimate is consistent: $pr(\hat{E}_n = E^*) \to 1$ as $n \to \infty$.

4.8 Extension of SpaCE JAM to high dimensions

In this section, we propose an approximation to SpaCE JAM that can speed up computations in high dimensions. Our proposal is motivated by recent work in the Gaussian setting by Witten et al. [2011] and Mazumder and Hastie [2012]. They showed that for the graphical lasso (4.4), the connected components of the estimated conditional independence graph are precisely the connected components of the estimated marginal independence graph, where the jth and kth variables are considered marginally independent when $|\mathbf{x}_j^T \mathbf{x}_k| < \lambda$. Consequently, one can obtain the exact solution to the graphical lasso problem in substantially reduced computational time by identifying the connected components of the marginal independence graph, and solving the graphical lasso optimization problem on the variables within each connected component.

We now apply the same principle to SpaCE JAM in order to quickly approximate the solution to (4.7) in high dimensions. Let $\rho_m^{(jk)} = \sup_{f,g \in \mathcal{F}} \rho(f(x_k), g(x_j))$ be the maximal correlation between x_j and x_k over the univariate functions in $\mathcal{F}$ such that $f(x_k)$ and $g(x_j)$ have finite variance. Define the marginal dependence graph $\Gamma_M = (V, E_M)$, where $(j,k) \in E_M$ when $\rho_m^{(jk)} \neq 0$. If the jth and kth variables are in different connected components of Γ_M, then they must be conditionally independent in the large-sample SpaCE JAM graph. Theorem 7, proven in the Appendix, makes this assertion precise.

Theorem 7. *Let $C_1, \ldots C_l$ be the connected components of Γ_M. Suppose the space of functions $\mathcal{F}$ contains linear functions. If $j \in C_u$ and $k \notin C_u$ for some $1 \leq u \leq l$, then $(j,k) \notin E^*$.*

Theorem 7 forms the basis for Algorithm 2. There, we approximate the maximal correlation using the canonical correlation [Mardia et al., 1980] between the basis expansions $\mathbf{\Psi}_{kj}$ and $\mathbf{\Psi}_{jk}$: $\hat{\rho}_m^{(jk)} = \max_{\mathbf{v},\mathbf{w} \in \mathbb{R}^r} \rho(\mathbf{\Psi}_{jk}\mathbf{v}, \mathbf{\Psi}_{jk}\mathbf{w})$.

1: For $(j,k) \in V \times V$, calculate $\hat{\rho}_m^{(jk)}$, the sample canonical correlation between $\mathbf{\Psi}_{kj}$ and $\mathbf{\Psi}_{jk}$.
2: Construct the marginal independence graph: $(j,k) \in \hat{\Gamma}_M$ when $|\hat{\rho}_m^{(jk)}| \geq \lambda_2$.
3: Find the connected components $C_1, \ldots C_l$ of $\hat{\Gamma}_M$.
4: Perform Algorithm 1 on each connected component.

Algorithm 2: A fast approximation for SpaCE JAM in high dimensions

In order to show that i) Algorithm 2 provides an accurate approximation to the original SpaCE JAM problem, ii) the resulting estimator outperforms methods that rely on Gaussian assumptions when those assumptions are violated, and iii) Algorithm 2 is indeed faster than Algorithm 1, we replicated the graph used in Section 4.5.1 five times. This gives $d = 500$ variables, broken into five components. We took $n = 250$, and set $\mathbf{\Psi}_{jk} = [\,\mathbf{x}_k, \mathbf{x}_k^2, \mathbf{x}_k^3\,]$.

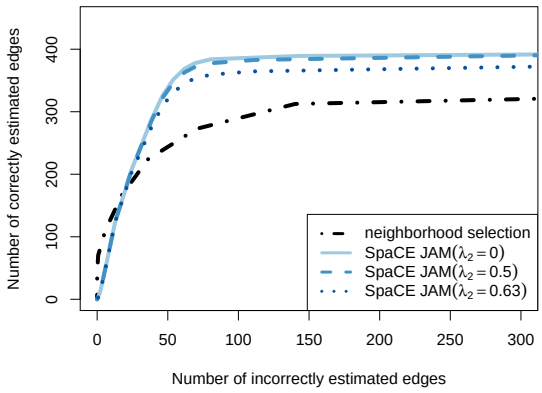

Figure 4.5: Performance of SpaCE JAM using Algorithm 2. The number of correctly and incorrectly estimated edges are averaged over 100 simulated data sets, for each of 100 tuning parameter values. SpaCE JAM was applied using cubic polynomials as basis functions. The competing method is that of Meinshausen and Bühlmann [2006].

In Figure 4.5 we see that when λ_2 in Algorithm 2 is small, there is little loss in statistical efficiency relative to the full SpaCE JAM algorithm (Algorithm 1), which

is a special case of Algorithm 2 with $\lambda_2 = 0$. Further, we see that SpaCE JAM outperforms neighborhood selection even when λ_2 is large. Using Algorithm 2 with $\lambda_2 = 0.5$ and $\lambda_2 = 0.63$ led to a reduction in computation time over Algorithm 1 by 25% and 70%, respectively.

We note here that Theorem 7 continues to hold if maximal correlation $\rho_m^{(jk)}$ is replaced with some other measure of marginal association $\rho_*^{(jk)}$, provided that $\rho_*^{(jk)}$ dominates maximal correlation in the sense that $\rho_*^{(jk)} = 0$ implies that $\rho_m^{(jk)} = 0$. That is, any measure of marginal association, such as mutual information, which detects the same associations as maximal correlation (i.e. $\rho_*^{(jk)} \neq 0$ if $\rho_m^{(jk)} \neq 0$) can be used in Algorithm 2.

Chapter 5

DISCUSSION

This dissertation focuses on the development of new statistical methods for detecting conditional associations in high dimensions. Rigorously evaluating evidence for conditional associations is challenging in the high dimensional setting, since it requires simultaneous estimation of many parameters with a limited amount of data.

In Chapter 2 we proposed the *penalized score test* for linear regression, which can be viewed as an approximation to classical tests which are based on multiple linear regression. The idea of our approach is simple: we estimate the effects of all but a single feature of interest, using penalized regression, and test the residuals for correlation with the held-out feature. Using penalized regression to estimate high-dimensional nuisance parameters is not new. For instance, in spatial statistics, ridge regression is commonly used to account for confounding by location. However, we show that the same principle applies much more broadly. In particular, we study lasso penalization in more detail, and show that variable selection with the lasso can be understood as a decision based on our proposed test. That is, the penalized score test can be used to formalize variable selection with the lasso as a statistical hypothesis test. In Chapter 3 we extended the penalized score test to the general M-estimation setting, where we explored inference in high-dimensional GLMs and Gaussian graphical models in more detail.

In Chapters 2 and 3, we argued that instead of measuring conditional dependence *per se*, the penalized score test strikes a compromise between conditional and marginal dependence, where the compromise is governed by a tuning parameter λ. We showed in simulations that this compromise is useful in practice, in that it controls type-I

error at a similar level to an unbiased test of conditional dependence, while enjoying substantially higher power than classical tests. However, it remains to rigorously evaluate the contexts in which the penalized score test is useful, and when it can be misleading. While we give asymptotic theory which allows the dimension to be much larger than the sample size, in practice we cannot expect that for an arbitrary dimension, sample size, and data structure, we can reliably distinguish the effect of a single feature from the joint effects of other features. Some rules of thumb regarding what is a reasonable statistical question to ask of a particular high-dimensional data set would be greatly useful.

There are a few interesting extensions of Chapters 2 and 3. Here, we focused on the problem of testing the effect of a single feature. However, the results should extend naturally to tests of multiple features. Also, we focused on hypothesis testing, and did not consider effect estimation in detail. Estimating effects, and their associated standard errors, follows in a straightforward way from the penalized score testing framework, but it remains to rigorously evaluate the performance of such estimators on real data. In addition, our theoretical results for the ℓ_1-penalized score test should apply more generally to other penalties. While penalty choice has been studied extensively for the purposes of estimation and prediction, it remains to evaluate which penalties are most effective for inference.

Lastly, the field of high-dimensional inference is rapidly growing, and many methods have recently been proposed [see e.g. Zhang and Zhang, 2011, Lockhart et al., 2013, van de Geer et al., 2013, Javanmard and Montanari, 2013, Lee et al., 2013b, Taylor et al., 2014]. It would be extremely useful to evaluate the relative performance of these methods empirically.

In Chapter 4, we explored modifications of available methods for estimating conditional dependence graphs, to allow for flexible estimation of the conditional effects. In particular, we noted the strong distributional assumptions that are implicit in methods that assume linearity, and proposed a semi-parametric alternative, called

SpaCEJAM, which allows for more flexible estimation of conditional independence relationships. In this chapter, we used the sparsity pattern of groups of coefficients in order to estimate the conditional dependence graph. However, we may be able to perform more formal inference about conditional dependence by applying the penalized score testing framework of Chapters 2 and 3. This would require extending the penalized score test to groups of variables, which we leave to future work.

Appendix A

APPENDIX: TECHNICAL PROOFS

A.1 Proofs for Chapter 2

In this section we prove Proposition 3. First, we state and prove a basic result. We then proceed by stating and proving lemmas needed in the proof of Proposition 3.

Lemma 1. *Let* $\{X_{ij} : i = 1, \ldots, n; j = 1, \ldots, d\}$ *be a set of random variables such that* $\{X_{1j}, \ldots, X_{nj}\}$ *are mutually independent. Assume* $\mathbb{E}(X_{ij}) = 0$*, and that there exist* $h, c > 0$*, not depending on* i *or* j *such that* $\Pr(|X_{ij}| \geq x) \leq 2\exp(-hx^2)$*,* $\forall x > c$*. Denote* $Z_j = \sum_{i=1}^{n} X_{ij}/\sqrt{n}$*. Then*

$$\max_{j=1,\ldots,d} |Z_j| = O_p\left(\log^{1/2}(d)\right).$$

Proof. First we state a well-known equivalent definition of a sub-Gaussian random variable, which follows from, e.g. Lemma 14.2 in Bühlmann and van de Geer [2011]. We have that $\Pr(|X_{ij}| \geq x) \leq 2\exp(-hx^2)$ for $x > c$ if and only if $M_{X_{ij}}(t) \leq \exp(-kt^2)$ for some $k > 0$, where $M_{X_{ij}}(t)$ is the moment generating function of X_{ij}. Using this fact, we know that Z_j is sub-Gaussian since $M_{Z_j}(t) = \prod_{i=1}^{n} M_{X_{ij}}(t/\sqrt{n}) \leq \exp(-kt^2)$. Applying the union bound, we get that

$$\begin{aligned}
\Pr\left[\max_{j=1,\ldots,d} |Z_j| > t\log^{1/2}(d)\right] &\leq \sum_{j=1}^{d} \Pr\left[\,|Z_j| > t\log^{1/2}(d)\right] \\
&\leq 2d\exp(-ht^2\log(d)) \\
&= 2\exp(\log(d)[1 - ht^2]) \\
&\leq 2\exp(\log(2)[1 - ht^2]),
\end{aligned}$$

where the last inequality holds when $ht^2 > 1$ and $d \geq 2$. Thus, we can choose a large value of t, not depending on d, such that $\Pr\left[\max_{j=1,\ldots,d} |Z_j| > t\log^{1/2}(d)\right]$ is arbitrarily small, which gives the result. □

Lemma 2. *Suppose conditions (**A1**),(**A5**) and (**A6**) hold. Then the estimator*

$$\tilde{\mathbf{b}} = \underset{\mathbf{b}:\|\mathbf{b}_{\mathcal{A}^c}\|=0}{\arg\min} \left\{ \frac{1}{2n}\|\mathbf{y} - a_\lambda \mathbf{x} - \mathbf{Z}\mathbf{b}\|_2^2 + \lambda\|\mathbf{b}\|_1 \right\}$$

satisfies $\|\tilde{\mathbf{b}}_{\mathcal{A}} - \mathbf{b}_{\lambda\mathcal{A}}\|_2 = O_p\left(\sqrt{q\log(q)/n}\right)$.

Proof. To simplify the notation, assume without loss of generality that $a_\lambda = 0$. Otherwise we can replace $\mathbf{y}$ with $\mathbf{y} - a_\lambda\mathbf{x}$, and the proof still holds.

Let $Q(\mathbf{c}) = \|\mathbf{y} - \mathbf{Z}_{\mathcal{A}}\mathbf{c}\|_2^2/(2n) + \lambda\|\mathbf{c}\|_1$, so that $\tilde{\mathbf{b}}_{\mathcal{A}} = \arg\min_{\mathbf{c}\in\mathbb{R}^q} Q(\mathbf{c})$. Note that $Q(\cdot)$ is strictly convex, and thus $\tilde{\mathbf{b}}_{\mathcal{A}}$ is unique, since $\mathbf{Z}_{\mathcal{A}}$ is full rank by (**A6**). We will show that for all $\xi > 0$ there exists a constant m, not depending on n, such that

$$\lim_{n\to\infty} \Pr\left[\inf_{\mathbf{c}:\|\mathbf{c}\|_2=m} Q(\mathbf{b}_{\lambda\mathcal{A}} + \mathbf{c}\sqrt{q\log(q)/n}) > Q(\mathbf{b}_{\lambda\mathcal{A}})\right] \geq 1 - \xi. \tag{A.1}$$

Convexity of Q then implies that $\tilde{\mathbf{b}}_{\mathcal{A}}$ is in the ball $\{\mathbf{b}_{\lambda\mathcal{A}} + \mathbf{c}\sqrt{q\log(q)/n} : \|\mathbf{c}\|_2 \leq m\}$ with probability at least $1 - \xi$. Thus, we have $\lim_{n\to\infty}\Pr[\|\tilde{\mathbf{b}}_{\mathcal{A}} - \mathbf{b}_{\lambda\mathcal{A}}\|_2 > m\sqrt{q\log(q)/n}] \leq \xi$, i.e. $\|\tilde{\mathbf{b}}_{\mathcal{A}} - \mathbf{b}_{\lambda\mathcal{A}}\|_2 = O_p(\sqrt{q\log(q)/n})$.

