# Spectral Deconfounding and Perturbed Sparse Linear Models

Domagoj Ćevid, Peter Bühlmann, Nicolai Meinshausen
ETH Zürich

November 14, 2018

## 1 Abstract

Standard high-dimensional regression methods assume that the underlying coefficient vector is sparse. This might not be true in some cases, in particular in presence of hidden, confounding variables. Such hidden confounding can be represented as a high-dimensional linear model where the sparse coefficient vector is perturbed. We develop and investigate a class of methods for such model. We propose some spectral transformations, which change the singular values of the design matrix, as a preprocessing step for the data which serves as input for the Lasso. We show that, under some assumptions, one can achieve the optimal $\ell_1$-error rate for estimating the underlying sparse coefficient vector. We also illustrate the performance on simulated data and a real-world genomic dataset.

## 2 Introduction

Many datasets nowadays include measurements from many variables. The corresponding models are typically high-dimensional with many more parameters than the sample size. For statistical estimation and inference, there is a vast literature which assumes sparsity. See, for example, the monographs [4, 9, 12]. Most often, sparsity is assumed in the strong $\ell_0$ sense, saying that only a small number of coefficients or parameters are non-zero. Extensions of such (strong) sparsity are mentioned below.

Our focus is on linear models. If we assume that the response is affected only by a small number of predictors, meaning that the coefficient vector is sparse, we can efficiently estimate the active set and the corresponding coefficients with the Lasso and related methods achieving the minimax optimal $\ell_1$ estimation error rate, see [4, 26, 2]. However, sometimes the sparsity assumption is not adequate and one needs to relax it. Instead of just a few predictors affecting our response, we might additionally have a small contribution from many predictors. Such situations are covered with (i) the notion of weak sparsity [24], where the parameter $\beta$ fulfils the condition that $\|\beta\|_q$ is small for some $0 < q < 1$ or (ii) assuming the structure that $\beta$ can be represented as a sum of a sparse and a dense vector. The case (i) does not call for a new method or algorithm: in fact, the Lasso still exhibits optimal convergence rate if $\|\beta\|_q$ is sufficiently small [24]. On the other hand, case (ii) requires a different method such as Lava [6]. We investigate in this paper how to deal with the case (ii) when the parameter is a sum of a sparse and a dense part.

We propose a simple spectral transformation, the so-called Trim transform, of the response $Y$ and the design matrix $X$ consisting of the values of the predictors. The transformed response and design matrix are then the input for a high-dimensional sparse regression technique: we consider the Lasso as a prime example. We investigate the theoretical properties and empirical performances for a class of spectral transformations. As a result, we conclude why our Trim transform, but also the Lava method [6], are favourable over a range of scenarios, pointing out also some advantages over other techniques and approaches, see also Section 2.1.

One of our main motivations to study the case (ii) with a sum of sparse and dense parts of the parameter is that it arises in presence of confounding variables which affect both the predictors and the response, thus introducing additional correlations. Confounding variables are additional variables that we did not account for by including them in our model. Some examples of confounding variables are the age or gender of patients or batch effects such as the equipment used or laboratory conditions such as temperature or humidity. If many predictors are affected by the confounding variables, we expect that the true underlying regression vector will be changed by some small, dense perturbation.

Confounding is a severe issue when interpreting regression parameters, often, but not necessarily, in connection with causal inference. A prime example are genetic studies where unobserved confounding can easily lead to spurious correlations and partial dependencies [20]. Adjusting for the confounding variables is very important in practice and several deconfounding methods have been suggested for various settings [8, 17, 7, 22, 27]. Many methods try to estimate the confounding variables directly from the data, usually by using some factor analysis technique and often ignoring other effects, which might be very hard to do accurately. Some methods require additional assumptions about the confounding structure which might not hold, for example that the batches are known [16]. In addition, there are not many theoretical results that justify the methods, especially since many are quite complicated and therefore difficult to analyse.

## 2.1 Relation to other work and our contribution

For adjusting for the effect of the confounding variables, the most prominent method in practice is to remove the top several principal components of the predictors, see for example [20]. Such PCA adjustment is a special case of a spectral transformation. Our presented theory explains when and why this works well and we conclude that our proposed Trim transform is often a better choice.

Chandrasekaran et al. [5] have addressed the problem of estimating the precision matrix, assumed to be sparse, if a few variables are unobserved. Then the observed precision matrix can be represented as a sum of the initial sparse precision matrix and a low-rank perturbation due to the confounding variables. This model is similar to the one we consider, but the assumptions and the goals differ. We aim to estimate just the regression coefficients instead of the whole precision matrix and the method we use is much simpler. Furthermore, the theoretical conclusions are substantially different: we establish optimal convergence rates in terms of the $\ell_1$-norm estimation error while they consider support recovery and $\ell_\infty$ bounds for the low-dimensional setting, assuming strong conditions.

The Puffer transform has been suggested for improving the variable selection properties of the Lasso for a sparse high-dimensional linear model [15]. Our theory gives a much more precise result about the Puffer transform: the Trim transform is at least as good as Puffer transform and substantially better when the number of samples is close to the number of predictors. Shah et al. [23] use the Puffer transform in combination with bootstrap aggregation in order to estimate the covariance matrix, a very different quantity than the precision matrix or regression coefficients, under the presence of confounding variables.

The Lava estimator [6] is the most similar to our Trim transform. The theory we develop gives a clean result for the $\ell_1$-norm estimation error for the sparse parameter vector and establishes the optimal minimax convergence rate. Such result has not been established in [6]. In addition, our developments suggests a simple rule for choosing the tuning parameter of the Trim transform; for the Lava, it also suggests the choice of the $\ell_2$-norm regularization parameter.

Our contribution can be seen as threefold. We propose a simple spectral transformation called Trim transform which is perhaps slightly easier to use than the Lava estimator. Furthermore, for the linear model where the underlying sparse parameter has been perturbed, we provide novel theory establishing optimal convergence rates for a class of spectral transformations for the $\ell_1$-norm estimation error of the true underlying sparse parameter. Finally, we use these results to show how the issue of confounding can be addressed by the Trim transform and using the Lasso afterwards: we establish the optimal convergence rate under some assumptions and illustrate the empirical performances on simulated and real genomic data. Our method is entirely modular and can be used in conjunction with any high-dimensional regression methods, including the Lasso and many other algorithms.

# 3 The models

We are going to consider two models: a perturbed linear model and a confounding model. The latter can be represented as a perturbed linear model with a special structure of the perturbation term.

## 3.1 Perturbed linear model

Let us consider the sparse linear regression model, where the sparse coefficient vector has been perturbed by some vector $b$:

$$Y = X(\beta + b) + \epsilon. \tag{3.1}$$

Here $Y \in \mathbb{R}^n$ is the response vector and $X \in \mathbb{R}^{n\times p}$ is the design matrix, fixed or random, $\beta \in \mathbb{R}^p$ is a sparse vector and typically dense $b \in \mathbb{R}^p$, which we think of as a perturbation of the sparse vector $\beta$; $n$ denotes the sample size and $p$ the number of predictor variables. The main interest is to recover the sparse part $\beta$ of the regression parameter: the dense perturbation is viewed as a nuisance which we want to get rid of. The support of $\beta$ is denoted by $S$ and its size by $s$. Finally, the errors are assumed to be independent and identically distributed sub-Gaussian variables with mean zero and parameter $\sigma^2$.

We are going to focus on the case where $X$ has a fixed design. The results for random design, where the rows of $X$ are independent and identically distributed random vectors, follow from the results for a fixed design by conditioning on $X$.

Our model is in general unidentifiable since we can only infer $\beta + b$ from the data. This makes the estimation of $\beta$ impossible, unless we impose some conditions on $b$. If $b$ has certain structure, we will be able to retrieve the sparse $\beta$, e.g. by assuming that $b$ is dense or converges to 0 in some norm. We investigate under which conditions we are able to infer the sparse part $\beta$ and how efficiently in terms of statistical accuracy.

## 3.2 Confounding model

A prominent case when a perturbation of $\beta$ occurs is caused by some unobserved, confounding variables affecting both the predictors and the response. In this way, we observe spurious regression coefficients between the predictors and the response without actual causation. The model is given by:

$$\begin{aligned} X &= H\Gamma + E \\ Y &= X\beta + H\delta + \eta. \end{aligned} \tag{3.2}$$

with random terms $H \in \mathbb{R}^{n\times q}$, $E \in \mathbb{R}^{n\times p}$ and $\eta \in \mathbb{R}^n$ having independent rows and being jointly independent of each other. We note that this is a model with random design matrix $X$ with i.i.d. rows.

Here $H \in \mathbb{R}^{n\times q}$ is the random matrix of the hidden confounding variables, where $q$ is their number. The matrix $\Gamma \in \mathbb{R}^{q\times p}$ and the vector $\delta \in \mathbb{R}^q$ contain the coefficients describing the linear effect of those confounding variables on $X$ and $Y$, respectively. The random term $E \in \mathbb{R}^{n\times p}$ can be seen as the unconfounded design matrix; without confounding ($\Gamma = 0$) it equals $X$. The columns of $E$ are allowed to be correlated; if the components of $E$ are uncorrelated, $X$ is generated from a factor model [1]. Here, in addition, the hidden variables do not encode a factor structure for $X$ alone, but also generate confounding effects. $\eta \in \mathbb{R}^n$ is a vector of additive errors

**Remark** (Structural Equation Model)**.** A main example of the model in (3.2) is a structural equation model (SEM) where $X \leftarrow H\Gamma + E$, $Y \leftarrow X\beta + H\delta + \eta$ and thus $\beta$ is the causal effect of $X$ on $Y$. In a standard SEM with no further hidden variables, the components of $E$ would be assumed independent.

We further require a Gaussianity assumption for the model in (3.2), primarily to make the theory in Section 5 more straightforward; see also the comment about the Gaussian assumption below. Assuming the rows of $H$ are multivariate normal random variables, without loss of generality we may assume that $H_{ij} \overset{i.i.d.}{\sim} N(0,1)$, because otherwise we can change $\Gamma$ and $\delta$ accordingly. The rows of $E$ are independent and identically distributed as $N_p(0, \Sigma_E)$ for some fixed covariance matrix $\Sigma_E \in \mathbb{R}^{p\times p}$. The components of the response error $\eta$ are i.i.d. sub-Gaussian random variables with parameter $\sigma_\eta^2$ assumed to be $\eta \sim N_n(0, \sigma_\eta^2 I_n)$.

Considering the distribution of the observed variables $X, Y$, we get that the model in (3.2) is equivalent to the perturbed model $Y = X(\beta + b) + \epsilon$ with the rows of $X$ being independent and identically distributed as $N_p(0, \Sigma)$, where now

$$\begin{aligned}
&\Sigma = \Gamma^T\Gamma + \Sigma_E \\
&\sigma^2 = \sigma_\eta^2 + \|\delta\|_2^2 - \delta^T\Gamma(\Gamma^T\Gamma + \Sigma_E)^{-1}\Gamma^T\delta \\
&b = (\Gamma^T\Gamma + \Sigma_E)^{-1}\Gamma^T\delta
\end{aligned} \tag{3.3}$$

Thus, in the confounding model the coefficient perturbation arises naturally and it has a complex relationship to the design matrix.

**Remark** (Gaussianity assumption)**.** In the confounding model we have

$$Y = X(\beta + b) + (H\delta - Xb) + \nu$$

and $b$ satisfies that $\mathrm{Cov}(X, H\delta - Xb) = 0$. The Gaussianity assumption in the confounding model gives us that $X$ and $H\delta - Xb$ are independent, so the design matrix $X$ is independent of the error term $\epsilon = H\delta - Xb + \nu$. We require such independence in the proof of Theorem 1 in order to bound the tail of $\|X^T\epsilon\|_\infty$. One might still be able to bound the latter when $\epsilon$ is only uncorrelated with $X$, but our theory does not cover this case.

# 4 Method

In the following, we propose and motivate some methods based on a class of spectral transformations.

## 4.1 Spectral transformations

Let $X = UDV^T$ be the singular value decomposition of $X$, where $U \in \mathbb{R}^{n\times r}, D \in \mathbb{R}^{r\times r}, V \in \mathbb{R}^{r\times p}$, where $r = \min(n, p)$ is the rank of $X$. We write $d_1 \geq d_2 \geq \ldots \geq d_r$ for the diagonal elements of $D$. We use the form of SVD which uses only non-zero singular values.

