Inference for High-Dimensional Instrumental Variables Regression

David A. Gold

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

Master of Science

University of Washington

2017

Committee:

Johannes C. Lederer

Jon A. Wellner

Program Authorized to Offer Degree:
Statistics

University of Washington

Abstract

Inference for High-Dimensional
Instrumental Variables Regression

David A. Gold

Chair of the Supervisory Committee:
Prof. Dr. Johannes C. Lederer
Statistics

This thesis concerns statistical inference for the components of a high-dimensional regression parameter despite possible endogeneity of each regressor. Given a first-stage linear model for the endogenous regressors and a second-stage linear model for the response variable, we develop a novel adaptation of the parametric one-step update to a generic second-stage estimator. We provide high-level conditions under which the scaled update is asymptotically normal. We introduce a two-stage Lasso procedure and show that, under a Gaussian noise regime, the second-stage Lasso estimator satisfies the aforementioned conditions. Using these results, we construct asympotitically valid confidence intervals for the components of the second-stage regression vector. We complement our asymptotic theory with empirical studies, which demonstrate the relevance of our method in finite samples.

LIST OF TABLES

Table Number Page

TABLE OF CONTENTS

ACKNOWLEDGMENTS

The time I have spent writing the present work has been one of immense academic and personal growth. There are too many people I must thank for this. I express my gratitude and apologies to those whom I have inevitably omitted but deserve to be mentioned. For the rest, I will try to be brief.

I am especially grateful to Johannes Lederer and Jing Tao for their technical insight, professional guidance, and belief and trust in my potential as a researcher. I have learned so much from them, and would not have been able to produce this thesis without their support and encouragement.

I would like to thank Jon Wellner, who graciously gives me his time and energy even though I never make appointments.

I would like to express my gratitude to the wonderful administrative staff of the Department of Statistics. I gratefully acknowledge financial support from the University of Washington's Royalty Research Fund and from Amazon Cloud Credits for Research.

I have gotten by with more than a little help from my friends — in particular Sean Jewell and Bryan Martin in the Department of Statistics and the beautiful humans of the dance, movement, and arts communities of Cascadia. You all know who you are.

I owe so much to Mo for her guidance and tireless compassion. I am grateful beyond words to Maya for the magic she brings to my life.

My grandparents, parents, and siblings continue to support me with their unconditional love. I will never come close to repaying them.

DEDICATION

to my loving grandparents

PREFACE

This thesis is a version of a manuscript available at https://arxiv.org/abs/1708.05499 and for which Johannes Lederer, Jing Tao and David Gold designed research and numerical studies. The latter work is simultaneously being prepared for publication as of writing this preface.

Chapter 1

INTRODUCTION

1.1 Overview

The problem of estimating a high-dimensional regression parameter appears ubiquitously in the data-intensive sciences and has been extensively studied [17, 27, 30]. Statistical procedures for quantifying the uncertainty of such estimates are, however, much less developed. In particular, although considerable progress has been made for inference in standard high-dimensional linear regression [32, 53, 58], much less is known for more complex models.

The main concern of this paper is to extend the aforementioned developments to the *linear instrumental variables (IV) model*. To motivate the latter, we consider the ordinary linear model

$$\boldsymbol{y} \;=\; \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}\,,$$

where $\boldsymbol{y} \in \mathbb{R}^n$ is the vector-valued response, $\boldsymbol{X} \in \mathbb{R}^{n\times p_{\boldsymbol{x}}}$ is the matrix of regressors with rows $\boldsymbol{x}_i \in \mathbb{R}^{p_{\boldsymbol{x}}}$, $\boldsymbol{\beta} \in \mathbb{R}^{p_{\boldsymbol{x}}}$ is the regression vector, and $\boldsymbol{u} \in \mathbb{R}^n$ is the noise vector with components u_i.

It is well-known that standard inference for $\boldsymbol{\beta}$ using least-squares estimation is invalid if $\mathrm{E}[u_i\boldsymbol{x}_i] \neq \mathbf{0}$. Unfortunately, such is the norm in many practical contexts. Selection biases, omitted variables, measurement error, and the interplay of mutually interdependent process that each exhibit random variation can each lead to the failure of the moment condition $\mathrm{E}[u_i\boldsymbol{x}_i] = 0$. In such cases, it is preferable to use inferential procedures that allow for the failure of this moment condition. The method of instrumental variables accomplishes as much by assuming that the observations $\boldsymbol{x}_i$ can in turn be modeled in terms of observable *instrumental variables* $\boldsymbol{z}_i \in \mathbb{R}^{p_{\boldsymbol{z}}}$ that satisfy $\mathrm{E}[u_i \,|\, \boldsymbol{z}_i] = 0$:

$$\boldsymbol{x}_i \;=\; \mathbf{A}^\top \boldsymbol{z}_i + \boldsymbol{v}_i\,,$$

where $\mathbf{A} \in \mathbb{R}^{p_{\boldsymbol{z}} \times p_{\boldsymbol{x}}}$ is a matrix of regression coefficients and $\boldsymbol{v}_i \in \mathbb{R}^{p_{\boldsymbol{x}}}$ is a noise vector whose components may have non-trivial correlation with u_i. Following the econometric literature, we call such regressors $\boldsymbol{x}_i$ *endogenous* and the instrumental variables $\boldsymbol{z}_i$ *exogenous* [49].

Inference for such models in the low-dimensional setting, that is in which the number of samples exceeds both the number of regressors in $\boldsymbol{X}$ and the number of regressors in $\boldsymbol{Z}$, has been extensively studied and put to use in many economic applications [5]. However, situations in which either the number of instrumental variables or both the number of instrumental variables and endogenous variables exceed the number of available observations are not uncommon. Apart from the situation in which the analyst collects data on a large number of initial covariates, relevant scenarios include those in which the endogenous variables are factors with many levels and in which the analyst wishes to expand a small number of variables into a large series of basis functions. We are therefore interested in inference for doubly high-dimensional settings in which the number of samples is dominated by both the number of endogenous regressors and the number of instrumental variables.

To develop a method for conducting inference in such settings, we follow an approach similar to those of [32, 53, 58], who "de-bias" the Lasso to obtain asymptotic pivots for the low-dimensional components of the high-dimensional regression vector $\boldsymbol{\beta}$ when endogeneity is absent. The de-biased estimator decomposes into a main term linear in the noise and a remainder term that, under certain growth conditions on the model parameters, is asymptotically negligible. We formulate the de-biasing procedure as a generic one-step update, which can be applied to any initial estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$. In parametric models, the one-step update $\tilde{\boldsymbol{\beta}}$ to an initial estimator $\hat{\boldsymbol{\beta}}$ is one Newton-Raphson step in the direction of a solution to the empirical analogue of the score equations. To adapt the one-step update to handle endogeneity of $\boldsymbol{X}$, we (i) choose the update as a step towards the solution to the empirical analogue of a valid moment condition and (ii) apply the update to a generic second-stage estimator $\hat{\boldsymbol{\beta}}$ that depends on the predicted conditional means $\hat{\mathrm{E}}[\boldsymbol{X} \mid \boldsymbol{Z}]$. The resultant estimator decomposes into a main term and four remainder terms. We present high-level conditions under which the updated estimator yields asymptotic pivots for the components of $\boldsymbol{\beta}$ and we show, as

an example, how these conditions may be satisfied by a two-stage Lasso estimation routine. The main challenges of establishing the example are due to the involved structure of the remainder terms, whose control in turn require a variety of probabilistic techniques and lead to extensive proofs.

1.2 Our contributions

Our primary contribution is to develop methods for conducting statistical inference for the low-dimensional components β_j of a high-dimensional regression vector $\boldsymbol{\beta}$ despite endogeneity of the respective regressors. We present a novel adaptation of the one-step update and high-level conditions under which the updated estimator yields asymptotic pivots for the β_j. A related contribution concerns sparse inverse covariance matrix estimation. The updated estimator $\tilde{\boldsymbol{\beta}}$ depends on an estimate of the inverse covariance matrix $\boldsymbol{\Theta}$ of the conditional means $\mathrm{E}[\boldsymbol{x}_i|\boldsymbol{z}_i]$. However, we do not observe these conditional means directly, and must base our estimate of $\boldsymbol{\Theta}$ on the predictions $\hat{\mathrm{E}}[\boldsymbol{x}_i|\boldsymbol{z}_i]$. We use essentially the CLIME estimator $\widehat{\boldsymbol{\Theta}}$ of [19] but must do additional, novel work to account for the prediction step in deriving probabilistic guarantees for the estimator's performance.

A third contribution is to show that the updated second-stage Lasso estimator studied in Chapter 4 satisfies the high-level conditions cited above and therefore supports inference for the $\tilde{\beta}_j$. To show as much, we develop probabilistic bounds for the second-stage ℓ_1 estimation error, and we use these bounds to show asymptotic negligibility of the four remainder terms described in the previous section. We also demonstrate the feasibility of the compatibility condition in the second-stage regression, thereby justifying the practical use of the second-stage rates.

Though our estimator is not a generalized method of moments (GMM) estimator [29], we suspect that efficiency results require prediction of the conditional means of the endogenous variables given the instrumental variables. Much of the present work is devoted to accounting for the prediction error when both the first- and second- stage regression models are high-dimensional. This contrasts with the methods of [26, 43], who do not account for the need

to predict the optimal instruments [1, 2, 31, 42].

Much of the proofs factor nicely into deterministic and stochastic components. A final contribution is to respect this structure in the intermediate proof steps, thereby maximizing the generality of our results. This allows future analysts easily to combine the generic bounds contained in Section 4.4 with concentration results for specific error and design matrix distribution regimes and thereby derive the growth conditions required for good asymptotic behavior of the updated second-stage Lasso estimator under a variety of models.

1.3 Related work

Our work relates to the classical research on inference for instrumental variable models. In particular, the estimator we propose is a high-dimensional generalization to the familiar two-stage least squares (2SLS) estimator for low-dimensional linear regression models with endogenous regressors back from some early work such as [1, 2, 25, 51].

Our work also relates to the more recent research on inference for high-dimensional linear instrumental variables models such as [9, 11, 23, 26]. In particular, [9, 11] use the Lasso to obtain representations of the optimal IVs of [1, 2, 29] for models in which the conditional mean of the response is linear in a small and fixed number of endogenous variables. In comparison, the estimation procedure we study in Chapter 4 uses the Lasso for both the first and second stages of estimation, and we obtain asymptotic normality and confidence intervals of our estimator by a one-step correction to the second stage. Our results therefore complement the analysis of [9, 11] by extending methods of inference to models with many endogenous regressors.

Works that develop inferential methods for cases in which the endogenous variables are high-dimensional include [23] and [26], who propose post-selection and combinatoric finite-sample bounds-based methods, respectively, for conducting inference for low-dimensional components of a high-dimensional regression vector under endogeneity. We notice that [43] also provides an inferential method for high-dimensional linear IV models by using a Dantzig selector. However, their method does not account for the first-stage of estimation; the latter

is required to predict the conditional means $\mathrm{E}[\boldsymbol{x}_i \,|\, \boldsymbol{z}_i]$, which we suspect are required for efficient inference of the components β_j.

1.4 Organization

The rest of this thesis is organized as follows. We introduce and motivate the instrumental variables model and a generic two-stage estimation procedure in Sections 2.1 and 2.2, respectively. In Chapter 3, we present an adaptation of the parametric one-step update to the generic second-stage estimator defined in Section 2.2. We show in Section 3.2 that the updated second-stage estimator $\tilde{\boldsymbol{\beta}}$ decomposes into (i) a main term linear in the noise $\boldsymbol{u}$ and (ii) four remainder terms. The one-step update requires an estimate of the population precision matrix of the conditional means $\mathrm{E}[\boldsymbol{x}_i \,|\, \boldsymbol{z}_i]$; we discuss an appropriate estimation procedure for this quantity in Section 3.3. Our main result concerns the asymptotic distribution of $\sqrt{n}(\tilde{\beta}_j - \beta_j)/\omega_j$, where ω_j is an appropriate scale factor, under the high-level assumption that the remainder terms are asymptotically negligible. We present this result in Section 3.4. This result allows us to construct asymptotically valid confidence intervals for the individual components of the second-stage regression parameter $\boldsymbol{\beta}$.

In Chapter 4, we introduce a two-stage Lasso estimator of the regression parameter $\boldsymbol{\beta}$ and show under a Gaussian noise regime that it is suitable for use with the one-step update developed in Chapter 3. To this end, we provide finite-sample bounds for the estimation error of the first- and second-stage Lasso estimators in Sections 4.2.3 and 4.2.4, respectively. To demonstrate the feasibility of the conditions required for the latter bounds, we present an analysis of the compatibility condition in the context of the second-stage estimation in Section 4.2.5. In Section 4.3, we show that the remainder terms are asymptotically negligible under the two-stage Lasso estimation routine.

Finally, in Chapter 5, we present the results of numerical studies that demonstrate the relevance of our theoretical results to finite samples. All proofs are contained in the Supplementary Materials.

As we discuss in Chapter 4, the finite-sample bounds for ℓ_1-regularized estimators factor

into deterministic and stochastic parts. The deterministic parts yield bounds for both the errors of the estimators presented in Section 4.1 and the remainder terms introduced in Section 3.1 under the two-stage Lasso. Such bounds are generic over various types of noise regimes. From these generic bounds, we derive specific results that concern the Gaussian noise regime introduced in Section 2.1. We present the specific estimation error bounds in Sections 4.2.3 and 4.2.4 and the generic bounds in Section 4.4; all results concerning the remainder terms are contained in Section 4.4. Furthermore, a number of results depend on probablistic statements concerning properties of various design matrices. We present such statements as generally as possible in the relevant results; specific treatments are given as examples, such as Examples 3.3.3, 3.4.2, and 4.3.4.

1.5 Basic notation and preliminaries

We adopt the following general notational conventions. For $p \in \mathbb{N}$, we let $[p] := \{1, \dots, p\}$. We let bold and non-bold lowercase letters denote vectors and scalars, respectively; we use bold uppercase letters to denote matrices. We typically denote the components of a vector (matrix) by the non-bold (lowercase) counterpart of the letter that denotes the vector (matrix). If $\boldsymbol{M} = (m_{ij})_{i,j \in [n] \times [p]}$, we use a superscript to refer to columns $\boldsymbol{m}^j = (m_{ij})_{i \in [n]}$ and a subscript to refer to rows $\boldsymbol{m}_i = (m_{ij})_{j \in [p]}$. We let $\| \cdot \|_q$ and $\langle \cdot, \cdot \rangle$ denote the usual ℓ_q norm and inner product over Euclidean spaces, respectively;

For $\boldsymbol{m} \in \mathbb{R}^p$, we let $\operatorname{supp}(\boldsymbol{m}) := \{j \in [p] : m_j \neq 0\}$ and $\|\boldsymbol{m}\|_0 = |\operatorname{supp}(\boldsymbol{m})|$; we let $\|\boldsymbol{m}\|_\infty = \max_{j \in [p]} |m_j|$. For matrices $\boldsymbol{M} \in \mathbb{R}^{n \times p}$, we let $\|\boldsymbol{M}\|_\infty = \max_{i,j \in [n] \times [p]} |m_{ij}|$ and $\|\boldsymbol{M}\|_{L_1} = \max_{j \in [p]} \|\boldsymbol{m}^j\|_1$. For matrices $\boldsymbol{M}_1, \boldsymbol{M}_2 \in \mathbb{R}^{n \times p}$, we write $\boldsymbol{M}_1 \succ \boldsymbol{M}_2$ if $\boldsymbol{M}_1 - \boldsymbol{M}_2$ is positive-definite.

If x is a (random or deterministic) quantity indexed by $i \in [n]$, we let $\mathbb{E}_n[x_i] = n^{-1} \sum_{i=1}^n x_i$. If X_n is a sequence of random variables, we write $X_n \rightsquigarrow X$ if X_n converges weakly to X. For $a, b \in \mathbb{R}$, we let $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. We write $a_n \lesssim b_n$ if $a_n \leq C_n b_n$ for a C_n that is of constant order. We say that a sequence of events $\mathcal{E} \equiv \mathcal{E}_n$ occurs with probability approaching one if $\lim_{n \to \infty} \mathrm{P}\, \mathcal{E} = 1$.

We recall the following definitions of the sub-Gaussian and sub-exponential norms.

Definition 1.5.1 (Sub-Gaussian and sub-exponential norms)**.** For $q \geq 1$ and a random variable X, we write

$$\|X\|_{\psi_q} := \inf\{t \in (0, \infty) : \mathrm{E}[\exp(|X|^q/t^q) - 1] \leq 1\}$$

if the infimum exists. The *sub-Gaussian norm* of a random variable X is given by $\|X\|_{\psi_2}$; the *sub-exponential norm* of a random variable X is given by $\|X\|_{\psi_1}$. The corresponding norms for a random p-vector $\boldsymbol{X}$ are given by

$$\|\boldsymbol{X}\|_{\psi_q} := \sup_{\boldsymbol{x} \in \mathbb{R}^p \,:\, \|\boldsymbol{x}\|_2 = 1} \|\langle \boldsymbol{X}, \boldsymbol{x} \rangle\|_{\psi_q} .$$

Chapter 2

TWO-STAGE ESTIMATION

To contend with endogeneity, the method of instrumental variables isolates variation in the endogenous regressors induced by the instrumental variables. In Section 2.1, we posit the two-stage linear IV model to describe such relationships. In Section 2.2, we discuss a generic two-stage estimation routine that respects the structure of the model.

2.1 Two-stage model

Our model of interest is

$$y_i \;=\; \boldsymbol{x}_i^\top \boldsymbol{\beta} + u_i \,, \tag{2.1}$$

$$x_{ij} \;=\; \boldsymbol{z}_i^\top \boldsymbol{\alpha}^j + v_{ij} \,, \tag{2.2}$$

where: i ranges from 1 to n (unless stated otherwise); j ranges from 1 to $p_{\boldsymbol{x}}$ (unless stated otherwise); the vectors $\boldsymbol{x}_i \in \mathbb{R}^{p_{\boldsymbol{x}}}$ consist of the *second-stage regressors* $x_{i1}, \ldots, x_{ip_{\boldsymbol{x}}}$; the vector $\boldsymbol{\beta} \in \mathbb{R}^{p_{\boldsymbol{x}}}$ is the parameter of interest; the vectors $\boldsymbol{z}_i \in \mathbb{R}^{p_{\boldsymbol{z}}}$ consist of the *first-stage regressors* $z_{i1}, \ldots, z_{ip_{\boldsymbol{z}}}$; the quantities u_i and $\boldsymbol{v}_i := (v_{i1}, \ldots, v_{ip_{\boldsymbol{x}}})^\top$ are random noise elements that satisfy

$$\mathrm{E}[u_i|\boldsymbol{z}_i] = 0 \,, \qquad \mathrm{E}[\boldsymbol{v}_i|\boldsymbol{z}_i] = \boldsymbol{0} \,; \tag{2.3}$$

and the vectors $\boldsymbol{\alpha}^j$ are regression parameters up to which the respective conditional means $d_{ij} := \mathrm{E}[x_{ij}|\boldsymbol{z}_i] = \boldsymbol{z}_i^\top \boldsymbol{\alpha}^j$ are specified. In matrix notation, we write

$$\boldsymbol{y} \;=\; \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$$

and

$$\boldsymbol{X} \;=\; \boldsymbol{D} + \boldsymbol{V} \;=\; \boldsymbol{Z}\mathbf{A} + \boldsymbol{V} \,,$$

where: the vectors $\boldsymbol{y}, \boldsymbol{u} \in \mathbb{R}^n$ consist of the responses y_i and the noise components u_i, respectively; the matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p_{\boldsymbol{x}}}$ has columns $\boldsymbol{x}^j$ given by $\boldsymbol{x}^j = (x_{1j}, \ldots, x_{nj})^\top$; the matrix $\boldsymbol{D} = \mathrm{E}[\boldsymbol{X}|\boldsymbol{Z}] \in \mathbb{R}^{n \times p_{\boldsymbol{x}}}$ has columns $\boldsymbol{d}^j$ given by $\boldsymbol{d}^j = (d_{1j}, \ldots, d_{nj})^\top$; the matrix $\boldsymbol{Z} \in \mathbb{R}^{n \times p_{\boldsymbol{z}}}$ has columns $\boldsymbol{z}^k$ given by $\boldsymbol{z}^k = (z_{1k}, \ldots, z_{nk})^\top$; and the matrix $\mathbf{A} \in \mathbb{R}^{p_{\boldsymbol{z}} \times p_{\boldsymbol{x}}}$ has columns given by $\boldsymbol{\alpha}^j$. We make the following assumption concerning the n-indexed sequence of regression parameters $\mathbf{A}, \boldsymbol{\beta}$.