We now proceed to prove (A.1). Let $\mathbf{w} = \arg\min_{\mathbf{c}\in\mathbb{R}^q:\|\mathbf{c}\|_2=m} Q(\mathbf{b}_{\lambda\mathcal{A}}+\mathbf{c}\sqrt{q\log(q)/n})$. Expanding terms, we can write

$$\begin{aligned} Q\left(\mathbf{b}_{\lambda\mathcal{A}} + \mathbf{w}\sqrt{q\log(q)/n}\right) - Q(\mathbf{b}_{\lambda\mathcal{A}}) = &-\frac{\sqrt{q\log(q)}}{n^{3/2}}\mathbf{w}^T\mathbf{Z}_{\mathcal{A}}^T(\mathbf{y} - \mathbf{Z}_{\mathcal{A}}\mathbf{b}_{\lambda\mathcal{A}}) + \frac{q\log(q)}{2n^2}\mathbf{w}^T\mathbf{Z}_{\mathcal{A}}^T\mathbf{Z}_{\mathcal{A}}\mathbf{w} \\ &+ \lambda\|\mathbf{b}_{\lambda\mathcal{A}} + \mathbf{w}\sqrt{q\log(q)/n}\|_1 - \lambda\|\mathbf{b}_{\lambda\mathcal{A}}\|_1. \end{aligned} \tag{A.2}$$

First note that, for $g, h \in \mathbb{R}$, $|g + h| = |g| + \text{sign}(g)h$ when $|h| \leq |g|$. Thus, we have $\left\|\mathbf{b}_{\lambda\mathcal{A}} + \mathbf{w}\sqrt{q\log(q)/n}\right\|_1 = \|\mathbf{b}_{\lambda\mathcal{A}}\|_1 + \boldsymbol{\tau}_{\mathcal{A}}^T\mathbf{w}\sqrt{q\log(q)/n}$ when $m\sqrt{q\log(q)/n} <$

$\min\{|b_{\lambda,1}|,\ldots,|b_{\lambda,q}|\} = b_{\min}$. Since $\sqrt{q\log(q)/n}/b_{\min} \to 0$ by (**A6**), for n large enough we can write

$$\begin{aligned} Q\left(\mathbf{b}_{\lambda\mathcal{A}} + \mathbf{w}\sqrt{q\log(q)/n}\right) - Q(\mathbf{b}_{\lambda\mathcal{A}}) &= -\frac{\sqrt{q\log(q)}}{n^{3/2}}\mathbf{w}^T\mathbf{Z}_{\mathcal{A}}^T(\mathbf{y} - \mathbf{Z}_{\mathcal{A}}\mathbf{b}_{\lambda\mathcal{A}}) + \frac{q\log(q)}{2n^2}\mathbf{w}^T\mathbf{Z}_{\mathcal{A}}^T\mathbf{Z}_{\mathcal{A}}\mathbf{w} \\ &\quad + \lambda\sqrt{\frac{q\log(q)}{n}}\boldsymbol{\tau}_{\mathcal{A}}^T\mathbf{w} \\ &= -\frac{\sqrt{q\log(q)}}{n^{3/2}}\mathbf{w}^T\left[\mathbf{Z}_{\mathcal{A}}^T(\mathbf{y} - \mathbf{Z}_{\mathcal{A}}\mathbf{b}_{\lambda\mathcal{A}}) - \lambda n\boldsymbol{\tau}_{\mathcal{A}}\right] \\ &\quad + \frac{q\log(q)}{2n^2}\mathbf{w}^T\mathbf{Z}_{\mathcal{A}}^T\mathbf{Z}_{\mathcal{A}}\mathbf{w} \end{aligned} \tag{A.3}$$

We now bound $\mathbf{w}^T\left[\mathbf{Z}_{\mathcal{A}}^T(\mathbf{y} - \mathbf{Z}_{\mathcal{A}}\mathbf{b}_{\lambda\mathcal{A}}) - \lambda n\boldsymbol{\tau}_{\mathcal{A}}\right]$ in (A.3). Note that $\mathbf{Z}_{\mathcal{A}}^T(\mathbf{y} - \mathbf{Z}_{\mathcal{A}}\mathbf{b}_{\lambda\mathcal{A}}) - \lambda n\boldsymbol{\tau}_{\mathcal{A}} = \mathbf{Z}_{\mathcal{A}}^T\boldsymbol{\epsilon}$, using (2.12). Thus we have

$$\begin{aligned} \left|\mathbf{w}^T\left[\mathbf{Z}_{\mathcal{A}}^T(\mathbf{y} - \mathbf{Z}_{\mathcal{A}}\mathbf{b}_{\lambda\mathcal{A}}) - \lambda n\boldsymbol{\tau}_{\mathcal{A}}\right]\right| &\leq \|\mathbf{w}\|_1\|\mathbf{Z}_{\mathcal{A}}^T\boldsymbol{\epsilon}\|_\infty \\ &\leq \sqrt{q}\|\mathbf{w}\|_2\|\mathbf{Z}_{\mathcal{A}}^T\boldsymbol{\epsilon}\|_\infty, \end{aligned} \tag{A.4}$$

Now note that $\mathbf{w}^T\mathbf{Z}_{\mathcal{A}}^T\mathbf{Z}_{\mathcal{A}}\mathbf{w}/n \geq \Lambda_{\min}^2(\boldsymbol{\Sigma}_{\mathcal{A}})\|\mathbf{w}\|_2^2$. Thus, we get that

$$\begin{aligned} Q\left(\mathbf{b}_{\lambda\mathcal{A}} + \mathbf{w}\sqrt{q\log(q)/n}\right) - Q(\mathbf{b}_{\lambda\mathcal{A}}) &\geq -\frac{q\log^{1/2}(q)}{n^{3/2}}\|\mathbf{w}\|_2\|\mathbf{Z}_{\mathcal{A}}^T\boldsymbol{\epsilon}\|_\infty + \frac{q\log(q)}{2n}\Lambda_{\min}^2(\boldsymbol{\Sigma}_{\mathcal{A}})\|\mathbf{w}\|_2^2 \\ &= \frac{q\log^{1/2}(q)m}{n}\left(m\log^{1/2}(q)\Lambda_{\min}^2(\boldsymbol{\Sigma}_{\mathcal{A}})/2 - \|\mathbf{Z}_{\mathcal{A}}^T\boldsymbol{\epsilon}\|_\infty/\sqrt{n}\right). \end{aligned}$$

We know $\|\mathbf{Z}_{\mathcal{A}}^T\boldsymbol{\epsilon}\|_\infty/\sqrt{n} = O_p(\log^{1/2}(q))$, by Lemma 1, which applies by (**A5**). Thus, we can choose m, not depending on n, such that (A.1) holds, provided that $\Lambda_{\min}(\boldsymbol{\Sigma}_{\mathcal{A}})$ is bounded below, which is guaranteed by (**A6**). □

Lemma 3. *Suppose conditions (**A1**) and (**A3**-**A6**) hold. Then any minimizer* $\hat{\mathbf{b}}_\lambda$ *of*

$$\|\mathbf{y} - a_\lambda\mathbf{x} - \mathbf{Z}\mathbf{b}\|_2^2/(2n) + \lambda\|\mathbf{b}\|_1 \tag{A.5}$$

satisfies $\|\hat{\mathbf{b}}_{\lambda\mathcal{A}} - \mathbf{b}_{\lambda\mathcal{A}}\|_2 = O_p\left(\sqrt{q\log(q)/n}\right)$ *and* $\lim_{n\to\infty}\Pr[\|\hat{\mathbf{b}}_{\lambda\mathcal{A}^c}\|_2 = 0] = 1$.

Proof. As in Lemma 2, assume without loss of generality that $a_\lambda = 0$.

It suffices to show that $\tilde{\mathbf{b}}$, from Lemma 2, is the unique minimizer of (A.5) with probability tending to 1, since this implies that $\lim_{n\to\infty} \Pr[\|\hat{\mathbf{b}}_{\lambda\mathcal{A}^c}\|_2 = 0] = 1$ and that $[n/(q\log q)]^{1/2}\|\hat{\mathbf{b}}_{\lambda\mathcal{A}} - \mathbf{b}_{\lambda\mathcal{A}}\|_2 = [n/(q\log q)]^{1/2}\|\tilde{\mathbf{b}}_{\mathcal{A}} - \mathbf{b}_{\lambda\mathcal{A}}\|_2 + o_p(1)$.

By the Karush-Kuhn-Tucker conditions, $\tilde{\mathbf{b}}$ is a minimizer of (A.5) if and only if $\mathbf{Z}^T(\mathbf{y} - \mathbf{Z}\tilde{\mathbf{b}})/n = \lambda\tilde{\boldsymbol{\tau}}$ for some $\tilde{\boldsymbol{\tau}}$ satisfying $\|\tilde{\boldsymbol{\tau}}\|_\infty \leq 1$ and $\tilde{\tau}_i = \text{sign}(\tilde{b}_i)$ for $\tilde{b}_i \neq 0$. Since we already know that $\|\mathbf{Z}_{\mathcal{A}}^T(\mathbf{y} - \mathbf{Z}\tilde{\mathbf{b}})/n\|_\infty \leq \lambda$ and $\mathbf{z}_i^T(\mathbf{y} - \mathbf{Z}\tilde{\mathbf{b}})/n = \lambda\text{sign}(\tilde{b}_i)$ for $\tilde{b}_i \neq 0$ (by the definition of $\tilde{\mathbf{b}}$), if we can show that

$$\lim_{n\to\infty} \Pr\left[\frac{1}{n}\left\|\mathbf{Z}_{\mathcal{A}^c}^T(\mathbf{y} - \mathbf{Z}\tilde{\mathbf{b}})\right\|_\infty < \lambda\right] = 1, \tag{A.6}$$

then we will have shown that $\tilde{\mathbf{b}}$ is a minimizer of (A.5) with probability tending to 1. Furthermore, when $\|\mathbf{Z}_{\mathcal{A}^c}^T(\mathbf{y} - \mathbf{Z}\tilde{\mathbf{b}})/n\|_\infty < \lambda$ holds then $\tilde{\mathbf{b}}$ is the unique minimizer of $\|\mathbf{y} - \mathbf{Z}\mathbf{b}\|_2^2/(2n) + \lambda\|\mathbf{b}\|_1$. To see this note that *all* minimizers of (A.5) produce the same fitted values [Tibshirani, 2013]. Thus, if $\|\mathbf{Z}_{\mathcal{A}^c}^T(\mathbf{y} - \mathbf{Z}\tilde{\mathbf{b}})/n\|_\infty < \lambda$, then $\|\mathbf{Z}_{\mathcal{A}^c}^T(\mathbf{y} - \mathbf{Z}\hat{\mathbf{b}}_\lambda)/n\|_\infty < \lambda$ for any minimizer $\hat{\mathbf{b}}_\lambda$ of (A.5), which implies that $\|\hat{\mathbf{b}}_{\lambda\mathcal{A}^c}\|_2 = 0$, by the Karush-Kuhn-Tucker conditions. When $\|\hat{\mathbf{b}}_{\lambda\mathcal{A}^c}\|_2 = 0$, then $\hat{\mathbf{b}}_\lambda = \tilde{\mathbf{b}}$, which is unique, as was argued in the proof of Lemma 2.

We now show that (A.6) holds. Adding and subtracting $\mathbf{Z}_{\mathcal{A}^c}^T\mathbf{Z}_{\mathcal{A}}\mathbf{b}_{\lambda\mathcal{A}}/n$ we get

$$\frac{1}{n}\|\mathbf{Z}_{\mathcal{A}^c}^T(\mathbf{y} - \mathbf{Z}\tilde{\mathbf{b}})\|_\infty \leq \frac{1}{n}\|\mathbf{Z}_{\mathcal{A}^c}^T(\mathbf{y} - \mathbf{Z}_{\mathcal{A}}\mathbf{b}_{\lambda\mathcal{A}})\|_\infty + \frac{1}{n}\|\mathbf{Z}_{\mathcal{A}^c}^T\mathbf{Z}_{\mathcal{A}}(\mathbf{b}_{\lambda\mathcal{A}} - \tilde{\mathbf{b}}_{\mathcal{A}})\|_\infty. \tag{A.7}$$

First, we bound $(1/n)\|\mathbf{Z}_{\mathcal{A}^c}^T\mathbf{Z}_{\mathcal{A}}(\mathbf{b}_{\lambda\mathcal{A}} - \tilde{\mathbf{b}}_{\mathcal{A}})\|_\infty$ in (A.7). By (**A4**) and Lemma 2, we have

$$\begin{aligned}
\frac{1}{n}\left\|\mathbf{Z}_{\mathcal{A}^c}^T\mathbf{Z}_{\mathcal{A}}(\tilde{\mathbf{b}}_{\mathcal{A}} - \mathbf{b}_{\lambda\mathcal{A}})\right\|_\infty &\leq \frac{1}{n}\left\|\mathbf{Z}_{\mathcal{A}^c}^T\mathbf{Z}_{\mathcal{A}}\right\|_\infty \|\tilde{\mathbf{b}}_{\mathcal{A}} - \mathbf{b}_{\lambda\mathcal{A}}\|_\infty \\
&\leq \frac{1}{n}\left\|\mathbf{Z}_{\mathcal{A}^c}^T\mathbf{Z}_{\mathcal{A}}\right\|_\infty \|\tilde{\mathbf{b}}_{\mathcal{A}} - \mathbf{b}_{\lambda\mathcal{A}}\|_2 \\
&= o_p(1).
\end{aligned} \tag{A.8}$$

We now bound $(1/n)\|\mathbf{Z}_{\mathcal{A}^c}^T(\mathbf{y}-\mathbf{Z}_{\mathcal{A}}\mathbf{b}_{\lambda\mathcal{A}})\|_\infty$ in (A.7). Recall that $\mathbf{Z}^T(\mathbf{y}-\mathbf{Z}\mathbf{b}_\lambda)-n\lambda\boldsymbol{\tau} = \mathbf{Z}^T\boldsymbol{\epsilon}$. Using Lemma 1 to bound $\|\mathbf{Z}_{\mathcal{A}^c}^T\boldsymbol{\epsilon}\|_\infty/n$ we get that

$$\begin{aligned}\frac{1}{n}\|\mathbf{Z}_{\mathcal{A}^c}^T(\mathbf{y}-\mathbf{Z}\mathbf{b}_\lambda)\|_\infty &= \frac{1}{n}\|\mathbf{Z}_{\mathcal{A}^c}^T(\mathbf{y}-\mathbf{Z}\mathbf{b}_\lambda) - n\lambda\boldsymbol{\tau}_{\mathcal{A}^c} + n\lambda\boldsymbol{\tau}_{\mathcal{A}^c}\|_\infty \\ &\leq \frac{1}{n}\|\mathbf{Z}_{\mathcal{A}^c}^T\boldsymbol{\epsilon}\|_\infty + \lambda\|\boldsymbol{\tau}_{\mathcal{A}^c}\|_\infty \\ &\leq O_p\left(\frac{\log^{1/2}(d-q)}{n^{1/2}}\right) + \lambda(1-\delta) \\ &= o_p(1) + \lambda(1-\delta), \end{aligned} \tag{A.9}$$

where we used $\|\boldsymbol{\tau}_{\mathcal{A}^c}\|_\infty < 1-\delta$, by (**A3**), and $\log^{1/2}(d-q)/n^{1/2} \to 0$, by (**A6**).