The idea is to first transform our data by applying some specific linear transformation $F : \mathbb{R}^n \to \mathbb{R}^n$ and then perform the Lasso algorithm:

$$\begin{aligned}
X &\to \tilde{X} := FX \\
Y &\to \tilde{Y} := FY \\
\hat{\beta} = \arg\min_\beta &\left\{ \frac{1}{n}\|\tilde{Y} - \tilde{X}\beta\|_2^2 + \lambda\|\beta\|_1 \right\}.
\end{aligned}$$

We restrict our attention to the class of spectral transformations, which transform the singular values of $X$ while keeping its singular vectors intact. Let $\tilde{D}$ be an arbitrary $r \times r$ diagonal matrix with diagonal elements $\tilde{d}_1, \dots, \tilde{d}_r$. Our spectral transformation matrix is given by

$$F = U \begin{bmatrix} \tilde{d}_1/d_1 & 0 & \dots & 0 \\ 0 & \tilde{d}_2/d_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \tilde{d}_r/d_r \end{bmatrix} U^T \tag{4.1}$$

and then we have

$$\tilde{X} = FX = U\tilde{D}V^T$$

In this paper we explore the question of what is a good choice of $F$ for the estimation of $\beta$. In general, the Lasso performs best when the predictors are uncorrelated and when the errors are independent. Therefore, a good choice of $F$ needs to find a good balance between a well behaved error term $\tilde{\epsilon} = F\epsilon$, well behaved design matrix $\tilde{X}$ and well behaved perturbation term $\tilde{X}b$. We show that we can, under some assumptions, achieve the optimal $\ell_1$-norm error rate for the estimation of the unknown sparse coefficients.

In order to do that, our spectral transformation must not significantly increase the small singular values, must ensure that there are no singular values which are much larger than the rest and that sufficiently many singular values stay reasonably large.

One such transformation is the **Trim transform** which limits all the singular values to be at most some constant $\tau$.

$$\tilde{d}_i = \min(d_i, \tau) \tag{4.2}$$

We also show that the median singular value is a good choice of $\tau$:

$$\tau = d_{\lfloor r/2 \rfloor}$$

## 4.2 Existing methods and motivation

We discuss some existing methods which are special cases of spectral transformations and provide further explanations and relations.

### 4.2.1 Examples of spectral transformations

Several existing methods consist of first transforming the data with a certain spectral transformation as in (4.1), for some choice of the matrix $\tilde{D}$, and then running some regression algorithm, such as the Lasso.

**Lava** One such example is the Lava estimator [6], designed for the linear model where the coefficient vector can be written as a sum of a dense and a sparse vector. It is originally given by

$$(\hat{\beta}, \hat{b}) = \arg\min_{\beta, b} \left\{ \frac{1}{n} \|Y - X(\beta + b)\|_2^2 + \lambda_2 \|b\|_2^2 + \lambda_1 \|\beta\|_1 \right\}$$

which can be seen as a combination of Lasso and Ridge regression. It is shown in [6] that the solution of this optimization problem is given by

$$\begin{aligned} F &= (I_p - X(X^T X + n\lambda_2 I_p)^{-1} X^T)^{1/2}, \\ \hat{\beta} &= \arg\min_{\beta} \left\{ \frac{1}{n} \|\tilde{Y} - \tilde{X}\beta\|_2^2 + \lambda_1 \|\beta\|_1 \right\}, \\ \hat{b} &= (X^T X + n\lambda_2 I_p)^{-1} X^T (Y - X\hat{\beta}). \end{aligned}$$

From here one can see that the estimator of the sparse part is just a Lasso estimator applied on the transformed data where

$$\tilde{d_i} = \sqrt{\frac{n\lambda_2 d_i^2}{n\lambda_2 + d_i^2}}.$$

This transformation is visualized in Figure 5.1.

**Puffer transform** Another example is the Puffer transform introduced in [15], which uses the Lasso after mapping all non-zero singular values to $\tilde{d_i} = 1$. The algorithm is analyzed in [15] as a pre-conditioning method for the variable selection problem when there is no coefficient perturbation present. This transformation decreases the correlations between the columns of the design matrix, but it can inflate the errors, especially when $p$ is close to $n$. It can also be thought of as a special case of the Lava transformation in the case when $\lambda_2 \to 0$, since then $\frac{\tilde{d_i}}{\sqrt{n\lambda_2}} \to 1$ (the denominator here is just a scaling factor). The transformation is displayed in Figure 5.1.

**PCA adjustment** Another example of a spectral transformation is given by PCA-based methods for adjusting for hidden confounders [21]. One adjusts for a first few principal components from the columns of the design matrix $X$ before further analysis in hope of removing the effect of the confounding variables [22], [13]. This procedure is in fact analogous to applying a spectral transformation, where the matrix $\tilde{D}$ is obtained from $D$ by mapping to 0 the singular values corresponding to the prinicipal components one wants to adjust for. See also Figure 5.1 for an illustration.

In the confounding model (3.2), the effect of the coefficient perturbation will approximately lie in the span of the first few principal components of $X$ (see Figure 4.1). One can reduce the effect of the confounding variables by removing those principal components. This also helps to decorrelate the columns of $X$. The problem with this approach is that we need to know how many principal components to remove in order to reduce the effect of the confounding variables, but still preserve the signal. This might be hard to do, unless the effect of the confounding is so strong that several singular values of $X$ are significantly larger than the rest.

### 4.2.2 Visual representation

We would like that the perturbation term $Xb$ is small compared to the signal $X\beta$. The more $b$ is aligned with the singular vectors of $X$ corresponding to the large singular values, the larger $\|Xb\|_2$ will be. This will especially be the case in the confounding model (See Figure 4.1).

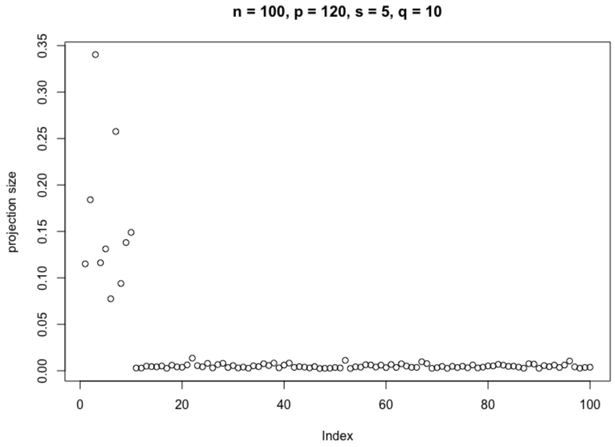

Figure 4.1: Size of the projection of $b$ onto $V_i$ for different $i$, for a random dataset drawn from the confounding model with $q = 10$ confounding variables, as described in section 6.1.1. We see that the projections of $b$ on the first 10 singular values are substantially larger than the rest.

Shrinking large singular values ensures that $\|\tilde{X}b\|$ stays small regardless of the direction $b$ is pointing to. On the other hand, we do not expect $\beta$ to be aligned with the large singular vectors, which is very unlikely. This is represented in Figure 4.2. Therefore, by shrinking large singular values, $\|X\beta\|_2$ will decrease much more compared to $\|Xb\|_2$.

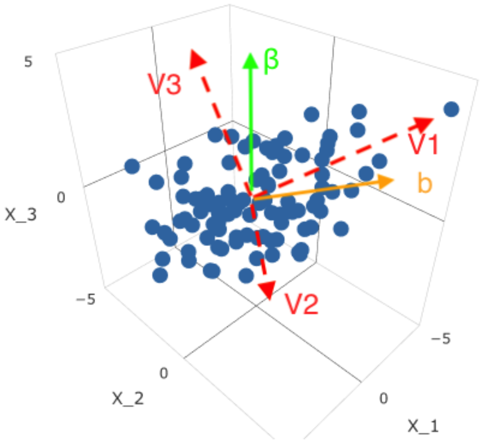

Figure 4.2: Visualisation of the relationship between the perturbation $b$, signal $\beta$ and singular vectors of $X$. In the confounding model $b$ will be much more aligned with the singular vectors corresponding to the large singular values than $\beta$.

### 4.2.3 Goodness of fit and a connection to the Lava method

There is a large literature on penalized least squares estimators for different penalty terms. Our method, on the other hand, keep the $\ell_1$ penalty from the Lasso, but change the goodness of fit term. Instead of using the $\ell_2$ distance between the measured response $Y$ and the fit $X\hat{\beta}$, we use a different metric:

$$\hat{\beta} = \arg\min_{\beta} \left\{ \frac{1}{n}(Y - X\beta)^T F^T F(Y - X\beta) + \lambda \|\beta\|_1 \right\}.$$

This penalizes the residual differently in different directions. This is sensible because the residuals have different variances in different directions due to the coefficient perturbation:

$$Y - X\beta = Xb + \epsilon = \sum_{i=1}^{n} \big( d_i(V_i^T b) + (U^T \epsilon) \big) U_i$$

Assuming that the residual has mean 0, the second moment of the residual in the direction of $U_i$ conditional on the design $X$ is $d_i^2 \mathbb{E}[(V_i^T b)^2 | X] + \sigma^2$. In order to penalize all directions equally, we would need to scale down the corresponding singular value by $\sqrt{d_i^2 \mathbb{E}[(V_i^T b)^2 | X] + \sigma^2}$. Therefore, we would need to have

$$\tilde{d}_i \propto \frac{d_i}{\sqrt{d_i^2 \mathbb{E}[(V_i^T b)^2 | X] + \sigma^2}}$$

This additionally justifies using the spectral transformations which keep $U$ fixed and transform only the singular values of $X$.

As an example, in the case when the perturbation vector is isotropic and independent of $X$, so that $\mathbb{E}[(V_i^T b)^2|X]$ does not depend on $i$, we get from above that

$$\tilde{d}_i \propto \frac{d_i}{\sqrt{d_i^2 + \frac{p\sigma^2}{\mathbb{E}\|b\|_2^2}}} \tag{4.3}$$

which is exactly the transformation of the Lava algorithm.

We note that generalised least squares also falls into the category of changing the goodness of fit measure to

$$(Y - X\beta)^T \Omega^{-1} (Y - X\beta).$$

Here, however, the matrix $\Omega^{-1}$ is decorrelating the error and has nothing to do with the spectrum of $X$. In fact, our presented theory in Section 5 analyses the effect of a spectral transformation on the error in terms of leading to correlated and inflated errors. It shows that we have to deal with a trade-off between error inflation and deconfounding or reducing the effect of the coefficient perturbation.

# 5 Theoretical Results

In this section we analyse how the $\ell_1$-estimation error for the sparse coefficient $\beta$ behaves depending on the spectral transformation we are using. We derive results for the perturbed linear model (3.1) and then we use the relationship (3.3) to establish results for the confounding model.

We show that, under the model assumptions given in Section 5.3.1 with the Trim transform (4.2), even in the presence of the coefficient perturbation, we achieve the minimax optimal rate of the Lasso in the case without perturbation. In addition, we give sufficient conditions for other spectral transformations to achieve this rate and explore when these conditions are satisfied. Finally, we discuss under which assumptions for the confounding model (3.2) we can apply our results for the perturbed linear model to get the optimal $\ell_1$-estimation rate of the underlying sparse parameter.

## 5.1 Notation

For any square matrix $M$ we define the compatibility constant which is a kind of smallest restricted eigenvalue for measuring the well-posedness of $M$

[4]:

$$\phi_M := \inf_{\|\alpha\|_1 \leq 5\|\alpha_S\|_1} \frac{\sqrt{\alpha^T M \alpha}}{\frac{1}{\sqrt{s}}\|\alpha_S\|_1},$$

where $S$ is the support set of $\beta$, $s$ is the size of $S$ and $\alpha_S$ is a vector consisting only of the components of $\alpha$ which are in $S$.

Let us also write $\tilde{\Sigma} := \frac{1}{n}\tilde{X}^T\tilde{X}$, and $\hat{\Sigma} = \frac{1}{n}X^TX$. We denote the $k$-th largest diagonal element of the transformed singular values $\tilde{D}$ by $\tilde{d}_{(k)}$. We write $V_{(k)}$ for the corresponding column of $V$, where $X = UDV^T$ is the SVD of $X$ and write also $M_k = [V_{(1)}, \ldots, V_{(k)}][V_{(1)}, \ldots, V_{(k)}]^T$. We denote the smallest (non-zero) singular value of any rectangular matrix $A$ by $\lambda_{\min}(A)$.

Finally, we use the notation $A = \Omega(B)$ if $\frac{B}{A} = \mathcal{O}(1)$, i.e. if $A$ has asymptotically at least the same rate as $B$. $A = \mathcal{O}_p(B)$ means that there exists a constant $c > 0$ such that $\mathbb{P}(A > cB) \to 0$ and $\Omega_p$ is defined analogously.

## 5.2 Upper bounds for the $\ell_1$-estimation error of $\beta$

The first result describes the effect of an arbitrary linear transformation on the $\ell_1$-estimation error of the Lasso:

**Theorem 1.** *Assume the model in (3.1) with fixed design $X$ and i.i.d. zero-mean sub-Gaussian errors $\epsilon_i$, for $i = 1, \ldots n$. Let $F \in \mathbb{R}^{n\times n}$ be an arbitrary linear transformation and $A > 0$ an arbitrary fixed constant. Then there exists $\lambda_{\min} \geq 0$ such that for any $\lambda \in [\lambda_{\min}, B\lambda_{\min}]$, where $B \geq 1$ is a fixed constant, and with probability at least $1 - 2p^{1-A^2/8}$ we have*

$$\|\hat{\beta} - \beta\|_1 \leq C_1 \frac{s\sigma}{\phi^2_{\tilde{\Sigma}}}\sqrt{\frac{\log p}{n}} \max_i \left(\frac{X^T(F^TF)^2X}{n}\right)^{1/2}_{ii} + C_2\sqrt{\frac{s}{n}}\frac{\|\tilde{X}b\|_2}{\phi_{\tilde{\Sigma}}},$$

*where $C_1, C_2$ are constants depending only on $A$ and $B$.*

This bound exhibits a relationship between the effects of our transformation on the error via the quantity $\max_i \left(X^T(F^TF)^2X\right)_{ii}$ which is the maximal variance of some component of $\tilde{X}^T\tilde{\epsilon}$, on the coefficient perturbation appearing as $\|\tilde{X}b\|_2$, and on the design matrix via the compatibility constant $\phi_{\tilde{\Sigma}}$. We need to balance the effect of these three terms.