Assumption 2.1.1 (Regularity of $\mathbf{A}, \boldsymbol{\beta}$)**.** The quantities $\|\mathbf{A}\|_{L_1}$ and $\|\boldsymbol{\beta}\|_1$ are bounded above by universal constants $m_{\mathbf{A}}, m_{\boldsymbol{\beta}} < \infty$, respectively.

We let $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}} = \boldsymbol{Z}^\top \boldsymbol{Z}/n$ denote the empirical Gram matrix of the instrumental variables.

Remark 2.1.2 (Mean of $\boldsymbol{z}_i$)*.* We require that the first-stage regressors $\boldsymbol{z}_i$ have mean zero in order to simplify the following exposition and to apply concentration results under more specific distributional assumptions, such as in Example 3.3.3. This assumption can be relaxed at the expense of brevity and given a sufficient reformulation of the required concentration results.

We call the models of (2.2) and (2.1) the *first-stage* and *second-stage models*. In general, the second-stage regression parameter $\boldsymbol{\beta}$ is the target of inference, and the first-stage model encodes additional information. We are primarily concerned with the case in which the second-stage regressors are endogenous — that is, when $\mathrm{E}[\boldsymbol{u}|\boldsymbol{X}] \neq \mathbf{0}$. In this case, the model described above is the linear instrumental variables model [3]. In the sequel, we refer to the first- and second-stage regressors as instrumental and endogenous variables, respectively. We note that the results of Sections 3 and 4 continue to hold if the $\boldsymbol{x}^j$ are exogenous, though this setting is not our focus.

As remarked in the Introduction, the linear IV model has been studied extensively in the low-dimensional setting, where the number $p_{\boldsymbol{x}}$ of endogenous variables $\boldsymbol{x}^j$ is fixed. We are particularly concerned with the high-dimensional regime in which both $p_{\boldsymbol{x}}$ and the number $p_{\boldsymbol{z}}$ of instrumental variables $\boldsymbol{z}^k$ increase with n. Our results generalize to the low-dimensional case in which $p_{\boldsymbol{z}}, p_{\boldsymbol{x}}$ are held fixed with respect to n, but we do not treat this case explicitly

in the present essay. Regardless of whether the model is high-dimensional, we require that $p_{\boldsymbol{x}} \leq p_{\boldsymbol{z}}$ in order to maintain identifiability of $\mathrm{E}[\boldsymbol{x}_i|\boldsymbol{z}_i]$.

We include distributional specifications for the noise vectors as a separate assumption, which we present below.

Assumption 2.1.3 (Homoscedastic Gaussian noise regime)**.** The noise elements $u_i, \boldsymbol{v}_i$ satisfy

$$(u_i, \boldsymbol{v}_i)|\boldsymbol{z}_i \sim \mathcal{N}_{1+p_{\boldsymbol{x}}}(\boldsymbol{0}, \boldsymbol{\Sigma}_{u\boldsymbol{v}}), \qquad \boldsymbol{\Sigma}_{u\boldsymbol{v}} := \begin{pmatrix} \sigma_u^2 & \boldsymbol{\sigma}_{u\boldsymbol{v}}^\top \\ \boldsymbol{\sigma}_{u\boldsymbol{v}} & \boldsymbol{\Sigma}_{\boldsymbol{v}} \end{pmatrix},$$

where $\boldsymbol{\sigma}_{u\boldsymbol{v}} := (\sigma_{uv^1}, \ldots, \sigma_{uv^{p_{\boldsymbol{x}}}})^\top$ consists of the noise covariances $\sigma_{uv^j} := \mathrm{cov}(\boldsymbol{u}, \boldsymbol{v}^j)$, and where $\boldsymbol{\Sigma}_{\boldsymbol{v}}$ is an unstructured covariance matrix with diagonal entries $\sigma_{v^j}^2 := \mathrm{var}(\boldsymbol{v}^j)$ for $j \in [p_{\boldsymbol{x}}]$. Further, considered as components of an n-indexed sequence of models, the variances $\sigma_u^2 \equiv \sigma_{u,n}^2$ and $\sigma_{v^j}^2 \equiv \sigma_{v^j,n}^2$ are bounded strictly away from zero and infinity for $j \in [p_{\boldsymbol{x}}]$.

2.2 Generic two-stage estimators

Our proposed method of inference for the components β_j of the second-stage regression parameter $\boldsymbol{\beta}$ involves estimators that reflect the structure of the model described above. The method itself can be described in terms of generic estimators given as functions of the data. We now introduce notation for such generic estimators that will be used in Chapter 3.

For each $j \in [p_{\boldsymbol{x}}]$, let $\hat{\boldsymbol{\alpha}}^j \equiv \hat{\boldsymbol{\alpha}}^j(\boldsymbol{x}^j, \boldsymbol{Z})$ denote a generic *first-stage estimator* of the first-stage regression vector $\boldsymbol{\alpha}^j$ based on the data $\boldsymbol{x}^j$ and $\boldsymbol{Z}$. We write $\widehat{\mathbf{A}} := (\hat{\boldsymbol{\alpha}}^1, \ldots, \hat{\boldsymbol{\alpha}}^{p_{\boldsymbol{x}}})$ for the matrix of estimated regression vectors. From such an estimator $\widehat{\mathbf{A}}$ we may predict the conditional means $\boldsymbol{d}_i = \mathrm{E}[\boldsymbol{x}_i|\boldsymbol{z}_i]$ for $i \in [n]$ with

$$\hat{\boldsymbol{d}}_i := \boldsymbol{z}_i^\top \widehat{\mathbf{A}};$$

we write $\widehat{\boldsymbol{D}}$ for the predicted conditional mean matrix whose rows are given by the $\hat{\boldsymbol{d}}_i$, and we write $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}} := \widehat{\boldsymbol{D}}^\top \widehat{\boldsymbol{D}}/n$. Our choice of the notation $\widehat{\boldsymbol{D}}$ reflects the fact that this quantity predicts and, under certain conditions, approaches in probability the conditional mean matrix $\boldsymbol{D}$; it does not approach the endogenous design matrix $\boldsymbol{X}$. We write $\hat{\boldsymbol{\beta}} \equiv \hat{\boldsymbol{\beta}}(\boldsymbol{y}, \widehat{\boldsymbol{D}})$ for

a generic *second-stage estimator* of the second-stage regression parameter $\boldsymbol{\beta}$ based on the response $\boldsymbol{y}$ and the predicted conditional means $\widehat{\boldsymbol{D}}$.

The present estimation scheme is similar to that of the two-stage least-squares (2SLS) estimator developed by [8, 9, 34, 46, 51] and studied in connection with the limited information maximum likelihood (LIML) estimator by [4]. In Chapter 3, we propose a method for conducting inference for the components β_j in the high-dimensional setting in which $p_{\boldsymbol{x}} > n$. We formulate the method in terms of the generic two-stage estimators introduced in the present section. In Chapter 4, we carry out the implementation of the method for ℓ_1-regularized two-stage estimators.

Chapter 3

A ONE-STEP UPDATE UNDER ENDOGENEITY

Our main contribution is to develop a method for statistical inference for the components β_j of the second-stage regression vector $\boldsymbol{\beta}$. In general, statistical inference for high-dimensional regression parameters is a difficult problem. Regularized estimators, such as the Lasso and ridge regression, are often used for the purpose of high-dimensional parameter estimation but generally do not have asymptotic distributions suitable for inference; see [35, 44]. In studying the model of Section 2.1, we must also account for the dependence of the second-stage estimator on the first-stage estimators.

The basis for our procedure is to adapt the parametric one-step update to the two-stage estimation procedure described in Section 2.2. In Section 3.1, we review the use of the one-step estimator in parametric models and its application to high-dimensional inference for the ordinary linear model. In Section 3.2, we adapt the one-step update to the two-stage estimation procedure described in Section 2.2. We describe an procedure to estimate the inverse of $\boldsymbol{\Sigma_d} = \mathrm{E}[\boldsymbol{d}_i \boldsymbol{d}_i^\top]$ in Section 3.3. Section 3.4 discusses high-level conditions under which the scaled updated estimator is asymptotically normal.

3.1 One-step update

The one-step update is a general method for constructing efficient estimators for parameters in parametric and semiparametric models [14, Sections 2.5, 7.3]. For our purposes, we review only the use in parametric models.

Recall that the Newton-Raphson method for finding the root in $\boldsymbol{b}$ to a *target system* of $p_{\boldsymbol{x}}$ equations

$$\boldsymbol{h}(y_i, \boldsymbol{x}_i; \boldsymbol{b}) \equiv (h_1(y_i, \boldsymbol{x}_i; \boldsymbol{b}), \ldots, h_{p_{\boldsymbol{x}}}(y_i, \boldsymbol{x}_i; \boldsymbol{b}))^\top = \mathbf{0}$$

is to update an approximation $\boldsymbol{b}^k$ by the rule

$$\boldsymbol{b}^{k+1} \;=\; \boldsymbol{b}^k - \left[\left.\frac{\partial \boldsymbol{h}}{\partial \boldsymbol{b}}\right|_{\boldsymbol{b}=\boldsymbol{b}^k}\right]^{-1} \boldsymbol{h}(y_i, \boldsymbol{x}_i; \boldsymbol{b}^k)\,,$$

where $\left.\frac{\partial \boldsymbol{h}}{\partial \boldsymbol{b}}\right|_{\boldsymbol{b}=\boldsymbol{b}^k}$ is the Jacobian matrix of $\boldsymbol{h}$ with respect to $\boldsymbol{b}$ evaluated at $\boldsymbol{b}^k$. In the ordinary Gaussian linear model

$$\boldsymbol{y} \;=\; \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}\,, \tag{3.1}$$

the score function

$$\boldsymbol{h}(y_i, \boldsymbol{x}_i; \boldsymbol{b}) \;=\; -\boldsymbol{x}_i(y_i - \boldsymbol{x}_i^\top \boldsymbol{b})$$

satisfies $\mathrm{E}[\boldsymbol{h}(y_i, \boldsymbol{x}_i; \boldsymbol{\beta})] = \boldsymbol{0}$ given the *orthogonality condition*

$$\mathrm{E}[\boldsymbol{x}_i u_i] \;=\; \boldsymbol{0}\,. \tag{3.2}$$

The *one-step update* $\tilde{\boldsymbol{\beta}}$ to an initial estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is given by

$$\tilde{\boldsymbol{\beta}} \;=\; \hat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\Theta}}\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})/n\,,$$

where $\widehat{\boldsymbol{\Theta}}$ denotes the inverse of

$$\left.\frac{\partial(-\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})/n)}{\partial \boldsymbol{b}}\right|_{\boldsymbol{b}=\hat{\boldsymbol{\beta}}} \;=\; \boldsymbol{X}^\top\boldsymbol{X}/n \;=\; \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\,.$$

The estimator is so-named because it is one Newton-Raphson step in the direction of the solution in $\boldsymbol{b}$ to the empirical analogue

$$\mathbb{E}_n[\boldsymbol{h}(y_i, \boldsymbol{x}_i; \boldsymbol{b})] \;=\; -\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})/n \;=\; \boldsymbol{0}\,.$$

of the score equation. In general, for parametric models fixed in n and under some regularity conditions, the one-step update in which the target system consists of the score equations yields an efficient estimator from a $\sqrt{n}$-consistent estimator [14, Sections 2.5]. Indeed, note that if $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}$ is invertible, as is generally assumed in the low-dimensional setting, then $\tilde{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$, which is just the maximum likelihood estimator for the Gaussian linear model.

The case when the model is high-dimensional is less well studied. When $p_{\boldsymbol{x}} > n$, the empirical covariance matrix $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}$ is not invertible. Letting $\widehat{\boldsymbol{\Theta}}$ instead denote an approximate inverse of the Jacobian matrix, one writes

$$\begin{aligned}\tilde{\boldsymbol{\beta}} &= \hat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\Theta}}\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})/n \\ &= \hat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\Theta}}\boldsymbol{X}^\top(\boldsymbol{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + \boldsymbol{u})/n \\ &= \boldsymbol{\beta} + \widehat{\boldsymbol{\Theta}}\boldsymbol{X}^\top\boldsymbol{u}/n + (\widehat{\boldsymbol{\Theta}}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}} - \boldsymbol{I})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\,.\end{aligned}$$

The latter term in the above display is the "remainder" after incomplete inversion of $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}$. Thus, in the high-dimensional one-stage linear model, the one-step update satisfies

$$\sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \widehat{\boldsymbol{\Theta}}\boldsymbol{X}^\top\boldsymbol{u}/\sqrt{n} + \underbrace{\sqrt{n}(\widehat{\boldsymbol{\Theta}}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}} - \boldsymbol{I})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})}_{\boldsymbol{f}}\,.$$

From the above we write

$$\begin{aligned}\sqrt{n}(\tilde{\beta}_j - \beta_j) &= \hat{\boldsymbol{\theta}}_j^\top\boldsymbol{X}^\top\boldsymbol{u}/\sqrt{n} + f_j \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\hat{\boldsymbol{\theta}}_j^\top\boldsymbol{x}_i u_i + f_j\,, \qquad (3.3)\end{aligned}$$

where $\hat{\boldsymbol{\theta}}_j$ is the j^{th} row of $\widehat{\boldsymbol{\Theta}}$. The structure of the main term on the right-hand side above suggests to use $\sqrt{n}(\tilde{\beta}_j - \beta_j)/\hat{\omega}_j$, where $\hat{\omega}_j$ is an appropriate estimate of $\omega_j = (\mathrm{E}[\langle\hat{\boldsymbol{\theta}}_j, \boldsymbol{x}_i\rangle^2 u_i^2])^{1/2}$, as an asymptotic pivot for β_j.

When the initial estimator $\hat{\boldsymbol{\beta}}$ is the Lasso, the updated estimator $\tilde{\boldsymbol{\beta}}$ is sometimes called the desparsified [53] or debiased [32] Lasso, though these authors obtain the form of $\tilde{\boldsymbol{\beta}}$ by means other than the one-step update. The general upshot of their results is that if $\|\boldsymbol{f}\|_\infty = o_{\mathrm{P}}(1)$, and if $\hat{\boldsymbol{\theta}}_j$ and $\boldsymbol{x}_i$ are independent of u_i, then the updated Lasso estimator satisfies

$$\sqrt{n}(\tilde{\beta}_j - \beta_j)/\hat{\omega}_j \rightsquigarrow Z_j \sim \mathcal{N}(0,1)\,,$$

where $\hat{\omega}_j$ is an appropriate estimate of ω_j. A key requirement of the latter two works is the control of the quantity $\|\widehat{\boldsymbol{\Theta}}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}} - \boldsymbol{I}\|_\infty$. The authors combine such bounds with ℓ_1 rates for the Lasso estimator to control $\|\boldsymbol{f}\|_\infty$. Thus, appropriate selection of $\widehat{\boldsymbol{\Theta}}$ is required to obtain a desirable weak limit of $\sqrt{n}(\tilde{\beta}_j - \beta_j)/\hat{\omega}_j$.

3.2 One-step with endogeneity

In this section, we develop a novel adaptation of the one-step update that, under suitable high-level conditions, yields asymptotic pivots for the second-stage components β_j of the two-stage model described in Section 2.1. We note that the present development is valid for any initial second-stage estimator $\hat{\boldsymbol{\beta}}$; demonstrating that the aforementioned high-level conditions are satisfied requires consideration of particular estimators.

Our motivation for the update to a generic second-stage estimator is largely informal and proceeds as follows. Recall that the one-step update of the parametric model is one Newton-Raphson step in the direction of a solution to the respective score equations. If the noise elements u_i are not Gaussian, then the quantity $\boldsymbol{h}(y_i, \boldsymbol{x}_i; \boldsymbol{b}) = -\boldsymbol{x}_i(y_i - \boldsymbol{x}_i^\top \boldsymbol{b})$ is not a score per se but still satisfies $\mathrm{E}[\boldsymbol{h}(y_i, \boldsymbol{x}_i; \boldsymbol{\beta})] = \mathbf{0}$ as a consequence of the orthogonality condition (3.2). However, in the case of the presently considered model, the condition in (3.2) does not hold. Instead, the conditional moment restriction $\mathrm{E}[u_i|\boldsymbol{z}_i] = 0$ in (2.3) entails the orthogonality condition $\mathrm{E}[\boldsymbol{z}_i u_i] = \mathbf{0}$ for the instrumental variables, and in turn that $\mathrm{E}[\boldsymbol{d}_i u_i] = \mathbf{0}$ for the conditional means $\boldsymbol{d}_i$. This suggests that, to develop a one-step update for a generic second-stage estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ of the present model, we ought to take the empirical analogue

$$\mathbb{E}_n[-\hat{\boldsymbol{d}}_i(y_i - \boldsymbol{x}_i^\top \boldsymbol{b})] \;=:\; \mathbb{E}_n[\tilde{\boldsymbol{h}}(y_i, \boldsymbol{x}_i, \hat{\boldsymbol{d}}_i; \boldsymbol{b})] = -\widehat{\boldsymbol{D}}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})/n \;=\; \mathbf{0}\,,$$

of $\mathrm{E}[\boldsymbol{d}_i u_i] = \mathbf{0}$ as the target system for which the root is sought via a Newton-Raphson update. We have elected to base the target system on the moment condition $\mathrm{E}[\boldsymbol{d}_i u_i] = \mathbf{0}$ in accordance with optimal weighting regimes for Generalized Method of Moments (GMM) estimators; see [1, 2, 29, 42]. Further, since the $\boldsymbol{d}_i$ are generally unavailable, we instead use the predicted conditional mean matrix $\widehat{\boldsymbol{D}}$ in the target system above. The one-step update $\tilde{\boldsymbol{\beta}}$ to a second-stage estimator $\hat{\boldsymbol{\beta}}$ is then given by

$$\tilde{\boldsymbol{\beta}} \;=\; \hat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\Theta}}\mathbb{E}_n[\tilde{\boldsymbol{h}}(y_i, \boldsymbol{x}_i, \hat{\boldsymbol{d}}_i; \boldsymbol{b})] \;=\; \hat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{D}}^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})/n\,, \tag{3.4}$$

where we continue to let $\widehat{\boldsymbol{\Theta}}$ denote an (approximate) inverse to the Jacobian matrix in $\boldsymbol{b}$ of the score $\tilde{\boldsymbol{h}}(y_i, \boldsymbol{x}_i, \hat{\boldsymbol{d}}_i; \boldsymbol{b})$. The following lemma characterizes a similar decomposition of the

updated estimator $\tilde{\boldsymbol{\beta}}$ as in the one-stage model.

Lemma 3.2.1 (Decomposition of one-step second-stage estimator)**.** *Consider the two-stage linear model described in Section 2.1. Let $\widehat{\boldsymbol{D}}$ be a prediction of the conditional mean matrix $\boldsymbol{D}$ from an estimate $\widehat{\mathbf{A}}$ of the first-stage regression matrix $\mathbf{A}$. Let $\hat{\boldsymbol{\beta}}$ be a second-stage estimator based on the predictions $\widehat{\boldsymbol{D}}$. Let $\widehat{\boldsymbol{\Theta}}$ denote a generic $p_{\boldsymbol{x}} \times p_{\boldsymbol{x}}$ matrix. The one-step second-stage estimator*

$$\tilde{\boldsymbol{\beta}} \,=\, \hat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{D}}^{\top}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})/n$$

satisfies

$$\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta} \,=\, \widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{D}}^{\top}\boldsymbol{u}/n + \widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{D}}^{\top}(\boldsymbol{X} - \widehat{\boldsymbol{D}})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})/n + (\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{I})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\,,$$

where $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}} = \widehat{\boldsymbol{D}}^{\top}\widehat{\boldsymbol{D}}/n$.