Altogether, applying the bounds (A.8) and (A.9) to (A.7), we have

$$\left\|(1/n)\mathbf{Z}_{\mathcal{A}^c}^T(\mathbf{y}-\mathbf{Z}\tilde{\mathbf{b}})\right\|_\infty \leq \lambda(1-\delta) + o_p(1),$$

which is smaller than λ with probability tending to 1. Thus, (A.6) holds. □

Lemma 4. *Suppose conditions (**A1**) and (**A3**-**A6**) hold. Then any minimizer $\hat{\mathbf{b}}_\lambda$ of $\|\mathbf{y}-a_\lambda\mathbf{x}-\mathbf{Z}\mathbf{b}\|_2^2/(2n)+\lambda\|\mathbf{b}\|_1$ satisfies*

$$\sqrt{n}(\hat{\mathbf{b}}_{\lambda\mathcal{A}}-\mathbf{b}_{\lambda\mathcal{A}}) = \frac{1}{\sqrt{n}}\boldsymbol{\Sigma}_{\mathcal{A}}^{-1}\mathbf{Z}_{\mathcal{A}}^T\boldsymbol{\epsilon} + o_p(1).$$

Proof. As in Lemma 2, assume without loss of generality that $a_\lambda = 0$.

By the stationary conditions defining $\hat{\mathbf{b}}_\lambda$ we have that

$$\mathbf{Z}^T(\mathbf{y}-\mathbf{Z}\hat{\mathbf{b}}_\lambda)/n = \lambda\hat{\boldsymbol{\tau}}, \tag{A.10}$$

for some $\hat{\boldsymbol{\tau}} \in [-1,1]^d$. First, we show that $\Pr[\hat{\boldsymbol{\tau}}_{\mathcal{A}} = \boldsymbol{\tau}_{\mathcal{A}}] \to 1$ as $n \to \infty$ (recall from (2.12) that $\boldsymbol{\tau}_{\mathcal{A}} = \operatorname{sign}(\mathbf{b}_{\lambda\mathcal{A}})$). By Lemma 3, given $\xi > 0$, there exists a $c > 0$ such that $\lim_{n\to\infty}\Pr\left[\|\hat{\mathbf{b}}_{\lambda\mathcal{A}}-\mathbf{b}_{\lambda\mathcal{A}}\|_2 < c\sqrt{q\log(q)/n}\right] > 1-\xi$. Further, by (**A6**),

we have $c\sqrt{q\log(q)/n} < b_{\min}$ for n large enough. Now, $b_{\min} > c\sqrt{q\log(q)/n} > \|\hat{\mathbf{b}}_{\lambda\mathcal{A}} - \mathbf{b}_{\lambda\mathcal{A}}\|_2 \geq \|\hat{\mathbf{b}}_{\lambda\mathcal{A}} - \mathbf{b}_{\lambda\mathcal{A}}\|_\infty$ implies that all elements of $\hat{\mathbf{b}}_{\lambda\mathcal{A}}$ are non-zero and that $\text{sign}(\hat{\mathbf{b}}_{\lambda\mathcal{A}}) = \text{sign}(\mathbf{b}_{\lambda\mathcal{A}})$. Thus, since $\hat{\boldsymbol{\tau}}_{\mathcal{A}} = \text{sign}(\hat{\mathbf{b}}_{\lambda\mathcal{A}})$ when all elements of $\hat{\mathbf{b}}_{\lambda\mathcal{A}}$ are non-zero, we have $\lim_{n\to\infty}\Pr[\hat{\boldsymbol{\tau}}_{\mathcal{A}} = \boldsymbol{\tau}_{\mathcal{A}}] > 1-\xi$. Since ξ is arbitrary, we have $\lim_{n\to\infty}\Pr[\hat{\boldsymbol{\tau}}_{\mathcal{A}} = \boldsymbol{\tau}_{\mathcal{A}}] = 1$.

Thus, from (A.10), we can write

$$\begin{aligned}
\mathbf{0} &= \mathbf{Z}_{\mathcal{A}}^T(\mathbf{y} - \mathbf{Z}\hat{\mathbf{b}}_\lambda)/\sqrt{n} - \sqrt{n}\lambda\hat{\boldsymbol{\tau}}_{\mathcal{A}} \\
&= \mathbf{Z}_{\mathcal{A}}^T(\mathbf{y} - \mathbf{Z}\hat{\mathbf{b}}_\lambda)/\sqrt{n} - \sqrt{n}\lambda\boldsymbol{\tau}_{\mathcal{A}} + o_p(1) \\
&= \mathbf{Z}_{\mathcal{A}}^T(\mathbf{y} - \mathbf{Z}\mathbf{b}_\lambda)/\sqrt{n} - \sqrt{n}\lambda\boldsymbol{\tau}_{\mathcal{A}} - \frac{1}{n}\mathbf{Z}_{\mathcal{A}}^T\mathbf{Z}_{\mathcal{A}}\sqrt{n}(\hat{\mathbf{b}}_{\lambda\mathcal{A}} - \mathbf{b}_{\lambda\mathcal{A}}) \\
&\quad - \frac{1}{n}\mathbf{Z}_{\mathcal{A}}^T\mathbf{Z}_{\mathcal{A}^c}\sqrt{n}\hat{\mathbf{b}}_{\lambda\mathcal{A}^c} + o_p(1). \qquad \text{(A.11)}
\end{aligned}$$

Now, from Lemma 3 we have that $\Pr[\|\hat{\mathbf{b}}_{\lambda\mathcal{A}^c}\|_2 = 0] \to 1$. Also, from (2.12) we know that $\mathbf{Z}_{\mathcal{A}}^T(\mathbf{y} - \mathbf{Z}\mathbf{b}_\lambda)/\sqrt{n} - \sqrt{n}\lambda\boldsymbol{\tau}_{\mathcal{A}} = \mathbf{Z}_{\mathcal{A}}^T\boldsymbol{\epsilon}/\sqrt{n}$. Thus, we can write

$$\mathbf{0} = \mathbf{Z}_{\mathcal{A}}^T\boldsymbol{\epsilon}/\sqrt{n} - \frac{1}{n}\mathbf{Z}_{\mathcal{A}}^T\mathbf{Z}_{\mathcal{A}}\sqrt{n}(\hat{\mathbf{b}}_{\lambda\mathcal{A}} - \mathbf{b}_{\lambda\mathcal{A}}) + o_p(1).$$

Multiplying through by $\boldsymbol{\Sigma}_{\mathcal{A}}^{-1}$ gives the result. □

With these lemmas, we can now prove our main result.

Proof of Proposition 3. Recall that $\hat{\mathbf{b}}_\lambda^0 = \arg\min_{\mathbf{b}}\{\|\mathbf{y} - \mathbf{Z}\mathbf{b}\|_2^2/(2n) + \lambda\|\mathbf{b}\|_1\}$. Since $a_\lambda = 0$, we have that $\hat{\mathbf{b}}_\lambda^0 = \hat{\mathbf{b}}_\lambda$, using the notation of Lemmas 3 and 4.

First, we have

$$\begin{aligned}
T_\lambda &= \frac{1}{\sqrt{n}}\mathbf{x}^T(\mathbf{y} - \mathbf{Z}\hat{\mathbf{b}}_\lambda) \\
&= \frac{1}{\sqrt{n}}\mathbf{x}^T(\mathbf{y} - \mathbf{Z}\mathbf{b}_\lambda) - \frac{1}{n}\mathbf{x}^T\mathbf{Z}_{\mathcal{A}}\sqrt{n}(\hat{\mathbf{b}}_{\lambda\mathcal{A}} - \mathbf{b}_{\lambda\mathcal{A}}) - \frac{1}{\sqrt{n}}\mathbf{x}^T\mathbf{Z}_{\mathcal{A}^c}\hat{\mathbf{b}}_{\lambda\mathcal{A}^c}.
\end{aligned}$$

Now, $\Pr[\|\hat{\mathbf{b}}_{\lambda\mathcal{A}^c}\|_2 = 0] \to 1$ by Lemma 3. Thus, we get

$$T_\lambda = \frac{1}{\sqrt{n}}\mathbf{x}^T(\mathbf{y} - \mathbf{Z}\mathbf{b}_\lambda) - \boldsymbol{\sigma}_{x\mathcal{A}}^T\sqrt{n}(\hat{\mathbf{b}}_{\lambda\mathcal{A}} - \mathbf{b}_{\lambda\mathcal{A}}) + o_p(1),$$

where $\boldsymbol{\sigma}_{x\mathcal{A}} = \mathbf{Z}_{\mathcal{A}}^T\mathbf{x}/n$.

Now, using Lemma 4, we know

$$\sqrt{n}\boldsymbol{\sigma}_{x\mathcal{A}}^T(\hat{\mathbf{b}}_{\lambda\mathcal{A}} - \mathbf{b}_{\lambda\mathcal{A}}) = \frac{1}{\sqrt{n}}\boldsymbol{\sigma}_{x\mathcal{A}}^T\boldsymbol{\Sigma}_{\mathcal{A}}^{-1}\mathbf{Z}_{\mathcal{A}}^T\boldsymbol{\epsilon} + o_p(1).$$

Also, since $H_{0\lambda} : a_\lambda = 0$ is true, we have $\mathbf{x}^T(\mathbf{y} - \mathbf{Z}\mathbf{b}_\lambda) = \mathbf{x}^T\boldsymbol{\epsilon}$, and so

$$T_\lambda = \frac{1}{\sqrt{n}}\left(\mathbf{x}^T\boldsymbol{\epsilon} - \boldsymbol{\sigma}_{x\mathcal{A}}^T\boldsymbol{\Sigma}_{\mathcal{A}}^{-1}\mathbf{Z}_{\mathcal{A}}^T\boldsymbol{\epsilon}\right) + o_p(1) = \frac{1}{\sqrt{n}}\mathbf{x}^T(\mathbf{I}_n - \mathbf{P}_{\mathcal{A}})\boldsymbol{\epsilon} + o_p(1).$$

Dividing by $\sigma_{\boldsymbol{\epsilon}}\sqrt{\mathbf{x}^T(\mathbf{I}_n - \mathbf{P}_{\mathcal{A}})\mathbf{x}/n}$ we get

$$\frac{T_\lambda}{\sigma_{\boldsymbol{\epsilon}}\sqrt{\mathbf{x}^T(\mathbf{I}_n - \mathbf{P}_{\mathcal{A}})\mathbf{x}/n}} = \frac{\mathbf{r}^T\boldsymbol{\epsilon}}{\sigma_{\boldsymbol{\epsilon}}\|\mathbf{r}\|_2} + o_p(1),$$

where $\mathbf{r}^T = \mathbf{x}^T(\mathbf{I}_n - \mathbf{P}_{\mathcal{A}})$. Now, the Lindeberg-Feller Central Limit Theorem guarantees that $T_\lambda / \left(\sigma_{\boldsymbol{\epsilon}}\sqrt{\mathbf{x}^T(\mathbf{I}_n - \mathbf{P}_{\mathcal{A}})\mathbf{x}/n}\right) \to_d N(0,1)$ if the Lindeberg condition holds:

$$\lim_{n\to\infty}\sum_{i=1}^n \mathbb{E}\left[\frac{(r_i\epsilon_i)^2}{\sigma_{\boldsymbol{\epsilon}}^2\|\mathbf{r}\|_2^2}1\left\{\frac{|r_i\epsilon_i|}{\sigma_{\boldsymbol{\epsilon}}\|\mathbf{r}\|_2} > \eta\right\}\right] = 0, \quad \forall\eta > 0.$$

Using that $|r_i| \le \|\mathbf{r}\|_\infty$, and that the ϵ_i's are identically distributed, we get

$$\begin{aligned}
&\sum_{i=1}^n \mathbb{E}\left[\frac{(r_i\epsilon_i)^2}{\sigma_{\boldsymbol{\epsilon}}^2\|\mathbf{r}\|_2^2}1\left\{\frac{|r_i\epsilon_i|}{\sigma_{\boldsymbol{\epsilon}}\|\mathbf{r}\|_2} > \eta\right\}\right] \\
&\quad\le \sum_{i=1}^n \frac{r_i^2}{\sigma_{\boldsymbol{\epsilon}}^2\|\mathbf{r}\|_2^2}\mathbb{E}\left[\epsilon_i^2 1\left\{\frac{|\epsilon_i|\|\mathbf{r}\|_\infty}{\sigma_{\boldsymbol{\epsilon}}\|\mathbf{r}\|_2} > \eta\right\}\right] \\
&\quad= \frac{1}{\sigma_{\boldsymbol{\epsilon}}^2}\mathbb{E}\left[\epsilon_1^2 1\left\{\frac{|\epsilon_1|\|\mathbf{r}\|_\infty}{\sigma_{\boldsymbol{\epsilon}}\|\mathbf{r}\|_2} > \eta\right\}\right].
\end{aligned}$$

Now, since $\|\mathbf{r}\|_\infty/\|\mathbf{r}\|_2 \to 0$ by (**A2**), we have that $\epsilon_1^2 1\{|\epsilon_1|\|\mathbf{r}\|_\infty/(\sigma_{\boldsymbol{\epsilon}}\|\mathbf{r}\|_2) > \eta\} \to_p 0$. Thus we can apply the Dominated Convergence Theorem, using the dominating random variable ϵ_1^2, which satisfies $\mathbb{E}[\epsilon_1^2] = \sigma_\epsilon^2 < \infty$ and $\epsilon_1^2 \geq \epsilon_1^2 1\{|\epsilon_1|\|\mathbf{r}\|_\infty/(\sigma_{\boldsymbol{\epsilon}}\|\mathbf{r}\|_2) > \eta\}$ with probability 1, to get that

$$\frac{1}{\sigma_{\boldsymbol{\epsilon}}^2}\mathbb{E}\left[\epsilon_1^2 1\left\{\frac{|\epsilon_1|\|\mathbf{r}\|_\infty}{\sigma_{\boldsymbol{\epsilon}}\|\mathbf{r}\|_2} > \eta\right\}\right] \to 0,$$

which in turn gives the result. □

A.2 Proofs for Chapter 3

In this section we prove Theorem 5. We begin by stating and prove a few lemmas needed in the proof of Theorem 5.