Theorem 1 holds for any linear transformation $F$. In the following, also motivated by the arguments in Section 4.2, we restrict ourselves to spectral transformations. We will describe a class of spectral transformations $F$ which leads to optimal rates as described in Theorem 3 which says that, under some assumptions,

$$\|\hat{\beta} - \beta\|_1 = \mathcal{O}_p\left(\frac{\sigma s}{\lambda_{\min}(\Sigma)}\sqrt{\frac{\log p}{n}}\right).$$

Thus, for rate-optimality, it suffices to consider spectral transformations.

In order to proceed, we need to understand how exactly $\phi_{\tilde{\Sigma}}$ depends on the transformed singular values in $\tilde{D}$. The answer to this question depends delicately on the singular vectors $V$ of $X$ as well. The following bound helps us to understand the behaviour of the compatibility constant depending on the transformed singular values.

**Lemma 1.** *Consider a spectral transformation $F$ as in (4.1). Let $1 \leq k < r = \min(n,p)$ be an arbitrary integer. Then:*

$$\phi^2_{\tilde{\Sigma}} \geq \sum_{i=1}^{r} \frac{1}{n} \tilde{d}^2_{(i)} (\phi^2_{M_i} - \phi^2_{M_{i-1}}) \geq \frac{1}{n} \tilde{d}^2_{(k)} \phi^2_{M_k}.$$

Using this Lemma and Theorem 1, we can describe the dependence of the $\ell_1$-estimation error on the transformed singular values $\tilde{d}_i$:

**Theorem 2.** *Under assumptions of Theorem 1, for any $k \leq r = \min(n,p)$ and any spectral transformation $F$ mapping $d_i$ to $\tilde{d}_i$, we get*

$$\|\hat{\beta} - \beta\|_1 \leq C_1 \frac{s\sigma}{\frac{1}{n}\tilde{d}^2_{(k)} \phi^2_{M_k}} \sqrt{\frac{\log p}{n}} \max_i \left(\frac{\tilde{d}_i}{d_i}\right)^2 + C_2 \sqrt{s} \frac{\tilde{d}_{(1)} \|V^T b\|_2}{\tilde{d}_{(k)} \phi_{M_k}}.$$

Since nothing changes in our bounds if we multiply all $\tilde{d}_i$ by a constant, we will assume without loss of generality that $\tilde{d}_i \leq d_i$, so we are only shrinking the singular values. This allows us to control the term $\max_i \left(\tilde{d}_i / d_i\right)^2$ in Theorem 2.

In order to control the error caused by the coefficient perturbation, we need to make $\tilde{d}_{(1)}$ small (see Theorem 2). However, we need to carefully shrink the singular values, since we need $\phi_{\tilde{\Sigma}}$ to stay large. One can easily show that under some mild conditions, the sample covariance matrix $\hat{\Sigma} = \frac{X^T X}{n}$ satisfies $\phi^2_{\hat{\Sigma}} \geq \frac{\lambda_{\min}(\Sigma)}{2}$ with high probability (see [4]) and we need to ensure that $\phi_{\tilde{\Sigma}}$ is bounded away from zero as well.

### 5.3 Optimal rates for $\ell_1$-estimation error of $\beta$

We develop here the theory for optimal convergence rate of $\|\hat{\beta} - \beta\|_1$. In what follows, we assume for simplicity that we have the high-dimensional case, where $p \geq n$. However, the theory developed in the rest of this section also holds for the case $n > p$ with small adjustments. We will discuss the case $n > p$ in more details in the section 5.5.

### 5.3.1 Model assumptions

We require the following assumptions:

**(A1)** (Coefficient perturbation) The perturbation vector $b$ satisfies

$$\|V^T b\|_2 = \mathcal{O}\left(\frac{\sigma}{\lambda_{\min}(\Sigma)}\sqrt{\frac{s\log p}{p}}\right).$$

**(A2)** (Singular vectors of $X$) For any $k = \Omega(n)$, we have

$$\phi^2_{M_k} = \Omega\left(\frac{n}{p}\right).$$

**(A3)** (Singular values of $X$) For any $k$ such that $\limsup \frac{k}{n} < 1$ it holds that

$$d_k^2 = \Omega\left(\lambda_{\min}(\Sigma)p\right).$$

We will justify these assumptions and give examples of the models for which they hold in the Section 5.4.

### 5.3.2 Rate-optimal spectral transformations

Assumption (**A1**) implies that the perturbation $b$ is not too big. Assumption (**A2**) states that $\phi_{M_k}$ is not too small for any $k$ of order $n$. If in addition a certain proportion of the transformed singular values $\tilde{d}_i$ is large enough, Lemma 1 ensures that $\phi_{\tilde{\Sigma}}$ is bounded away from zero. In addition, assumption (**A3**) ensures that enough of the singular values of $X$ are large enough, so it is indeed possible to choose such $\tilde{d}_i$.

**Theorem 3.** *Assume that the model assumptions **(A1)**, **(A2)** and **(A3)** hold. Consider a spectral transformation $F = U\tilde{D}D^{-1}U^T$ with $\tilde{d}_i \leq d_i$ which satisfies: there exists $k = \Omega(n)$ such that*

***(B1)*** $\tilde{d}_{(k)} = \Omega\left(\tilde{d}_{(1)}\right)$

***(B2)*** $\tilde{d}^2_{(k)} = \Omega\left(\lambda_{\min}(\Sigma)\, p\right)$

*Then we can choose $\lambda$ so that, despite the coefficient perturbation, the $\ell_1$-estimation error achieves the minimax optimal rate:*

$$\|\hat{\beta} - \beta\|_1 = \mathcal{O}_p\left(\frac{\sigma s}{\lambda_{\min}(\Sigma)}\sqrt{\frac{\log p}{n}}\right)$$

*In addition, the Trim transform (4.2) with $\tau = d_{\lfloor tn \rfloor}$, where $t \in (0,1)$ is an arbitrary constant, satisfies the conditions **(B1)** and **(B2)**. Other examples of such spectral transformations are discussed below.*

**Remark.** If we just require consistency $\|\hat{\beta} - \beta\|_1 = o_P(1)$ instead of the optimal rate, we can relax the assumption (**A1**) to $\|V^T b\|_2 = o\left(\sqrt{n/(ps)}\right)$. Furthermore, if the assumptions **(A3)** and **(A2)** do not hold for all $k$, but there only exists a constant $k$ satisfying $d_k^2 = \Omega(\lambda_{\min}(\Sigma)p)$ and $\phi_{M_k}^2 = \Omega\left(n/p\right)$ then the Trim transform with parameter $\tau = d_k$ still achieves the minimax optimal rate of the $\ell_1$-estimation error.

The assumption (**B1**) states that, after applying the transformation, the largest singular value is not too large, i.e. it is of the same order as some fraction of other singular values. This controls the effect of the coefficient perturbation. The assumption **(B2)** ensures that after transformation, some fraction of the singular values is large enough so that we still keep enough of the signal.

There are several different possibilities, in addition to the Trim transform, for choosing a spectral transformation that satisfies the assumptions (**B1**) and (**B2**). We will list some possibilities below. However, in order to achieve the optimal rate, some of them, such as the Lasso or PCA adjustment, need further assumptions on the distribution of singular values. The visual representations of the spectral transformations discussed below is given in the Figure 5.1.

**Lasso** The simplest option is to take $\tilde{d}_i = d_i$, i.e. the usual Lasso algorithm. Assumption (**B2**) is trivially satisfied if our model satisfies the assumption (**A3**). However, (**B1**) requires that the largest singular values does not depart from the others. This is true if the predictors are i.i.d. which typically does not hold, for example, in the presence of confounding variables.

**Trim transform** The Trim transform (4.2) with $\tau = d_{\lfloor tn \rfloor}$ for some $t \in (0,1)$ fixes the problems with the Lasso, as we have seen in Theorem 3. If we take $k = \lfloor tn \rfloor$, we see that $\tilde{d}_{(k)} = \tilde{d}_{(1)}$ so **(B1)** holds. Furthermore, the assumption **(A3)** implies the assumption **(B2)**.

**Step function** The assumptions **(B1), (B2)** will still be true even if we in addition shrink any singular value $d_i \leq \tau$ arbitrarily. For example, we can map them to 0 so that our mapping is a step function: $\tilde{d}_i = \tau \mathbb{1}(d_i \geq \tau)$. However, it might be better not to shrink those singular values, since this makes the compatibility constant larger.

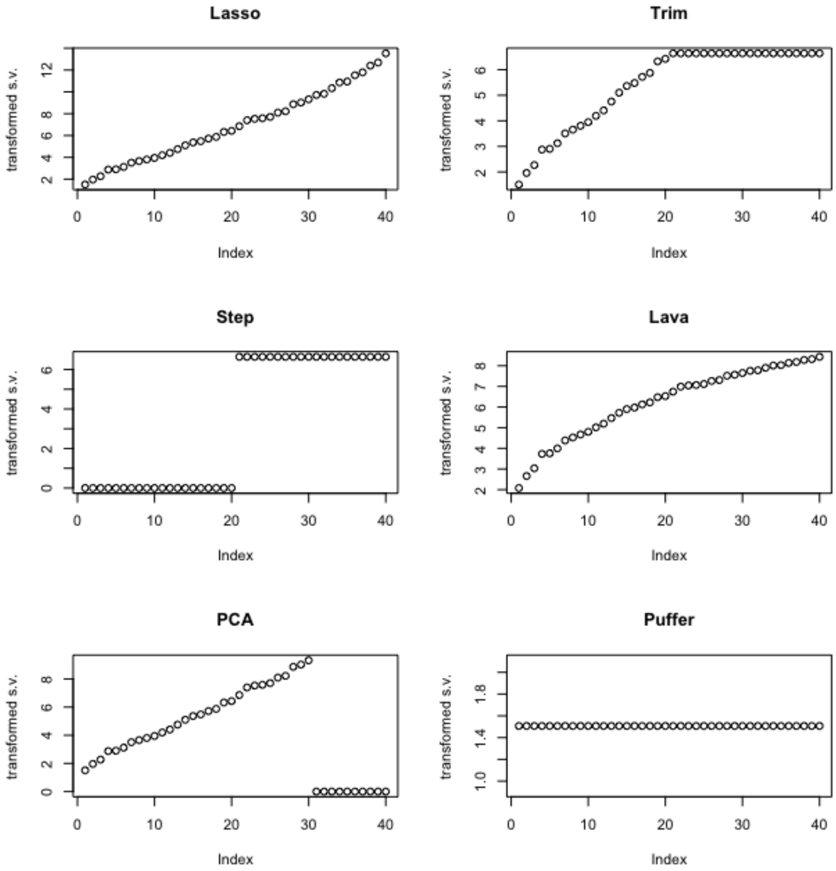

Figure 5.1: Singular values of $\tilde{X}$ after applying spectral transformations corresponding to different methods to $40 \times 60$ matrix $X$ with i.i.d. standard normal entries.

**PCA adjustment** The method which removes some principal components by mapping the corresponding singular values to 0 will satisfy **(B1)** if we remove the large principal components. However, we need to make sure that we keep a proportion of the principal components with large singular values in order to satisfy **(B2)**. This is only possible if only a few of the principal components of $X$ are much larger than the rest, which is the case if the number of the confounding variables is small.

**Puffer transformation** For the Puffer transform [15], where we map all singular values to a constant $d_n$ (because of homogeneity it does not matter to which constant we map it, but we have assumed w.l.o.g. that $\tilde{d}_i \leq d_i$, so we need to map them to $d_n$), the assumption **(B1)** is trivially satisfied.

However, for **(B2)** we need to have $d_n^2 = \Omega\left(\lambda_{\min}(\Sigma)\, p\right)$. From [25], we have that this will hold if and only if $\liminf \frac{p}{n} > 1$.

**Lava** The mapping $d_i \to cd_i/\sqrt{c^2 + d_i^2}$ used in the Lava algorithm [6] behaves similarly as the Trim transform $\tilde{d}_i = \min(d_i, \tau)$. It leaves the small singular values almost unchanged and approximately maps the large singular values to a constant $c$. If we take $k = \lfloor tn \rfloor$ and $c = d_{\lfloor tn \rfloor}$ where $t \in (0, 1)$, the assumption **(B1)** is satisfied since $\tilde{d}_1 \leq c = \sqrt{2}\tilde{d}_k$. Furthermore, by assumption **(A2)** we have $\tilde{d}_k^2 = \frac{1}{2}d_k^2 = \Omega(\lambda_{\min}(\Sigma)p)$, so the assumption **(B2)** is satisfied. This transformation has the property that it is smoother than the Trim transform. We note that with this comment and Theorem 3, we have established the rate optimality of Lava for estimating the sparse parameter $\beta$ in a high-dimensional regression model: such an optimality result of Lava is not given in [6].