If one were to follow strictly the prescription of the Newton-Raphson method for selection of $\widehat{\boldsymbol{\Theta}}$ for the updated second-stage estimator $\tilde{\boldsymbol{\beta}}$, one would select

$$\begin{aligned}\widehat{\boldsymbol{\Theta}} \,\approx\, \left[\frac{\partial \mathbb{E}_n[\tilde{\boldsymbol{h}}(\boldsymbol{b})]}{\partial \boldsymbol{b}}(\hat{\boldsymbol{\beta}})\right]^{-1} \,&=\, \left[\frac{\partial(-\widehat{\boldsymbol{D}}^{\top}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b})/n)}{\partial \boldsymbol{b}}(\hat{\boldsymbol{\beta}})\right]^{-1} \\ &=\, \left[\widehat{\boldsymbol{D}}^{\top}\boldsymbol{X}/n\right]^{-1}.\end{aligned}$$

However, the decomposition obtained in Lemma 3.2.1 suggests that $\widehat{\boldsymbol{\Theta}}$ ought to control, say, the sup-norm of $\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{I}$, and hence aim to invert $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}}$ rather than $\widehat{\boldsymbol{D}}^{\top}\boldsymbol{X}/n$. We emphasize that the one-step formulation, insofar as it follows the Newton-Raphson method, is merely a vehicle for producing an updated estimator $\tilde{\boldsymbol{\beta}}$; in particular, Lemma 3.2.1 is valid regardless of what convergence properties an actual Newton-Raphson algorithm incorporating a specific choice of $\widehat{\boldsymbol{\Theta}}$ may exhibit. We may choose $\widehat{\boldsymbol{\Theta}}$ in whatever manner is most appropriate for achieving our goal, which is to obtain a tractable limiting distribution for $\sqrt{n}(\tilde{\beta}_j - \beta_j)$. That said, the two suggestions for how to choose $\widehat{\boldsymbol{\Theta}}$ may be reconciled somewhat by noting that both $\widehat{\boldsymbol{D}}^{\top}\boldsymbol{Z}/n$ and $\widehat{\boldsymbol{D}}^{\top}\widehat{\boldsymbol{D}}/n$ are equal to the empirical Gram matrix $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}}$ modulo additional terms whose sup-norms can be controlled given the rate $\|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1}$ and appropriate

concentration results for $\|\boldsymbol{Z}^\top \boldsymbol{v}^j/n\|_\infty$. Under a variety of distributions for the $\boldsymbol{z}_i$, such as a sub-Gaussian regime, one finds $\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{\Sigma}_{\boldsymbol{d}}\|_\infty = o_{\mathrm{P}}(1)$ under appropriate growth restrictions on $p_{\boldsymbol{x}}$; see Example 3.4.2.

For our purposes, we consider the matrix $\widehat{\boldsymbol{\Theta}}$ primarily as an estimator of the population quantity $\boldsymbol{\Theta} := \mathrm{E}[\boldsymbol{d}_i \boldsymbol{d}_i^\top]^{-1}$. In particular, we require good behavior of $\widehat{\boldsymbol{\Theta}}$ as such an estimator to derive the asymptotic distribution of $\sqrt{n}(\tilde{\beta}_j - \beta_j)$. Note that the estimator $\tilde{\boldsymbol{\beta}}$ of (3.4) satisfies

$$\begin{aligned}\sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) \;&=\; \widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{D}}^\top \boldsymbol{u}/\sqrt{n} \\ &\quad + \underbrace{\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{D}}^\top(\boldsymbol{X} - \widehat{\boldsymbol{D}})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})/\sqrt{n}}_{\boldsymbol{f}_3} + \underbrace{\sqrt{n}(\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{I})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})}_{\boldsymbol{f}_4}\,.\end{aligned}$$

Our choice of subscripts is motivated by the fact that the term $\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{D}}^\top \boldsymbol{u}/n$ in the display above must be decomposed further in order to handle the non-trivial covariance of $\widehat{\boldsymbol{\Theta}}$ and $\widehat{\boldsymbol{D}}$ with $\boldsymbol{u}$ given $\boldsymbol{Z}$. To this end, we write

$$\begin{aligned}\sqrt{n}(\tilde{\beta}_j - \beta_j) \;&=\; \hat{\boldsymbol{\theta}}_j^\top \widehat{\boldsymbol{D}}^\top \boldsymbol{u}/\sqrt{n} + f_{3,j} + f_{4,j} \\ &= \hat{\boldsymbol{\theta}}_j^\top \boldsymbol{D}^\top \boldsymbol{u}/\sqrt{n} + \hat{\boldsymbol{\theta}}_j^\top(\widehat{\boldsymbol{D}} - \boldsymbol{D})^\top \boldsymbol{u}/\sqrt{n} + f_{3,j} + f_{4,j} \\ &= \boldsymbol{\theta}_j^\top \boldsymbol{D}^\top \boldsymbol{u}/\sqrt{n} \\ &\quad + \underbrace{(\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j)^\top \boldsymbol{D}^\top \boldsymbol{u}/\sqrt{n}}_{f_{1,j}} + \underbrace{\hat{\boldsymbol{\theta}}_j^\top(\widehat{\boldsymbol{D}} - \boldsymbol{D})^\top \boldsymbol{u}/\sqrt{n}}_{f_{2,j}} + f_{3,j} + f_{4,j}\,, \qquad (3.5)\end{aligned}$$

where: (i) $\boldsymbol{\theta}_j, \hat{\boldsymbol{\theta}}_j$ denote the j^{th} rows of $\boldsymbol{\Theta}, \widehat{\boldsymbol{\Theta}}$, respectively, (ii) $f_{\ell,j}$ denotes the j^{th} component of the remainder term $\boldsymbol{f}_\ell$, and (iii) we implicitly let the index ℓ range from 1 to 4 unless noted otherwise. Since $\boldsymbol{\theta}_j$ is deterministic and the conditional distribution of $\boldsymbol{u}$ given $\boldsymbol{D}$ is tractable by assumption, the primary term $\boldsymbol{\theta}_j^\top \boldsymbol{D}^\top \boldsymbol{u}/\sqrt{n}$ on the right-hand side above is suitable for analysis. As in the case of the main term for the ordinary one-step update discussed Section 3.1, this term can be written as

$$\boldsymbol{\theta}_j^\top \boldsymbol{D}^\top \boldsymbol{u}/\sqrt{n} \;=\; \frac{1}{\sqrt{n}} \sum_{i=1}^n \langle \boldsymbol{\theta}_j, \boldsymbol{d}_i \rangle u_i \;=\; \sqrt{n}\mathbb{E}_n[\langle \boldsymbol{\theta}_j, \boldsymbol{d}_i \rangle u_i]\,.$$

This observation is similar to that of [43], who derive a similar asymptotic linearization but do not account for prediction of the conditional means $\boldsymbol{d}_i$. In Section 3.4, we show that the quantity

$$W_{j,n} := \sqrt{n}(\tilde{\beta}_j - \beta_j)/\hat{\omega}_j \,, \tag{3.6}$$

where $\hat{\omega}_j^2$ is an appropriate estimator of

$$\omega_j^2 := \mathrm{E}\big[\mathbb{E}_n[\langle \boldsymbol{\theta}_j, \boldsymbol{d}_i\rangle^2 u_i^2]\big] \,, \tag{3.7}$$

converges weakly to a $\mathcal{N}(0,1)$ random variable under high level conditions on the remainder terms $f_{\ell,j}$. From this result one may construct asymptotically valid confidence intervals for the regression components β_j.

We have described a strategy for inference for the components $\tilde{\beta}_j$. To implement the strategy for a specific choice of first- and second-stage estimators, one must identify the conditions under which the remainder terms $\boldsymbol{f}_\ell$ vanish in probability. We demonstrate such an implementation in Chapter 4. The conditions in turn depend on the properties of the estimator $\widehat{\boldsymbol{\Theta}}$. In the following section, we introduce an estimator suitable for our purposes.

3.3 Estimating $\boldsymbol{\Theta}$

The one-step second-stage estimator $\tilde{\boldsymbol{\beta}}$ depends on an estimator $\widehat{\boldsymbol{\Theta}}$ of $\boldsymbol{\Theta} = \boldsymbol{\Sigma}_{\boldsymbol{d}}^{-1}$, where $\boldsymbol{\Sigma}_{\boldsymbol{d}} := \mathrm{E}[\boldsymbol{d}_i \boldsymbol{d}_i^\top]$ is the population covariance matrix of the conditional means $\boldsymbol{d}_i$. In general, estimating the population precision matrix incurs two main difficulties in the high-dimensional setting. First, the empirical covariance matrix $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}}$ is singular when $p_{\boldsymbol{x}} > n$ and cannot be inverted to produce an estimator of $\boldsymbol{\Theta}$. Second, even if an inverse is available, one cannot naïvely use the continuous mapping theorem to derive asymptotic guarantees if $p_{\boldsymbol{x}} \to \infty$, since the sequence of population covariance matrices $\boldsymbol{\Sigma}_{\boldsymbol{d}} \equiv \boldsymbol{\Sigma}_{\boldsymbol{d},n}$ does not itself have a limit if $p_{\boldsymbol{x}} \to \infty$. In addition to these general difficulties, we must further contend with the fact that the conditional mean matrix $\boldsymbol{D}$ is unknown. Hence any estimator of $\widehat{\boldsymbol{\Theta}}$ will depend on the prediction $\widehat{\boldsymbol{D}}$, and guarantees for such an estimator must account for such dependence.

We use a slight modification of the CLIME estimator of [19] to contend with the challenges described above. The rows $\hat{\boldsymbol{\theta}}_j$ of the estimator $\widehat{\boldsymbol{\Theta}}$ are obtained as solutions to the CLIME program codified below.

Program 3.3.1 (Program for $\hat{\boldsymbol{\theta}}_j$)**.**

$$\begin{aligned} \underset{\boldsymbol{m} \in \mathbb{R}^{p_{\boldsymbol{x}}}}{\text{minimize:}} \quad & Q(\boldsymbol{m}) \;:=\; \|\boldsymbol{m}\|_1 \,, \\ \text{subject to:} \quad & \|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}} \boldsymbol{m} - \boldsymbol{e}_j\|_\infty \;\leq\; \mu \,, \end{aligned}$$

where $\boldsymbol{e}_j$ denotes the j^{th} canonical basis vector in $p_{\boldsymbol{x}}$ dimensions and $\mu > 0$ is a regularization parameter.

The present estimator $\widehat{\boldsymbol{\Theta}}$ differs in only one respect from that of the CLIME estimator of [19]. The latter symmetrize the matrix $\boldsymbol{\Theta}$ with rows obtained as solutions to the aforementioned optimization problem, whereas we use the raw solutions. We omit the symmetrization step for simplicity; the ℓ_∞ and ℓ_1 guarantees that [19] obtain for the estimation error of the CLIME estimator continue to hold. We include the requisite guarantees for the unsymmetrized estimator in the Supplementary Materials.

The present estimator $\widehat{\boldsymbol{\Theta}}$ also differs in an important respect from that of [32]. The latter also obtain an inverse Gram matrix approximation as a solution to a convex program with identical constraints as in Program 3.3.1 but with objective function $Q(\boldsymbol{m}) = \mathbb{E}_n[\langle \boldsymbol{m}, \boldsymbol{x}_i \rangle^2]$. To our knowledge, however, it is currently unknown whether the choice of Q in [32] yields guarantees comparable to those of the CLIME estimator.

The ℓ_∞ bound for $\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j$ may be derived as a direct consequence of the optimality of $\hat{\boldsymbol{\theta}}$ under the assumption that the respective row $\boldsymbol{\theta}_j$ of the population precision matrix is feasible for Program 3.3.1. However, the ℓ_1 bound depends on a further requirement on the population quantity $\boldsymbol{\Theta}$. We express this requirement in the following definition, in which we adopt the nomenclature and notation of [19] in the following definition.

Definition 3.3.2 (Uniformity class)**.** Following [19], we define the *uniformity class* of population precision matrices $\boldsymbol{\Theta} = \boldsymbol{\Sigma}_{\boldsymbol{d}}^{-1}$ relative to the controlled tolerance $q \in [0, 1)$ and the

generalized sparsity level $s_{\boldsymbol{\Theta}} > 0$ by

$$\mathcal{U}(m_{\boldsymbol{\Theta}}, q, s_{\boldsymbol{\Theta}}) := \left\{ \boldsymbol{\Theta} = (\theta_{jk})_{j,k=1}^{p_{\boldsymbol{x}}} \succ \mathbf{0} : \|\boldsymbol{\Theta}\|_{L_1} \leq m_{\boldsymbol{\Theta}}; \max_{j \in [p_{\boldsymbol{x}}]} \sum_{k \in [p_{\boldsymbol{x}}]} |\theta_{jk}|^q \leq s_{\boldsymbol{\Theta}} \right\}. \tag{3.8}$$

In the sequel, we assume as part of high-level regularity conditions that $\boldsymbol{\Theta} \in \mathcal{U}(m_{\boldsymbol{\Theta}}, q, s_{\boldsymbol{\Theta}})$ and that the model parameters $m_{\boldsymbol{\Theta}}$ and $s_{\boldsymbol{\Theta}}$ are well-behaved as functions of n. These parameters appear in the rates for the remainder terms in our analysis of the two-stage Lasso of Chapter 4. For generic results, we also make similar high-level assumptions that the rows $\boldsymbol{\theta}_j$ of the population precision matrix are feasible for Program 3.3.1. We also provide probabilistic guarantees that the latter condition holds under specific model assumptions. For a given tolerance $\mu > 0$, we define the set

$$\mathcal{T}_{\boldsymbol{\Theta}}(\mu) := \left\{ \|\boldsymbol{\Theta}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}} - \boldsymbol{I}\|_{\infty} \leq \mu \right\}. \tag{3.9}$$

Given the event $\mathcal{T}_{\boldsymbol{\Theta}}(\mu)$, the rows $\boldsymbol{\theta}_j$ of $\boldsymbol{\Theta}_j$ are each feasible for the respective Program 3.3.1. Example 3.3.3, which is adapted from [32] and [57], demonstrates the use concentration inequalities to achieve such probabilistic guarantees that $\mathcal{T}_{\boldsymbol{\Theta}}(\mu)$ occurs for sub-Gaussian designs $\boldsymbol{D}$ and suitably chosen tolerances μ.

Example 3.3.3. [Probability of $\mathcal{T}_{\boldsymbol{\Theta}}(\mu)$] Consider $\boldsymbol{D} = \boldsymbol{Z}\mathbf{A}$ with i.i.d. rows $\boldsymbol{d}_i$ satisfying $\mathrm{E}[\boldsymbol{d}_i] = \mathbf{0}$ and $\mathrm{E}[\boldsymbol{d}_i \boldsymbol{d}_i^{\top}] = \boldsymbol{\Sigma}_{\boldsymbol{d}}$. Let $\boldsymbol{\Theta} = \boldsymbol{\Sigma}_{\boldsymbol{d}}^{-1}$, and let $\overline{\boldsymbol{\Sigma}}_{\boldsymbol{d}} = \mathbb{E}_n[\boldsymbol{d}_i \boldsymbol{d}_i^{\top}]$. Suppose that the rows of $\boldsymbol{D}\boldsymbol{\Theta}^{1/2}$ are sub-Gaussian with sub-Gaussian norm τ; and (ii) the minimum and maximum singular values of $\boldsymbol{\Sigma}_{\boldsymbol{d}}$ are bounded below and above, respectively, by universal constants $\sigma_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{d}}), \sigma_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{d}})$. Cite [32, Lemma 23] and [57] to write

$$\mathrm{P}\left\{ \|\boldsymbol{\Theta}\overline{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{I}\|_{\infty} > a\sqrt{\log p_{\boldsymbol{x}}/n} \right\} \leq 2p_{\boldsymbol{x}}^{-c_{\mathrm{inv}}},$$

where $a > 0$ is a controlled quantity,

$$c_{\mathrm{inv}} := \frac{a^2 s_{\mathbf{A}}^2 \sigma_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{d}})}{24 e^2 \tau^4 \sigma_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{d}})} - 2,$$

and $n \geq (a^2 s_{\mathbf{A}}^2 \sigma_{\min}(\mathbf{\Sigma_d}) \log p_{\boldsymbol{x}})/(4e^2 \sigma_{\max}(\mathbf{\Sigma_d})\tau^4)$. It then follows from Lemma 3.5.3 that, for $\epsilon < m_{\mathbf{A}}$

$$\begin{aligned}
&\mathrm{P}\big\{\|\mathbf{\Theta}\widehat{\mathbf{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{I}\|_\infty > a\sqrt{\log p_{\boldsymbol{x}}/n} + 3m_{\mathbf{\Theta}} m_{\mathbf{A}} \|\widehat{\mathbf{\Sigma}}_{\boldsymbol{z}}\|_\infty \epsilon\big\} \\
&\qquad \leq \mathrm{P}\big\{\|\mathbf{\Theta}\overline{\mathbf{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{I}\|_\infty > a\sqrt{\log p_{\boldsymbol{x}}/n}\big\} + \mathrm{P}\{\|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1} > \epsilon\} \\
&\qquad \leq 2p_{\boldsymbol{x}}^{-c_{\mathrm{inv}}} + \mathrm{P}\{\|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1} > \epsilon\}\,.
\end{aligned}$$

Under the conditions of Example 3.3.3, if (i) $\|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1} = o_{\mathrm{P}}(1)$, (ii) c_{inv} is positive, and (iii) $\|\widehat{\mathbf{\Sigma}}_{\boldsymbol{z}}\|_\infty = O_{\mathrm{P}}(1)$, then the population precision matrix $\mathbf{\Theta}$ is feasible for Program 3.3.1 with high probability.

To specify a in the tolerance μ such that c_{inv} is positive would require knowledge of quantities such as the sub-Gaussian norm τ of $\boldsymbol{d}_i$, which is not feasible. We note that we are not concerned with the feasibility of such quantities in this essay; we explain this position in Section 4.2, where we discuss tuning parameter selection. In Chapter 5, we discuss a practical scheme for selecting the quantity μ that gives good empirical results.

3.4 Asymptotic normality

In Section 3.2, we showed that the updated second-stage estimator $\tilde{\boldsymbol{\beta}}$ satisfies $\sqrt{n}(\tilde{\beta}_j - \beta_j) = \sqrt{n}\mathbb{E}_n[\langle \boldsymbol{\theta}_j, \boldsymbol{d}_i\rangle u_i] + \sum_{\ell=1}^4 f_{\ell,j}$. If the remainder terms vanish in probability, then $\sqrt{n}(\tilde{\beta}_j - \beta_j)$ shares the same weak limit as $\sqrt{n}\mathbb{E}_n[\langle \boldsymbol{\theta}_j, \boldsymbol{d}_i\rangle u_i]$, if it exists. If the model were fixed in n, the central limit theorem would entail that the latter quantity converges weakly to a $\mathcal{N}(0, \omega_j^2)$, where we recall $\omega_j^2 = \mathrm{E}[\langle \boldsymbol{\theta}_j, \boldsymbol{d}_i\rangle^2 u_i^2]$. In the high-dimensional setting, in which the model varies with n, the quantity ω_j^2 need not converge. Nonetheless, we show in Theorem 3.4.1 that $\sqrt{n}(\tilde{\beta}_j - \beta_j)/\omega_j$ converges weakly to a standard Normal random variable and that the limit continues to hold if ω_j is replaced by an estimator $\hat{\omega}_j$ that satisfies $|\hat{\omega}_j - \omega_j| = o_{\mathrm{P}}(1)$. Note that Theorem 3.4.1 gives conditions under which the limit holds given homoscedastic Gaussian noise (Condition 5) as well as conditions under which the limit holds given generic homoscedastic noise (Condition 6).