Lemma 5. *Suppose conditions (***B1***), (***B3***), (***B5-6***) and (***B8***) hold. Then*

$$\tilde{\mathbf{b}} = \underset{\mathbf{b}:\|\mathbf{b}_{\mathcal{A}^c}\|_2=0}{\arg\min}\{l(a_\lambda, \mathbf{b})/n + \lambda\|\mathbf{b}\|_1\}$$

satisfies $\|\tilde{\mathbf{b}}_{\mathcal{A}} - \mathbf{b}_{\lambda\mathcal{A}}\|_2 = O_p(\sqrt{(q\log q)/n})$.

Proof. To simplify notation, suppress the dependence on a, e.g. $l(\mathbf{b})/n = l(a_\lambda, \mathbf{b})/n$, $\dot{\mathbf{l}}(\mathbf{b}) = \dfrac{\partial}{\partial \mathbf{b}} l(a_\lambda, \mathbf{b})$ and $\ddot{\mathbf{l}}(\mathbf{b}) = \dfrac{\partial^2}{\partial \mathbf{b}^2} l(a_\lambda, \mathbf{b})$. Further, let $h_n = \sqrt{(q\log q)/n}$.

Let $Q(\mathbf{c}) = l([\mathbf{c}, \mathbf{0}])/n + \lambda\|\mathbf{c}\|_1$, where $\mathbf{c} \in \mathbb{R}^q$, so that $\tilde{\mathbf{b}}_{\mathcal{A}} = \arg\min_{\mathbf{c}} Q(\mathbf{c})$. To prove this lemma, we will show that for all $\xi > 0$, there exists a constant m, not depending on n, such that

$$\lim_{n\to\infty} \Pr\left[\inf_{\mathbf{c}:\|\mathbf{c}\|_2=m} Q\left(\mathbf{b}_{\lambda\mathcal{A}} + h_n\mathbf{c}\right) > Q\left(\mathbf{b}_{\lambda\mathcal{A}}\right)\right] \geq 1 - \xi. \tag{A.12}$$

Convexity of Q then implies that $\tilde{\mathbf{b}}_{\mathcal{A}}$ is in the region $\{\mathbf{b}_{\lambda\mathcal{A}} + h_n\mathbf{w} : \|\mathbf{w}\|_2 < m\}$ with probability at least $1-\xi$. Thus we have that $\lim_{n\to\infty} \Pr[\, h_n^{-1}\|\tilde{\mathbf{b}}_{\mathcal{A}} - \mathbf{b}_{\lambda\mathcal{A}}\|_2 < m] < 1-\xi$, i.e. $\|\tilde{\mathbf{b}}_{\mathcal{A}} - \mathbf{b}_{\lambda\mathcal{A}}\|_2 = O_p(h_n)$, where $h_n = \sqrt{(q\log q)/n}$.

We now proceed to prove (A.12). Let $\mathbf{w} = \arg\min_{\mathbf{c}:\|\mathbf{c}\|_2=m} Q\left(\mathbf{b}_{\lambda\mathcal{A}} + h_n\mathbf{c}\right)$. Using Taylor's expansion, we can write

$$Q\left(\mathbf{b}_{\lambda\mathcal{A}} + h_n\mathbf{w}\right) - Q\left(\mathbf{b}_{\lambda\mathcal{A}}\right) = \frac{1}{n}\left[h_n\dot{\mathbf{l}}_{\mathcal{A}}(\mathbf{b}_\lambda)^T\mathbf{w} + \frac{h_n^2}{2}\mathbf{w}^T\ddot{\mathbf{l}}_{\mathcal{A}\mathcal{A}}(\bar{\mathbf{b}})\mathbf{w}\right] + \lambda\left(\|\mathbf{b}_{\lambda\mathcal{A}} + h_n\mathbf{w}\|_1 - \|\mathbf{b}_{\lambda\mathcal{A}}\|_1\right), \tag{A.13}$$

for some $\bar{\mathbf{b}}$ between $\mathbf{b}_{\lambda\mathcal{A}}$ and $\mathbf{b}_{\lambda\mathcal{A}} + h_n\mathbf{w}$.

First note that, for $g, t \in \mathbb{R}$, $|g + t| = |g| + \text{sign}(g)t$ when $|t| \leq |g|$. Thus, we have $\|\mathbf{b}_{\lambda\mathcal{A}} + h_n\mathbf{w}\|_1 = \|\mathbf{b}_{\lambda\mathcal{A}}\|_1 + h_n\boldsymbol{\tau}_{\mathcal{A}}^T\mathbf{w}$ when $mh_n < \min\{|b_{\lambda,1}|, \ldots, |b_{\lambda,q}|\} = b_{\min}$. Since $h_n/b_{\min} \to 0$ by (**B6**), for n large enough we can write

$$\begin{aligned} Q\left(\mathbf{b}_{\lambda\mathcal{A}} + h_n\mathbf{w}\right) - Q\left(\mathbf{b}_{\lambda\mathcal{A}}\right) &= \frac{1}{n}\left[h_n\dot{\mathbf{l}}_{\mathcal{A}}(\mathbf{b}_\lambda)^T\mathbf{w} + \frac{h_n^2}{2}\mathbf{w}^T\ddot{\mathbf{l}}_{\mathcal{A}\mathcal{A}}(\bar{\mathbf{b}})\mathbf{w}\right] \\ &\quad + \lambda h_n\boldsymbol{\tau}_{\mathcal{A}}^T\mathbf{w} \\ &= \frac{h_n}{n}\left[\dot{\mathbf{l}}_{\mathcal{A}}(\mathbf{b}_\lambda) + n\lambda\boldsymbol{\tau}_{\mathcal{A}}\right]^T\mathbf{w} + \frac{h_n^2}{2n}\mathbf{w}^T\ddot{\mathbf{l}}_{\mathcal{A}\mathcal{A}}(\bar{\mathbf{b}})\mathbf{w}. \end{aligned} \tag{A.14}$$

We first bound the term $\left[\dot{\mathbf{l}}_{\mathcal{A}}(\mathbf{b}_\lambda) + n\lambda\boldsymbol{\tau}_{\mathcal{A}}\right]^T\mathbf{w}$ in (A.14). Using the Cauchy-Schwartz inequality, we have

$$\begin{aligned} \left[\dot{\mathbf{l}}_{\mathcal{A}}(\mathbf{b}_\lambda) + n\lambda\boldsymbol{\tau}_{\mathcal{A}}\right]^T\mathbf{w} &\leq \left\|\dot{\mathbf{l}}_{\mathcal{A}}(\mathbf{b}_\lambda) + n\lambda\boldsymbol{\tau}_{\mathcal{A}}\right\|_\infty \|\mathbf{w}\|_1 \\ &\leq \sqrt{q}\left\|\dot{\mathbf{l}}_{\mathcal{A}}(\mathbf{b}_\lambda) + n\lambda\boldsymbol{\tau}_{\mathcal{A}}\right\|_\infty \|\mathbf{w}\|_2. \end{aligned} \tag{A.15}$$

Now, we bound the term $\mathbf{w}^T\ddot{\mathbf{l}}_{\mathcal{A}\mathcal{A}}(\bar{\mathbf{b}})\mathbf{w}$ in (A.14). Adding and subtracting $\ddot{\mathbf{l}}_{\mathcal{A}\mathcal{A}}(\mathbf{b}_\lambda)-$

$n\mathbf{U}_{\mathcal{A}\mathcal{A}}$ we get

$$\begin{aligned}\mathbf{w}^T\ddot{\mathbf{l}}_{\mathcal{A}\mathcal{A}}(\bar{\mathbf{b}})\mathbf{w} &= n\mathbf{w}^T\left[\ddot{\mathbf{l}}_{\mathcal{A}\mathcal{A}}(\bar{\mathbf{b}})/n - \ddot{\mathbf{l}}_{\mathcal{A}\mathcal{A}}(\mathbf{b}_\lambda)/n\right]\mathbf{w} + n\mathbf{w}^T\left[\ddot{\mathbf{l}}_{\mathcal{A}\mathcal{A}}(\mathbf{b}_\lambda)/n - \mathbf{U}_{\mathcal{A}\mathcal{A}}(\mathbf{b}_\lambda)\right]\mathbf{w} \\ &\quad + n\mathbf{w}^T\mathbf{U}_{\mathcal{A}\mathcal{A}}\mathbf{w} \\ &\geq -n\left\|\ddot{\mathbf{l}}_{\mathcal{A}\mathcal{A}}(\bar{\mathbf{b}})/n - \ddot{\mathbf{l}}_{\mathcal{A}\mathcal{A}}(\mathbf{b}_\lambda)/n\right\|_2^2 - n\left\|\ddot{\mathbf{l}}_{\mathcal{A}\mathcal{A}}(\mathbf{b}_\lambda)/n - \mathbf{U}_{\mathcal{A}\mathcal{A}}\right\|_2^2 \\ &\quad + n\mathbf{w}^T\mathbf{U}_{\mathcal{A}\mathcal{A}}\mathbf{w}.\end{aligned}$$

Now, applying (**B8**) to the above, and noting that $\mathbf{w}^T\mathbf{U}_{\mathcal{A}\mathcal{A}}\mathbf{w} \geq \Lambda^2_{\min}(\mathbf{U}_{\mathcal{A}\mathcal{A}})$, we get

$$\begin{aligned}\frac{h_n^2}{2n}\mathbf{w}^T\ddot{\mathbf{l}}_{\mathcal{A}\mathcal{A}}(\bar{\mathbf{b}})\mathbf{w} &\geq \frac{h_n^2}{2}O_p(\|\bar{\mathbf{b}} - \mathbf{b}_\lambda\|_2^2) + \frac{h_n^2}{2}O_p(h_n^2) + \frac{h_n^2}{2}\Lambda_{\min}(\mathbf{U}_{\mathcal{A}\mathcal{A}})\|\mathbf{w}\|_2^2 \\ &\geq o_p(h_n^2) + \frac{(mh_n)^2}{2}\Lambda^2_{\min}(\mathbf{U}_{\mathcal{A}\mathcal{A}}). \qquad \text{(A.16)}\end{aligned}$$

Altogether, using the bounds (A.15) and (A.16) in (A.14) we have that

$$\begin{aligned}Q\left(\mathbf{b}_{\lambda\mathcal{A}} + h_n\mathbf{w}\right) - Q\left(\mathbf{b}_{\lambda\mathcal{A}}\right) &\geq \frac{(mh_n)^2}{2}\Lambda^2_{\min}(\mathbf{U}_{\mathcal{A}\mathcal{A}}) - \frac{mh_n\sqrt{q}}{n}\left\|\dot{\mathbf{l}}_{\mathcal{A}}(\mathbf{b}_\lambda) + n\lambda\boldsymbol{\tau}_{\mathcal{A}}\right\|_\infty + o_p(h_n^2) \\ &= mh_n\left(mh_n\Lambda^2_{\min}(\mathbf{U}_{\mathcal{A}\mathcal{A}})/2 - \frac{\sqrt{q}}{n}\left\|\dot{\mathbf{l}}_{\mathcal{A}}(\mathbf{b}_\lambda) + n\lambda\boldsymbol{\tau}_{\mathcal{A}}\right\|_\infty\right) + o_p(h_n^2).\end{aligned}$$

Now, applying Lemma 1 and (**B5**) we have that $(\sqrt{q}/n)\left\|\dot{\mathbf{l}}_{\mathcal{A}}(\mathbf{b}_\lambda) + n\lambda\boldsymbol{\tau}_{\mathcal{A}}\right\|_\infty = O_p(\sqrt{(q\log q)/n}) = O_p(h_n)$. Thus, using the rates in (**B6**), we know we can choose m, such that for n large enough $mh_n\Lambda^2_{\min}(\mathbf{U}_{\mathcal{A}\mathcal{A}})/2 \geq (\sqrt{q}/n)\left\|\dot{\mathbf{l}}_{\mathcal{A}}(\mathbf{b}_\lambda) + n\lambda\boldsymbol{\tau}_{\mathcal{A}}\right\|_\infty$ with probability greater than $1-\xi$. With this choice of m, we get that $\lim_{n\to\infty}\Pr\left[Q\left(\mathbf{b}_{\lambda\mathcal{A}} + h_n\mathbf{w}\right) - Q\left(\mathbf{b}_{\lambda\mathcal{A}}\right.\right.$ $1-\xi$, and thus (A.12) holds. □

Lemma 6. *Suppose conditions (***B1***) and (***B3-8***) hold. Then there exists a minimizer* $\hat{\mathbf{b}}_\lambda$ *of* $Q(\mathbf{b}) = l(a_\lambda, \mathbf{b})/n + \lambda\|\mathbf{b}\|_1$ *that satisfies* $\|\hat{\mathbf{b}}_{\lambda\mathcal{A}} - \mathbf{b}_{\lambda\mathcal{A}}\|_2 = O_p(\sqrt{(q\log q)/n})$ *and* $\Pr[\|\hat{\mathbf{b}}_{\lambda\mathcal{A}^c}\|_2 = 0] \to 1$ *as* $n \to \infty$.

Proof. Let $\tilde{\mathbf{b}}$ be as in the Lemma 5, and let $h_n = \sqrt{(q\log q)/n}$. It suffices to show that $\tilde{\mathbf{b}}$ minimizes $Q(\mathbf{b})$ with probability tending to 1, since this implies the existence

of a minimizer $\hat{\mathbf{b}}_\lambda$ satisfying $\lim_{n\to\infty}\Pr[\|\hat{\mathbf{b}}_{\lambda\mathcal{A}^c}\|_2 = 0] = 1$ and $h_n^{-1}\|\hat{\mathbf{b}}_{\lambda\mathcal{A}} - \mathbf{b}_{\lambda\mathcal{A}}\|_2 = h_n^{-1}\|\tilde{\mathbf{b}}_{\mathcal{A}} - \mathbf{b}_{\lambda\mathcal{A}}\|_2 + o_p(1) = O_p(1)$.