There are plenty of other mappings which one can use and with the same properties. For example, the arctan mapping $d_i \to \frac{2c}{\pi}\arctan\left(\frac{\pi d_i}{2c}\right)$ or the exponential mapping $d_i \to c - c\exp\left(-\frac{d_i}{c}\right)$. However, such transformations are somewhat artificial and do not have the interpretation as the Lava, which arises as the solution of the modified Lasso optimization problem with an $\ell_1$- and $\ell_2$-norm regularization.

## 5.4 Validity of the assumptions

In this section we will justify the assumptions from Section 5.3.1. We will especially focus on the confounding model (3.2) in order to investigate under which assumptions on this model our method achieves the optimal error rate.

**Assumption (A1)** This condition says that the coefficient perturbation can not be too large. Since in general it is impossible to distinguish the true coefficient vector $\beta$ from the perturbed coefficient vector $\beta + b$, our model is unidentifiable. To address this, we assume for the perturbation that $\|V^T b\| \to 0$. The rate $\mathcal{O}(\sqrt{\log p/p})$ may seem too strict, but this is the rate with respect to the $\ell_2$-norm, so if the perturbation vector is dense, this becomes approximately $\|b\|_1 = \mathcal{O}(\sqrt{\log p})$.

For the worst case coefficient perturbation, where we want to be able to ensure that the estimation error is small regardless of the direction of the perturbation vector $b$, we need to assume $\|b\|_2 = \mathcal{O}(\sqrt{\log p/p})$. However, if our coefficient perturbation is drawn uniformly from a ball of fixed radius in $\mathbb{R}^p$ and independently of $X$, we have $\mathbb{E}\|V^T b\|_2^2 = \frac{n}{p}\|b\|_2^2$, so we just need $\|b\|_2 = \mathcal{O}(\sqrt{\log p/n})$. Finally, if the coefficient perturbation is caused by the confounding variables as in (3.2), it is given by $b = (\Gamma^T\Gamma + \Sigma_E)^{-1}\Gamma^T\delta$ and it satisfies the following:

**Lemma 2.** *Assume that the coefficients in the confounding model (3.2) satisfy that* $\|\delta\|_2 = \mathcal{O}(\sqrt{\log p})$ *and* $\lambda_{\min}(\Gamma) = \Omega\left(\sqrt{p}\right)$. *Assume also that* $\lambda_{\min}(\Sigma_E)$ *is bounded from below. Then we have:*

$$\|b\|_2 = \|(\Gamma^T\Gamma + \Sigma_E)^{-1}\Gamma^T\delta\|_2 = \mathcal{O}\left(\sqrt{\frac{\log p}{p}}\right)$$

*The condition* $\lambda_{\min}(\Gamma) = \Omega_p(\sqrt{p})$ *is satisfied, for example, if* $\liminf \frac{p}{q} > 1$ *and either the rows or columns of* $\Gamma$ *are independent, identically distributed* $N(0, \Omega)$ *random variables with* $\lambda_{\min}(\Omega)$ *bounded away from zero. If the components of* $\delta$ *are i.i.d. we have* $\|\delta\|_2 = \mathcal{O}_p(\sqrt{q})$, *so we require that the number of latent variables is* $q = \mathcal{O}(\log p)$.

From this we see that it is important that the effect of the latent variables is spread over many predictors. If this is not true, $\lambda_{\min}(\Gamma)$ and thus $\|b\|$ will be too large.

**Assumption (A2)** This assumption ensures that the singular vectors of $X$ are pointing in the favourable direction. This always holds under the uniformity condition described in the next lemma.

**Lemma 3.** *If* $V$ *has a uniform distribution on the Stiefel manifold independently of* $D$, *then for any* $k = \Omega(n)$, *we have*

$$\phi^2_{M_k} = \Omega_p\left(\frac{n}{p}\right).$$

This uniformity assumption is sensible to make since it will be true under any of the two following scenarios: the first is that $\Sigma$ is a multiple of the identity matrix, i.e. that the components of $X$ are i.i.d. normal random variables; the second is that the singular vectors of $\Sigma$ have the uniform distribution on the space of orthogonal matrices themselves. This, for example, might happen in the confounding model (3.2), when $\Sigma_E = \sigma_E^2 I_p$ and $\Gamma Q$ has the same distribution as $\Gamma$ for any orthogonal matrix $Q$, for example if the components of $\Gamma$ are i.i.d. normal variables. The uniformity assumption is sufficient, but not necessary for the assumption (**A2**) to hold. However, the distribution of $\phi_{M_k}$ is not tractable otherwise and the assumption **A2** is hard to verify.

**Assumption (A3)** This assumption implies that a certain proportion of singular values is large enough. The following lemma shows that the assumption (**A3**) is satisfied for the random Gaussian design, such as in confounding model (3.2).

**Lemma 4.** *Assume that $X$ is a random design matrix with rows being drawn i.i.d. from the $N_p(0, \Sigma)$ distribution. If* $\limsup \frac{k}{n} < 1$ *or if* $\liminf \frac{p}{n} > 1$*, we have*

$$d_k^2 = \Omega_p(\lambda_{\min}(\Sigma)p).$$

## 5.5 Low dimensional case: $n > p$

In the previous sections we have considered the high-dimensional case where the number of predictors is higher than the sample size. When $n > p$, the statement of the main Theorem 3, namely that we can achieve the optimal error rate $\mathcal{O}\left(\frac{s\sigma}{\lambda_{\min}(\Sigma)}\sqrt{\frac{\log p}{n}}\right)$, and its proof are still valid if we modify the assumptions as follows:

**(A1')** The perturbation vector $b$ satisfies $\|b\|_2 = \mathcal{O}\left(\frac{\sigma}{\lambda_{\min}(\Sigma)}\sqrt{\frac{s\log p}{n}}\right)$

**(A2')** For any $k = \Omega(p)$, we have $\phi^2_{M_k} = \Omega(1)$

**(A3')** For any $k$ such that $\limsup \frac{k}{p} < 1$, it holds that $d_k^2 = \Omega(\lambda_{\min}(\Sigma)p)$

and then our spectral transformation needs to satisfy

**(B1') and (B2')** $\exists k = \Omega(p)$ such that $\tilde{d}_{(k)} = \Omega(\tilde{d}_{(1)})$ and $\tilde{d}_{(k)} = \Omega(\lambda_{\min}(\Sigma)p)$

which remains true for Lava or Trim transform.

The assumptions **(A2')**, **(A3')** can easily be justified for the random design by the analogues of Lemma 3 and 4. The only substantially stronger assumption is **(A1')**, which will not hold for the confounding model (3.2), since from Lemma 2, which holds for low-dimensional setting as well, we only have that

$$\|b\|_2 = \mathcal{O}\left(\sqrt{\frac{\log p}{p}}\right).$$

This is because $b$ only depends on how the confounding variables affect the predictors and not on the number of data points. The more predictors we have, the more is the effect of the confounding variables spread out.

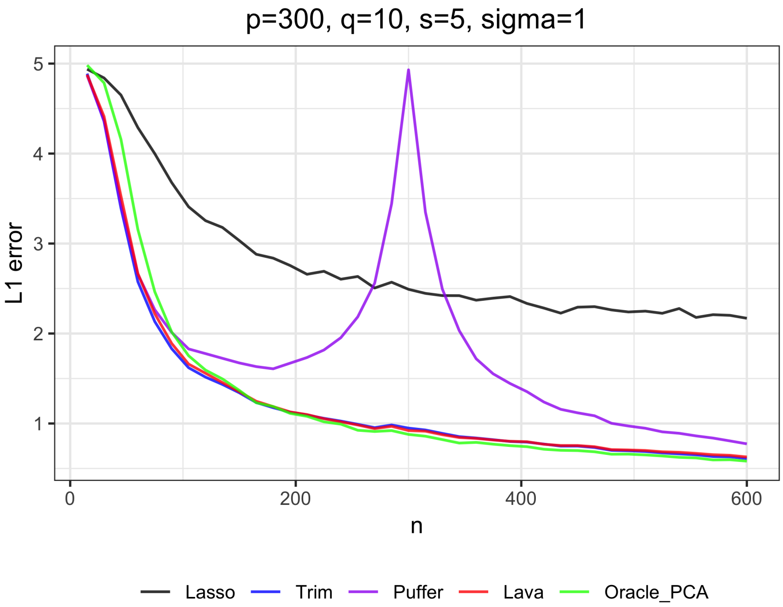

Figure 5.2: Dependence of the estimation error $\|\hat{\beta} - \beta\|_1$ on the sample size, including $p < n$, for different spectral transformations in the confounding model, as described in Section 6.1.1.

From the proof of Theorem 3, we obtain for the low-dimensional setting:

$$\|\hat{\beta} - \beta\|_1 = \mathcal{O}\left( \frac{s\sigma}{\lambda_{\min}(\Sigma)} \sqrt{\frac{\log p}{n}} + \sqrt{s}\|b\|_2 \right).$$

When $n > p$, the first term will not dominate the second term anymore, and our theory only guarantees the error rate $\mathcal{O}\left(\sqrt{\frac{s \log p}{p}}\right)$. This would still converge to zero if $p$ (with $p < n$) diverges to $\infty$ and the sparsity $s$ is sufficiently small. One can not expect the same error rate as in the high-dimensional setting, since this would imply that, for fixed $p$, the error converges to 0 as $n \to \infty$ and this can not happen because of the perturbation $b$. Since for the coefficients in the active set we can not distinguish the signal from the perturbation, we expect that the error rate can not be smaller than $\approx s\|b\|_\infty$. It would be interesting to find the optimal error rate with respect to the perturbation $b$.

This is also illustrated in Figure 5.2, where we can see that even though the error decreases as we increase the number of data points, it still has a nonzero limit. However, the error is small, especially in the comparison with the standard Lasso, and there is a benefit in using our method.

# 6 Empirical Results

We present here empirical results for simulated and real data.

## 6.1 Simulations

We demonstrate the performance of various spectral transformations for estimating the coefficient vector $\beta$ with the Lasso: Trim transform, Lava, Puffer and PCA adjustment. We investigate the cases when the perturbation $b$ is randomly sampled and when it arises from the hidden confounding.

### 6.1.1 Setting

We generate the data from the confounding model (3.2). We take $\Sigma_E = I_p$ and $\beta = (1, 1, 1, 1, 1, 0, \dots, 0)$, so $s = 5$. For various numbers $q$ of hidden confounders, we sample the coefficients $\Gamma_{ij}$ and $\delta_i$ independently as standard normal random variables. Finally, we consider different noise levels $\sigma$ for the standard deviation of $\epsilon$. Unless stated otherwise, the sample size is set to be $n = 100$ and the dimensionality of the predictors is $p = 200$. All results are based on 500 independent simulations.

It is also interesting to consider the perturbed linear model (3.1). We do not generate data from this model directly, but we will modify the perturbation term $b$ obtained from the confounding model. This way we can compare the results obtained in the confounding model and the perturbed linear model directly with each other. We replace $b = (\Gamma^T\Gamma + \Sigma_E)^{-1}\Gamma^T\delta$ by $Qb$ where $Q$ is a random rotation matrix so that the new perturbation has the same size, but with uniformly random direction. We note that the resulting distribution is the same as of the perturbed linear model (3.1), where rows of $X$ are drawn from $N(0, \Sigma)$, where $\Sigma = \Gamma^T\Gamma + I_p$, and $b$ is drawn uniformly from a ball of radius $(\Gamma^T\Gamma + I_p)^{-1}\Gamma^T\delta$.

### 6.1.2 Choosing $\lambda$

In practice we encounter the problem of choosing the penalty level $\lambda$ for the Lasso after applying the spectral transformation. Usually this is done by cross-validation. However, in our case $\beta + b$ describes the data much better than $\beta$, which we are trying to recover. Therefore, cross-validation tends to choose much smaller value of $\lambda$ than the one we want. For this reason, and for simplicity, in our simulations we have used the oracle value of $\lambda$, i.e. the one for which $\|\hat{\beta}_\lambda - \beta\|_1$ is smallest.

Even though this might seem to be a problem, this method can still be efficiently used as a screening tool, e.g. we can choose the smallest $\lambda$ for which the corresponding set of selected variables will have the wanted size. We could use such this screened set of variables in the second stage, for example, by using OLS on the $\tilde{X}_{\hat{S}}$ and $\tilde{Y}$. Using other methods, such as OLS, in combination with the Lasso is common in practice [19].

### 6.1.3 Results

Here we present the results of the simulations for both the confounding model and the perturbed linear model. A fundamental difference between them is that the coefficient perturbation arising from the confounding model is pointing towards the singular vectors of $X$ corresponding to the large singular values (see Figure 4.1). This makes $\|Xb\|_2$ larger for a fixed $\|b\|_2$, and in this case the estimation error will be larger. On the other hand, in this case we can improve our accuracy more by shrinking large singular values, as will be shown below.