Theorem 3.4.1 (Weak limits)**.** *Suppose that*

1. *there exists a sequence* $c_n = o(1)$ *such that* $\mathrm{P}\{\|\overline{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{\Sigma}_{\boldsymbol{d}}\|_\infty > c_n\} = o(n)$*, where* $\overline{\boldsymbol{\Sigma}}_{\boldsymbol{d}} = \boldsymbol{D}^\top \boldsymbol{D}/n$*,*

2. $\|\boldsymbol{f}_\ell\|_\infty = o_{\mathrm{P}}(1)$ *for each* $1 \leq \ell \leq 4$*,*

3. $\boldsymbol{\Theta}_{jj} > \vartheta$ *for some universal* $\vartheta > 0$*,*

4. $\max_{j \in [p_{\boldsymbol{x}}]} \|\boldsymbol{\theta}_j\|_1 \leq m_{\boldsymbol{\Theta}}$ *for some universal* $m_{\boldsymbol{\Theta}} < \infty$*.*

If either

5. *Assumption 2.1.3 holds or*

6. *the instrumental variables* $\boldsymbol{z}_i$ *and noise elements* u_i *are iid with* $\mathrm{E}[u_i|\boldsymbol{z}_i] = \sigma_u^2$*, where* σ_u *is bounded away from zero and infinity uniformly in* n*, and there exist* $0 < \zeta < 1/2$ *and* $\nu > (1/2 - \zeta)^{-1}$ *such that*

$$\mathrm{P}\big\{|\langle \boldsymbol{\theta}_j, \boldsymbol{d}_i \rangle| > n^\zeta\big\} \;=\; o(n)\,, \qquad \mathrm{E}[|u_i|^{2+\nu}] \;\lesssim\; \sigma_u^{2+\nu}\,, \tag{3.10}$$

then

$$\sqrt{n}(\tilde{\beta}_j - \beta_j)/\omega_j \;\rightsquigarrow\; Z_j \sim \mathcal{N}(0,1)\,.$$

Furthermore, the limit continues to hold if ω_j *is replaced by an estimator* $\hat{\omega}_j$ *that satisfies* $|\hat{\omega}_j - \omega_j| = o_{\mathrm{P}}(1)$*.*

The proof of Theorem 3.4.1 is similar to those of [32, Lemma 3.2, Theorem 4.1]. It can be found in Section 3.5.3 of the Supplementary Materials.

Conditions 1 and 6 depend on the distribution of the conditional means $\boldsymbol{d}_i$ and hence of the instrumental variables $\boldsymbol{z}_i$. We show in the following example that both conditions are satisfied for sub-Gaussian $\boldsymbol{z}_i$.

Example 3.4.2 (Feasibility of Conditions for Theorem 3.4.1, sub-Gaussian $\boldsymbol{z}_i$)**.** Suppose that $\boldsymbol{z}_i$ is sub-Gaussian with $\|\boldsymbol{z}_i\|_{\psi_2} = \tau$. We claim that Condition 1 holds. To this end, note

that $\overline{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{\Sigma}_{\boldsymbol{d}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^\top - \boldsymbol{\Sigma}_{\boldsymbol{d}}$ and hence that $\frac{1}{n}\sum_{i=1}^n x_{ij}x_{ik} - \sigma_{jk}$, where $\sigma_{jk} := \boldsymbol{\Sigma}_{\boldsymbol{d},jk}$. Now, for any two random variables X and Y, it holds that $\|XY\|_{\psi_1} \leq 2\|X\|_{\psi_2}\|Y\|_{\psi_2}$. Further, if $\mu \in \mathbb{R}$ is a constant, then $\|\mu\|_{\psi_1} = |\mu|/\log 2$. Thus,

$$\|x_{ij}x_{ik} - \sigma_{jk}\|_{\psi_1} \;\leq\; 2\|x_{ij}\|_{\psi_2}\|x_{ik}\|_{\psi_2} + |\sigma_{jk}|/\sqrt{2}\,.$$

Note that $x_{ij} = \boldsymbol{z}_i^\top\boldsymbol{\alpha}^j = \sum_{k=1}^{p_{\boldsymbol{z}}}\alpha_k^j z_{ik}$ and that

$$\|\sum_{k=1}^{p_{\boldsymbol{z}}}\alpha_k^j z_{ik}\|_{\psi_2} \;\leq\; \sum_{k=1}^{p_{\boldsymbol{z}}}\|\alpha_k^j z_{ik}\|_{\psi_2} \;\leq\; \|\boldsymbol{z}_i\|_{\psi_2}\sum_{k=1}^{p_{\boldsymbol{z}}}|\alpha_k^j| \;\leq\; \tau m_{\mathbf{A}}$$

and similarly for $\|x_{ik}\|_{\psi_2}$. Thus

$$\|x_{ij}x_{ik} - \sigma_{jk}\|_{\psi_1} \;\leq\; 2\tau^2 m_{\mathbf{A}}^2 + |\sigma_{jk}|/\sqrt{2} \;=:\; \tau'\,.$$

Now apply the Bernstein-type inequality of [57, Proposition 5.16] to conclude that

$$\mathrm{P}\Big\{\frac{1}{n}\Big|\sum_{i=1}^n x_{ij}x_{ik} - \sigma_{jk}\Big| > t\Big\} \;\leq\; 2\exp\Big(-\frac{n}{6}\min\Big\{\big(\frac{t}{e\tau'}\big)^2, \frac{t}{e\tau'}\Big\}\Big)\,.$$

Choose $t = a\sqrt{(\log(p_{\boldsymbol{x}} \vee n))/n}$. If $n \geq (a/(e\tau'))^2\log(p_{\boldsymbol{x}} \vee n)$, then

$$\mathrm{P}\Big\{\frac{1}{n}\Big|\sum_{i=1}^n x_{ij}x_{ik} - \sigma_{jk}\Big| > a\sqrt{(\log(p_{\boldsymbol{x}} \vee n))/n}\Big\} \;\leq\; 2(p_{\boldsymbol{x}} \vee n)^{-a^2/(6e^2\tau'^2)}\,,$$

and hence

$$\mathrm{P}\big\{\|\overline{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{\Sigma}_{\boldsymbol{d}}\|_\infty > a\sqrt{(\log(p_{\boldsymbol{x}} \vee n))/n}\big\} \;\leq\; 2(p_{\boldsymbol{x}} \vee n)^{2-a^2/(6e^2\tau'^2)}\,,$$

which follows from taking the union bound over $j, k \in [p_{\boldsymbol{x}}]$. Condition 1 then follows under the assumption that $\sqrt{(\log(p_{\boldsymbol{x}} \vee n))/n} = o(1)$ and by choosing a large enough so that $a^2/(6e^2\tau'^2) > 3$.

The requirement that $\mathrm{P}\big\{|\langle\boldsymbol{\theta}_j, \boldsymbol{d}_i\rangle| > n^\zeta\big\} = o(n)$ of Condition 6 can be shown according to the reasoning in the proof of [32, Lemma 6.3]. Indeed, the latter authors show a slightly stronger result, namely that for any $0 < \zeta < 1/2$, it holds that

$$\mathrm{P}\big\{\|\boldsymbol{D}^\top\boldsymbol{\theta}_j\|_\infty > n^\zeta\big\} \;\leq\; n\exp(-Cn^{2\zeta})$$

for an appropriate constant $C > 0$.

The main application of Theorem 3.4.1 is the construction of asymptotically valid confidence intervals under a wide variety of nosie regimes. Given $j \in [p_{\boldsymbol{x}}]$ and a confidence level α, an asymptotic $(1-\alpha)\%$ confidence interval $\hat{\mathcal{I}}_{\alpha,j}$ is given by

$$\hat{\mathcal{I}}_{\alpha,j} := \left[\tilde{\beta}_j - z_\alpha \hat{\omega}_j,\, \tilde{\beta}_j + z_\alpha \hat{\omega}_j\right], \tag{3.11}$$

where $z_\alpha = \Phi^{-1}(1-\alpha/2)$ and $\hat{\omega}_j$ satisfies the conditions of Theorem 3.4.1. We present an empirical study of the finite-sample properties of this procedure for the updated two-stage Lasso estimator in Chapter 5.

3.5 Materials required for Chapter 3

3.5.1 Materials required for Section 3.2

Proof of Lemma 3.2.1. Note that

$$\begin{aligned}
\tilde{\boldsymbol{\beta}} &= \hat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\Theta}}\mathbb{E}_n[\tilde{\boldsymbol{h}}(\hat{\boldsymbol{\beta}})] \\
&= \hat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{D}}^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})/n \\
&= \hat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{D}}^\top(\boldsymbol{X}[\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}] + \boldsymbol{u})/n \\
&= \hat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{D}}^\top(\widehat{\boldsymbol{D}}[\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}] + [\boldsymbol{X} - \widehat{\boldsymbol{D}}][\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}] + \boldsymbol{u})/n \\
&= \boldsymbol{\beta} + \widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{D}}^\top\boldsymbol{u}/\sqrt{n} + \widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{D}}^\top(\boldsymbol{X} - \widehat{\boldsymbol{D}})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})/n + (\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{I})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}),
\end{aligned}$$

as claimed. □

3.5.2 Materials required for Section 3.3

We present the properties of the estimators $\hat{\boldsymbol{\theta}}_j$ required for the results of Sections 4.3 and 3.4. Lemma 3.5.1, which gives an ℓ_∞ bound for the estimation error $\hat{\boldsymbol{\theta}}_j - \hat{\boldsymbol{\theta}}$, is comparable to [19, Theorem 4]; the proofs are similar but depend on different conditions on the covariance estimator $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}}$. The proof of Lemma 3.5.2 follows that of [19, Theorem 6].

Lemma 3.5.1. *Suppose that: (i) the quantity* $\|\boldsymbol{\Theta}\|_{L_1}$ *is bounded above by a constant* $m_{\boldsymbol{\Theta}} < \infty$*; and (ii)* $\widehat{\boldsymbol{\Theta}}$ *is an estimate of* $\boldsymbol{\Theta} = \boldsymbol{\Sigma}_{\boldsymbol{d}}^{-1} = \mathrm{cov}(\boldsymbol{d}_i)^{-1}$ *with rows* $\hat{\boldsymbol{\theta}}_j$ *obtained as solutions to Program 3.3.1. Then, on the set* $\mathcal{T}_{\boldsymbol{\Theta}}(\mu)$*,*

$$\|\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j\|_\infty \leq 2m_{\boldsymbol{\Theta}}\mu$$

for each $j \in [p_{\boldsymbol{x}}]$*.*

Proof of Lemma 3.5.1. First, observe that the conditions of the present lemma entail that

$$\|\boldsymbol{\Theta}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{I}\|_\infty \leq \mu\,, \qquad \|\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{I}\|_\infty \leq \mu\,.$$

Now, on the set $\mathcal{T}_{\boldsymbol{\Theta}}(\mu)$, each row $\boldsymbol{\theta}_j$ is feasible for the respective Specific Program 3.3.1. It then follows from the optimality of $\hat{\boldsymbol{\theta}}_j$ that $\|\hat{\boldsymbol{\theta}}_j\|_1 \leq \|\boldsymbol{\theta}_j\|_1$ for each $j \in [p_{\boldsymbol{x}}]$ and hence that $\max_{j\in[p_{\boldsymbol{x}}]} \|\hat{\boldsymbol{\theta}}_j\|_1 \leq \|\boldsymbol{\Theta}\|_{L_1}$. Next, note that

$$\begin{aligned}
\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta} = (\widehat{\boldsymbol{\Theta}}\boldsymbol{\Sigma}_{\boldsymbol{d}} - \boldsymbol{I})\boldsymbol{\Theta} &= (\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}} + \widehat{\boldsymbol{\Theta}}(\boldsymbol{\Sigma}_{\boldsymbol{d}} - \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}}) - \boldsymbol{I})\boldsymbol{\Theta} \\
&= (\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{I})\boldsymbol{\Theta} + \widehat{\boldsymbol{\Theta}}(\boldsymbol{\Sigma}_{\boldsymbol{d}} - \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}})\boldsymbol{\Theta} \\
&= (\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{I})\boldsymbol{\Theta} + \widehat{\boldsymbol{\Theta}}(\boldsymbol{I} - \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}}\boldsymbol{\Theta})\,.
\end{aligned}$$

Combine the previous display with the previously discussed bounds to obtain

$$\begin{aligned}
\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_\infty &\leq \|(\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{I})\boldsymbol{\Theta}\|_\infty + \|\widehat{\boldsymbol{\Theta}}(\boldsymbol{I} - \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}}\boldsymbol{\Theta})\|_\infty \\
&\leq \|\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{I}\|_\infty\|\boldsymbol{\Theta}\|_{L_1} + \max_{j\in[p_{\boldsymbol{x}}]} \|\hat{\boldsymbol{\theta}}_j\|_1\|\boldsymbol{I} - \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}}\boldsymbol{\Theta}\|_\infty \\
&= \|\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{I}\|_\infty\|\boldsymbol{\Theta}\|_{L_1} + \max_{j\in[p_{\boldsymbol{x}}]} \|\hat{\boldsymbol{\theta}}_j\|_1\|\boldsymbol{I} - \boldsymbol{\Theta}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}}\|_\infty \\
&\leq 2m_{\boldsymbol{\Theta}}\mu\,,
\end{aligned}$$

where the inference to the penultimate line above follows from that $\boldsymbol{I}, \boldsymbol{\Theta}$, and $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}}$ are each symmetric and that the ℓ_∞-norm is invariant under transposition of its argument. The result follows immediately from the previous display. □

Lemma 3.5.2. *Suppose in addition to the conditions of Lemma 3.5.1 that* $\boldsymbol{\Theta} \in \mathcal{U}(m_{\boldsymbol{\Theta}}, q, s_{\boldsymbol{\Theta}})$. *Then,*

$$\|\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j\|_1 \;\leq\; 2c_q(2m_{\boldsymbol{\Theta}}\mu)^{1-q}s_{\boldsymbol{\Theta}}$$

for each $j \in [p_{\boldsymbol{x}}]$, *where* $c_q := 1 + 2^{1-q} + 3^{1-q}$.

Proof of Lemma 3.5.2. See the proof of line (14) of [19, Theorem 6]. □

The following Lemma is required for Example 3.3.3

Lemma 3.5.3. *For* $\epsilon < m_{\mathbf{A}}$, *it holds that*

$$\begin{aligned}\mathrm{P}\big\{\|\boldsymbol{\Theta}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{I}\|_\infty > \mu + 3m_{\boldsymbol{\Theta}}m_{\mathbf{A}}\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty\epsilon\big\} \;\leq\; &\mathrm{P}\big\{\|\boldsymbol{\Theta}\overline{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{I}\|_\infty > \mu\big\} \\ &+ \mathrm{P}\big\{\|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1} \leq \epsilon\big\}\,.\end{aligned}$$

where $\overline{\boldsymbol{\Sigma}}_{\boldsymbol{d}} = \mathbb{E}_n[\boldsymbol{d}_i\boldsymbol{d}_i^\top]$.

Proof of Lemma 3.5.3. Note that

$$\begin{aligned}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}} \;=\; \widehat{\boldsymbol{D}}^\top\widehat{\boldsymbol{D}}/n \;&=\; \boldsymbol{D}^\top\boldsymbol{D}/n + (\widehat{\mathbf{A}} - \mathbf{A})^\top\boldsymbol{Z}^\top\boldsymbol{Z}\mathbf{A}/n + \widehat{\mathbf{A}}^\top\boldsymbol{Z}^\top\boldsymbol{Z}(\widehat{\mathbf{A}} - \mathbf{A})/n \\ &=\; \overline{\boldsymbol{\Sigma}}_{\boldsymbol{d}} + (\widehat{\mathbf{A}} - \mathbf{A})^\top\boldsymbol{Z}^\top\boldsymbol{Z}\mathbf{A}/n + \mathbf{A}^\top\boldsymbol{Z}^\top\boldsymbol{Z}(\widehat{\mathbf{A}} - \mathbf{A})/n \\ &\quad + (\widehat{\mathbf{A}} - \mathbf{A})^\top\boldsymbol{Z}^\top\boldsymbol{Z}(\widehat{\mathbf{A}} - \mathbf{A})/n\,,\end{aligned}$$

so that

$$\begin{aligned}\|\boldsymbol{\Theta}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{I}\|_\infty \;\leq\; &\|\boldsymbol{\Theta}(\boldsymbol{D}^\top\boldsymbol{D}/n) - \boldsymbol{I}\|_\infty + \|\boldsymbol{\Theta}(\widehat{\mathbf{A}} - \mathbf{A})^\top\boldsymbol{Z}^\top\boldsymbol{Z}\mathbf{A}/n\|_\infty \\ &+ \|\boldsymbol{\Theta}\mathbf{A}^\top\boldsymbol{Z}^\top\boldsymbol{Z}(\widehat{\mathbf{A}} - \mathbf{A})/n\|_\infty + \|\boldsymbol{\Theta}(\widehat{\mathbf{A}} - \mathbf{A})^\top\boldsymbol{Z}^\top\boldsymbol{Z}(\widehat{\mathbf{A}} - \mathbf{A})/n\|_\infty \\ :=\; &\|\boldsymbol{\Theta}(\boldsymbol{D}^\top\boldsymbol{D}/n) - \boldsymbol{I}\|_\infty + \mathrm{I}_1 + \mathrm{I}_2 + \mathrm{I}_3\,.\end{aligned}$$

Note that

$$\begin{aligned}\mathrm{I}_1 \;=\; \|\boldsymbol{\Theta}(\widehat{\mathbf{A}} - \mathbf{A})^\top\boldsymbol{Z}^\top\boldsymbol{Z}\mathbf{A}/n\|_\infty \;&\leq\; \|\boldsymbol{\Theta}\|_{L_1}\|(\widehat{\mathbf{A}} - \mathbf{A})^\top\boldsymbol{Z}^\top\boldsymbol{Z}\mathbf{A}/n\|_\infty \\ &\leq\; m_{\boldsymbol{\Theta}}\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty\|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1}\|\mathbf{A}\|_{L_1} \\ &=\; m_{\boldsymbol{\Theta}}m_{\mathbf{A}}\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty\|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1}\,.\end{aligned}$$

The same bound holds for I_2 by symmetry of the ℓ_∞ norm under transposition of its argument.