By the Karush-Kuhn-Tucker conditions, $\tilde{\mathbf{b}}$ is a minimizer of $l(\mathbf{b})/n + \lambda\|\mathbf{b}\|_1$ when $\dot{\mathbf{l}}(\tilde{\mathbf{b}})/n = -\lambda\tilde{\boldsymbol{\tau}}$ for some $\tilde{\boldsymbol{\tau}}$ satisfying $\|\tilde{\boldsymbol{\tau}}\|_\infty \leq 1$ and $\tilde{\tau}_j = \text{sign}(\tilde{b}_j)$ for $\tilde{b}_j \neq 0$. We already know that $\|\dot{\mathbf{l}}_{\mathcal{A}}(\tilde{\mathbf{b}})/n\|_\infty \leq \lambda$ and $\dot{l}_j(\tilde{\mathbf{b}})/n = -\lambda\text{sign}(\tilde{b}_j)$ for $\tilde{b}_j \neq 0$, i.e. the conditions hold for $\dot{\mathbf{l}}_{\mathcal{A}}(\tilde{\mathbf{b}})$. Thus, if we can show that

$$\lim_{n\to\infty}\Pr\left[\frac{1}{n}\left\|\dot{\mathbf{l}}_{\mathcal{A}^c}(\tilde{\mathbf{b}})\right\|_\infty < \lambda\right] = 1, \tag{A.17}$$

then we will have shown that $\tilde{\mathbf{b}}$ minimizes $Q(\mathbf{b})$ with probability tending to 1.

We now show that (A.17) holds. Adding and subtracting $\dot{\mathbf{l}}_{\mathcal{A}^c}(\mathbf{b}_\lambda)$ we get

$$\frac{1}{n}\left\|\dot{\mathbf{l}}_{\mathcal{A}^c}(\tilde{\mathbf{b}})\right\|_\infty \leq \frac{1}{n}\left\|\dot{\mathbf{l}}_{\mathcal{A}^c}(\tilde{\mathbf{b}}) - \dot{\mathbf{l}}_{\mathcal{A}^c}(\mathbf{b}_\lambda)\right\|_\infty + \frac{1}{n}\left\|\dot{\mathbf{l}}_{\mathcal{A}^c}(\mathbf{b}_\lambda)\right\|_\infty \tag{A.18}$$

We begin by bounding $(1/n)\left\|\dot{\mathbf{l}}_{\mathcal{A}^c}(\tilde{\mathbf{b}}) - \dot{\mathbf{l}}_{\mathcal{A}^c}(\mathbf{b}_\lambda)\right\|_\infty$ in (A.18). Using Taylor's expansion we can write

$$\begin{aligned}
\frac{1}{n}\left\|\dot{\mathbf{l}}_{\mathcal{A}^c}(\tilde{\mathbf{b}}) - \dot{\mathbf{l}}_{\mathcal{A}^c}(\mathbf{b}_\lambda)\right\|_\infty &= \frac{1}{n}\left\|\ddot{\mathbf{l}}_{\mathcal{A}^c\mathcal{A}}(\bar{\mathbf{b}})(\tilde{\mathbf{b}}_{\mathcal{A}} - \mathbf{b}_{\lambda\mathcal{A}}) + \ddot{\mathbf{l}}_{\mathcal{A}^c\mathcal{A}^c}(\bar{\mathbf{b}})(\tilde{\mathbf{b}}_{\mathcal{A}^c} - \mathbf{b}_{\lambda\mathcal{A}^c})\right\|_\infty \\
&= \frac{1}{n}\left\|\ddot{\mathbf{l}}_{\mathcal{A}^c\mathcal{A}}(\bar{\mathbf{b}})(\tilde{\mathbf{b}}_{\mathcal{A}} - \mathbf{b}_{\lambda\mathcal{A}})\right\|_\infty \\
&\leq \frac{1}{n}\left\|\ddot{\mathbf{l}}_{\mathcal{A}^c\mathcal{A}}(\bar{\mathbf{b}})\right\|_\infty\left\|\tilde{\mathbf{b}}_{\mathcal{A}} - \mathbf{b}_{\lambda\mathcal{A}}\right\|_\infty \\
&\leq \frac{1}{n}\left\|\ddot{\mathbf{l}}_{\mathcal{A}^c\mathcal{A}}(\bar{\mathbf{b}})\right\|_\infty\left\|\tilde{\mathbf{b}}_{\mathcal{A}} - \mathbf{b}_{\lambda\mathcal{A}}\right\|_2 = o_p(1),
\end{aligned} \tag{A.19}$$

where we used that $\frac{1}{n}\left\|\ddot{\mathbf{l}}_{\mathcal{A}^c\mathcal{A}}(\bar{\mathbf{b}})\right\|_\infty = o_p(1/h_n)$ by (**B7**) while $\|\tilde{\mathbf{b}}_{\mathcal{A}} - \mathbf{b}_{\lambda\mathcal{A}}\|_2 = O_p(h_n)$ by Lemma 5

We now bound $(1/n)\left\|\dot{\mathbf{l}}_{\mathcal{A}^c}(\mathbf{b}_\lambda)\right\|_\infty$ in (A.18). Adding and subtracting $\lambda n\boldsymbol{\tau}_{\mathcal{A}^c}$ we get

$$\begin{aligned}\frac{1}{n}\left\|\dot{\mathbf{l}}_{\mathcal{A}^c}(\mathbf{b}_\lambda)\right\| &= \frac{1}{n}\left\|\dot{\mathbf{l}}_{\mathcal{A}^c}(\mathbf{b}_\lambda) + \lambda n\boldsymbol{\tau}_{\mathcal{A}^c} - \lambda n\boldsymbol{\tau}_{\mathcal{A}^c}\right\|_\infty \\ &\leq \frac{1}{n}\left\|\dot{\mathbf{l}}_{\mathcal{A}^c}(\mathbf{b}_\lambda) + \lambda n\boldsymbol{\tau}_{\mathcal{A}^c}\right\|_\infty + \lambda\|\boldsymbol{\tau}_{\mathcal{A}^c}\|_\infty \qquad \text{(A.20)} \\ &\leq \frac{1}{n}\left\|\dot{\mathbf{l}}_{\mathcal{A}^c}(\mathbf{b}_\lambda) + \lambda n\boldsymbol{\tau}_{\mathcal{A}^c}\right\|_\infty + \lambda(1-\delta),\end{aligned}$$

where we used (**B4**) to bound $\|\boldsymbol{\tau}_{\mathcal{A}^c}\|_\infty$. Now, applying Lemma 1 to $(1/n)\left\|\dot{\mathbf{l}}_{\mathcal{A}^c}(\mathbf{b}_\lambda) + \lambda n\boldsymbol{\tau}_{\mathcal{A}^c}\right\|_\infty$, which applies by (**B5**), we get that

$$\begin{aligned}\frac{1}{n}\left\|\dot{\mathbf{l}}_{\mathcal{A}^c}(\mathbf{b}_\lambda) + \lambda n\boldsymbol{\tau}_{\mathcal{A}^c}\right\| &\leq O_p\left(\frac{\log^{1/2}(d-q)}{n^{1/2}}\right) + \lambda(1-\delta) \\ &= o_p(1) + \lambda(1-\delta), \qquad \text{(A.21)}\end{aligned}$$

where we used the rates in (**B6**) to get $O_p\left(\log^{1/2}(d-q)/(n^{1/2})\right) = o_p(1)$. Altogether, applying the bounds (A.19) and (A.21) to (A.18), we have

$$\frac{1}{n}\left\|\dot{\mathbf{l}}_{\mathcal{A}^c}(\tilde{\mathbf{b}})\right\|_\infty \leq \lambda(1-\delta) + o_p(1),$$

which is smaller than λ with probability tending to 1. Thus, (A.17) holds. □

Lemma 7. *Suppose conditions (***B1***) and (***B3-8***) hold. Let $\hat{\mathbf{b}}_\lambda$ be the minimizer of $l(a_\lambda, \mathbf{b})/n + \lambda\|\mathbf{b}\|_1$ satisfying Lemma 6, and let $\mathbf{w} \in \mathbb{R}^q$ be a vector with $\|\mathbf{w}\|_2 = 1$. Then*

$$\sqrt{n}\mathbf{w}^T(\hat{\mathbf{b}}_\lambda - \mathbf{b}_\lambda) = -\frac{1}{\sqrt{n}}\mathbf{w}^T\mathbf{U}_{\mathcal{A}\mathcal{A}}^{-1}\left(\dot{\mathbf{l}}_{\mathcal{A}}(\mathbf{b}_\lambda) + \lambda n\boldsymbol{\tau}_{\mathcal{A}}\right) + o_p(1).$$

In other words $\hat{\mathbf{b}}_\lambda$ is asymptotically linear with influence function $-\mathbf{U}_{\mathcal{A}\mathcal{A}}^{-1}\left(\dot{\mathbf{l}}_{\mathcal{A}}(a_\lambda, \mathbf{b}_\lambda; X_i) + \lambda\boldsymbol{\tau}_{\mathcal{A}}\right)$.

Proof. As in Lemma 5, let $h_n = \sqrt{(q\log q)/n}$. By the stationary conditions defining $\hat{\mathbf{b}}_\lambda$ we have that

$$\mathbf{0} = \dot{\mathbf{l}}(\hat{\mathbf{b}}_\lambda)/n + \lambda\hat{\boldsymbol{\tau}}, \qquad \text{(A.22)}$$

for some $\hat{\boldsymbol{\tau}} \in [-1,1]^d$. First, we show that $\Pr[\hat{\boldsymbol{\tau}}_{\mathcal{A}} = \boldsymbol{\tau}_{\mathcal{A}}] \to 1$ as $n \to \infty$ (recall from (**B1** that $\boldsymbol{\tau}_{\mathcal{A}} = \text{sign}(\mathbf{b}_{\lambda\mathcal{A}})$). By Lemma 6, given a $\xi > 0$, there exists a $m > 0$ such that $\lim_{n\to\infty} \Pr\left[\|\hat{\mathbf{b}}_{\lambda\mathcal{A}} - \mathbf{b}_{\lambda\mathcal{A}}\|_2 < mh_n\right] > 1 - \xi$. Further, by (**B6**), we have $mh_n < b_{\min}$ for n large enough. Now, $b_{\min} > mh_n > \|\hat{\mathbf{b}}_{\lambda\mathcal{A}} - \mathbf{b}_{\lambda\mathcal{A}}\|_2 \geq \|\hat{\mathbf{b}}_{\lambda\mathcal{A}} - \mathbf{b}_{\lambda\mathcal{A}}\|_\infty$ implies that all elements of $\hat{\mathbf{b}}_{\lambda\mathcal{A}}$ are non-zero and that $\text{sign}(\hat{\mathbf{b}}_{\lambda\mathcal{A}}) = \text{sign}(\mathbf{b}_{\lambda\mathcal{A}})$. Thus, since $\hat{\boldsymbol{\tau}}_{\mathcal{A}} = \text{sign}(\hat{\mathbf{b}}_{\lambda\mathcal{A}})$ when all elements of $\hat{\mathbf{b}}_{\lambda\mathcal{A}}$ are non-zero, we have $\lim_{n\to\infty} \Pr[\hat{\boldsymbol{\tau}}_{\mathcal{A}} = \boldsymbol{\tau}_{\mathcal{A}}] > 1 - \xi$. Since ξ is arbitrary, we have $\lim_{n\to\infty} \Pr[\hat{\boldsymbol{\tau}}_{\mathcal{A}} = \boldsymbol{\tau}_{\mathcal{A}}] = 1$.

Now, given an $\mathbf{r} \in \mathbb{R}^q$ with $\|\mathbf{r}\|_2 = O(1)$ to be specified later, (A.22) implies that

$$\begin{aligned}
0 &= \mathbf{r}^T \left(\dot{\mathbf{l}}_{\mathcal{A}}(\hat{\mathbf{b}}_\lambda)/\sqrt{n} + \lambda\sqrt{n}\boldsymbol{\tau}_A\right) + o_p(1) \\
&= \frac{1}{\sqrt{n}}\mathbf{r}^T \left(\dot{\mathbf{l}}_{\mathcal{A}}(\mathbf{b}_\lambda) + \lambda n \boldsymbol{\tau}_A\right) + \frac{1}{n}\mathbf{r}^T\ddot{\mathbf{l}}_{\mathcal{A}\mathcal{A}}(\bar{\mathbf{b}}_\lambda)\sqrt{n}(\hat{\mathbf{b}}_{\lambda\mathcal{A}} - \mathbf{b}_{\lambda\mathcal{A}}) \\
&\quad + \frac{1}{n}\mathbf{r}^T\ddot{\mathbf{l}}_{\mathcal{A}\mathcal{A}^c}(\bar{\mathbf{b}}_\lambda)\sqrt{n}\hat{\mathbf{b}}_{\lambda\mathcal{A}^c} + o_p(1) \\
&= \frac{1}{\sqrt{n}}\mathbf{r}^T \left(\dot{\mathbf{l}}_{\mathcal{A}}(\mathbf{b}_\lambda) + \lambda n \boldsymbol{\tau}_A\right) + \frac{1}{n}\mathbf{r}^T\ddot{\mathbf{l}}_{\mathcal{A}\mathcal{A}}(\bar{\mathbf{b}}_\lambda)\sqrt{n}(\hat{\mathbf{b}}_{\lambda\mathcal{A}} - \mathbf{b}_{\lambda\mathcal{A}}) + o_p(1). \qquad \text{(A.23)}
\end{aligned}$$

Adding and subtracting $\ddot{\mathbf{l}}_{\mathcal{A}\mathcal{A}}(\bar{\mathbf{b}}_\lambda) - \mathbf{U}_{\mathcal{A}\mathcal{A}}$ we get

$$\begin{aligned}
\frac{1}{n}\mathbf{r}^T\ddot{\mathbf{l}}_{\mathcal{A}\mathcal{A}}(\bar{\mathbf{b}}_\lambda)\sqrt{n}(\hat{\mathbf{b}}_{\lambda\mathcal{A}} - \mathbf{b}_{\lambda\mathcal{A}}) &= \frac{1}{n}\mathbf{r}^T \left[\ddot{\mathbf{l}}_{\mathcal{A}\mathcal{A}}(\bar{\mathbf{b}}_\lambda) - \ddot{\mathbf{l}}_{\mathcal{A}\mathcal{A}}(\mathbf{b}_\lambda)\right]\sqrt{n}(\hat{\mathbf{b}}_{\lambda\mathcal{A}} - \mathbf{b}_{\lambda\mathcal{A}}) \\
&\quad + \mathbf{r}^T \left[\frac{1}{n}\ddot{\mathbf{l}}_{\mathcal{A}\mathcal{A}}(\mathbf{b}_\lambda) - \mathbf{U}_{\mathcal{A}\mathcal{A}}\right]\sqrt{n}(\hat{\mathbf{b}}_{\lambda\mathcal{A}} - \mathbf{b}_{\lambda\mathcal{A}}) \\
&\quad + \mathbf{r}^T\mathbf{U}_{\mathcal{A}\mathcal{A}}\sqrt{n}(\hat{\mathbf{b}}_{\lambda\mathcal{A}} - \mathbf{b}_{\lambda\mathcal{A}}). \qquad \text{(A.24)}
\end{aligned}$$