**Noise versus perturbation** In the left plot in Figure 6.1 we can see how the estimation error changes depending on the size of the noise $\sigma$ in the confounding model. When $\sigma$ is small, the perturbation $b$ has the biggest effect on the error. On the other hand, if $\sigma$ is large, then the influence of the perturbation $b$ becomes less important.

We can see that the standard Lasso is affected a lot by the coefficient perturbation, whereas the Puffer transformation is affected the most by additive noise, since the slope of its corresponding curve is the steepest. When $n, p$ are close to each other, some of the singular values of $X$ become quite small and thus mapping them to a constant can inflate the error $\epsilon$ a lot in the corresponding directions, this is evident in Figure 5.2.

We can observe that the oracle PCA adjustment, which removes exactly the $q$ largest singular values from $X$, works well, especially when $\sigma$ is small. For larger $\sigma$, we see that the Trim transform and Lava work slightly better since they do not remove that much of the signal.

In the right plot of Figure 6.1, we have randomized the direction of $b$ while keeping everything else constant, as described in Section 6.1.1. This then corresponds to a model with random perturbation $b$ but no specific further structure in terms of confounding. We can see a substantial improvement of the standard Lasso: in hindsight this shows that the Lasso is very sensitive to confounding variables but much less so to perturbation of sparsity. Also, it is worth noting that the PCA adjustment method is now consistently worse than the Trim transform or Lava, since the projection of $b$ onto the span of the first $q$ singular vectors is not that large anymore.

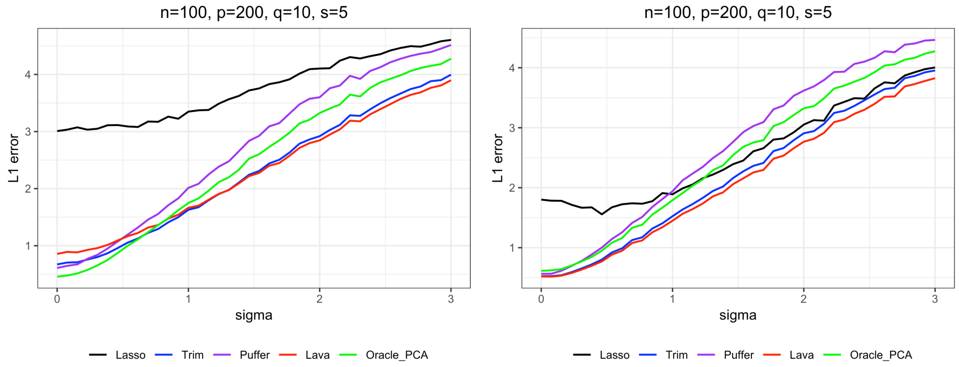

Figure 6.1: Dependence of the estimation error $\|\hat{\beta} - \beta\|_1$ on the size of the noise for different spectral transformation for confounding model (left) and the perturbed linear model (right), as described in Section 6.1.1.

**Number of confounding variables** In Figure 6.2 we can see how the estimation error depends on the number $q$ of confounding variables. As above, we see that the Lasso is severely affected by the presence of the confounding variables. The Puffer transform does not work very well since $n$ and $p$ are of similar size, whereas the Trim transform and Lava exhibit similar and good performance in all cases.

PCA adjustment works well only for the confounding model and only if we correctly guess the number of confounding variables. In the left plot in Figure 6.2 we can clearly see how the estimation error is affected by the misspecification of the number of the principal components we remove. The oracle PCA method, which removes exactly $q$ principal components, performs reasonably well, particularly for smaller values of $q$. However, if we overestimate or especially if we underestimate the number of confounding variables, the estimation error will become significantly worse compared to the Trim transform or Lava.

**Method robustness** We are interested in whether there are any disadvantages in using the spectral transformations if we wrongly think that the sparse coefficient has been perturbed or that there is some hidden confounding.

In Figure 6.3 we display the estimation error for the confounding model as in Figure 6.2, but where the coefficient bias $b$ has been set to 0. i.e. this is a standard sparse linear model with $X$ being generated from the spiked covariance model.

There is no indication for relevant differences between the performances of the Trim transform, Lava and the Lasso. The Lasso performs slightly better

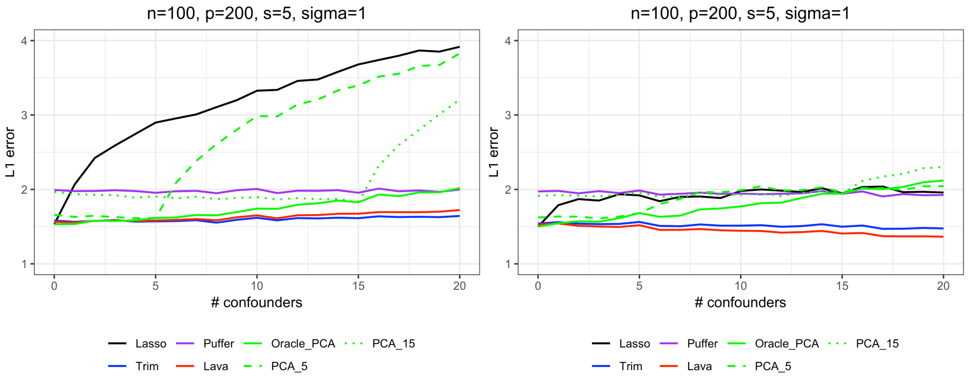

Figure 6.2: Dependence of the estimation error $\|\hat{\beta} - \beta\|_1$ on the number of confounding variables for different spectral transformation for confounding model (left) and the perturbed linear model (right) as described in section 6.1.1.

for larger values of $q$ and slightly worse for smaller $q$. It is worth noting that on this plot the estimation error starts to decrease as $q$ increases, which is due to a scaling issue. This happens because the variance of $X$ increases as $q$ increases, since $\Sigma = \Gamma^T\Gamma + \Sigma_E$, thus effectively increasing the signal to noise ratio.

Our empirical results support theoretical evidence which showed that it is safe to use wisely chosen spectral transformations such as the Trim transform or the Lava. If there are any confounding variables present, there is a large improvement over the standard Lasso. On the other hand, if there are no confounding variables, the Trim transform or Lava will have about the same performance as the Lasso. Therefore, our method can be thought of as an easy to use modification of the Lasso which is robust to hidden confounding.

## 6.2 Application to genomic dataset

In this section we demonstrate the robustness of our method against hidden confounders for a real genomic dataset where we have certain knowledge about the confounding variables. We inspect various spectral transformations followed by the Lasso and evaluate the differences between the estimates for the original data and the one where the confounding variables have been adjusted for.

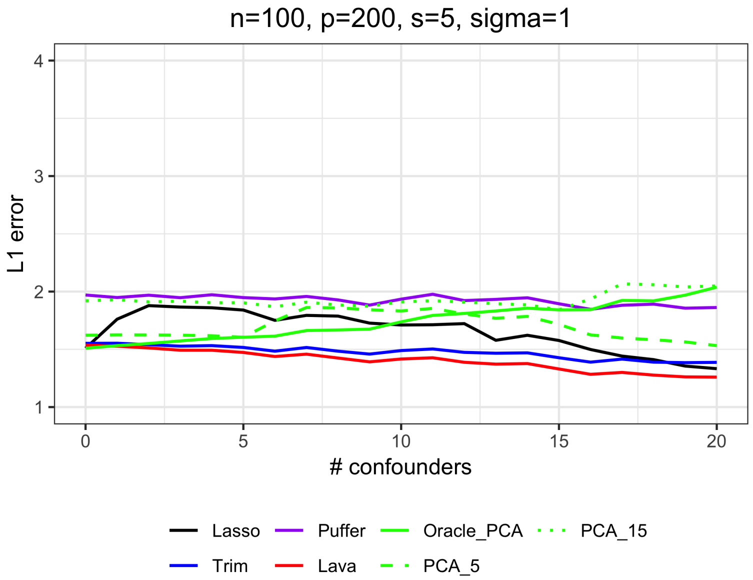

Figure 6.3: Size of the estimation error $\|\hat{\beta} - \beta\|_1$ for a sparse linear model where $\Sigma = \Gamma^T\Gamma + I_p$, i.e. the confounding model with the induced perturbation $b$ set to $b = 0$.

### 6.2.1 Gene expression dataset

We have obtained data from the GTEx Portal (http://gtexportal.org). The GTEx project provides large-scale data with an aim to help the scientific community to study gene expression, gene regulation and their relationship to genetic variation. It provides gene expression data from 11688 samples collected postmortem from 53 different tissues of 714 human donors.

Gene expression is a process in the cell in which the information stored in a certain gene is used for the synthesis of gene products such as proteins. In the GTEx Project it was quantified by the amount of the mRNA in the cell which was created from this gene. Gene expression differs among different people and even among different cells within the human body. The type of the cells is determined by the gene expression within them; even though the DNA in all cell nuclei is the same, cells in different tissues behave and look differently and perform significantly different tasks. Gene expression is also affected by the genetic variation and determining the expression quantitative trait loci (eQTL), which are parts of genome which explain the variation in the gene expression, is a very important problem which will help to understand the relationship between genetic variation and organismal phenotypes.

### 6.2.2 Setting

We use the fully processed, filtered and normalized gene expression matrix for the skeletal muscle tissue. We consider the gene expression of $p = 14713$ protein-coding genes measured from $n = 491$ samples. For our purpose, an important aspect of this dataset is that there are also $q = 65$ different covariates provided, which are proxys for hidden confounding variables. They include genotyping principal components and PEER factors. We thus obtain the deconfounded data by regressing out these given covariates.

The left panel of Figure 6.4 displays the singular values of the initial data matrix. We see that the first several singular values are substantially larger than the rest which suggests a possible existence of hidden confounders. In the right part of Figure 6.4 we can see the singular values of the deconfounded data matrix where we have regressed out all of the $q = 65$ covariates which are provided as confounding proxies.

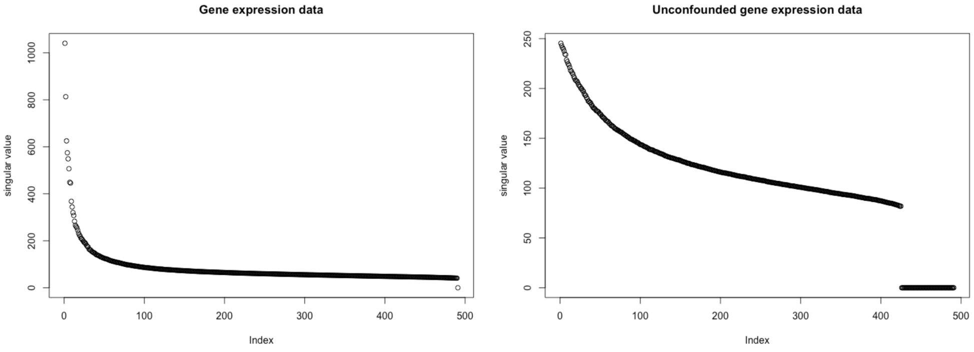

Figure 6.4: Singular values of the gene expression data matrix for skeletal muscle tissue before (left) and after (right) regressing out the provided $q = 65$ confounding covariates.

We are going to explore now the robustness of the Lasso, Trim transform, and Lava against hidden confounders by comparing the estimates based on the original and the deconfounded data. For a fixed value of $k$, we regress out first $k$ provided proxy confounders from the original gene expression data matrix $X$ in order to get the matrix $X^{(k)}$ and we randomly choose one column to represent the response $Y$. We are thus trying to explain the expression of one gene by the expressions of other genes.

For every $s = 1, \dots, 20$, we apply the given method on $X$ and $X^{(k)}$ with the regularization $\lambda$ chosen as the largest value such that the support size of $\hat{\beta}$ equals a pre-specified value $s$. This leads to estimates $\hat{\beta}_s$ and $\hat{\beta}_s^{(k)}$ We measure the similarity of the corresponding supports by $J(\operatorname{supp}\hat{\beta}_s, \operatorname{supp}\hat{\beta}_s^{(k)})$, where $J$ is the Jaccard distance:

$$J(A, B) = \frac{A \triangle B}{A \cup B}$$

### 6.2.3 Results

In the top left image in Figure 6.5, we can see the difference of the estimates for the original and the deconfounded data, where 5 arbitrarily chosen confounding variables have been removed. We can see that the Jaccard distance for the Lasso is close to 1, indicating that the estimated support sets are very different and almost disjoint; The Trim transform and Lava are much more robust to the hidden confounders and we see that the Jaccard distance between the estimates based on confounded and deconfounded data is much smaller.

In order to make sure that the choice of response $Y$ did not affect the results, we have repeated this experiment for 500 randomly chosen genes and averaged the obtained results. The results are also displayed in the Figure 6.5. We can see that, as we increase the number $k$ of confounding variables which we regress out, the Jaccard distance for all methods is increasing. This is to be expected since $X^{(k)}$ and $X$ are becoming more different as we increase $k$. However, we can infer that the Trim transform and Lava are consistently better than the Lasso, exhibiting also in this real dataset the robustness against confounding variables.