For the term I_3, similar reasoning yields

$$\begin{aligned}\mathrm{I}_3 &= \|\boldsymbol{\Theta}(\widehat{\mathbf{A}}-\mathbf{A})^\top \boldsymbol{Z}^\top \boldsymbol{Z}(\widehat{\mathbf{A}}-\mathbf{A})/n\|_\infty \\ &\leq m_{\boldsymbol{\Theta}}\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty \|\widehat{\mathbf{A}}-\mathbf{A}\|_{L_1}^2 = m_{\boldsymbol{\Theta}}\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty \|\widehat{\mathbf{A}}-\mathbf{A}\|_{L_1}^2 .\end{aligned}$$

If $\epsilon < m_{\mathbf{A}}$, then, on the set $\{\|\widehat{\mathbf{A}}-\mathbf{A}\|_{L_1} \leq \epsilon\}$, it holds that $\mathrm{I}_3 \leq m_{\boldsymbol{\Theta}} m_{\mathbf{A}} \|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty \|\widehat{\mathbf{A}}-\mathbf{A}\|_{L_1}$. Conclude that

$$\mathrm{P}\big\{\mathrm{I}_1+\mathrm{I}_2+\mathrm{I}_3 > 3m_{\boldsymbol{\Theta}} m_{\mathbf{A}} \|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty \epsilon\big\} \leq \mathrm{P}\{\|\widehat{\mathbf{A}}-\mathbf{A}\|_{L_1} > \epsilon\}$$

and hence that

$$\begin{aligned}&\mathrm{P}\big\{\|\boldsymbol{\Theta}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{I}\|_\infty > \mu + 3m_{\boldsymbol{\Theta}} m_{\mathbf{A}} \|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty \|\widehat{\mathbf{A}}-\mathbf{A}\|_{L_1}\big\} \\ \leq\ &\mathrm{P}\big\{\|\boldsymbol{\Theta}\mathbb{E}_n[\boldsymbol{d}_i^{\otimes 2}] - \boldsymbol{I}\|_\infty > \mu\big\} + \mathrm{P}\{\|\widehat{\mathbf{A}}-\mathbf{A}\|_{L_1} > \epsilon\},\end{aligned}$$

as claimed. □

3.5.3 Materials required for Section 3.4

Proof of Theorem 3.4.1. The proof of the first claim consists of two steps. First, we show that the quantity

$$Z_{j,n} := \frac{1}{\sqrt{n}} \sum_{i=1}^n \langle \boldsymbol{\theta}_j, \boldsymbol{d}_i \rangle u_i/\omega_j .$$

and $\sqrt{n}(\tilde{\beta}_j - \beta_j)/\omega_j$ share the same weak limit. Second, we show that $Z_{j,n} \rightsquigarrow \mathcal{N}(0,1)$. To establish the first step, we claim that

$$\lim_{n\to\infty} \mathrm{P}\{\sqrt{n}(\tilde{\beta}_j - \beta_j)/\omega_j \leq t\} \leq \lim_{n\to\infty} \mathrm{P}\{Z_{j,n} \leq t\}, \tag{3.12}$$

for all $t \in \mathbb{R}$. An analogous lower bound follows by a matching argument, which shows that $\sqrt{n}(\tilde{\beta}_j - \beta_j)/\omega_j$ and $Z_{j,n}$ share the same weak limit. To show the claim above, let $t \in \mathbb{R}$

be given, fix a controlled $\epsilon > 0$, and note that, by the decomposition of (3.5), we have

$$\begin{aligned} \mathrm{P}\big\{\sqrt{n}(\tilde{\beta}_j - \beta_j)/\omega_j \leq t\big\} \;&\leq\; \mathrm{P}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\langle \boldsymbol{\theta}_j, \boldsymbol{d}_i\rangle u_i/\omega_j + \sum_{\ell=1}^{4} f_{\ell,j}/\omega_j \leq t + 4\epsilon\right\} \\ &\leq\; \mathrm{P}\big\{Z_{j,n} \leq t + \epsilon\big\} + \sum_{\ell=1}^{4}\mathrm{P} f_{\ell,j}/\omega_j > \epsilon\,. \end{aligned}$$

By specification of σ_u and $\boldsymbol{\Theta}_{jj}$ in Assumption 2.1.3 and Condition 3 of the present theorem, it follows that ω_j is bounded strictly away from 0 uniformly in n. The assumptions of the present theorem then entail that $\mathrm{P} f_{\ell,j}/\omega_j > \epsilon = o(1)$ for all $\epsilon > 0$ and each ℓ. Letting ϵ tend to zero shows the claim of (3.12). It follows from the analogous lower bound that

$$\lim_{n\to\infty} \mathrm{P}\{\sqrt{n}(\tilde{\beta}_j - \beta_j)/\omega_j \leq t\} \;=\; \lim_{n\to\infty} \mathrm{P}\{Z_{j,n} \leq t\}$$

for all $t \in \mathbb{R}$, thus completing the first step.

Next, we show that, under each of Conditions 5 and 6 in the statement of the present theorem, $Z_{j,n} \rightsquigarrow Z_j \sim \mathcal{N}(0,1)$. To this end, we define the quantity

$$w_j^2 \;:=\; \boldsymbol{\theta}_j^\top \overline{\boldsymbol{\Sigma}}_{\boldsymbol{d}} \boldsymbol{\theta}_j \;=\; \frac{1}{n}\sum_{i=1}^{n}\langle \boldsymbol{\theta}_j, \boldsymbol{d}_i\rangle^2\,.$$

We claim first that

$$\tilde{Z}_{j,n} \;:=\; \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\langle \boldsymbol{\theta}_j, \boldsymbol{d}_i\rangle u_i}{w_j \sigma_u} \;\rightsquigarrow\; Z_j \sim \mathcal{N}(0,1)$$

under each of Conditions 5 and 6 and second that $\sigma_u w_j/\omega_j \to_{\mathrm{P}} 1$. Since $Z_{j,n} = \frac{w_j\sigma_u}{\omega_j}\tilde{Z}_{j,n}$, the desired limit follows from an application Slutsky's Lemma.

To show the first claim under Condition 5, note that, by specification of w_j, we have under Assumption 2.1.3 that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\langle \boldsymbol{\theta}_j, \boldsymbol{d}_i\rangle u_i}{w_j \sigma_u} \,|\, \boldsymbol{Z} \;\sim\; \mathcal{N}(0,1)\,.$$

Thus

$$\lim_{n\to\infty} \mathrm{P}\{\tilde{Z}_{j,n} \leq t\} \;=\; \lim_{n\to\infty} \mathrm{E}\big[\mathrm{P}\{\tilde{Z}_{j,n} \leq t \,|\, \boldsymbol{Z}\}\big] \;=\; \lim_{n\to\infty} \mathrm{E}[\Phi(t) \,|\, \boldsymbol{Z}] \;=\; \Phi(t)$$

for all $t \in \mathbb{R}$, where Φ denotes the cdf of a standard Normal random variable. This shows the desired weak limit under Condition 5.

We use the Lindeberg-Feller central limit theorem to show the limit under Condition 6. To begin, write

$$\tilde{Z}_{j,n} \;=\; \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \xi_i \,, \qquad \xi_i \;:=\; \langle \boldsymbol{\theta}_j, \boldsymbol{d}_i \rangle u_i / (w_j \sigma_u) \,.$$

Note that

$$\mathrm{E}[\xi_i] \;=\; \mathrm{E}\big[\langle \boldsymbol{\theta}_j, \boldsymbol{d}_i \rangle / (w_j \sigma_u) \mathrm{E}[u_i | \boldsymbol{z}_i] \big] \;=\; 0$$

and that

$$\sigma_n^2 \;:=\; \sum_{i=1}^{n} \mathrm{E}[\xi_i^2] \;=\; \mathrm{E}\left[\frac{1}{w_j^2} \sum_{i=1}^{n} \langle \boldsymbol{\theta}_j, \boldsymbol{d}_i \rangle^2 \mathrm{E}[(u_i/\sigma_u)^2 | \boldsymbol{z}_i] \right] \;=\; n \,.$$

To demonstrate Lindeberg's condition, let $\delta > 0$ be arbitrary and write

$$\begin{aligned}
\sigma_n^{-2} \sum_{i=1}^{n} \mathrm{E}[\xi_i^2 \mathbb{1}\{|\xi_i| > \delta \sigma_n\}] \;&=\; \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}\big[\mathrm{E}[\xi_i^2 \mathbb{1}\{|\xi_i| > \delta \sigma_n\} \,|\, \boldsymbol{z}_i] \big] \\
&=\; \mathrm{E}\left[\frac{1}{n} \frac{1}{w_j^2} \sum_{i=1}^{n} \langle \boldsymbol{\theta}_j, \boldsymbol{d}_i \rangle^2 \mathrm{E}[(u_i/\sigma_u)^2 \mathbb{1}\{|\xi_i| > \delta \sqrt{n}\} \,|\, \boldsymbol{z}_i] \right] \\
&=\; n \mathrm{E}[(u_1/\sigma_u)^2 \mathbb{1}\{|\xi_1| > \delta \sqrt{n}\}] \,.
\end{aligned}$$

For brevity, we write $\tilde{u}_1 := u_1/\sigma_u$. Introduce the set $\mathcal{T} := \{|\langle \boldsymbol{\theta}_j, \boldsymbol{d}_1 \rangle \leq n^{\zeta}\}$ and note, since $|\xi_1| \leq w_j^{-1} |\langle \boldsymbol{\theta}_j, \boldsymbol{d}_i \rangle| |\tilde{u}_1|$, that

$$\big\{ |\xi_1| > \delta \sqrt{n} \big\} \cap \mathcal{T} \;\subseteq\; \big\{ |\tilde{u}_1| > \delta w_j n^{1/2 - \zeta} \big\} \,,$$

and hence that $\mathbb{1}\{|\xi_1| > \delta \sqrt{n}\} \cap \mathcal{T} \leq \mathbb{1}\{|\tilde{u}_1| > \delta w_j n^{1/2-\zeta}\}$. Combine this inequality with the result of two displays previous to obtain

$$\begin{aligned}
\sigma_n^{-2} \sum_{i=1}^{n} \mathrm{E}[\xi_i^2 \mathbb{1}\{|\xi_i| > \delta \sigma_n\}] \;&=\; n \mathrm{E}\big[\tilde{u}_1^2 (\mathbb{1}\{|\xi_1| > \delta \sqrt{n}\} \cap \mathcal{T} + \mathbb{1}\{|\xi_1| > \delta \sqrt{n}\} \cap \mathcal{T}^{\complement}) \big] \\
&=\; n \underbrace{\mathrm{E}[\tilde{u}_1^2 \mathbb{1}\{|\tilde{u}_1| > \delta w_j n^{1/2-\zeta}\}]}_{\mathrm{I}_1} + n \underbrace{\mathrm{E}[\tilde{u}_1^2 \mathbb{1}_{\mathcal{T}^{\complement}}]}_{\mathrm{I}_2} \,,
\end{aligned}$$

where the substitution of indicators in the final line above is permitted since $\tilde{u}_1^2 \geq 0$. To show Lindeberg's condition, it sufficies to show that I_1 and I_2 are each $o(n)$.

To treat I_1, consider the event $\{w_j \leq \vartheta^{1/2}/\sqrt{2}\}$ and write

$$\begin{aligned}\mathbb{1}\{|\tilde{u}_1| > \delta w_j n^{1/2-\zeta}\} &= \mathbb{1}\{|\tilde{u}_1| > \delta w_j n^{1/2-\zeta}\} \cap \{w_j \leq \vartheta^{1/2}/\sqrt{2}\} \\ &\quad + \mathbb{1}\{|\tilde{u}_1| > \delta w_j n^{1/2-\zeta}\} \cap \{w_j > \vartheta^{1/2}/\sqrt{2}\} \\ &\leq \mathbb{1}\{w_j \leq \vartheta^{1/2}/\sqrt{2}\} + \mathbb{1}\{|\tilde{u}_1| > \delta\vartheta^{1/2} n^{1/2-\zeta}\}\end{aligned}$$

so that

$$\mathrm{I}_1 \leq \underbrace{\mathrm{E}[\tilde{u}_1^2 \mathbb{1}\{w_j \leq \vartheta^{1/2}/\sqrt{2}\}]}_{\mathrm{I}_{1\mathrm{a}}} + \underbrace{\mathrm{E}[\tilde{u}_1^2 \mathbb{1}\{|\tilde{u}_1| > \delta\vartheta^{1/2} n^{1/2-\zeta}\}]}_{\mathrm{I}_{1\mathrm{b}}}.$$

Observe that

$$\begin{aligned}\mathrm{I}_{1\mathrm{a}} &= \mathrm{E}\big[\mathbb{1}\{w_j \leq \vartheta^{1/2}/\sqrt{2}\}\mathrm{E}[\tilde{u}_1^2 \,|\, \boldsymbol{z}_i]\big] \\ &= \mathrm{P}\{w_j \leq \vartheta^{1/2}/\sqrt{2}\} \\ &\leq \mathrm{P}\{w_j^2 - \boldsymbol{\Theta}_{jj} \leq \vartheta/2 - \boldsymbol{\Theta}_{jj}\} \\ &\leq \mathrm{P}\{w_j^2 - \boldsymbol{\Theta}_{jj} \leq -\vartheta/2\} \\ &\leq \mathrm{P}\{|w_j^2 - \boldsymbol{\Theta}_{jj}| \geq \vartheta/2\}.\end{aligned}$$

Now note that

$$\begin{aligned}|w_j^2 - \boldsymbol{\Theta}_{jj}| &= |\boldsymbol{\theta}_j^\top \overline{\boldsymbol{\Sigma}}_{\boldsymbol{d}} \boldsymbol{\theta}_j - \boldsymbol{\Theta}_{jj}| \\ &= |\boldsymbol{\theta}_j^\top (\overline{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{\Sigma}_{\boldsymbol{d}}) \boldsymbol{\theta}_j| \leq \|\boldsymbol{\theta}_j\|_1^2 \|\overline{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{\Sigma}_{\boldsymbol{d}}\|_\infty.\end{aligned}$$

Thus

$$\begin{aligned}\mathrm{I}_{1\mathrm{a}} &\leq \mathrm{P}\{|w_j^2 - \boldsymbol{\Theta}_{jj}| \geq \vartheta/2\} \\ &\leq \mathrm{P}\big\{\|\overline{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{\Sigma}_{\boldsymbol{d}}\|_\infty \geq \vartheta/(2\|\boldsymbol{\theta}_j\|_1^2)\big\} = o(n)\end{aligned} \tag{3.13}$$

by Conditions 1 and 4 of the present theorem.

To treat $\mathrm{I}_{1\mathrm{b}}$, note that

$$\begin{aligned}\big\{|\tilde{u}_1| > \delta\vartheta^{1/2} n^{1/2-\zeta}\big\} &\subseteq \big\{|\tilde{u}_1|^\nu > (\delta\vartheta^{1/2})^\nu n^{\nu(1/2-\zeta)}/\sqrt{2}\big\} \\ &= \big\{\sqrt{2}(\delta\vartheta^{1/2})^\nu n^{\nu(1/2-\zeta)} |\tilde{u}_1|^\nu > 1\big\}\end{aligned}$$

and hence that

$$\mathbb{1}\{|\tilde{u}_1| > \delta\vartheta^{1/2}n^{1/2-\zeta}\} \;\leq\; \mathbb{1}\{|\tilde{u}_1| > \delta\vartheta^{1/2}n^{1/2-\zeta}\}\big(\sqrt{2}(\delta\vartheta^{1/2})^{\nu}n^{\nu(1/2-\zeta)}|\tilde{u}_1|^{\nu}\big)\,.$$

Noting that $\tilde{u}_1^2 \geq 0$, substitute the right-hand side above into the expression for $\mathrm{I}_{1\mathrm{b}}$ and drop the indicator to obtain

$$\lim_{n\to\infty} n\mathrm{I}_{1\mathrm{b}} \;\leq\; \lim_{n\to\infty} \sqrt{2}(\delta\vartheta^{1/2})^{\nu}n^{1-\nu(1/2-\zeta)}\mathrm{E}[|\tilde{u}_1|^{2+\nu}] \;=\; 0\,,$$

where the above limit depends on the specification of ν, ζ in Condition 6 of the present theorem. Combine the above display with (3.13) to conclude that $\mathrm{I}_1 = o(n)$.

To show as much for I_2, observe that

$$\lim_{n\to\infty} n\mathrm{I}_2 \;=\; \lim_{n\to\infty} n\mathrm{E}\big[\mathbb{1}\{|\langle\boldsymbol{\theta}_j, \boldsymbol{d}_i\rangle| > n^{\zeta}\}\mathrm{E}[\tilde{u}_1^2 \,|\, \boldsymbol{z}_i]\big]$$
$$\lim_{n\to\infty} n\mathrm{P}\{|\langle\boldsymbol{\theta}_j, \boldsymbol{d}_i\rangle| > n^{\zeta}\} \;=\; 0$$

by Condition (3.10) of the present theorem. This concludes the demonstration of Lindeberg's condition. It follows that $\tilde{Z}_{j,n} \rightsquigarrow \mathcal{N}(0,1)$. To show as much for $Z_{j,n}$ and hence for $\sqrt{n}(\tilde{\beta}_j - \beta_j)$, it suffices to show that $w_j\sigma_u/\omega_j \to_{\mathrm{P}} 1$. Note that $\omega_j = \sigma_u\boldsymbol{\Theta}_{jj}$ and hence that $w_j\sigma_u/\omega_j = w_j/\boldsymbol{\Theta}_{jj}$. Since $\boldsymbol{\Theta}_{jj}$ is bounded strictly away from zero uniformly in n, we have $|w_j/\boldsymbol{\Theta}_{jj} - 1| = |w_j - \boldsymbol{\Theta}_{jj}|/\boldsymbol{\Theta}_{jj}$ and hence that it suffices to show that $|w_j - \boldsymbol{\Theta}_{jj}| = o_{\mathrm{P}}(1)$. But, as we established above, we have for arbitrary $\epsilon > 0$

$$\mathrm{P}\{|w_j - \boldsymbol{\Theta}_{jj}| > \epsilon\} \;\leq\; \mathrm{P}\{\|\overline{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{\Sigma}_{\boldsymbol{d}}\|_{\infty} > \epsilon/m_{\boldsymbol{\Theta}}^2\} \;=\; o(1)$$

by Condition 1 of the present theorem. Thus $w_j\sigma_u/\omega_j \to_{\mathrm{P}} 1$ and hence $Z_{j,n} \rightsquigarrow \mathcal{N}(0,1)$ under each Condition 5 and 6 of the present theorem.

It remains to show that the limit holds when ω_j is replaced by an estimator $\hat{\omega}_j$ that satisfies $|\hat{\omega}_j - \omega_j| = o_{\mathrm{P}}(1)$. Suppose that $\hat{\omega}_j$ is such an estimator. We claim that $\hat{\omega}_j/\omega_j - 1 = o_{\mathrm{P}}(1)$. To see as much, note that $|\hat{\omega}_j/\omega_j - 1| = |\hat{\omega}_j - \omega_j|/\omega_j = o_{\mathrm{P}}(1)$ by the specification of $\hat{\omega}_j$ and that $\omega_j = \sigma_u\boldsymbol{\Theta}_{jj}$ is strictly bounded away from zero uniformly in n. It then follows from the continuous mapping theorem that $\omega_j/\hat{\omega}_j \to_{\mathrm{P}} 1$ and then from Slutksky's lemma that $\sqrt{n}(\tilde{\beta}_j - \beta_j)/\hat{\omega}_j = (\omega_j/\hat{\omega}_j)Z_{j,n} \rightsquigarrow \mathcal{N}(0,1)$, as claimed. □

Chapter 4

EXAMPLE: TWO-STAGE LASSO

Theorem 3.4.1 depends on high-level assumptions that ensure good behavior of the remainder terms $\boldsymbol{f}_\ell$ and standard error estimate $\hat{\omega}_j$. In this chapter, we demonstrate how such conditions may be satisfied in the high-dimensional setting. In particular, we introduce in Section 4.1 a two-stage Lasso estimation procedure, for which we provide theoretical bounds in Section 4.2. The rates for the second-stage estimation error are particularly involved due to the dependence on the predicted conditional means from the first-stage estimation. In Section 4.3, we identify conditions under which the remainder terms $\boldsymbol{f}_\ell$ vanish in probability under the two-stage Lasso procedure.