By Lipschitz continuity in (**B8**) and Lemma 6 we have

$$\begin{aligned}
\frac{1}{n}\mathbf{r}^T \left[\ddot{\mathbf{l}}_{\mathcal{A}\mathcal{A}}(\bar{\mathbf{b}}_\lambda) - \ddot{\mathbf{l}}_{\mathcal{A}\mathcal{A}}(\mathbf{b}_\lambda)\right]\sqrt{n}(\hat{\mathbf{b}}_{\lambda\mathcal{A}} - \mathbf{b}_{\lambda\mathcal{A}}) &\leq \|\mathbf{r}\|_2 \left\|\ddot{\mathbf{l}}_{\mathcal{A}\mathcal{A}}(\bar{\mathbf{b}}_\lambda)/n - \ddot{\mathbf{l}}_{\mathcal{A}\mathcal{A}}(\mathbf{b}_\lambda)/n\right\|_2 \sqrt{n}\|\hat{\mathbf{b}}_{\lambda\mathcal{A}} - \mathbf{b}_{\lambda\mathcal{A}}\|_2 \\
&= O_p\left(\sqrt{n}\|\hat{\mathbf{b}}_{\lambda\mathcal{A}} - \mathbf{b}_{\lambda\mathcal{A}}\|_2^2\right) \\
&= O_p\left(\frac{q\log(q)}{\sqrt{n}}\right) = o_p(1). \qquad \text{(A.25)}
\end{aligned}$$

Also by (**B8**) and Lemma 6 we have that

$$\mathbf{r}^T \left[\frac{1}{n}\ddot{\mathbf{l}}_{\mathcal{A}\mathcal{A}}(\mathbf{b}_\lambda) - \mathbf{U}_{\mathcal{A}\mathcal{A}}\right] \sqrt{n}(\hat{\mathbf{b}}_{\lambda\mathcal{A}} - \mathbf{b}_{\lambda\mathcal{A}}) \leq \|\mathbf{r}\|_2 \left\|\frac{1}{n}\ddot{\mathbf{l}}_{\mathcal{A}\mathcal{A}}(\mathbf{b}_\lambda) - \mathbf{U}_{\mathcal{A}\mathcal{A}}\right\|_2 \sqrt{n}\|\hat{\mathbf{b}}_{\lambda\mathcal{A}} - \mathbf{b}_{\lambda\mathcal{A}}\|_2$$
$$= O_p\left(\frac{q\log(q)}{\sqrt{n}}\right) = o_p(1). \tag{A.26}$$

Thus, applying the bounds (A.25) and (A.26) to (A.24), we get that

$$\frac{1}{n}\mathbf{r}^T\ddot{\mathbf{l}}_{\mathcal{A}\mathcal{A}}(\bar{\mathbf{b}}_\lambda)\sqrt{n}(\hat{\mathbf{b}}_{\lambda\mathcal{A}} - \mathbf{b}_{\lambda\mathcal{A}}) = \mathbf{r}^T\mathbf{U}_{\mathcal{A}\mathcal{A}}\sqrt{n}(\hat{\mathbf{b}}_{\lambda\mathcal{A}} - \mathbf{b}_{\lambda\mathcal{A}}) + o_p(1).$$

Applying this to (A.23) we get that

$$0 = \frac{1}{\sqrt{n}}\mathbf{r}^T\left(\dot{\mathbf{l}}_{\mathcal{A}}(\mathbf{b}_\lambda) + \lambda n\boldsymbol{\tau}_A\right) + \mathbf{r}^T\mathbf{U}_{\mathcal{A}\mathcal{A}}\sqrt{n}(\hat{\mathbf{b}}_{\lambda\mathcal{A}} - \mathbf{b}_{\lambda\mathcal{A}}) + o_p(1).$$

Choosing $\mathbf{r} = \mathbf{U}_{\mathcal{A}\mathcal{A}}^{-1}\mathbf{w}$ with $\|\mathbf{w}\|_2 = 1$, and noting that $\|\mathbf{r}\|_2 \leq 1/\Lambda_{\min}(\mathbf{U}_{\mathcal{A}\mathcal{A}}) = O(1)$ by (**B6**), we get the result. □

Proof of Theorem 5. Recall that $\hat{\mathbf{b}}_\lambda^0 = \arg\min_{\mathbf{b}}\{l(0,\mathbf{b}_\lambda)/n + \lambda\|\mathbf{b}\|_1\}$, and that when $a_\lambda = 0$, we have $\hat{\mathbf{b}}_\lambda^0 = \hat{\mathbf{b}}_\lambda$, using the notation of Lemmas 5 and 6. If $\hat{\mathbf{b}}_\lambda^0$ is not unique, let it be the minimizer satisfying Lemmas 6 and 7.

First, for some $\bar{\mathbf{b}}$ between $\mathbf{b}_\lambda$ and $\hat{\mathbf{b}}_\lambda^0$ we have

$$T_\lambda = \frac{1}{\sqrt{n}}\dot{l}_a(\hat{\mathbf{b}}_\lambda^0) = \frac{1}{\sqrt{n}}\dot{l}_a(\mathbf{b}_\lambda) + \frac{1}{n}\ddot{\mathbf{l}}_{\mathcal{A}a}^T(\bar{\mathbf{b}})\sqrt{n}(\hat{\mathbf{b}}_{\lambda\mathcal{A}}^0 - \mathbf{b}_{\lambda\mathcal{A}}) + \frac{1}{\sqrt{n}}\ddot{\mathbf{l}}_{\mathcal{A}^c a}^T(\bar{\mathbf{b}})\hat{\mathbf{b}}_{\lambda\mathcal{A}^c}^0.$$

Now, $\Pr[\|\hat{\mathbf{b}}_{\lambda\mathcal{A}^c}^0\|_2 = 0] \to 0$ by Lemma 6, and so

$$T_\lambda = \frac{1}{\sqrt{n}}\dot{l}_a(\mathbf{b}_\lambda) + \frac{1}{n}\ddot{\mathbf{l}}_{\mathcal{A}a}^T(\bar{\mathbf{b}})\sqrt{n}(\hat{\mathbf{b}}_{\lambda\mathcal{A}}^0 - \mathbf{b}_{\lambda\mathcal{A}}) + o_p(1).$$

Adding and subtracting $\ddot{\mathbf{l}}_{\mathcal{A}a}(\mathbf{b}_\lambda)/n - \mathbf{u}_{\mathcal{A}a}$ we can write

$$\begin{aligned}\ddot{\mathbf{l}}_{\mathcal{A}a}(\bar{\mathbf{b}})/n\sqrt{n}(\hat{\mathbf{b}}^0_{\lambda\mathcal{A}} - \mathbf{b}_{\lambda\mathcal{A}}) = \frac{1}{n}\left[\ddot{\mathbf{l}}_{\mathcal{A}a}(\bar{\mathbf{b}}) - \ddot{\mathbf{l}}_{\mathcal{A}a}(\mathbf{b}_\lambda)\right]^T \sqrt{n}(\hat{\mathbf{b}}^0_{\lambda\mathcal{A}} - \mathbf{b}_{\lambda\mathcal{A}}) \\ + \left[\ddot{\mathbf{l}}_{\mathcal{A}a}(\mathbf{b}_\lambda)/n - \mathbf{u}_{\mathcal{A}a}\right]^T \sqrt{n}(\hat{\mathbf{b}}^0_{\lambda\mathcal{A}} - \mathbf{b}_{\lambda\mathcal{A}}) + \mathbf{u}^T_{\mathcal{A}a}\sqrt{n}(\hat{\mathbf{b}}^0_{\lambda\mathcal{A}} - \mathbf{b}_{\lambda\mathcal{A}}).\end{aligned}$$

By (**B8**) and Lemma 6 we have that

$$\frac{1}{n}\left[\ddot{\mathbf{l}}_{\mathcal{A}a}(\bar{\mathbf{b}}) - \ddot{\mathbf{l}}_{\mathcal{A}a}(\mathbf{b}_\lambda)\right]^T \sqrt{n}(\hat{\mathbf{b}}^0_{\lambda\mathcal{A}} - \mathbf{b}_{\lambda\mathcal{A}}) = o_p(1),$$

and that

$$\left[\ddot{\mathbf{l}}_{\mathcal{A}a}(\mathbf{b}_\lambda)/n - \mathbf{u}_{\mathcal{A}a}\right]^T \sqrt{n}(\hat{\mathbf{b}}^0_{\lambda\mathcal{A}} - \mathbf{b}_{\lambda\mathcal{A}}) = o_p(1).$$

Thus, we get

$$T_\lambda = \frac{1}{\sqrt{n}}\dot{l}_a(\mathbf{b}_\lambda) + \sqrt{n}\mathbf{u}^T_{\mathcal{A}a}(\hat{\mathbf{b}}^0_{\lambda\mathcal{A}} - \mathbf{b}_{\lambda\mathcal{A}}) + o_p(1).$$

Now, using Lemma 7 we get that

$$\begin{aligned}T_\lambda &= \frac{1}{\sqrt{n}}\dot{l}_a(\mathbf{b}_\lambda) - \frac{1}{\sqrt{n}}\mathbf{u}^T_{\mathcal{A}a}\mathbf{U}^{-1}_{\mathcal{A}\mathcal{A}}\left(\dot{\mathbf{l}}_{\mathcal{A}}(\mathbf{b}_\lambda) + \lambda n\boldsymbol{\tau}_{\mathcal{A}}\right) + o_p(1) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^n Z_i + o_p(1),\end{aligned}$$

where $Z_i \equiv \dot{l}_a(\mathbf{b}_\lambda; X_i) - \mathbf{u}^T_{\mathcal{A}a}\mathbf{U}^{-1}_{\mathcal{A}\mathcal{A}}\left(\dot{\mathbf{l}}_{\mathcal{A}}(\mathbf{b}_\lambda; X_i) + \lambda\boldsymbol{\tau}_{\mathcal{A}}\right)$. Note that Z_i's are independent and identically distributed, with $\mathbb{E}_{\mathcal{P}}[Z_i] = 0$ and $\mathrm{var}_{\mathcal{P}}[Z_i] \equiv v = v_{aa} + \mathbf{u}^T_{\mathcal{A}a}\mathbf{U}^{-1}_{\mathcal{A}\mathcal{A}}(\mathbf{V}_{\mathcal{A}\mathcal{A}}\mathbf{U}^{-1}_{\mathcal{A}\mathcal{A}}\mathbf{u}_{\mathcal{A}a} - 2\mathbf{v}_{\mathcal{A}a})$. Thus, applying the Central Limit Theorem to T_λ/v, we get the result. □

A.3 Proofs for Chapter 4

Proof of Theorem 6

Proof. First, $\hat{\boldsymbol{\beta}}$ is a solution to (4.7) if and only if

$$-\frac{1}{n}\boldsymbol{\Psi}_{jk}^T\left(\mathbf{x}_j - \sum_{l\neq j}\boldsymbol{\Psi}_{jl}\hat{\boldsymbol{\beta}}_{jl}\right) + \lambda \mathbf{g}_{jk}(\hat{\boldsymbol{\beta}}) = 0 \quad \text{for } (j,k) \in V \times V, \tag{A.27}$$

where $\mathbf{g}_{jk}(\hat{\boldsymbol{\beta}})$ is the vector satisfying

$$\begin{aligned} \mathbf{g}_{jk}(\boldsymbol{\beta}) = \frac{\boldsymbol{\Psi}_{jk}\boldsymbol{\beta}_{jk}}{(\|\boldsymbol{\Psi}_{jk}\boldsymbol{\beta}_{jk}\|_2^2 + \|\boldsymbol{\Psi}_{kj}\boldsymbol{\beta}_{kj}\|_2^2)^{1/2}} \quad &\text{when} \quad \|\boldsymbol{\beta}_{jk}\|_2 + \|\boldsymbol{\beta}_{kj}\|_2 \neq 0 \\ \|\mathbf{g}_{jk}(\boldsymbol{\beta})\|_2^2 + \|\mathbf{g}_{kj}(\boldsymbol{\beta})\|_2^2 \leq 1 \quad &\text{when} \quad \|\boldsymbol{\beta}_{jk}\|_2 + \|\boldsymbol{\beta}_{kj}\|_2 = 0. \end{aligned}$$

We base our proof on the primal-dual witness method of Wainwright [2009]. That is, we construct a coefficient-subgradient pair $(\hat{\boldsymbol{\beta}}, \hat{\mathbf{g}})$, and show that they solve (4.7) and produce the correct sparsity pattern, with probability tending to 1. For $(j,k) \in E^*$, we construct $\hat{\boldsymbol{\beta}}_{jk}$ and the corresponding sub-gradients $\hat{\mathbf{g}}_{jk}$ using SpaCE JAM, restricted to edges in E^*:

$$\underset{\boldsymbol{\beta}_{jk}:(j,k)\in E^*}{\arg\min}\left\{\frac{1}{2n}\sum_{j=1}^{d}\|\mathbf{x}_j - \sum_{k\in S_j}\boldsymbol{\Psi}_{jk}\boldsymbol{\beta}_{jk}\|_2^2 + \lambda\sum_{(j,k)\in E^*}\left(\|\boldsymbol{\Psi}_{jk}\boldsymbol{\beta}_{jk}\|_2^2 + \|\boldsymbol{\Psi}_{kj}\boldsymbol{\beta}_{jk}\|_2^2\right)^{1/2}\right\}. \tag{A.28}$$

For $(j,k) \in E^{*c}$, we set $\hat{\boldsymbol{\beta}}_{jk} = 0$, and use (A.27) to solve for the remaining $\hat{\mathbf{g}}_{jk}$ when $k \notin S_j$. Now, $\hat{\boldsymbol{\beta}}$ is a solution to (4.7) if

$$\|\mathbf{g}_{jk}(\hat{\boldsymbol{\beta}}_{jk})\|_2^2 + \|\mathbf{g}_{kj}(\hat{\boldsymbol{\beta}}_{kj})\|_2 \leq 1 \text{ for } (j,k) \notin E^*. \tag{A.29}$$

In addition, $\hat{E}_n = E^*$ when

$$\hat{\boldsymbol{\beta}}_{S_j} \neq \mathbf{0} \text{ for } j = 1, \ldots, d. \tag{A.30}$$

Thus, it suffices to show that that Equations (A.29) and (A.30) hold with high probability.