## 7 Discussion

We propose to add robustness against hidden confounding variables by pre-employing a wisely chosen spectral transformation before using the Lasso or other high-dimensional sparse regression techniques. There is essentially nothing to lose but much to be gained which is in line with the typical argument of robustness [14], [11]. Besides the robustness issue, we can also take the viewpoint of deconfounding before further analysis: this would be the typical approach for problems and applications where hidden confounding is expected to happen, a prime example being genetics [21].

The confounding issue in the context of linear models can be represented as a regression problem with coefficient $\beta + b$; the coefficient $\beta$ is the true underlying parameter in absence of confounding variables, while the perturbation $b$ is due to the confounding. We develop theory for a linear model with regression parameters $\beta + b$ where $\beta$ is sparse and the perturbation $b$

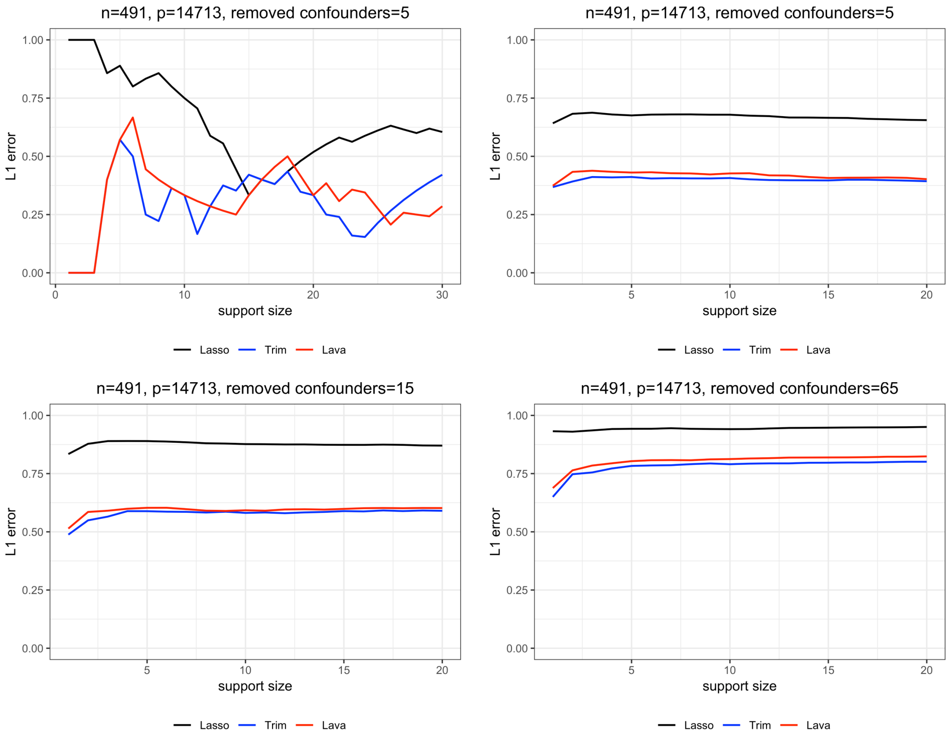

Figure 6.5: Jaccard distance of the supports of the estimates based on the original and deconfounded data for one randomly chosen response (top left). Jaccard distance, averaged over 500 randomly chosen responses, of the supports of estimates based on the original data and data with 5 (top right), 15 (bottom left) and 65 (bottom right) confounder proxies removed.

sufficiently small or of a special structure. We show for a class of spectral transformations, in conjunction with using the Lasso afterwards, that the method achieves the minimax optimal convergence rate of $\|\hat{\beta} - \beta\|_1$, that is, for estimating the sparse parameter part of the problem; see Section 5 and Theorem 3. Such a theoretical result is entirely new and covers also the Lava method [6]. In particular, the theoretical result also establishes spectral deconfounding as an optimal method for removing the effect of dense hidden confounders in high-dimensional settings.

Another advantage of our approach is its simplicity: it is just one simple pre-transformation step before using the Lasso. It just requires the computation of the SVD of the design matrix which has computational complexity of $\mathcal{O}(\min(n^2p, np^2))$ and can be done in just a few lines of code.

The topic of deconfounding has not received too much attention, despite its practical importance [3, 10]. Here we have shown that it is possible to protect against hidden dense confounding in the case of linear regression. Similar ideas might be powerful as well for more complicated models.

# References

[1] Theodore Wilbur Anderson, Theodore Wilbur Anderson, Theodore Wilbur Anderson, Theodore Wilbur Anderson, and Etats-Unis Mathématicien. *An introduction to multivariate statistical analysis*, volume 2. Wiley New York, 1958.

[2] Peter J Bickel, Ya'acov Ritov, Alexandre B Tsybakov, et al. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

[3] M Alan Brookhart, Til Stürmer, Robert J Glynn, Jeremy Rassen, and Sebastian Schneeweiss. Confounding control in healthcare database research: challenges and potential approaches. *Medical care*, 48(6 0):S114, 2010.

[4] Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

[5] Venkat Chandrasekaran, Pablo A Parrilo, and Alan S Willsky. Latent variable graphical model selection via convex optimization. In *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*, pages 1610–1613. IEEE, 2010.

[6] Victor Chernozhukov, Christian Hansen, Yuan Liao, et al. A lava attack on the recovery of sums of dense and sparse signals. *The Annals of Statistics*, 45(1):39–76, 2017.

[7] Johann A Gagnon-Bartsch and Terence P Speed. Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, 13(3):539–552, 2012.

[8] David Gerard and Matthew Stephens. Empirical bayes shrinkage and false discovery rate estimation, allowing for unwanted variation. *arXiv preprint arXiv:1709.10066*, 2017.

[9] Christophe Giraud. *Introduction to high-dimensional statistics*. Chapman and Hall/CRC, 2014.

[10] Sander Greenland, James M Robins, and Judea Pearl. Confounding and collapsibility in causal inference. *Statistical science*, pages 29–46, 1999.

[11] Frank R Hampel, Elvezio M Ronchetti, Peter J Rousseeuw, and Werner A Stahel. *Robust statistics: the approach based on influence functions*, volume 196. John Wiley & Sons, 2011.

[12] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. CRC press, 2015.

[13] Jim C Huang and Nebojsa Jojic. Variable selection through correlation sifting. In *International Conference on Research in Computational Molecular Biology*, pages 106–123. Springer, 2011.

[14] Peter J Huber. Robust statistics. In *International Encyclopedia of Statistical Science*, pages 1248–1251. Springer, 2011.

[15] Jinzhu Jia, Karl Rohe, et al. Preconditioning the lasso for sign consistency. *Electronic Journal of Statistics*, 9(1):1150–1172, 2015.

[16] Iain M Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Annals of statistics*, pages 295–327, 2001.

[17] Jeffrey T Leek and John D Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS genetics*, 3(9):e161, 2007.

[18] Vladimir Alexandrovich Marchenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik*, 114(4):507–536, 1967.

[19] Nicolai Meinshausen. Relaxed lasso. *Computational Statistics & Data Analysis*, 52(1):374–393, 2007.

[20] John Novembre, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R Boyko, Adam Auton, Amit Indap, Karen S King, Sven Bergmann, Matthew R Nelson, et al. Genes mirror geography within europe. *Nature*, 456(7218):98, 2008.

[21] John Novembre and Matthew Stephens. Interpreting principal component analyses of spatial population genetic variation. *Nature genetics*, 40(5):646, 2008.

[22] Debashis Paul, Eric Bair, Trevor Hastie, and Robert Tibshirani. "preconditioning" for feature selection and regression in high-dimensional problems. *The Annals of Statistics*, pages 1595–1618, 2008.

[23] Rajen Shah and Nicolai Meinshausen. Rsvp-graphs: Fast high-dimensional covariance matrix estimation under latent confounding. *arXiv preprint arXiv:1811.01076*, 2018.

[24] Sara Van de Geer. Estimation and testing under sparsity. *Lecture Notes in Mathematics*, 2159, 2016.

[25] Roman Vershynin. High dimensional probability. *An Introduction with Applications*, 2016.

[26] JM Wainwright. High-dimensional statistics: A non-asymptotic viewpoint. *preparation. University of California, Berkeley*, 2015.

[27] Yixin Wang and David M Blei. The blessings of multiple causes. *arXiv preprint arXiv:1805.06826*, 2018.

# 8 Acknowledgements

We would like to thank Gian Thanei and Benjamin Frot for providing processed data for our biological applications.

# 9 Proofs

**Theorem 1.** *Assume the model in (3.1) with fixed design $X$ and i.i.d. zero-mean sub-Gaussian errors $\epsilon_i$, for $i = 1, \dots n$. Let $F \in \mathbb{R}^{n \times n}$ be an arbitrary linear transformation and $A > 0$ an arbitrary fixed constant. Then there exists $\lambda_{\min} \geq 0$ such that for any $\lambda \in [\lambda_{\min}, B\lambda_{\min}]$, where $B \geq 1$ is a fixed constant, and with probability at least $1 - 2p^{1-A^2/8}$ we have*

$$\|\hat{\beta} - \beta\|_1 \leq C_1 \frac{s\sigma}{\phi^2_{\tilde{\Sigma}}} \sqrt{\frac{\log p}{n}} \max_i \left( \frac{X^T (F^T F)^2 X}{n} \right)^{1/2}_{ii} + C_2 \sqrt{\frac{s}{n}} \frac{\|\tilde{X} b\|_2}{\phi_{\tilde{\Sigma}}},$$

*where $C_1, C_2$ are constants depending only on $A$ and $B$.*

**Proof.** Denote by $\beta^0$ the true coefficient vector.

Since $\hat{\beta}$ minimizes $\frac{1}{n}\|\tilde{Y}-\tilde{X}\beta\|_2^2+\lambda\|\beta\|_1$, we have:

$$\begin{aligned}
\frac{1}{n}\|\tilde{Y}-\tilde{X}\hat{\beta}\|_2^2+\lambda\|\hat{\beta}\|_1 &\leq \frac{1}{n}\|\tilde{Y}-\tilde{X}\beta^0\|_2^2+\lambda\|\beta^0\|_1 \\
\frac{1}{n}\|\tilde{X}(\hat{\beta}-\beta^0)\|_2^2+\lambda\|\hat{\beta}\|_1 &\leq \frac{2}{n}(\tilde{Y}-\tilde{X}\beta^0)^T\tilde{X}(\hat{\beta}-\beta^0)+\lambda\|\beta^0\|_1 \\
&\leq \frac{2}{n}\tilde{\epsilon}^T\tilde{X}(\hat{\beta}-\beta^0)+\frac{2}{n}b^T\tilde{X}^T\tilde{X}(\hat{\beta}-\beta^0)+\lambda\|\beta^0\|_1 \\
\frac{1}{n}\|\tilde{X}(\hat{\beta}-\beta^0-b)\|_2^2+\lambda\|\hat{\beta}\|_1 &\leq \frac{2}{n}\tilde{\epsilon}^T\tilde{X}(\hat{\beta}-\beta^0)+\frac{1}{n}\|\tilde{X}b\|_2^2+\lambda\|\beta^0\|_1
\end{aligned}$$

Let us work on the event $\{\|\frac{2}{n}\tilde{X}^T\tilde{\epsilon}\|_\infty \leq \tau\}$, which has probability at least $1-2p^{1-A^2/8}$ for $\tau = A\sigma\sqrt{\frac{\log(p)}{n}}\max_{i\leq n}\left(\frac{X^T(F^TF)^2X}{n}\right)_{ii}^{1/2}$, as it is shown in the Lemma 5. On this event we have

$$\frac{2}{n}\tilde{\epsilon}^T\tilde{X}(\hat{\beta}-\beta^0) \leq \frac{2}{n}\|\tilde{X}^T\tilde{\epsilon}\|_\infty\|\hat{\beta}-\beta^0\|_1 \leq \tau\|\hat{\beta}-\beta^0\|_1$$

from Hölder's inequality. We now have:

$$\frac{1}{n}\|\tilde{X}(\hat{\beta}-\beta^0-b)\|_2^2+\lambda\|\hat{\beta}\|_1 \leq \tau\|\hat{\beta}-\beta^0\|_1+\frac{1}{n}\|\tilde{X}b\|_2^2+\lambda\|\beta^0\|_1$$

By using that $\beta^0_{S^c}=0$, we get that

$$\begin{aligned}
&\frac{1}{n}\|\tilde{X}(\hat{\beta}-\beta^0-b)\|_2^2+(\lambda-\tau)\|\hat{\beta}_{S^c}\|_1 \\
&\qquad\leq \tau\|\hat{\beta}_S-\beta^0_S\|_1+\lambda\|\beta^0_S\|_1-\lambda\|\hat{\beta}_S\|_1+\frac{1}{n}\|\tilde{X}b\|_2^2 \\
&\qquad\leq (\lambda+\tau)\|\hat{\beta}_S-\beta^0_S\|_1+\frac{1}{n}\|\tilde{X}b\|_2^2
\end{aligned}$$

Let us now write

$$\phi_{\tilde{\Sigma}}(L,S) = \min_{\beta\in R(L,S)}\frac{\sqrt{\beta^T\tilde{\Sigma}\beta}}{\frac{1}{\sqrt{s}}\|\beta_S\|_1} > 0$$

We need to consider two cases:

- Case 1: $\frac{1}{n}\|\tilde{X}b\|_2^2 \leq \lambda\|\hat{\beta}_S-\beta^0_S\|_1$
- Case 2: $\frac{1}{n}\|\tilde{X}b\|_2^2 \geq \lambda\|\hat{\beta}_S-\beta^0_S\|_1$

In the first case we have

$$\frac{1}{n}\|\tilde{X}(\hat{\beta}-\beta^0-b)\|_2^2+(\lambda-\tau)\|\hat{\beta}_{S^c}-\beta_{S^c}\|_1\leq(2\lambda+\tau)\|\hat{\beta}_S-\beta^0_S\|_1$$

From this we see that the error $\hat{\beta}-\beta\in R(L,S)=\{x:\|x_{S^c}\|_1\leq L\|x_S\|_1\}$ for $L=\frac{2\lambda+\tau}{\lambda-\tau}$ (we take $\lambda>\tau$), so we have:

$$\begin{aligned}\frac{1}{n}\|\tilde{X}(\hat{\beta}-\beta^0-b)\|_2^2+(\lambda-\tau)\|\hat{\beta}-\beta^0\|_1&\leq 3\lambda\|\hat{\beta}_S-\beta^0_S\|_1\\&\leq\frac{3\lambda\sqrt{s}\|\tilde{X}(\hat{\beta}-\beta^0)\|_2}{\sqrt{n}\phi_{\tilde{\Sigma}}(L,S)}\\&\leq\frac{3\lambda\sqrt{s}\|\tilde{X}(\hat{\beta}-\beta^0-b)\|_2}{\sqrt{n}\phi_{\tilde{\Sigma}}(L,S)}+\frac{3\lambda\sqrt{s}\|\tilde{X}b\|_2}{\sqrt{n}\phi_{\tilde{\Sigma}}(L,S)}\\&\leq\frac{9\lambda^2 s}{2\phi_{\tilde{\Sigma}}(L,S)^2}+\frac{1}{n}\|\tilde{X}(\hat{\beta}-\beta^0-b)\|_2^2+\frac{1}{n}\|\tilde{X}b\|_2^2\end{aligned}$$

by using the inequality $xy\leq\frac{x^2}{4}+y^2$ twice, which finally gives us

$$(\lambda-\tau)\|\hat{\beta}-\beta^0\|_1\leq\frac{9\lambda^2 s}{2\phi_{\tilde{\Sigma}}(L,S)^2}+\frac{1}{n}\|\tilde{X}b\|_2^2$$

In the second case we have

$$\frac{1}{n}\|\tilde{X}(\hat{\beta}-\beta^0-b)\|_2^2+(\lambda-\tau)\|\hat{\beta}-\beta^0\|_1\leq\frac{3}{n}\|\tilde{X}b\|_2^2$$

So, regardless whether we are in the Case 1 or the Case 2, we get that

$$(\lambda-\tau)\|\hat{\beta}-\beta^0\|_1\leq\frac{9\lambda^2 s}{2\phi_{\tilde{\Sigma}}(L,S)^2}+\frac{3}{n}\|\tilde{X}b\|_2^2$$

By dividing by $(\lambda-\tau)$ and minimizing over $\lambda>\tau$, we get that the minimum value of the RHS of the bound is:

$$\frac{9s\tau}{\phi_{\tilde{\Sigma}}(L,S)^2}+\sqrt{\left(\frac{9s\tau}{\phi_{\tilde{\Sigma}}(L,S)^2}\right)^2+\frac{54s\|\tilde{X}b\|_2^2}{\phi_{\tilde{\Sigma}}(L,S)^2 n}}$$

which is achieved for

$$\lambda_{\min}=\tau+\sqrt{\tau^2+\frac{2\phi_{\tilde{\Sigma}}(L,S)^2\|\tilde{X}b\|_2^2}{3sn}}$$

Therefore, for $\lambda \in [\lambda_{\min}, B\lambda_{\min}]$, we then get

$$\|\hat{\beta} - \beta\|_1 \leq \frac{9B^2 s\tau}{\phi_{\tilde{\Sigma}}(L,S)^2} + B^2 \sqrt{\left(\frac{9s\tau}{\phi_{\tilde{\Sigma}}(L,S)^2}\right)^2 + \frac{54s\|\tilde{X}b\|_2^2}{\phi_{\tilde{\Sigma}}(L,S)^2 n}}$$

In the case when $b = 0$ and $F = I_n$ (the usual Lasso regression), we indeed take $\lambda = 2\tau$. We can see that, when the coefficient perturbation is present, it is better to penalize more as this will remove the effect of the perturbation to some extent.

Since $L = \frac{2\lambda+\tau}{\lambda-\tau}$ and $\lambda \geq 2\tau$, we have $L \leq 5$ and then

$$\phi_{\tilde{\Sigma}}(L,S) \geq \phi_{\tilde{\Sigma}}(5,S) = \phi_{\tilde{\Sigma}}$$

Finally, by using this and the inequality $\sqrt{x^2+y^2} \leq x + y$ where $x, y > 0$, we get

$$\|\hat{\beta} - \beta^0\|_1 \leq \frac{18B^2 s\tau}{\phi_{\tilde{\Sigma}}^2} + B^2 \sqrt{\frac{27s\|\tilde{X}b\|_2^2}{\phi_{\tilde{\Sigma}}^2 n}}$$

which is what we wanted to show. We see that $C_1 = 18AB^2$ and $C_2 = 3\sqrt{3}B^2$. □

**Lemma 5.** *Let $A > 0$ be arbitrary constant. Let us define*

$$\tau = A\sigma\sqrt{\frac{\log(p)}{n}} \max_{i \leq n} \left(\frac{X^T(F^TF)^2X}{n}\right)_{ii}^{1/2}$$

*If the components of $\epsilon$ are i.i.d. sub-Gaussian random variables with mean 0 and parameter $\sigma$, we have*

$$\mathbb{P}\left(\frac{2}{n}\|\tilde{X}^T\tilde{\epsilon}\|_\infty \leq \tau\right) \geq 1 - 2p^{1-A^2/8}$$

**Proof.** Let us write $\nu = \frac{2}{n}\tilde{X}^T\tilde{\epsilon} = \frac{2}{n}X^TF^TF\epsilon$. We have that $\nu_i$ is a sub-Gaussian random variable with mean 0 and parameter $\sigma_i = \frac{2\sigma}{n}\|(F^TF)Xe_i\|_2$ From the union bound we have:

$$\mathbb{P}(\|\nu\|_\infty > \tau) \leq \sum_i \mathbb{P}(|\nu_i| > \tau) \leq p \max_i \mathbb{P}(|\nu_i| > \tau)$$

From the tail bound for sub-Gaussian random variables we now get:

$$\mathbb{P}(\|\nu\|_\infty \leq \tau) \geq 1 - 2\exp\left(-\frac{\tau^2}{2\max_i \sigma_i^2} + \log p\right)$$

Therefore, choosing

$$\tau = A\sigma\sqrt{\frac{\log(p)}{n}} \max_i \left(\frac{X^T(F^TF)^2X}{n}\right)_{ii}^{1/2}$$

ensures that

$$\mathbb{P}(\|\nu\|_\infty \leq \tau) \geq 1 - 2p^{1-A^2/8}$$

as required. □

**Lemma 1.** *Consider a spectral transformation $F$ as in (4.1). Let $1 \leq k < r = \min(n,p)$ be an arbitrary integer. Then:*

$$\phi^2_{\tilde{\Sigma}} \geq \sum_{i=1}^{r} \frac{1}{n}\tilde{d}^2_{(i)}(\phi^2_{M_i} - \phi^2_{M_{i-1}}) \geq \frac{1}{n}\tilde{d}^2_{(k)}\phi^2_{M_k}.$$

**Proof.** We have

$$\alpha^T\tilde{\Sigma}\alpha = \sum_{i\leq n}\tilde{d}^2_i(V_i^T\alpha)^2 = \sum_{i\leq n}(\tilde{d}^2_{(i)} - \tilde{d}^2_{(i+1)})\sum_{j\leq i}(V_{(j)}^T\alpha)^2$$

where we define $\tilde{d}_{n+1} = 0$ for convenience. Now using the fact that the infimum of the sum is not smaller than the sum of the infimums, we get

$$\phi^2_{\tilde{\Sigma}} \geq \sum_{i\leq n}\frac{1}{n}(\tilde{d}^2_{(i)} - \tilde{d}^2_{(i+1)})\phi^2_{M_i} = \sum_{i\leq n}\frac{1}{n}\tilde{d}^2_{(i)}(\phi^2_{M_i} - \phi^2_{M_{i-1}})$$

where $M_0$ is defined as the null matrix for convenience. Let us now fix $k \leq n$. By using that the sequence $\tilde{d}_{(i)}$ is decreasing, we have

$$\sum_{i\leq n}\frac{1}{n}\tilde{d}^2_{(i)}(\phi^2_{M_i} - \phi^2_{M_{i-1}}) \geq \sum_{i\leq k}\frac{1}{n}\tilde{d}^2_{(k)}(\phi^2_{M_i} - \phi^2_{M_{i-1}}) = \frac{1}{n}\tilde{d}^2_{(k)}\phi^2_{M_k}$$

which finishes the proof. □

**Theorem 2.** *Under assumptions of Theorem 1, for any $k \leq r = \min(n,p)$ and any spectral transformation $F$ mapping $d_i$ to $\tilde{d}_i$, we get*

$$\|\hat{\beta} - \beta\|_1 \leq C_1\frac{s\sigma}{\frac{1}{n}\tilde{d}^2_{(k)}\phi^2_{M_k}}\sqrt{\frac{\log p}{n}}\max_i\left(\frac{\tilde{d}_i}{d_i}\right)^2 + C_2\sqrt{s}\frac{\tilde{d}_{(1)}\|V^Tb\|_2}{\tilde{d}_{(k)}\phi_{M_k}}.$$

**Proof.** By the Lemma 1, we have

$$\phi^2_{\tilde{\Sigma}} \geq \frac{1}{n}\tilde{d}^2_{(k)}\phi^2_{M_k}$$

From the facts that $X$ is scaled so that its columns have norm $\sqrt{n}$ and that $F$ is a spectral transform mapping $d_i$ to $\tilde{d}_i$, we get

$$\max_{i\leq n}\left(\frac{X^T(F^TF)^2X}{n}\right)_{ii}^{1/2} = \max_{i\leq n}\frac{1}{\sqrt{n}}\|(F^TF)Xe_i\|_2$$
$$\leq \max_{i\leq n}\frac{1}{\sqrt{n}}\|F^TF\|_2\|Xe_i\|_2 \leq \max_{i\leq n}\left(\frac{\tilde{d}_i}{d_i}\right)^2$$

Finally, we have that

$$\|\tilde{X}b\|_2 = \|U\tilde{D}V^Tb\|_2 \leq \|\tilde{D}\|_2\|V^Tb\|_2 = \tilde{d}_{(1)}\|V^Tb\|_2$$

Combining those inequalities with the bound from Theorem 1, we get the required bound for spectral transformation $F$. □

**Theorem 3.** *Assume that the model assumptions **(A1)**, **(A2)** and **(A3)** hold. Consider a spectral transformation $F = U\tilde{D}D^{-1}U^T$ with $\tilde{d}_i \leq d_i$ which satisfies: there exists $k = \Omega(n)$ such that*

***(B1)*** $\tilde{d}_{(k)} = \Omega\left(\tilde{d}_{(1)}\right)$

***(B2)*** $\tilde{d}^2_{(k)} = \Omega\left(\lambda_{\min}(\Sigma)\,p\right)$

*Then we can choose $\lambda$ so that, despite the coefficient perturbation, the $\ell_1$-estimation error achieves the minimax optimal rate:*

$$\|\hat{\beta}-\beta\|_1 = \mathcal{O}_p\left(\frac{\sigma s}{\lambda_{\min}(\Sigma)}\sqrt{\frac{\log p}{n}}\right)$$

*In addition, the Trim transform (4.2) with $\tau = d_{\lfloor tn\rfloor}$, where $t\in(0,1)$ is an arbitrary constant, satisfies the conditions **(B1)** and **(B2)**. Other examples of such spectral transformations are discussed below.*

**Proof.** Let us first show that if $\tilde{D}$ satisfies assumptions **(B1)** and **(B2)**, we have

$$\|\hat{\beta}-\beta\|_1 = \mathcal{O}_p\left(\frac{\sigma s}{\lambda_{\min}(\Sigma)}\sqrt{\frac{\log p}{n}}\right)$$

Since $\tilde{d}_i \leq d_i$, by Theorem 2, we can choose $\lambda$ such that with probability going to 1 exponentially fast in $p$, we have:

$$\|\hat{\beta}-\beta\|_1 \leq C_1\frac{s\sigma}{\frac{1}{n}\tilde{d}^2_{(k)}\phi^2_{M_k}}\sqrt{\frac{\log p}{n}} + C_2\sqrt{s}\frac{\tilde{d}_{(1)}\|V^Tb\|_2}{\tilde{d}_{(k)}\phi_{M_k}}$$

The assumption **(B1)** gives us that $\frac{\tilde{d}_{(1)}}{\tilde{d}_{(k)}} = \mathcal{O}(1)$, whereas the assumption **(A2)** and the fact that $k = \Omega(n)$ give us that $\phi^2_{M_k} = \Omega\left(\frac{n}{p}\right)$. Therefore, the second term is of order

$$C_2\sqrt{s}\frac{\tilde{d}_{(1)}\|V^T b\|_2}{\tilde{d}_{(k)}\phi_{M_k}} = \mathcal{O}\left(\sqrt{\frac{sp}{n}}\|V^T b\|_2\right)$$

The assumption **(B2)** gives us $\tilde{d}^2_{(k)} = \Omega\left(\lambda_{\min}(\Sigma)p\right)$, which together with $\phi^2_{M_k} = \Omega\left(\frac{n}{p}\right)$ gives us that the first term is of order

$$C_1\frac{s\sigma}{\frac{1}{n}\tilde{d}^2_{(k)}\phi^2_{M_k}}\sqrt{\frac{\log p}{n}} = \mathcal{O}\left(\frac{\sigma s}{\lambda_{\min}(\Sigma)}\sqrt{\frac{\log p}{n}}\right)$$

So if we want that the sum of those two terms is of that order, we need to have

$$\|V^T b\|_2 = \mathcal{O}\left(\frac{\sigma}{\lambda_{\min}(\Sigma)}\sqrt{\frac{s\log p}{p}}\right)$$

which is the assumption (**A1**).