4.1 Two-stage estimator

For $j \in [p_{\boldsymbol{x}}]$, we let $\hat{\boldsymbol{\alpha}}^j$ denote the *first-stage Lasso estimator*

$$\hat{\boldsymbol{\alpha}}^j \in \operatorname*{arg\,min}_{\boldsymbol{a} \in \mathbb{R}^{p_{\boldsymbol{z}}}} \left\{ \|\boldsymbol{x}^j - \boldsymbol{Z}\boldsymbol{a}\|_2^2/(2n) + r_j \|\boldsymbol{a}\|_1 \right\}. \tag{4.1}$$

We let $\boldsymbol{r} := (r_1, \ldots, r_{p_{\boldsymbol{x}}})$ denote the tuple of first-stage tuning parameters, and we write

$$r_{\mathbf{A}} := \max_{j \in [p_{\boldsymbol{x}}]} r_j \,. \tag{4.2}$$

We let $\hat{d}_{ij} := \boldsymbol{z}_i^\top \hat{\boldsymbol{\alpha}}^j$ denote the predicted conditional mean of x_{ij} given $\boldsymbol{z}_i$ based on the estimates $\hat{\boldsymbol{\alpha}}^j$ and write $\widehat{\boldsymbol{D}} = \boldsymbol{Z}\widehat{\mathbf{A}}$, where the matrix $\widehat{\boldsymbol{D}}$ has columns given by $\hat{\boldsymbol{d}}^j := (\hat{d}_{1j}, \ldots, \hat{d}_{nj})^\top$ and the matrix $\widehat{\mathbf{A}} \in \mathbb{R}^{p_{\boldsymbol{z}} \times p_{\boldsymbol{x}}}$ has columns given by the $\hat{\boldsymbol{\alpha}}^j$.

We define the *second-stage Lasso estimator* to be

$$\hat{\boldsymbol{\beta}} \in \operatorname*{arg\,min}_{\boldsymbol{b} \in \mathbb{R}^{p_{\boldsymbol{x}}}} \left\{ \|\boldsymbol{y} - \widehat{\boldsymbol{D}}\boldsymbol{b}\|_2^2/(2n) + r_{\boldsymbol{\beta}} \|\boldsymbol{b}\|_1 \right\}. \tag{4.3}$$

In Sections 4.2.3 and 4.2.4, we develop sparsity-based results that require the following quantities. We write $S_j = \operatorname{supp} \boldsymbol{\alpha}^j$ for the *active sets* of the first-stage regression parameters, and we write $s_{\boldsymbol{\alpha}^j} := |S_j|$ and $s_{\mathbf{A}} := \max_{j \in [p_{\boldsymbol{x}}]} s_{\boldsymbol{\alpha}^j}$; we write $S_{\boldsymbol{\beta}} := \operatorname{supp} \boldsymbol{\beta}$ for the active set of the second-stage regression parameter and $s_{\boldsymbol{\beta}} := |S_{\boldsymbol{\beta}}|$. Finally, we note that ℓ_0 sparsity is not a limitation in principle and that more general regression vectors may be considered at the price of additional complexity [17, Sections 6.2.3-4], [11, 9].

4.2 Estimation error bounds

In this section, we present estimation error bounds for the first- and second- stage estimators described in Section 4.1. Both such bounds depend on the same fundamental strategy for proving finite-sample guarantees for ℓ_1-regularized estimators. This strategy consists of two parts. The first part is the *oracle inequality*, which establishes a deterministic bound for the estimation and prediction performance of the Lasso on a particular set of interest. The second part is the control of the *empirical process term*, which defines the set of interest. We include such prerequisites in Section 4.4.1 of the Supplementary Materials.

Before we present the estimation error bounds for the first- and second-stage Lasso estimators, we first discuss two important ingredients for such results: (i) the compatibility condition (Section 4.2.1), which is required in the proof of the oracle inequality, and (ii) appropriate tuning parameter selection (Section 4.2.2), which is required for the control of the empirical process terms.

4.2.1 Compatibility condition

The oracle bounds rely on the good behavior of certain moduli of continuity of the empirical Gram matrices $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}} = \boldsymbol{Z}^\top \boldsymbol{Z}/n$ and $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}} = \widehat{\boldsymbol{D}}^\top \widehat{\boldsymbol{D}}/n$. We codify this requirement in the following definition.

Definition 4.2.1 (Compatibility condition)**.** For a given index set $S \subseteq [p]$, $p \in \mathbb{N}$, define

the double-cone

$$\mathcal{C}(S) := \left\{ \boldsymbol{\delta} \in \mathbb{R}^p \setminus \mathbf{0} : \|\boldsymbol{\delta}_{S^\complement}\|_1 \leq 3\|\boldsymbol{\delta}_S\|_1 \right\}. \tag{4.4}$$

We say that the *compatibility condition* holds for the matrix $\boldsymbol{M} \in \mathbb{R}^{n \times p}$ relative to the index set S and the constant $\phi^2 > 0$ if

$$\phi^2 \leq \inf_{\boldsymbol{\delta} \in \mathcal{C}(S)} \frac{|S| \|\boldsymbol{M}\boldsymbol{\delta}\|_2^2}{n \|\boldsymbol{\delta}\|_1^2} \tag{4.5}$$

holds. We call such a quantity ϕ^2 the *compatibility constant*.

The compatibility condition is so named because it interfaces between the ℓ_1 norm of the estimation error and the ℓ_2 prediction error of the Lasso estimator. It is instrumental in bounding the estimation and prediction error of the Lasso and is a standard assumption in ℓ_1-regularized estimation literature. For this purpose, the index set S is taken to be the active set of the target regression parameter [17, Chapter 6]. The constant 3 is arbitrary; alternative choices require adjustment of other constants that appear in the bounds [15].

A related, slightly stronger condition known as the *restricted eigenvalue condition* is elsewhere used for the same end; see [15], as well as [17, Chapter 6] and [52] for discussion of the compatibility, restricted eigenvalue, and other related conditions.

The compatibility and restricted eigenvalue conditions are sometimes defined more generally in terms of the cardinality s of the index set S rather than a specific index set. For instance, [15, Assumption RE(s, c_0)] require for their restricted eigenvalue condition that the quantity

$$\kappa(s, c_0) := \min_{S \subseteq [p] : |S| \leq s} \min_{\substack{\boldsymbol{\delta} \neq \mathbf{0}: \\ \|\boldsymbol{\delta}_{S^\complement}\|_1 \leq c_0 \|\boldsymbol{\delta}_S\|_1}} \|\boldsymbol{M}\boldsymbol{\delta}\|_2 / (\sqrt{n} \|\boldsymbol{\delta}_S\|_2) \tag{4.6}$$

be bounded away from zero. The rationale for taking the minimum over all such index sets S is that the true support of $\boldsymbol{\beta}$ is unknown. See also the discussion of [45]. We note also that the compatibility and restricted eigenvalue conditions can be replaced by slightly weaker assumptions at the cost of more involved definitions [22].

4.2.2 Tuning parameters

Practical use of the first- and second-stage Lasso estimators requires selection of the tuning parameter r_j and $r_{\boldsymbol{\beta}}$ for $j \in [p_{\boldsymbol{x}}]$. A number of proposals for theoretical choices of tuning parameters [10, 11, 15, 17] for the ordinary (one-stage) linear model exist in the regularized regression literature. When the noise is homoscedastic and Gaussian, the optimal choice of tuning parameter depends on the noise level, and hence cannot be feasibly implemented without estimation of the latter. A second vein of literature concerns data-adaptive methods for the Lasso. Examples include (i) [41], who study the asymptotic properties of optimally tuned (with respect to mean square error) Lasso estimators in connection with solutions to adaptively tuned approximate message passing algorithms; (ii) AV_∞ of [20], who provide ℓ_∞ estimation error guarantees; (iii) $\mathrm{AV}_{\mathrm{Pr}}$ of [21], who provide guarantees for a post-lasso procedure under prediction error loss; (iv) stability selection for variable selection in [39] and subsequent work by [47]; (v) LinSelect of [7, 28]; and (vi) [13], who cite the oracle bounds of [36, 55, 56] to establish the asymptotic guarantees in prediction loss for the cross-validated highly adaptive Lasso (HAL) estimator.

A number of authors have also proposed modifications to the Lasso program that eliminate the need for some or all tuning of the penalty level with respect to model components. Examples include: (i) the square-root Lasso, group square-root Lasso, and scaled Lasso [6, 12, 18, 48, 50], for which calibration of the tuning parameter is pivotal with respect to the noise level; (ii) the TREX estimator [37, 16], which attempts to eliminate the need for tuning altogether; and (iii) [9], who derive oracle results for a weighted Lasso problem under general noise regimes by using concentration inequalities for self-normalized sums from [33].

In general, the bounds of Sections 4.2.3 and 4.2.4 are based on oracle choices of the tuning parameters $r_j, r_{\boldsymbol{\beta}}$, which depend explicitly on inestimable quantities. It would be preferable to give results for *data-adaptive* tuning parameters for which the respective Lasso problems can feasibly be implemented. For the present work, we are content to demonstrate that there exist sequences of oracle tuning parameters that tend to 0 sufficiently fast to ensure that the

remainder terms $\boldsymbol{f}_\ell$ are asymptotically negligible. In practice, cross-validated choices of Lasso tuning parameters and μ chosen according to the scheme described in Section 3.3 suffice in favorable parameter configurations. We provide evidence for this claim in Chapter 5.

We now present finite-sample bounds for the first- (Section 4.2.3) and second-stage (Section 4.2.4) Lasso estimators $\widehat{\mathbf{A}}$ and $\hat{\boldsymbol{\beta}}$.

4.2.3 First stage

The first stage is a high-dimensional multi-task regression problem. Recall that the model for $\boldsymbol{x}^j$ is given by

$$\boldsymbol{x}^j \;=\; \boldsymbol{Z}\boldsymbol{\alpha}^j + \boldsymbol{v}^j\,,$$

where $\boldsymbol{v}^j$ has nontrivial covariance with the noise $\boldsymbol{u}$. It suffices for our purposes to take a naïve approach to bounding the quantity $\|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1} = \max_{j\in[p_{\boldsymbol{x}}]} \|\hat{\boldsymbol{\alpha}}^j - \boldsymbol{\alpha}^j\|_1$. That is, we simultaneously bound the estimation error of each individual task. One could use a more complex approach such as [38] to treat different patterns of joint sparsity amongst the first-stage regression vectors.

The bounds of the present section require the following assumption.

Assumption 4.2.2 (First-stage compatibility conditions)**.** For each active set $S_j = \operatorname{supp} \boldsymbol{\alpha}^j$ of the first-stage model, there exists a constant $\phi_j > 0$ such that $\boldsymbol{Z}$ satisfies the compatibility condition with respect to S_j and ϕ_j. We write $\phi_{\mathbf{A}} := \max_{j\in[p_{\boldsymbol{x}}]} \phi_j$.

As stated in Definition 4.2.1, whether $\boldsymbol{Z}$ satisfies the compatibility condition with respect to one active set S_{j_1} does not bear directly on whether it satisfies the compatibility condition with respect to another active set S_{j_2} for $j_1, j_2 \in [p_{\boldsymbol{x}}]$. As such, it is non-trivial to assume that the compatibility condition as specified in Definition 4.2.1 holds for each active set S_j for $j \in [p_{\boldsymbol{x}}]$ when $p_{\boldsymbol{x}}$ tends to infinity. However, the condition that $\boldsymbol{Z}$ satisfies the compatibility condition with respect to each active set S_j is entailed by requiring that $\kappa(s, c_0)$ of (4.6) for $s = \max_{j\in[p_{\boldsymbol{x}}]} s_{\boldsymbol{\alpha}^j}$ and $c_0 = 3$ be bounded away from 0. Thus, the discussion of [15, Section 4] concerning the feasibility of their Assumption RE(s, c_0) applies as well to Assumption 4.2.2.

We will refer to the following specific choices of first-stage tuning parameters throughout the sequel.

Definition 4.2.3 (First-stage tuning parameters, Gaussian noise)**.** The following choices of the first-stage tuning parameters are appropriate for the Gaussian noise regime of Assumption 2.1.3:

$$r_j \;:=\; \sigma_{v^j} c \sqrt{\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty \log p_{\boldsymbol{z}}/n}$$

for each $j \in [p_{\boldsymbol{x}}]$, where c is a controlled quantity.

The following Lemma provides finite sample guarantees for $\|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1}$ under the choice of tuning parameters in Definition 4.2.3.

Lemma 4.2.4 (Bound for $\|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1}$, Gaussian noise)**.** *Suppose that Assumptions 4.2.2 and 2.1.3 hold. Set r_j according to Definition 4.2.3. Then,*

$$\mathrm{P}\big\{\|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1} > 4\frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2}\sigma_v c\sqrt{\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty (\log p_{\boldsymbol{z}})/n}\big\} \;\leq\; 2p_{\boldsymbol{z}}^{1-c_{\mathrm{ep}}}\,,$$

where $\sigma_v := \max_{j\in[p_{\boldsymbol{x}}]}\sigma_{v^j}$, $c_{\mathrm{ep}} := c^2/32 - 1$ and $c > 0$ is as specific in Definition 4.2.3.

Assuming that $\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty = O_{\mathrm{P}}(1)$ and that $\sigma_v,\ \phi_{\mathbf{A}}$ are bounded strictly away from zero and infinity uniformly in n, Lemma 4.2.4 entails that $\|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1} = O_{\mathrm{P}}(s_{\mathbf{A}}\sqrt{\log(p_{\boldsymbol{z}})/n})$, essentially identical to the Lasso rate for single task regression problems. The condition that $\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty = O_{\mathrm{P}}(1)$ follows from tail bounds for $\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}} - \boldsymbol{\Sigma}_{\boldsymbol{z}}\|_\infty$; such bounds can be derived under, say, a sub-Gaussianity assumption for the $\boldsymbol{z}_i$ according to the reasoning of Example 3.4.2.

We are primarily interested in controlling the maximum estimation error $\|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1}$ in order to establish bounds for the second-stage estimation error $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1$ in the following section.

4.2.4 Second stage

Simultaneous control of the first-stage estimation errors $\hat{\boldsymbol{\alpha}}^j - \boldsymbol{\alpha}^j$ is a straightforward consequence of the standard theoretical results for the Lasso. On the other hand, the bounds

for the second-stage estimation error $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$, which we study in the present section, are more involved due to the dependence of $\hat{\boldsymbol{\beta}}$ on the predicted conditional means $\hat{\boldsymbol{d}}_i$. Our strategy, which can be found in the Supplementary Materials, is to write $\boldsymbol{y} = \widehat{\boldsymbol{D}}\boldsymbol{\beta} + \tilde{\boldsymbol{u}}$, where $\tilde{\boldsymbol{u}} := \boldsymbol{u} + [(\boldsymbol{D} - \widehat{\boldsymbol{D}}) + \boldsymbol{V}]\boldsymbol{\beta}$ and apply concentration results to bound the probability of the event $\{4\|\widehat{\boldsymbol{D}}^\top \tilde{\boldsymbol{u}}\|_\infty \leq r_{\boldsymbol{\beta}}\}$, allowing us to adapt oracle inequality arguments for the Lasso to the present case.

As for the first-stage bounds, we require that $\widehat{\boldsymbol{D}}$ satisfy the compatibility condition:

Assumption 4.2.5 (Second-stage compatibility condition)**.** There exists a constant $\phi_{\boldsymbol{\beta}}$ such that $\widehat{\boldsymbol{D}}$ satisfies the compatibility condition with respect to $S_{\boldsymbol{\beta}}$ and $\phi_{\boldsymbol{\beta}}$.

We will refer to the following specific choices of first-stage tuning parameters throughout the sequel.

Definition 4.2.6 (Second-stage tuning parameter)**.** Define the second-stage tuning parameter $r_{\boldsymbol{\beta}}$ according to

$$r_{\boldsymbol{\beta}} := 16\frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2} r_{\mathbf{A}} \|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty \big(4m_{\boldsymbol{\beta}}\frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2} r_{\mathbf{A}} + m_{\mathbf{A}}\big) + \big(4\frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2} r_{\mathbf{A}} + m_{\mathbf{A}}\big)\big(m_{\boldsymbol{\beta}}\lambda_{\boldsymbol{V}} + \lambda_{\boldsymbol{u}}\big)\,, \tag{4.7}$$

where $\lambda_{\boldsymbol{V}}, \lambda_{\boldsymbol{u}} > 0$ are chosen according to the noise assumptions, $r_{\mathbf{A}} = \max_{j\in[p_{\boldsymbol{x}}]} r_j$, r_j is a tuning parameter for the respective first-stage problem, and $m_{\mathbf{A}}, m_{\boldsymbol{\beta}}$ are as defined in Assumption 2.1.1.

If the noise vectors u_i and $\boldsymbol{v}^j$ are Gaussian, the following choice of $\lambda_{\boldsymbol{V}}$ and $\lambda_{\boldsymbol{u}}$ lead to good behavior of $\hat{\boldsymbol{\beta}}$.

Definition 4.2.7 ($\lambda_{\boldsymbol{u}}, \lambda_{\boldsymbol{V}}$, Gaussian noise)**.** From Lemmas 4.4.3 and 4.4.4, the following choices of $\lambda_{\boldsymbol{V}}, \lambda_{\boldsymbol{u}}$ are appropriate under the noise regime of Assumption 2.1.3:

$$\lambda_{\boldsymbol{V}} := \sigma_{\boldsymbol{v}} c\sqrt{\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty \log p_{\boldsymbol{z}}/n}\,, \qquad \lambda_{\boldsymbol{u}} := \sigma_u c\sqrt{\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty \log p_{\boldsymbol{z}}/n}\,,$$

where $\sigma_{\boldsymbol{v}} := \max_{j\in[p_{\boldsymbol{x}}]} \sigma_{v^j}$ and c is as specified in Definition 4.2.3.

We now present probabilistic bounds for the ℓ_1 estimation error for the second-stage Lasso estimator $\hat{\boldsymbol{\beta}}$.

Lemma 4.2.8 (Bound for $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1$, Gaussian noise)**.** *Suppose that Assumptions 4.2.5 and 2.1.3 hold. For each $j \in [p_{\boldsymbol{x}}]$, set r_j according to Definition 4.2.3; set $\lambda_{\boldsymbol{V}}, \lambda_{\boldsymbol{u}}$ according to Definition 4.2.7; finally, set $r_{\boldsymbol{\beta}}$ according to Definition 4.2.6. Then,*

$$\begin{aligned}
&\mathrm{P}\Big\{ \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 > \\
&\quad 4\frac{s_{\boldsymbol{\beta}}}{\phi^2_{\boldsymbol{\beta}}}\Big(4\frac{s_{\mathbf{A}}}{\phi^2_{\mathbf{A}}}\sigma_v\big(m_{\boldsymbol{\beta}}\sigma_v[16\frac{s_{\mathbf{A}}}{\phi^2_{\mathbf{A}}}\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty + 1] + \sigma_u\big)c^2\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty \log(p_{\boldsymbol{z}})/n \\
&\qquad + m_{\mathbf{A}}\big(16\frac{s_{\mathbf{A}}}{\phi^2_{\mathbf{A}}}\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty\sigma_v + \sigma_u[m_{\boldsymbol{\beta}} + 1]\big)c\sqrt{\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty \log(p_{\boldsymbol{z}})/n}\Big)\Big\} \\
&\leq\ 2p_{\boldsymbol{z}}^{1-c_{\mathrm{ep}}} + 2p_{\boldsymbol{z}}^{-c_{\mathrm{ep}}}\,.
\end{aligned}$$

Assuming that $\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty = O_{\mathrm{P}}(1)$ and that $\sigma_v, \sigma_u, \phi_{\mathbf{A}}, \phi_{\boldsymbol{\beta}}$ are bounded strictly away from zero and infinity uniformly in n, Lemma 4.2.8 entails that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 = O_{\mathrm{P}}\big(s_{\boldsymbol{\beta}}s^2_{\mathbf{A}}\log(p_{\boldsymbol{z}})/n + s_{\boldsymbol{\beta}}s_{\mathbf{A}}\sqrt{\log(p_{\boldsymbol{z}})/n}\big)$. Thus, we see that the convergence rate of the second-stage Lasso estimator is slower than the typical rate of $s_{\boldsymbol{\beta}}\sqrt{\log(p_{\boldsymbol{x}})/n}$ in the ordinary Gaussian linear model. Whether the present rate can be significantly improved is a direction for future work.

4.2.5 Second-stage compatibility condition

Since Lemma 4.2.8 requires that Assumption 4.2.5 holds, it behooves us to demonstrate the latter's feasibility. The following lemma provides such a guarantee. For other approaches to studying the empirical compatibility constants and restricted eigenvalues of random matrices, see [45, 54]. Unlike the extant literature on the compatibility condition, however, we must account for the prediction error of $\widehat{\boldsymbol{D}}$.