Condition (A.30): We start with the 'primal' problem. The stationary condition for $\hat{\boldsymbol{\beta}}_{S_j}$ is given by

$$-\frac{1}{n}\boldsymbol{\Psi}_{S_j}^T(\mathbf{x}_j - \boldsymbol{\Psi}_{S_j}\hat{\boldsymbol{\beta}}_{S_j}) + \lambda \hat{\mathbf{g}}_{S_j} = 0.$$

Denote by $\sum_{k\in S_j}\left[f_{jk}(\mathbf{x}_j) - f_{jk}^{(r)}(\mathbf{x}_j)\right] = \mathbf{w}_j$ the truncation error from including only r basis terms. We can write $\mathbf{x}_j = \boldsymbol{\Psi}_{S_j}\boldsymbol{\beta}_{S_j}^{*(r)} + \mathbf{w}_j + \boldsymbol{\epsilon}_j$. And so

$$\frac{1}{n}\boldsymbol{\Psi}_{S_j}^T\left(\boldsymbol{\Psi}_{S_j}(\hat{\boldsymbol{\beta}}_{S_j} - \boldsymbol{\beta}_{S_j}^{*(r)}) - \mathbf{w}_j - \boldsymbol{\epsilon}_j\right) + \lambda\hat{\mathbf{g}}_{S_j} = 0,$$

or

$$(\hat{\boldsymbol{\beta}}_{S_j} - \boldsymbol{\beta}_{S_j}^{*(r)}) = \left(\frac{1}{n}\boldsymbol{\Psi}_{S_j}^T\boldsymbol{\Psi}_{S_j}\right)^{-1}\left(\frac{1}{n}\boldsymbol{\Psi}_{S_j}^T\mathbf{w}_j + \frac{1}{n}\boldsymbol{\Psi}_{S_j}^T\boldsymbol{\epsilon}_j - \lambda\hat{\mathbf{g}}_{S_j}\right), \tag{A.31}$$

using the assumption that $\frac{1}{n}\boldsymbol{\Psi}_{S_j}^T\boldsymbol{\Psi}_{S_j}$ is invertible. We will now show that the inequality

$$\max_j \|\hat{\boldsymbol{\beta}}_{S_j} - \boldsymbol{\beta}_{S_j}^{*(r)}\|_\infty < \min_j \min_{k\in S_j} \|\boldsymbol{\beta}_{jk}^{*(r)}\|_\infty/2 \equiv \rho^*/2 \tag{A.32}$$

holds with high probability. This implies that $\|\hat{\boldsymbol{\beta}}_{jk}\|_2 \neq 0$ if $\|\boldsymbol{\beta}_{jk}^{*(r)}\|_2 \neq 0$.

From (A.31) we have that

$$\begin{aligned}\max_j \|\hat{\boldsymbol{\beta}}_{S_j} - \boldsymbol{\beta}_{S_j}^{*(r)}\|_\infty &\leq \max_j \left\|\boldsymbol{\Sigma}_{S_j,S_j}^{-1}\frac{1}{n}\boldsymbol{\Psi}_{S_j}^T\mathbf{w}_j\right\|_\infty + \max_j \left\|\boldsymbol{\Sigma}_{S_j,S_j}^{-1}\frac{1}{n}\boldsymbol{\Psi}_{S_j}^T\epsilon_j\right\|_\infty + \max_j \lambda\left\|\boldsymbol{\Sigma}_{S_j,S_j}^{-1}\hat{\mathbf{g}}_{S_j}\right\|_\infty \\ &\equiv T_1 + T_2 + T_3.\end{aligned}$$

Thus, to show (A.32) it suffices to bound T_1, T_2, and T_3.

- Bounding T_1:

  By assumption, we have that $|f_{jk}^{(r)}(x_k) - f_{jk}(x_k)| = O_p(1/r^m)$ uniformly in k. Thus, $n^{-1/2}\|\mathbf{w}_j\|_2 = \left\|1/n \sum_{k\in S_j}\left[f_{jk}^{(r)}(\mathbf{x}_k) - f_{jk}(\mathbf{x}_k)\right]\right\|_2 = O_p(s_j/r^m)$ uniformly in j.

  This implies that

$$\begin{aligned} T_1 &\leq \max_j \left\|\boldsymbol{\Sigma}_{S_j,S_j}^{-1}\frac{1}{n}\boldsymbol{\Psi}_{S_j}^T\mathbf{w}_j\right\|_2 \leq \max_j \left\|\boldsymbol{\Sigma}_{S_j,S_j}^{-1}\frac{1}{\sqrt{n}}\boldsymbol{\Psi}_{S_j}^T\right\|_2 \frac{1}{\sqrt{n}}\left\|\mathbf{w}_j\right\|_2 \\ &\leq C_{\min}^{-1/2}\max_j O_p\left(s_j/r^m\right) = O_p\left(\frac{\max_j s_j}{r^m}\right). \end{aligned}$$

  In the above, we used that $\Lambda_{\max}\left(\boldsymbol{\Sigma}_{S_j,S_j}^{-1}\frac{1}{\sqrt{n}}\boldsymbol{\Psi}_{S_j}^T\right) = \left(\Lambda_{\min}\left(\boldsymbol{\Sigma}_{S_j,S_j}\right)\right)^{1/2}$.

- Bounding T_2:

  Here, we use lemma 1 which bounds the ℓ_∞ norm of the average of high-dimensional i.i.d. vectors. First, by the definition of ϵ_j we must have that $\mathbb{E}[\,\psi_{jkt}(x_k)\epsilon_j\,] = 0$, i.e. the residuals are uncorrelated with the covariates.

  Let $z_{jkt} \equiv \psi_{jkt}(\mathbf{x}_k)^T\epsilon_j$, which is sum of n independent random variables with exponential tails. We have that

$$\max_j \|\boldsymbol{\Psi}_{S_j}^T\boldsymbol{\epsilon}_j\|_\infty/n = \max_j \max_{k\in S_j}\max_{t=1,\ldots,r}|z_{jkt}|/n \leq \max_{(j,k)\in E^*}\max_{t=1,\ldots,r}\left\{|z_{jkt}| \vee |z_{kjt}|/n\right\},$$

  the maximum of $2r|E^*|$ elements. We can thus apply lemma 1 to obtain

$$\begin{aligned} T_2 &= \max_j \left\|\boldsymbol{\Sigma}_{S_j,S_j}^{-1}\frac{1}{n}\boldsymbol{\Psi}_{S_j}^T\boldsymbol{\epsilon}_j\right\|_\infty \leq \max_j \left\|\boldsymbol{\Sigma}_{S_j,S_j}^{-1}\right\|_\infty \left\|\frac{1}{n}\boldsymbol{\Psi}_{S_j}^T\boldsymbol{\epsilon}_j\right\|_\infty \\ &\leq \max_j (rs_j)^{1/2} C_{\min}^{-1} O_p\left(\left(\frac{\log(2r|E^*|)}{n}\right)^{1/2}\right) = O_p\left(\left(\frac{\max_j s_j r\log(r|E^*|)}{n}\right)^{1/2}\right). \end{aligned}$$

- Bounding T_3:

We have that $\|\hat{\mathbf{g}}_{jk}\|_2^2 \leq 1$ for $(j,k) \in E^*$, so

$$T_3 \leq \lambda \max_j \left\| \mathbf{\Sigma}^{-1}_{S_j,S_j} \right\|_\infty \leq \lambda \max_j (rs_j)^{1/2} \left\| \mathbf{\Sigma}^{-1}_{S_j,S_j} \right\|_2 \leq \lambda \max_j \frac{(rs_j)^{1/2}}{C_{\min}}.$$

Altogether, we have shown that

$$\max_j \|\hat{\boldsymbol{\beta}}_{S_j} - \boldsymbol{\beta}^{*(r)}_{S_j}\|_\infty \leq O_p\left(\frac{\max_j s_j}{r^m}\right) + O_p\left(\left(\frac{(\max_j s_j) r \log(r|E^*|)}{n}\right)^{1/2}\right) + \lambda \max_j \frac{(rs_j)^{1/2}}{C_{\min}}.$$

By assumption,

$$\frac{1}{\rho^*} \max_j \left[\left(\frac{s_j r \log(r|E^*|)}{n} \right)^{1/2} + \frac{s_j}{r^m} + \lambda (rs_j)^{1/2} \right] \to 0$$

which implies that $\max_j \|\hat{\beta}_{S_j} - \boldsymbol{\beta}^{*(r)}_{S_j}\|_\infty < \rho^*/2$ with probability tending to 1 as $n \to \infty$.

Condition (A.29): We now consider the 'dual' problem. That is, we must show that $\|\hat{\mathbf{g}}_{jk}\|_2 + \|\hat{\mathbf{g}}_{kj}\|_2 \leq 1$ for each $(j,k) \notin E^*$. From the discussion of Condition (A.30), we know that

$$\begin{aligned}
\hat{\mathbf{g}}_{jk} &= \frac{1}{\lambda n} \mathbf{\Psi}^T_{jk} \left(\mathbf{\Psi}_{S_j} (\hat{\boldsymbol{\beta}}_{S_j} - \boldsymbol{\beta}^{*(r)}_{S_j}) - \mathbf{w}_j - \boldsymbol{\epsilon}_j \right) \\
&= \frac{1}{\lambda n} \mathbf{\Psi}^T_{jk} \left(\mathbf{\Psi}_{S_j} \mathbf{\Sigma}^{-1}_{S_j,S_j} \left(\frac{1}{n} \mathbf{\Psi}^T_{S_j} \mathbf{w}_j + \frac{1}{n} \mathbf{\Psi}^T_{S_j} \boldsymbol{\epsilon}_j - \lambda \hat{\mathbf{g}}_{S_j} \right) - \mathbf{w}_j - \boldsymbol{\epsilon}_j \right) \\
&= -\frac{1}{\lambda n} \mathbf{\Psi}^T_{jk} \left(\mathbf{I} - \frac{1}{n} \mathbf{\Psi}_{S_j} \mathbf{\Sigma}^{-1}_{S_j,S_j} \mathbf{\Psi}^T_{S_j} \right) \mathbf{w}_j - \frac{1}{\lambda n} \mathbf{\Psi}^T_{jk} \left(\mathbf{I} - \frac{1}{n} \mathbf{\Psi}_{S_j} \mathbf{\Sigma}^{-1}_{S_j,S_j} \mathbf{\Psi}^T_{S_j} \right) \boldsymbol{\epsilon}_j \\
&\quad - \frac{1}{n} \mathbf{\Psi}^T_{jk} \mathbf{\Psi}_{S_j} \mathbf{\Sigma}^{-1}_{S_j,S_j} \hat{\mathbf{g}}_{S_j} \\
&\equiv M_1^{jk} + M_2^{jk} + M_3^{jk}.
\end{aligned}$$

We will proceed by bounding $\|M_1^{jk}\|_2 + \|M_1^{kj}\|_2$, $\|M_2^{jk}\|_2 + \|M_2^{kj}\|_2$ and $\|M_3^{jk}\|_2 + \|M_3^{kj}\|_2$, which will give us a bound for the quantity of interest, $\|\hat{\mathbf{g}}_{jk}\|_2 + \|\hat{\mathbf{g}}_{kj}\|_2$.

- Bounding M_1: When bounding T_1 earlier, we saw that $n^{-1/2}\|\mathbf{w}_j\|_2 = O_p(s_j/r^m)$. Now $\left(\mathbf{I} - \mathbf{\Psi}_{S_j}\mathbf{\Sigma}^{-1}_{S_j,S_j}\mathbf{\Psi}^T_{S_j}/n\right)$ is a projection matrix, and by design $n^{-1/2}\mathbf{\Psi}_{jk}$ is orthogonal, so that all the eigenvalues of $n^{-1/2}\mathbf{\Psi}_{jk}$ are 1. Therefore

$$\|M_1^{jk}\|_2 \leq \frac{1}{\lambda} n^{-1/2}\|\mathbf{\Psi}_{jk}\|_2\, n^{-1/2}\,\|\mathbf{w}_j\|_2 = O_p\left(\frac{s_j}{\lambda r^m}\right),$$

and

$$\|M_1^{jk}\|_2 + \|M_1^{kj}\|_2 \leq O_p\left(\frac{s_j \vee s_k}{\lambda r^m}\right),$$

which tends to zero because $s_j/(\lambda r^m) \to 0$ uniformly in j.

- Bounding M_2:

First, note that

$$\begin{aligned}
\lambda\|M_2^{jk}\|_2 &\leq n^{-1}\|\mathbf{\Psi}^T_{jk}\boldsymbol{\epsilon}_j\|_2 + n^{-1/2}\|\mathbf{\Psi}_{jk}\|_2 \left\|n^{-1/2}\mathbf{\Psi}_{S_j}\mathbf{\Sigma}^{-1}_{S_j,S_j}\right\|_2 \|\mathbf{\Psi}^T_{S_j}\boldsymbol{\epsilon}_j\|_2/n \\
&\leq n^{-1}\|\mathbf{\Psi}^T_{jk}\boldsymbol{\epsilon}_j\|_2 + C_{\min}^{-1/2}\|\mathbf{\Psi}^T_{S_j}\boldsymbol{\epsilon}_j\|_2/n \\
&\leq \sqrt{r}\|\mathbf{\Psi}^T_{jk}\epsilon_j\|_\infty/n + (rs_j/C_{\min})^{1/2}\,\|\mathbf{\Psi}^T_{S_j}\boldsymbol{\epsilon}_j\|_\infty/n.
\end{aligned}$$

Then, applying lemma 1, as in the bound for T_2, we get

$$\lambda \max_{(j,k)\in E^{*c}} \|M_2^{jk}\|_2 \leq O_p\left(\left(\frac{r\log(r|E^{*c}|)}{n}\right)^{1/2}\right) + O_p\left(\left(\frac{r\max_j s_j \log(r|E^*|)}{n}\right)^{1/2}\right).$$

Thus, $\max_{(j,k)\in E^{*c}}\left\{\|M_2^{jk}\|_2 + \|M_2^{jk}\|_2\right\} \to 0$ when

$$\frac{r\log(r|E^{*c}|)}{\lambda^2 n} \to 0 \quad \text{and} \quad \max_j \frac{rs_j\log(r|E^*|)}{\lambda^2 n} \to 0.$$

- Bounding M_3:

By the irrepresentability assumption, we have that $\|M_3^{jk}\|_2^2 + \|M_3^{kj}\|_2^2 \leq 1 - \delta$

with probability tending to 1.