If we need just the consistency, we need both terms to tend to zero. This will be true whenever

$$\|V^T b\|_2 = o\left(\sqrt{\frac{n}{ps}}\right)$$

Let us now assume that the assumption **(A3)** holds. Then for $t \in (0,1)$ we have

$$d^2_{\lfloor tn\rfloor} = \Omega(\lambda_{\min}p)$$

so the Trim transform satisfies the assumption **(B2)** for $k = \lfloor tn\rfloor = \Omega(n)$. Also, the assumption **(B1)** is trivially satisfied since $\tilde{d}_{(1)} = \tilde{d}_{(k)} = d_k$. □

**Lemma 6.** *Let $B \in \mathbb{R}^{p\times p}$ be a symmetric positive definite matrix and let $A \in \mathbb{R}^{n\times p}$ be arbitrary matrix, $n < p$. Assume that the smallest singular value of $B$ is at least 1. Let $\sigma_i(A)$ and $\sigma_i(AB)$ be the i-th largest singular values of $A$ and $AB$ respectively. For $i \le n$, we have*

$$\sigma_i(A) \le \sigma_i(AB)$$

**Proof.** Let $e_1, \ldots, e_n$ i $f_1, \ldots, f_n$ be the left singular vectors of $A$ and $AB$ corresponding to the singular values in a decreasing order. For $i = 1$, since $\sigma_{\min}(B) \ge 1$, we have:

$$\sigma_1(AB) \ge \|(AB)^T e_1\|_2 \ge \|BA^T e_1\|_2 \ge \|A^T e_1\|_2 = \sigma_1(A)$$

Let us proceed by induction. Since $\dim(U \cap V) \geq \dim(U) + \dim(V) - n$, we conclude that $F_k = \text{span}\{f_1, \dots, f_k\}^\perp$ and $\text{span}\{e_1, \dots e_{k+1}\}$ have a non-trivial intersection, so we can choose a unit vector $v = \sum_{j=1}^{k+1} \alpha_j e_j \in F_k$. Since $\sigma_{k+1}(AB) = \max\{(AB)^T x : x \in F_k, \|x\|_2 = 1\}$, we have:

$$\sigma_{k+1}(AB) \geq \|(AB)^T v\|_2 \geq \|A^T v\|_2 = \sqrt{\sum_{j=1}^{k+1} \alpha_j^2 \sigma_j(A)^2} \geq \sigma_{k+1}(A)$$

The second inequality holds because $\sigma_{\min}(B) \geq 1$ and the last because $\sum \alpha_j^2 = 1$ and $\sigma_i(A)$ are decreasing. □

**Lemma 4.** *Assume that $X$ is a random design matrix with rows being drawn i.i.d. from the $N_p(0, \Sigma)$ distribution. If $\limsup \frac{k}{n} < 1$ or if $\liminf \frac{p}{n} > 1$, we have*

$$d_k^2 = \Omega_p(\lambda_{\min}(\Sigma) p).$$

**Proof.** Let $Z \in \mathbb{R}^{n \times p}$ be a matrix whose elements are independent standard normal variables: $Z_{ij} \overset{i.i.d.}{\sim} N(0,1)$. Let $\zeta_1 \leq \ldots \leq \zeta_n$ be the singular values of $Z$.

Since we can write $X = Z\Sigma^{1/2}$, by Lemma 6, we have $d_k \geq \lambda_{\min}(\Sigma)^{1/2} \zeta_k$, so it suffices to show that $\zeta_k = \Omega_p(p^{1/2})$.

From [25], we have that

$$\zeta_n \geq \sqrt{p} - \sqrt{n} - C\sqrt{\log p}$$

with probability at least $1 - 2p^{-C^2/2}$, so if $\liminf \frac{p}{n} > 1$, we have $\zeta_k \geq \zeta_n = \Omega_p(p^{1/2})$.

In the other case, we can assume $\frac{p}{n} \to 1$ (we know $\liminf \frac{p}{n} = 1$, but because of the bound above, we can w.l.o.g. consider only the subsequence converging to 1). The empirical distribution of the nonzero singular values of $\frac{1}{n} Z^T Z$ converges to the Marchenko-Pastur density supported on $[0, 4]$ [18], which is given by

$$\frac{1}{2\pi} \sqrt{\frac{4-x}{x}} \mathbb{1}\{x \in [0,4]\}$$

Let $t = \limsup \frac{k}{n} < 1$. Then we can choose $0 < \delta < 1 - t$ and $z > 0$ such that $\mathbb{P}(\zeta > z) = t + \delta$, where $\zeta$ is drawn from the Marchenko-Pastur density given above.

We have

$$\frac{\#\text{ singular values of } \frac{Z^T Z}{n} \text{ larger than } z}{n} \to t + \delta$$

so the number of singular values of $\frac{Z^TZ}{n}$ which are larger than $z$ will eventually be larger than $nt > k$, therefore $\frac{d_k^2}{n} > z > 0$ eventually, so $d_k = \Omega_p(n^{1/2}) = \Omega_p(p^{1/2})$, as required. □

**Lemma 3.** *If $V$ has a uniform distribution on the Stiefel manifold independently of $D$, then for any $k = \Omega(n)$, we have*

$$\phi^2_{M_k} = \Omega_p\left(\frac{n}{p}\right).$$

**Proof.** Let $Z \in \mathbb{R}^{k\times p}$ be a random matrix whose components are $Z_{ij} \overset{i.i.d.}{\sim} N(0,1)$. Let $Z = U_Z D_Z V_Z^T$ be its SVD and let $\zeta_1 \geq \ldots \geq \zeta_k$ be its singular values.

Since $V$ is independent of $D$, $[V_{(1)}, \ldots, V_{(k)}]$ is uniform on the Stiefel manifold as well. This matrix has the same distribution as the matrix $V_Z$ and thus $M_k$ has same distribution as $V_Z V_Z^T$

From the Lasso theory [4] we know that $\phi^2_{\frac{1}{k}Z^TZ} \geq 1/2$ with high probability. On the other hand we have $\phi^2_{\frac{1}{k}Z^TZ} \leq \frac{1}{k}\zeta_1^2 \phi^2_{V_Z V_Z^T}$.

From Corollary 5.35. of [25] we know

$$\zeta_1 \leq \sqrt{p} + \sqrt{k} + C\sqrt{\log p}$$

with probability at least $1 - 2p^{-C^2/2}$, for any $C > 0$. This implies that $\zeta_1 = \mathcal{O}_p(\sqrt{p})$. By combining those results, we have that $\phi^2_{V_Z V_Z^T} = \Omega_p\left(\frac{k}{p}\right) = \Omega_p\left(\frac{n}{p}\right)$, which finishes the proof. □

**Lemma 7.** *Let $A, B \in \mathbb{R}^{p\times p}$ be two symmetric, positive definite matrices such that $A \succeq B$ and $A, B$ commute. Let $C \in \mathbb{R}^{q\times p}$ be an arbitrary matrix. Then we have for an arbitrary $v \in \mathbb{R}^p$*

$$\|(C^TC + A)^{-1}C^Tv\|_2 \leq \|(C^TC + B)^{-1}C^Tv\|_2$$

**Proof.** Since $A$ and $B$ commute, we have that

$$A^2 - B^2 = (A - B)(A + B)$$

which is positive semidefinite, since it is a product of two positive semidefinite matrices, as $A \succeq B$.

Furthermore, $A(C^TC) \succeq B(C^TC)$ since $(A - B)(C^TC) \succeq 0$ is a product of two positive semidefinite matrices. Analogously, $(C^TC)A \succeq (C^TC)B$.

All this gives us that

$$\begin{aligned}(C^TC+A)^2 &= (C^TC)^2+(C^TC)A+A(C^TC)+A^2\\ &\succeq (C^TC)^2+(C^TC)B+B(C^TC)+B^2=(C^TC+B)^2\end{aligned}$$

which in turn implies that

$$(C^TC+A)^{-2}\preceq(C^TC+A)^{-2}$$

Finally, this gives us:

$$\begin{aligned}\|(C^TC+A)^{-1}C^Tv\|_2^2 &= v^TC(C^TC+A)^{-2}C^Tv\\ &\leq v^TC(C^TC+B)^{-2}C^Tv=\|(C^TC+B)^{-1}C^Tv\|_2^2\end{aligned}$$

which finishes the proof. □

**Lemma 2.** *Assume that the coefficients in the confounding model (3.2) satisfy that $\|\delta\|_2=\mathcal{O}(\sqrt{\log p})$ and $\lambda_{\min}(\Gamma)=\Omega\left(\sqrt{p}\right)$. Assume also that $\lambda_{\min}(\Sigma_E)$ is bounded from below. Then we have:*

$$\|b\|_2=\|(\Gamma^T\Gamma+\Sigma_E)^{-1}\Gamma^T\delta\|_2=\mathcal{O}\left(\sqrt{\frac{\log p}{p}}\right)$$

*The condition $\lambda_{\min}(\Gamma)=\Omega_p(\sqrt{p})$ is satisfied, for example, if $\liminf\frac{p}{q}>1$ and either the rows or columns of $\Gamma$ are independent, identically distributed $N(0,\Omega)$ random variables with $\lambda_{\min}(\Omega)$ bounded away from zero. If the components of $\delta$ are i.i.d. we have $\|\delta\|_2=\mathcal{O}_p(\sqrt{q})$, so we require that the number of latent variables is $q=\mathcal{O}\left(\log p\right)$.*

**Proof.** Assume $q<p$. Let us write $\Gamma=U_\Gamma CV_\Gamma^T$ for the SVD of $\Gamma$ and let $c_1\geq\ldots\geq c_q\geq 0$ be the singular values of $\Gamma$. Assume that the smallest singular value of $\Sigma_E$ is bounded from below by $\sigma_E^2$.

Since $\Sigma_E\succeq\sigma_E^2I$ and $\Sigma_E$ and $\sigma_E^2I$ commute, the Lemma 7 gives us that it suffices to show that

$$\|(\Gamma^T\Gamma+\sigma_E^2I_p)^{-1}\Gamma^T\delta\|_2=\mathcal{O}\left(\sqrt{\frac{\log p}{p}}\right)$$

We can write now

$$\begin{aligned}\|(\Gamma^T\Gamma+\sigma_E^2I_p)^{-1}\Gamma^T\delta\|_2^2 &= \sum_{i=1}^q\frac{c_i^2}{(c_i^2+\sigma_E^2)^2}((U_\Gamma)_i^T\delta)^2\leq\max_i\frac{c_i^2}{(c_i^2+\sigma_E^2)^2}\|\delta\|_2^2\\ &\leq\frac{\|\delta\|_2^2}{\lambda_{\min}(\Gamma)^2}=\mathcal{O}\left(\frac{\log p}{p}\right)\end{aligned}$$

which finishes the proof. From the proof we see that we can also allow to have $c_i \leq \frac{1}{\sqrt{p}}$. Small singular values of $\Gamma$ imply that we can ignore some confounding variables as they can be written as linear combination of others.

If the rows or columns of $\Gamma$ are i.i.d. $N(0, \Omega)$ random variables, we can write $\Gamma$ as $Z\Omega^{1/2}$ or $\Omega^{1/2}Z$ respectively, where $Z \in \mathbb{R}^{q\times p}$ has i.i.d. $N(0,1)$ components. In both cases we have

$$\lambda_{\min}(\Gamma) \geq \lambda_{\min}(\Omega^{1/2})\lambda_{\min}(Z) = \Omega_p(\sqrt{p})$$

since $\lambda_{\min}(\Omega)$ is bounded from below and $\lambda_{\min}(Z) = \Omega_p(\sqrt{p})$ (see Corollary 5.35 from [25]).

$\|\delta\|_2^2 = \mathcal{O}_p(\sqrt{q})$ follows from central limit theorem when the components of $\delta$ are i.i.d. □