Lemma 4.2.9 (Second-stage compatibility constant)**.** *Let $S \subseteq [p]$ be an arbitrary index set with $s = |S|$. For a given matrix $\boldsymbol{M} \in \mathbb{R}^{n\times p}$, define the quantity*

$$\phi^2_\dagger(\boldsymbol{M}, S)\ =\ \inf_{\delta\in\mathcal{C}(S)}\ \frac{s\|\boldsymbol{M}\boldsymbol{\delta}\|_2^2}{n\|\boldsymbol{\delta}_S\|_1^2}\,.$$

Let $\epsilon_1, \epsilon_2 > 0$ be arbitrary. Then,

$$\begin{aligned}
&\mathrm{P}\left\{\phi_\dagger^2(\widehat{\boldsymbol{D}}, S) < \Lambda_{\min}(\boldsymbol{\Sigma_d}) - \epsilon_2 - \epsilon_1\right\} \\
&\qquad \leq \ \mathrm{P}\big\{16s(2m_{\mathbf{A}}\|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1} + \|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1}^2)\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty > \epsilon_1\big\} \\
&\qquad\quad + \mathrm{P}\big\{16s\|\overline{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{\Sigma_d}\|_\infty > \epsilon_2\big\}\,,
\end{aligned}$$

where $\overline{\boldsymbol{\Sigma}}_{\boldsymbol{d}} = \boldsymbol{D}^\top \boldsymbol{D}/n$.

Lemma 4.2.9 may be combined with results for the maximum first-stage estimation error $\|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1}$ and the maximum entry-wise difference $\|\overline{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{\Sigma_d}\|_\infty$ to obtain specific bounds for $\phi_\dagger^2(\widehat{\boldsymbol{D}}, S_{\boldsymbol{\beta}})$ under different error and design matrix regimes. We present an example below.

Example 4.2.10 (Second-stage compatibility constant, Gaussian noise)**.** Suppose that Assumptions 4.2.2 and 2.1.3 hold. Set r_j according to Definition 4.2.3 for each $j \in [p_{\boldsymbol{x}}]$, and set $r_{\mathbf{A}} = \max_{j \in [p_{\boldsymbol{x}}]} r_j$. Set ϵ_1 in the statement of Lemma 4.2.9 as

$$\epsilon_1 \ = \ 128 s_{\boldsymbol{\beta}}(m_{\mathbf{A}} \frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2} r_{\mathbf{A}} + 2\frac{s_{\mathbf{A}}^2}{\phi_{\mathbf{A}}^4} r_{\mathbf{A}}^2)\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty\,,$$

so that

$$\mathrm{P}\big\{16 s_{\boldsymbol{\beta}}(2m_{\mathbf{A}}\|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1} + \|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1}^2)\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty > \epsilon_1\big\} \ \leq \ \mathrm{P}\big\{\|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1} > 4\frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2} r_{\mathbf{A}}\big\}\,.$$

Note that if $\frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2} r_{\mathbf{A}} = o(1)$ then, for n sufficiently large, we have $m_{\mathbf{A}} \frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2} r_{\mathbf{A}} + 2\big(\frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2}\big)^2 r_{\mathbf{A}}^2 \leq 3m_{\mathbf{A}} \frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2} r_{\mathbf{A}}$, which simplifies the foregoing display. Applying this bound and substituting the present choice of tuning parameters into the displays above, we conclude from Lemmas 4.2.4 and 4.2.9 that

$$\begin{aligned}
&\mathrm{P}\Big\{\phi_\dagger^2(\widehat{\boldsymbol{D}}, S_{\boldsymbol{\beta}}) < \Lambda_{\min}(\boldsymbol{\Sigma_d}) - \epsilon_2 - 384 m_{\mathbf{A}} s_{\boldsymbol{\beta}} \frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2} \sigma_v c \|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty^{3/2} \sqrt{\log(p_{\boldsymbol{z}})/n}\Big\} \\
&\qquad\qquad \leq 2p_{\boldsymbol{z}}^{1-c_{\mathrm{ep}}} + \mathrm{P}\big\{16 s_{\boldsymbol{\beta}}\|\overline{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{\Sigma_d}\|_\infty > \epsilon_2\big\}\,.
\end{aligned}$$

Thus, if (i) $s_{\boldsymbol{\beta}} \frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2} \sqrt{\log(p_{\boldsymbol{z}})/n} = o(1)$, (ii) $\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty = O_{\mathrm{P}}(1)$, and (iii) $s_{\boldsymbol{\beta}}\|\overline{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{\Sigma_d}\|_\infty = o_{\mathrm{P}}(1)$, we may conclude that the sequence $\phi_\dagger^2(\widehat{\boldsymbol{D}}, S_{\boldsymbol{\beta}})$ is bounded strictly away from zero with

high probability. We note again that the latter two conditions may be derived from, say, sub-Gaussian rates for $\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}} - \boldsymbol{\Sigma}_{\boldsymbol{z}}\|_\infty$ combined with bounds on the minimal and maximal eigenvalues of $\boldsymbol{\Sigma}_{\boldsymbol{z}}$ and growth restrictions on $s_{\boldsymbol{\beta}}\sqrt{\log(p_{\boldsymbol{z}})/n}$.

4.3 Remainder terms

The asymptotic results of Section 3.4 depend on the high-level assumption that the remainder terms $\boldsymbol{f}_\ell$ satisfy $\|\boldsymbol{f}_\ell\|_\infty = o_{\mathrm{P}}(1)$. In this section, we identify the specific conditions under which this assumption is satisfied for the two-stage Lasso. The primary end of these conditions, which we present in Assumption 4.3.1 below, is to ensure the ℓ_1 consistency of the first- and second-stage estimators and of the estimator $\widehat{\boldsymbol{\Theta}}$ specified in Section 3.3.

Assumption 4.3.1 (Model regularity)**.** The following regularity conditions are used in various combinations throughout Lemmas 4.4.10-4.4.16. The combinations are made explicit in the statement of each lemma.

1. The instrumental variable design matrices $\boldsymbol{Z} \equiv \boldsymbol{Z}_n$ satisfy the compatibility condition relative to the first-stage active sets $S_j \equiv S_{j,n} = \operatorname{supp}(\boldsymbol{\alpha}_n^j)$ and sequences of compatibility constants $\phi_j \equiv \phi_{j,n}$ strictly bounded away from zero and infinity uniformly in n;

2. The predicted conditional mean matrices $\widehat{\boldsymbol{D}} \equiv \widehat{\boldsymbol{D}}_n$ satisfy the compatibility condition relative to the second-stage active sets $S_{\boldsymbol{\beta}} \equiv S_{\boldsymbol{\beta},n} = \operatorname{supp}(\boldsymbol{\beta}_n)$ and a sequence of compatibility constants $\phi_{\boldsymbol{\beta}} \equiv \phi_{\boldsymbol{\beta},n}$ strictly bounded away from zero and infinity uniformly in n;

3. The growth condition $\max_{j\in[p_{\boldsymbol{x}}]} s_{\boldsymbol{\alpha}^j} r_j = o(1)$ holds;

4. The sequence of minimal and maximal eigenvalues of $\boldsymbol{\Sigma}_{\boldsymbol{d}} \equiv \boldsymbol{\Sigma}_{\boldsymbol{d},n}$, denoted respectively by $\Lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{d}})$ and $\Lambda_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{d}})$, are bounded away from zero and infinity uniformly in n;

5. The sequence of population quantities $\boldsymbol{\Theta} \equiv \boldsymbol{\Theta}_n$ satisfies $\boldsymbol{\Theta} \in \mathcal{U}(m_{\boldsymbol{\Theta}}, q, s_{\boldsymbol{\Theta}})$ for a universal constant $m_{\boldsymbol{\Theta}}$, $s_{\boldsymbol{\Theta}} \equiv s_{\boldsymbol{\Theta},n} > 0$ and controlled $q \in [0, 1)$;

6. The condition $\mathrm{P}\,\mathcal{T}_{\boldsymbol{\Theta}}(\mu)^{\complement} = o(1)$ holds;

7. The quantity $\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty = \|\mathbb{E}_n[\boldsymbol{z}_i \boldsymbol{z}_i^\top]\|_\infty$ satisfies $\lim_{n\to\infty} \mathrm{P}\{\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty > m_{\boldsymbol{Z}}\} = 0$ for a universal constant $m_{\boldsymbol{Z}}$;

8. The following growth conditions hold:

   (a) $\mu^{1-q} s_{\boldsymbol{\Theta}} \sqrt{\log p_{\boldsymbol{z}}} = o(1)$;

   (b) $s_{\mathbf{A}}^3 s_{\boldsymbol{\beta}} (\log p_{\boldsymbol{z}})^{3/2}/n + s_{\mathbf{A}}^2 s_{\boldsymbol{\beta}} (\log p_{\boldsymbol{z}})/\sqrt{n} = o(1)$;

   (c) $\mu s_{\boldsymbol{\beta}} \big(s_{\mathbf{A}}^2 \log p_{\boldsymbol{z}}/\sqrt{n} + s_{\mathbf{A}} \sqrt{\log p_{\boldsymbol{z}}}\big) = o(1)$.

Conditions 1 and 2 are prerequisites for the bounds on the first- and second-stage estimation errors of Sections 4.2.3 and 4.2.4; we discuss the feasibility of these assumptions there. Condition 3 is required for asymptotic control of the remainder terms. Condition 4 is not directly required for the bounds of Lemmas 4.4.10-4.4.16, but it is required for control of $\mathrm{P}\,\mathcal{T}_{\boldsymbol{\Theta}}(\mu)$ under the assumptions of Example 3.3.3; as noted in [11], it is a standard assumption in econometric models, for example. Condition 5 is required to control $\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j$ under ℓ_∞ and ℓ_1 norms as discussed in Section 3.3. Condition 6 is a high-level requirement for asymptotic negligibility of the remainder terms $\boldsymbol{f}_\ell$; it can be obtained as a consequence of specific model assumptions as in Example 3.3.3 with consequences for the required growth rates of model parameters as in Example 4.3.4. Condition 7 is similarly a high-level condition required for asymptotic negligibility of the remainder terms: it ensures that the empirical quantity $\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty = \|\boldsymbol{Z}^\top \boldsymbol{Z}/n\|_\infty$ behaves in probability as of constant order. It can be derived as a consequence of Condition 4 if $\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}} - \boldsymbol{\Sigma}_{\boldsymbol{z}}\|_\infty = o_{\mathrm{P}}(1)$; the latter condition can in turn be derived from distributional assumptions on the $\boldsymbol{z}_i$ as in Example 3.4.2. Condition 8 lists the model parameter growth conditions required for asymptotic negligibility of the remainder terms under the Gaussian noise regime of Assumption 2.1.3; these conditions should be

compared with the requirement $s \log p / \sqrt{n} = o(1)$ in [32, 53] for negligibility of the single remainder term that occurs under the ordinary linear model.

Note that while the quantity $m_{\boldsymbol{Z}}$ of Condition 7 appears in the bounds of Lemmas 4.4.10-4.4.16 of the Supplementary Materials, which give the rates for the remainder terms, we do not include it in the growth conditions of Condition 8. This is because, under the presently studied regime, $m_{\boldsymbol{Z}}$ is assumed of constant order. One could consider more general scenarios where the maximum entry of $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}$ is not bounded in probability and include the quantity $m_{\boldsymbol{Z}}$ in the aforementioned growth conditions. Doing so would in turn affect the rate at which $s_{\mathbf{A}}, s_{\boldsymbol{\beta}}$, and $p_{\boldsymbol{z}}$ may be allowed to grow with n while maintaining asymptotic negligibility of the remainder terms $\boldsymbol{f}_\ell$. Similar considerations come to bear on the quantities $m_{\mathbf{A}}, m_{\boldsymbol{\beta}}, m_{\boldsymbol{\Theta}}$, which we assume constant in the present essay for simplicity but could be allowed to vary and therefore figure into the growth conditions requisite for asymptotic negligibility of the remainder terms.

In addition to the model regularity conditions of Assumption 4.3.1, we require appropriate choices of the first- and second-stage Lasso tuning parameters and the estimator $\widehat{\boldsymbol{\Theta}}$. We codify such choices in the following Assumption.

Assumption 4.3.2 (Specification of estimators)**.** Let $\widehat{\mathbf{A}}$ and $\hat{\boldsymbol{\beta}}$ be the first- and second-stage Lasso estimators, respectively. The tuning parameters under the Gaussian noise regime of Assumption 2.1.3 are chosen according to

1. Definition 4.2.3 for the first-stage tuning parameters $\boldsymbol{r} = (r_j)_{j=1}^{p_{\boldsymbol{x}}}$ and (4.2) for the quantity $r_{\mathbf{A}}$;

2. Definition 4.2.7 for the quantities $\lambda_{\boldsymbol{u}}$ and $\lambda_{\boldsymbol{V}}$;

3. Definition 4.2.6 for the second-stage tuning parameter $r_{\boldsymbol{\beta}}$,

and let $\widehat{\boldsymbol{\Theta}}$ be an estimator of $\boldsymbol{\Theta}$ with rows $\hat{\boldsymbol{\theta}}_j$ given by solutions to Program 3.3.1.

We can now conclude the asymptotic negligibility of the remainder terms $\boldsymbol{f}_\ell$.

Lemma 4.3.3 (Negligibility of remainders)**.** *Suppose that Assumption 2.1.3 and Conditions 1-8 of Assumption 4.3.1 hold and that the estimators* $\widehat{\mathbf{A}}, \hat{\boldsymbol{\beta}}$, *and* $\widehat{\boldsymbol{\Theta}}$ *are chosen according to Assumption 4.3.2. Then,* $\|\boldsymbol{f}_\ell\|_\infty = o_{\mathrm{P}}(1)$ *for* $\ell \in \{1, 2, 3, 4\}$.

The primary use of Lemma 4.3.3 is to verify Condition 2 of Theorem 3.4.1. Indeed, the result justifies the use of the one-step update to the second-stage Lasso estimator to construct asymptotically valid confidence intervals for the components β_j according to (3.11).

We note that the quantity μ, which we recall is the tolerance parameter for Program 3.3.1, must be given careful consideration. Consider the growth rates specified in Condition 8 of Assumption 4.3.1. After excluding terms of constant order, we see that Conditions 8a and 8b require μ to be of small order $(s_{\boldsymbol{\Theta}}\sqrt{\log p_{\boldsymbol{z}}})^{\frac{1}{q-1}}$ and $(s_{\boldsymbol{\beta}} s_{\mathbf{A}} \log p_{\boldsymbol{z}})^{-1}$, respectively. However, μ must not tend to 0 so fast that the probability that $\boldsymbol{\Theta}$ is feasible for Program 3.3.1, which we recall is formally denoted by $\mathrm{P}\,\mathcal{T}_{\boldsymbol{\Theta}}(\mu)$, becomes bounded away from zero. The growth of μ must therefore balance two competing objectives. Since this feat is non-trivial, we present the following example of a model specification for which a feasible choice of μ can be theoretically justified.

Example 4.3.4 (Example 3.3.3, cont'd, Gaussian noise)**.** We continue the consideration of sub-Gaussian conditional mean matrix $\boldsymbol{D}$. Suppose that: (i) the conditions of Lemma 4.3.3 hold; (ii) the matrices $\boldsymbol{D}$ and $\boldsymbol{\Theta}$ are as specified in Example 3.3.3. As in that example, it holds for $\epsilon > 0$ and $n \geq (a^2 s_{\mathbf{A}}^2 \sigma_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{d}}) \log p_{\boldsymbol{x}})/(4e^2 \sigma_{\max}(\boldsymbol{\Sigma}_{\boldsymbol{d}})\tau^4)$ that

$$\begin{aligned} &\mathrm{P}\big\{\|\boldsymbol{\Theta}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{I}\|_\infty > a\sqrt{\log p_{\boldsymbol{x}}/n} + 3m_{\boldsymbol{\Theta}} m_{\mathbf{A}} \|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty \epsilon\big\} \\ &\quad \leq 2p_{\boldsymbol{x}}^{-c_{\mathrm{inv}}} + \mathrm{P}\big\{\|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1} > \epsilon\big\}, \end{aligned}$$

where c_{inv} and the controlled quantity a are as in Example 3.3.3. Thus, if $s_{\mathbf{A}}\sqrt{\log p_{\boldsymbol{x}}/n} =$

$o(1)$, then Lemma 4.2.4 entails that

$$\begin{aligned}
&\mathrm{P}\big\{\|\boldsymbol{\Theta}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{I}\|_\infty > a\sqrt{\log p_{\boldsymbol{x}}/n} + 12 m_{\boldsymbol{\Theta}} m_{\mathbf{A}} c(\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty)^{\frac{3}{2}} \sigma_v \frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2}\sqrt{\log p_{\boldsymbol{z}}/n}\big\} \\
&\quad \leq\ 2p_{\boldsymbol{x}}^{-c_{\mathrm{inv}}} + \mathrm{P}\big\{\|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1} > \sigma_v \frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2}\sqrt{\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty \log p_{\boldsymbol{z}}/n}\big\} \\
&\quad \leq\ 2p_{\boldsymbol{x}}^{-c_{\mathrm{inv}}} + 2p_{\boldsymbol{z}}^{1-c_{\mathrm{ep}}}
\end{aligned}$$

for n sufficiently large, where c_{ep} is as in Lemma 4.4.2. If there exists $m_{\boldsymbol{Z}} = O(1)$ that satisfies $\mathrm{P}\{\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty > m_{\boldsymbol{Z}}\} = o(1)$, then setting the tolerance μ for Program 3.3.1 according to

$$\begin{aligned}
\mu\ &=\ a\sqrt{\log p_{\boldsymbol{x}}/n} + 12 m_{\boldsymbol{\Theta}} m_{\mathbf{A}} c(m_{\boldsymbol{Z}})^{\frac{3}{2}} \sigma_v \frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2}\sqrt{\log p_{\boldsymbol{z}}/n} \\
&\lesssim\ s_{\mathbf{A}}\sqrt{\log p_{\boldsymbol{z}}/n}
\end{aligned}$$

under the noise regime of Assumption 2.1.3 yields

$$\mathrm{P}\big\{\|\boldsymbol{\Theta}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{I}\|_\infty > \mu\big\}\ \leq\ 2p_{\boldsymbol{x}}^{-c_{\mathrm{inv}}} + 2p_{\boldsymbol{z}}^{1-c_{\mathrm{ep}}} + \mathrm{P}\{\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty > m_{\boldsymbol{Z}}\}\,,$$

and hence that the population quantity $\boldsymbol{\Theta}$ is feasible for Program 3.3.1 with probability approaching one. Condition 8a of Assumption 4.3.1 then becomes

$$s_{\boldsymbol{\Theta}} s_{\mathbf{A}}^{1-q} (\log p_{\boldsymbol{z}})^{1-\frac{q}{2}} / n^{\frac{1-q}{2}}\ =\ o(1)\,, \tag{4.8}$$

and Condition 8c becomes

$$s_{\boldsymbol{\beta}} s_{\mathbf{A}}^3 (\log p_{\boldsymbol{z}})^{3/2}/n + s_{\boldsymbol{\beta}} s_{\mathbf{A}}^2 \log(p_{\boldsymbol{z}})/\sqrt{n}\ =\ o(1)\,, \tag{4.9}$$

which is identical to Condition 8b.

4.4 Materials required for Chapter 4

4.4.1 Materials required for Section 4.2

Our guarantees for estimating $\mathbf{A}$ and $\boldsymbol{\beta}$ consist of two parts. The first is the *oracle inequality*, which bounds the ℓ_1 estimation error of a generic Lasso estimator conditional on the occurrence of a special set $\mathcal{T}$. The oracle inequality is a fixture of the ℓ_1 regularized estimation literature; see for instance [17, Chapter 6]. We present it for the sake of completeness.