Thus, since $\|M_1^{jk}\|_2 + \|M_1^{kj}\|_2 + \|M_2^{jk}\|_2 + \|M_2^{kj}\|_2 = o_p(1)$, we have that for each $(j,k) \in E^{*c}$

$$\max_{(j,k)\in E^{*c}} \{\|\hat{\mathbf{g}}_{jk}\|_2 + \|\hat{\mathbf{g}}_{kj}\|_2\} \leq 1 - \delta$$

with probability tending to 1. Further, since we have strict dual feasibility, i.e. $\|\hat{\mathbf{g}}_{jk}\|_2 + \|\hat{\mathbf{g}}_{kj}\|_2 < 1$ for $(j,k) \in E^{*c}$, with probability tending to 1, the estimated graph is unique. □

Proof of Theorem 7

Proof. Consider a variable j, with $j \in C_u$. Our large-sample model requires minimizing $\mathbb{E}|x_j - \sum_{k\neq j}\sum_{t=1}^{\infty}\psi_{jkt}(x_k)\beta_{jkt}|^2$ with respect to the β_{jkt}, or equivalently, minimizing

$$\mathbb{E}|x_j - \sum_{k\neq j} f_{jk}(x_k)|^2$$

over functions $f_{jk} \in \mathcal{F}$. We have that

$$\begin{aligned}
\mathbb{E}|x_j - \sum_{k\neq j} f_{jk}(x_k)|^2 &= \mathbb{E}x_j^2 - 2\sum_{k\neq j}\mathbb{E}[\, x_j f_{jk}(x_k)\,] + \sum_{k\neq j}\sum_{l\neq j}\mathbb{E}[\, f_{jk}(x_k) f_{jl}(x_l)\,] \\
&= \mathbb{E}x_j^2 - 2\sum_{k\in C_u}\mathbb{E}[\, x_j f_{jk}(x_k)\,] - 2\sum_{k\notin C_u}\mathbb{E}[\, x_j f_{jk}(x_k)\,] + \sum_{k\in C_u}\sum_{l\in C_u}\mathbb{E}[\, f_{jk}(x_k) f_{jl}(x_l)\,] \\
&+ \sum_{k\notin C_u}\sum_{l\notin C_u}\mathbb{E}[\, f_{jk}(x_k) f_{jl}(x_l)\,] + 2\sum_{k\notin C_u}\sum_{l\in C_u}\mathbb{E}[\, f_{jk}(x_k) f_{jl}(x_l)\,].
\end{aligned}$$

By assumption $\sum_{k \notin C_u} \mathbb{E}[\, x_j f_{jk}(x_k)\,] = \sum_{k \notin C_u} \sum_{l \in C_u} \mathbb{E}[\, f_{jk}(x_k) f_{jl}(x_l)\,] = 0$. Thus, collecting terms, we get

$$\mathbb{E}|x_j - \sum_{k \neq j} f_{jk}(x_k)|^2 = \mathbb{E}|x_j - \sum_{k \in C_u} f_{jk}(x_k)|^2 + \mathbb{E}|\sum_{k \notin C_u} f_{jk}(x_k)|^2.$$

Minimization of this quantity with respect to $\{f_{jk} \in \mathcal{F}, k \notin C_u\}$ only involves the last term, which achieves its minimum at zero when $f_{jk}(\cdot) = 0$ almost everywhere for each $k \notin C_u$.

□

BIBLIOGRAPHY

Donald WK Andrews and Patrik Guggenberger. Hybrid and size-corrected subsampling methods. *Econometrica*, 77(3):721–762, 2009.

F.R. Bach. Bolasso: Model consistent lasso estimation through the bootstrap. In *Proceedings of the 25th International Conference on Machine Learning*, pages 33–40. ACM, 2008.

O. Banerjee, L. El Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *The Journal of Machine Learning Research*, 9:485–516, 2008. ISSN 1532-4435.

Katia Basso, Adam A Margolin, Gustavo Stolovitzky, Ulf Klein, Riccardo Dalla-Favera, and Andrea Califano. Reverse engineering of regulatory networks in human B cells. *Nature Genetics*, 37(4):382–390, 2005.

Richard Berk, Lawrence Brown, Andreas Buja, Kai Zhang, Linda Zhao, et al. Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837, 2013.

J. Besag. Statistical analysis of non-lattice data. *The Statistician*, 24(3):179–195, 1975.

Julian Besag. Spatial interaction and the statistical analysis of lattice systems. *J. R. Statist. Soc. B*, 36(2):192–236, 1974.

P. Breheny and J. Huang. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The Annals of Applied Statistics*, 5(1):232, 2011.

Peter Bühlmann and Sara van de Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer, 2011.

Florentina Bunea et al. Honest variable selection in linear and logistic regression models via ℓ_1 and $\ell_1+ \ell_2$ penalization. *Electronic Journal of Statistics*, 2:1153–1194, 2008.

A. Chatterjee and SN Lahiri. Bootstrapping lasso estimators. *Journal of the American Statistical Association*, 106(494):608–625, 2011.

Y Ann Chen, Jonas S Almeida, Adam J Richards, Peter Müller, Raymond J Carroll, and Baerbel Rohrer. A nonparametric approach to detect nonlinear correlation in gene expression. *Journal of Computational and Graphical Statistics*, 19(3):552–568, 2010.

Robert G Cowell, Philip Dawid, Steffen L Lauritzen, and David J Spiegelhalter. *Probabilistic Networks and Expert Systems: Exact Computational Methods for Bayesian Networks*, chapter 3.2.1. Springer, New York, 2007.

A.P. Dempster. Covariance selection. *Biometrics*, 28(1):157–175, 1972. ISSN 0006-341X.

Adrian Dobra and Alex Lenkoski. Copula Gaussian graphical models and their application to modeling functional disability data. *The Annals of Applied Statistics*, 5(2A):969–993, 2011.

Norman R Draper and R Craig Van Nostrand. Ridge regression and james-stein estimation: review and comments. *Technometrics*, 21(4):451–466, 1979.

M. Drton and M.D. Perlman. Model selection for gaussian concentration graphs. *Biometrika*, 91 (3):591–602, 2004.

M. Drton and M.D. Perlman. A sinful approach to gaussian graphical model selection. *Journal of Statistical Planning and Inference*, 138 (4):1179–1200, 2007.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

J. Fan, S. Guo, and N. Hao. Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of the Royal Statistical Society, Series B*, 74(1):37–65, 2012.

Bernd Fellinghauer, Peter Bühlmann, Martin Ryffel, Michael Von Rhein, and Jan D Reinhardt. Stable graphical model estimation with random forests for discrete, continuous, and mixed variables. *Computational Statistics & Data Analysis*, 64: 132–152, 2013.

H. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33, 2010.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. *glasso: Graphical lasso- estimation of Gaussian graphical models*, 2011. URL `http://www-stat.stanford.edu/~tibs/glasso`. R package version 1.4.

Sander Greenland. When should epidemiologic regressions use random coefficients? *Biometrics*, 56(3):915–921, 2000.

T.J. Hastie and R.J. Tibshirani. *Generalized Additive Models*. Chapman & Hall/CRC, Boca Raton, 1990.

Jean Hausser and Korbinian Strimmer. Entropy inference and the james–stein estimator, with application to nonlinear gene association networks. *The Journal of Machine Learning Research*, 10:1469–1484, 2009.

James S Hodges and Brian J Reich. Adding spatially-correlated errors can mess up the fixed effect you love. *The American Statistician*, 64(4):325–334, 2010.

P. Huber. Robust estimation of a location parameter. *Annals of Math. Stat.*, 53: 73–101, 1964.

P.J. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 221–33, 1967.

Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *arXiv preprint arXiv:1306.3171*, 2013.

CG Khatri and C.R. Rao. Characterizations of multivariate normality. I. through independence of some statistics. *Journal of Multivariate Analysis*, 6(1):81–94, 1976.

K. Knight and W. Fu. Asymptotics for lasso-type estimators. *Annals of Statistics*, 28(5):1356–1378, 2000.

Jason Lee, Yuekai Sun, and Jonathan E Taylor. On model selection consistency of penalized m-estimators: a geometric theory. In *Advances in Neural Information Processing Systems*, pages 342–350, 2013a.

Jason D Lee, Dennis L Sun, Yuekai Sun, and Jonathan E Taylor. Exact inference after model selection via the lasso. *arXiv preprint arXiv:1311.6238*, 2013b.

H. Leeb and B.M. Pötscher. Model selection and inference: Facts and fiction. *Econometric Theory*, 21(1):21–59, 2005.

E.L. Lehmann and George Casella. *Theory of Point Estimation*. Springer-Verlag, New York, NY, 1998.

Kuo-Ching Liang and Xiaodong Wang. Gene regulatory network reconstruction using conditional mutual information. *EURASIP Journal on Bioinformatics and Systems Biology*, (1):253894–253894, 2008.

H. Liu, J. Lafferty, and L. Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *The Journal of Machine Learning Research*, 10:2295–2328, 2009.

H. Liu, M. Xu, H. Gu, A. Gupta, J. Lafferty, and L. Wasserman. Forest density estimation. *Journal of Machine Learning Research*, 12:907–951, 2011.

Han Liu, Fang Han, Ming Yuan, John Lafferty, and Larry Wasserman. High-dimensional semiparametric Gaussian copula graphical models. *The Annals of Statistics*, 40(4):2293–2326, 2012.

Richard Lockhart, Jonathan Taylor, Ryan Tibshirani, and Robert Tibshirani. A significance test for the lasso. *arXiv preprint arXiv:1301.7161*, 2013.

James G MacKinnon and Halbert White. Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, 29(3):305–325, 1985.

K.V. Mardia, J.T. Kent, and J.M. Bibby. *Multivariate Analysis*. Academic press, Waltham, 1980.

R. Mazumder and T. Hastie. Exact covariance thresholding into connected components for large-scale graphical lasso. *The Journal of Machine Learning Research*, 13:781–794, 2012.

Rahul Mazumder, Jerome H Friedman, and Trevor Hastie. SparseNet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association*, 106(495):1125–1138, 2011.

P. McCullagh and John A. Nelder. *Generalized Linear Models*. Chapman and Hall/CRC, 2 edition, August 1989. ISBN 0412317605.

N. Meinshausen and P. Buhlmann. High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34(3):1436–1462, 2006.

N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006. ISSN 0090-5364.

N. Meinshausen and P. Bühlmann. Stability selection. *Journal of the Royal Statistical Society, Series B*, 72(4):417–473, 2010.

Patrick E Meyer, Frederic Lafitte, and Gianluca Bontempi. Minet: A R/Bioconductor package for inferring large transcriptional networks using mutual information. *BMC Bioinformatics*, 9(1):461–471, 2008.

Sahand Negahban, Pradeep D Ravikumar, Martin J Wainwright, and Bin Yu. A unified framework for the analysis of regularized m-estimators. *Statistical Science*, 26:538–557, 2012.

J. Pearl. *Causality: Models, Reasoning, and Inference*, volume 47, chapter 1.4 Functional Causal Models, pages 27–38. Cambridge Univ Press, 2000.

Thomas A Pearson and Teri A Manolio. How to interpret a genome-wide association study. *Jama*, 299(11):1335–1344, 2008.

J. Peng, P. Wang, N. Zhou, and J. Zhu. Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 104(486):735–746, 2009. ISSN 0162-1459.

Benedikt M Pötscher and Hannes Leeb. On the distribution of penalized maximum likelihood estimators: The lasso, scad, and thresholding. *Journal of Multivariate Analysis*, 100(9):2065–2082, 2009.

P. Ravikumar, J. Lafferty, H. Liu, and L. Wasserman. Sparse additive models. *J. R. Statist. Soc. B*, 71(5):1009–1030, 2009.

A.J. Rothman, P.J. Bickel, E. Levina, and J. Zhu. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

K. Sachs, O. Perez, D. Pe'er, D.A. Lauffenburger, and G.P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308 (5721):523–529, 2005.

A. Shojaie and G. Michailidis. Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. *Biometrika*, 97(3):519–538, 2010.

N. Simon and R. Tibshirani. Standardization and the group lasso penalty. *Statistica Sinica*, 22(3):983–1001, 2012a.

N. Simon and R. Tibshirani. Standardization and the group lasso penalty. *Statistica Sinica*, 22(3):983–1001, 2012b.

Jonathan Taylor, Richard Lockhart, Ryan J Tibshirani, and Robert Tibshirani. Post-selection adaptive inference for least angle regression and the lasso. *arXiv preprint arXiv:1401.3889*, 2014.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.

Ryan J Tibshirani. The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7:1456–1490, 2013.

Ryan J Tibshirani and Jonathan Taylor. Degrees of freedom in lasso problems. *The Annals of Statistics*, 40(2):1198–1232, 2012.

Sara van de Geer, Peter Bühlmann, and Ya'acov Ritov. On asymptotically optimal confidence regions and tests for high-dimensional models. *arXiv preprint arXiv:1303.0518*, 2013.

Sara A van de Geer. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, pages 614–645, 2008.

Aad W van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

C. Varin and P. Vidoni. A note on composite likelihood inference and model selection. *Biometrika*, 92(3):519–528, 2005.

C. Varin, N. Reid, and D. Firth. An overview of composite likelihood methods. *Statistica Sinica*, 21(1):5–42, 2011.

Arend Voorman. *lassoscore: high-dimensional inference with the penalized score test*, 2014. R package version 0.2.

M.J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1 constrained quadratic programming. *IEEE Transactions on Information Theory*, 55(5):2183–2202, 2009.

Y.J. Wang and E.H. Ip. Conditionally specified continuous distributions. *Biometrika*, 95(3):735–746, 2008.

Larry Wasserman. Low assumptions, high dimensions. *Rationality, Markets and Morals*, 2(49), 2011.

Halbert White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838, 1980. doi: 10.2307/1912934. URL `http://dx.doi.org/10.2307/1912934`.

D. Witten, J. Friedman, and N. Simon. New insights and faster computations for the graphical lasso. *Journal of Computational and Graphical Statistics*, 20(4):892–900, 2011.

Lingzhou Xue and Hui Zou. Regularized rank-based estimation of high-dimensional nonparanormal graphical models. *The Annals of Statistics*, 40(5):2541–2571, 2012.

Yan Yu and David Ruppert. Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association*, 97(460):1042–1054, 2002.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc. B*, 68(1):49–67, 2006.

M. Yuan and Y. Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35, 2007a.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68(1):49–67, 2007b.

M. Yuan and Y. Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35, 2007c. ISSN 0006-3444.

Cun-Hui Zhang and S Zhang. Confidence intervals for low-dimensional parameters with high-dimensional data. *arXiv preprint arXiv:1110.2563*, 2011.

P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.

Hui Zou and Runze Li. One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics*, 36(4):1509, 2008.

Hui Zou, Trevor Hastie, and Robert Tibshirani. On the "degrees of freedom" of the lasso. *The Annals of Statistics*, 35(5):2173–2192, 2007.