The oracle inequality itself is specific to neither the first- nor the second- stage estimators of the present work. Indeed, we require the result to derive bounds for both estimators. As such, we present the theorem in terms of a generic model that shares notation with neither the first- nor second- stage models described in Section 2.1 except for the number of observations n.

Theorem 4.4.1 (Oracle inequality)**.** *Consider the generic linear model*

$$\boldsymbol{g} \;=\; \boldsymbol{W}\boldsymbol{\gamma} + \boldsymbol{h}\,,$$

where $\boldsymbol{g} \in \mathbb{R}^n$ is a vector of univariate responses, $\boldsymbol{W} \in \mathbb{R}^{n\times p}$ is a design matrix with rows $\boldsymbol{w}_i$, $\boldsymbol{\gamma} \in \mathbb{R}^n$ is a noise vector with arbitrary distribution. Let $\hat{\boldsymbol{\gamma}}$ denote the Lasso estimator given by

$$\hat{\boldsymbol{\gamma}} \;\in\; \operatorname*{arg\,min}_{\boldsymbol{a}\in\mathbb{R}^p} \left\{ \|\boldsymbol{g} - \boldsymbol{W}\boldsymbol{a}\|_2^2/(2n) + r\|\boldsymbol{a}\|_1 \right\}.$$

Let $S_{\boldsymbol{\gamma}} := \operatorname{supp}\boldsymbol{\gamma}$, and let $s_{\boldsymbol{\gamma}} := |S_{\boldsymbol{\gamma}}|$. Suppose that $\boldsymbol{W} \in \mathcal{G}(S_{\boldsymbol{\gamma}}, \phi_{\boldsymbol{\gamma}})$ for a $\phi_{\boldsymbol{\gamma}} > 0$. Then, on the set

$$\mathcal{T}(r) \;:=\; \left\{ 4\|\boldsymbol{W}^\top \boldsymbol{h}/n\|_\infty \le r \right\}, \tag{4.10}$$

the bound

$$\|\boldsymbol{W}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})\|_2^2/n + r\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_1 \;\le\; 4s_{\boldsymbol{\gamma}} r^2/\phi_{\boldsymbol{\gamma}}^2$$

holds.

Proof of Theorem 4.4.1. The proof is algebra. To begin, note that the specification of $\hat{\boldsymbol{\gamma}}$ in the statement of the present theorem entails that

$$\|\boldsymbol{g} - \boldsymbol{W}\hat{\boldsymbol{\gamma}}\|_2^2/n + r\|\hat{\boldsymbol{\gamma}}\|_1 \;\le\; \|\boldsymbol{g} - \boldsymbol{W}\boldsymbol{\gamma}\|_2^2/n + r\|\boldsymbol{\gamma}\|_1\,.$$

Substituting $\boldsymbol{W}\boldsymbol{\gamma} + \boldsymbol{h}$ for $\boldsymbol{g}$ yields

$$\|\boldsymbol{W}(\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}}) + \boldsymbol{h}\|_2^2/n + r\|\hat{\boldsymbol{\gamma}}\|_1 \;\le\; \|\boldsymbol{h}\|_2^2/n + r\|\boldsymbol{\gamma}\|_1\,.$$

On the left-hand side of the above we have

$$\|\boldsymbol{W}(\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}}) + \boldsymbol{h}\|_2^2 \;=\; \|\boldsymbol{W}(\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}})\|_2^2 + 2\langle \boldsymbol{W}(\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}})\,, \boldsymbol{h}\rangle + \|\boldsymbol{h}\|_2^2\,,$$

where $\langle \cdot\,, \cdot \rangle$ denotes the standard inner product over Euclidean space. As the first term on the right-hand side above is squared, we may swap the order of $\boldsymbol{\gamma}$ and $\hat{\boldsymbol{\gamma}}$ therein. Combining the two displays above and simplifying then yields

$$\left[\|\boldsymbol{W}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})\|_2^2 + 2\langle \boldsymbol{W}(\boldsymbol{\gamma} - \hat{\boldsymbol{\gamma}})\,, \boldsymbol{h}\rangle\right]/n + r\|\hat{\boldsymbol{\gamma}}\|_1 \;\leq\; r\|\boldsymbol{\gamma}\|_1\,.$$

Rearranging terms in the above display leads to the following *basic inequality*:

$$\|\boldsymbol{W}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})\|_2^2/n + r\|\hat{\boldsymbol{\gamma}}\|_1 \;\leq\; 2\big\langle \boldsymbol{W}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}),\, \boldsymbol{h}\big\rangle/n + r\|\boldsymbol{\gamma}\|_1\,. \tag{4.11}$$

Apply Hölder's inequality to the modulus of the first term on the right-hand side above to find

$$\begin{aligned} 2\big\langle \boldsymbol{W}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}),\, \boldsymbol{h}\big\rangle/n \;&\leq\; 2\|\boldsymbol{W}^\top \boldsymbol{h}/n\|_\infty \|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_1 \\ &\leq\; \frac{1}{2} r\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_1\,, \end{aligned}$$

where the latter inference holds on the set $\mathcal{T}(r)$. Substitute the above display into the basic inequality of (4.11) and double the result to find that

$$2\|\boldsymbol{W}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})\|_2^2/n + 2r\|\hat{\boldsymbol{\gamma}}\|_1 \;\leq\; r\|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_1 + 2r\|\boldsymbol{\gamma}\|_1\,.$$

Now rearrange terms in the reverse triangle inequality

$$\|(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})_{S_{\boldsymbol{\gamma}}}\|_1 \;=\; \|\hat{\boldsymbol{\gamma}}_{S_{\boldsymbol{\gamma}}} - \boldsymbol{\gamma}_{S_{\boldsymbol{\gamma}}}\|_1 \;\geq\; \|\boldsymbol{\gamma}_{S_{\boldsymbol{\gamma}}}\|_1 - \|\hat{\boldsymbol{\gamma}}_{S_{\boldsymbol{\gamma}}}\|_1$$

to obtain that

$$\|\hat{\boldsymbol{\gamma}}_{S_{\boldsymbol{\gamma}}}\|_1 \;\geq\; \|\boldsymbol{\gamma}_{S_{\boldsymbol{\gamma}}}\|_1 - \|(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})_{S_{\boldsymbol{\gamma}}}\|_1$$

and hence that

$$\begin{aligned} \|\hat{\boldsymbol{\gamma}}\|_1 \;&=\; \|\hat{\boldsymbol{\gamma}}_{S_{\boldsymbol{\gamma}}}\|_1 + \|\hat{\boldsymbol{\gamma}}_{S_{\boldsymbol{\gamma}}^{\complement}}\|_1 \\ &\geq\; \|\boldsymbol{\gamma}_{S_{\boldsymbol{\gamma}}}\|_1 - \|(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})_{S_{\boldsymbol{\gamma}}}\|_1 + \|\hat{\boldsymbol{\gamma}}_{S_{\boldsymbol{\gamma}}^{\complement}}\|_1\,. \end{aligned}$$

Substitute the above display into the most recent derivation from the basic inequality to find

$$\begin{aligned} & 2\|\boldsymbol{W}(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})\|_2^2 + 2r\big[\|\boldsymbol{\gamma}_{S_{\boldsymbol{\gamma}}}\|_1 - \|(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})_{S_{\boldsymbol{\gamma}}}\|_1 + \|\hat{\boldsymbol{\gamma}}_{S_{\boldsymbol{\gamma}}^{\mathfrak{C}}}\|_1\big] \\ \leq\ & r\|\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}\|_1 + 2r\|\boldsymbol{\gamma}\|_1 \\ =\ & r\big[\|(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})_{S_{\boldsymbol{\gamma}}}\|_1 + \|(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})_{S_{\boldsymbol{\gamma}}^{\mathfrak{C}}}\|_1\big] + 2r\big[\|\boldsymbol{\gamma}_{S_{\boldsymbol{\gamma}}}\|_1 + \|\boldsymbol{\gamma}_{S_{\boldsymbol{\gamma}}^{\mathfrak{C}}}\|\big] \\ =\ & r\big[\|(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})_{S_{\boldsymbol{\gamma}}}\|_1 + \|\hat{\boldsymbol{\gamma}}_{S_{\boldsymbol{\gamma}}^{\mathfrak{C}}}\|_1\big] + 2r\|\boldsymbol{\gamma}_{S_{\boldsymbol{\gamma}}}\|_1\,, \end{aligned}$$

where, to infer the final line, recall that $\|\boldsymbol{\gamma}_{S_{\boldsymbol{\gamma}}^{\mathfrak{C}}}\|_1 = 0$ by specification of $S_{\boldsymbol{\gamma}}$. Simplify the above display by consolidating terms to find

$$2\|\boldsymbol{W}(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})\|_2^2/n + r\|\hat{\boldsymbol{\gamma}}_{S_{\boldsymbol{\gamma}}^{\mathfrak{C}}}\|_1 \leq 3r\|(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})_{S_{\boldsymbol{\gamma}}}\|_1\,.$$

From the fact that the components of $\boldsymbol{\gamma}$ vanish on $S_{\boldsymbol{\gamma}}$ and from the decomposability of the ℓ_1 norm, write

$$\begin{aligned} \|\hat{\boldsymbol{\gamma}}_{S_{\boldsymbol{\gamma}}^{\mathfrak{C}}}\|_1 &= \|(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})_{S_{\boldsymbol{\gamma}}^{\mathfrak{C}}}\|_1 \\ &= \|\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}\|_1 - \|(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})_{S_{\boldsymbol{\gamma}}}\|_1\,. \end{aligned}$$

Now substitute $\|\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}\|_1 - \|(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})_{S_{\boldsymbol{\gamma}}}\|_1$ for $\|\hat{\boldsymbol{\gamma}}_{S_{\boldsymbol{\gamma}}^{\mathfrak{C}}}\|_1$ two displays previous and consolidate terms to obtain

$$2\|\boldsymbol{W}(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})\|_2^2/n + r\|(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})\|_1 \leq 4r\|(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})_{S_{\boldsymbol{\gamma}}}\|_1\,.$$

Write

$$\|(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})\|_1 = \|(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})_{S_{\boldsymbol{\gamma}}}\|_1 + \|(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})_{S_{\boldsymbol{\gamma}}^{\mathfrak{C}}}\|_1$$

and combine with two displays previous to conclude that

$$\|(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})_{S_{\boldsymbol{\gamma}}^{\mathfrak{C}}}\|_1 \leq 3\|(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})_{S_{\boldsymbol{\gamma}}}\|_1\,.$$

The above display is the hypothesis of the compatibility condition, which holds for $\boldsymbol{W}$ relative to $S_{\boldsymbol{\gamma}}$ and $\phi_{\boldsymbol{\gamma}} > 0$ by assumption. Invoke the compatibility condition to find

$$\|(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})_{S_{\boldsymbol{\gamma}}}\|_1 \leq \sqrt{s_{\boldsymbol{\gamma}}}\|\boldsymbol{W}(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})\|_2/(\sqrt{n}\phi_{\boldsymbol{\gamma}})\,.$$

Next, write

$$\begin{aligned} 2\|\boldsymbol{W}(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})\|_2^2/n + r\|(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})\|_1 &\leq 4r\sqrt{s_{\boldsymbol{\gamma}}}\|\boldsymbol{W}(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})\|_2/(\sqrt{n}\phi_{\boldsymbol{\gamma}}) \\ &\leq \|\boldsymbol{W}(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})\|_2^2/n + 4s_{\boldsymbol{\gamma}}r^2/\phi_{\boldsymbol{\gamma}}^2\,, \end{aligned}$$

where the last step follows from the inequality $4xy \leq x^2 + 4y^2$ for $x, y \in \mathbb{R}$. Subtract $\|\boldsymbol{W}(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})\|_2^2/n$ from both sides to find

$$\|\boldsymbol{W}(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})\|_2^2/n + r\|\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}\|_1 \leq 4s_{\boldsymbol{\gamma}}r^2/\phi_{\boldsymbol{\gamma}}^2\,,$$

as claimed. □

Theorem 4.4.1 provides a deterministic guarantee for the ℓ_1 estimation error of a generic Lasso estimator $\hat{\boldsymbol{\gamma}}$ on the set $\mathcal{T}(r) = \big\{4\|\boldsymbol{W}^\top\boldsymbol{h}/n\|_\infty \leq r\big\}$. Consequently, it holds that

$$\mathrm{P}\big\{\|\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}\|_1 > 4s_{\boldsymbol{\gamma}}r^2/\phi_{\boldsymbol{\gamma}}^2\big\} \leq \mathrm{P}\,\mathcal{T}(r)^{\complement}\,.$$

The quantity $\|\boldsymbol{W}^\top\boldsymbol{h}/n\|_\infty$ is sometimes called the *empirical process term*; for instance, [17, Chapter 6].

It follows from the previous display that upper bounds for $\mathrm{P}\,\mathcal{T}(r)^{\complement}$ yield probabilistic guarantees for the ℓ_1 estimation error.

The probabilistic guarantees for Lasso estimation performance require that the tuning parameter dominate the empirical process term. In our consideration of both the first- and second- stage Lasso estimators, we encounter a number of such terms, each of the form $\|\boldsymbol{Z}^\top\boldsymbol{h}/n\|_\infty$ for various noise vectors $\boldsymbol{h}$. As such, we formulate the following lemma, which is used throughout our consideration of the homoscedastic Gaussian error regime, in terms of a generic Gaussian vector $\boldsymbol{h}$ with i.i.d. components. The lemma itself is a standard application of basic concentration results for Gaussian random variables. In the subsequent corollaries, we derive bounds for various empirical process terms by taking $\boldsymbol{h}$ to be, for instance, $\boldsymbol{u}$ and $\boldsymbol{v}^j$ for $j \in [p_{\boldsymbol{x}}]$. Such bounds are key ingredients of the results presented in Sections 4.2.3 and 4.2.4.

Lemma 4.4.2 (Control of $\|\boldsymbol{Z}^\top \boldsymbol{h}/n\|_\infty$, Gaussian noise)**.** *Let*

$$\boldsymbol{h} \,|\, \boldsymbol{Z} \sim \mathcal{N}_n(\boldsymbol{0}, \tau^2 \boldsymbol{I}) \,.$$

Then,

$$\mathrm{P}\big\{4\|\boldsymbol{Z}^\top \boldsymbol{h}/n\|_\infty > \tau c \sqrt{\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty \log p_{\boldsymbol{z}}/n}\big\} \;\le\; 2p_{\boldsymbol{z}}^{-c_{\mathrm{ep}}} \,,$$

where $c_{\mathrm{ep}} := c^2/32 - 1$ *and* $c > 0$ *is a controlled constant.*

Proof of Lemma 4.4.2. The proof is similar to the treatments of [32, Theorem 6] and [17, Lemma 6.2]. For each $j \in [p_{\boldsymbol{x}}]$, write

$$H_j := \langle \boldsymbol{z}^j, \boldsymbol{h} \rangle / (\|\boldsymbol{z}^j\|_2 \tau) \,.$$

It follows from the specification of $\boldsymbol{h}$ that $H_j \,|\, \boldsymbol{z}^j \sim \mathcal{N}(0,1)$. Observe that

$$\begin{aligned}
&\mathrm{P}\big\{|\langle \boldsymbol{z}^j, \boldsymbol{h} \rangle/n| > \tau \sqrt{\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty (t^2 + 2\log p_{\boldsymbol{z}})/n}\big\} \\
&\qquad \le \mathrm{P}\big\{|\langle \boldsymbol{z}^j, \boldsymbol{h} \rangle/n| > \|\boldsymbol{z}^j/\sqrt{n}\|_2 \tau \sqrt{(t^2 + 2\log p_{\boldsymbol{z}})/n}\big\} \\
&\qquad = \mathrm{P}\big\{|H_j| > \sqrt{t^2 + 2\log p_{\boldsymbol{z}}}\big\} \\
&\qquad = \mathrm{E}\big[\mathrm{P}\big\{|H_j| > \sqrt{t^2 + 2\log p_{\boldsymbol{z}}} \,\big|\, \boldsymbol{z}^j\big\}\big] \\
&\qquad = \mathrm{E}\big[2\mathrm{P}\big\{H_j > \sqrt{t^2 + 2\log p_{\boldsymbol{z}}} \,\big|\, \boldsymbol{z}^j\big\}\big] \\
&\qquad \le \mathrm{E}\big[2e^{-(t^2 + 2\log p_{\boldsymbol{z}})/2}\big] \;=\; 2e^{-(t^2 + 2\log p_{\boldsymbol{z}})/2} \,,
\end{aligned}$$

where the final line is just the Chernoff bound for $\mathcal{N}(0,1)$ random variables. Take the union bound over $j \in [p_{\boldsymbol{x}}]$ to find

$$\begin{aligned}
&\mathrm{P}\Big\{ \max_{j \in [p_{\boldsymbol{x}}]} |\langle \boldsymbol{z}^j, \boldsymbol{u} \rangle/n| > \tau \sqrt{\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty (t^2 + 2\log p_{\boldsymbol{z}})/n} \Big\} \\
&\qquad \le 2p_{\boldsymbol{x}} e^{-(t^2 + 2\log p_{\boldsymbol{z}})/2} \;=\; 2e^{-t^2/2} \,.
\end{aligned}$$

To conclude, choose t so that $e^{t^2/2} = p_{\boldsymbol{z}}^{c_{\mathrm{ep}}}$ and observe that

$$\begin{aligned}
4\sqrt{(t^2 + 2\log p_{\boldsymbol{z}})/n} \;&=\; 4\tau \sqrt{2\log(p_{\boldsymbol{z}}^{c_{\mathrm{ep}}+1})/n} \\
&= 4\tau \sqrt{(c^2/16)\log(p_{\boldsymbol{z}})/n} \;=\; \tau c \sqrt{\log p_{\boldsymbol{z}})/n} \,.
\end{aligned}$$

□

Lemmas 4.4.3-4.4.5 follow from Lemma 4.4.2 and are required throughout the results for the first- and second-stage estimation errors.

Lemma 4.4.3 (Control of $\|\boldsymbol{Z}^\top \boldsymbol{u}/n\|_\infty$, Gaussian noise)**.** *Suppose that $\boldsymbol{u}$ satisfies Assumption 2.1.3. Define the set*

$$\mathcal{T}_{\boldsymbol{u}}(\lambda) \,:=\, \{\|\boldsymbol{Z}^\top \boldsymbol{u}/n\|_\infty \leq \lambda\} \,=\, \{\|\boldsymbol{Z}^\top \boldsymbol{u}/n\|_\infty \leq \lambda\}\,, \tag{4.12}$$

where λ is a controlled quantity. Then

$$\mathrm{P}\,\mathcal{T}_{\boldsymbol{u}}\big(c\sigma_u\sqrt{\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty \log p_{\boldsymbol{z}}/(2n)}\big)^{\complement} \,\leq\, 2p_{\boldsymbol{z}}^{-c_{\mathrm{ep}}}\,.$$

Proof of Lemma 4.4.3. The result follows immediately from Lemma 4.4.2. □

Lemma 4.4.4 (Control of $\|\boldsymbol{Z}^\top \boldsymbol{V}/n\|_\infty$, Gaussian noise)**.** *Suppose that $\boldsymbol{V}$ satisfies Assumption 2.1.3. Define the set*

$$\mathcal{T}_{\boldsymbol{V}}(\lambda) \,:=\, \big\{\|\boldsymbol{Z}^\top \boldsymbol{V}/n\|_\infty \leq \lambda\big\} \,=\, \big\{\|\boldsymbol{Z}^\top \boldsymbol{V}/n\|_\infty \leq \lambda\big\}\,, \tag{4.13}$$

where $\lambda > 0$ is a controlled quantity. Then,

$$\mathrm{P}\,\mathcal{T}_{\boldsymbol{V}}\big(c\sigma_{\boldsymbol{v}}\sqrt{\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty \log p_{\boldsymbol{z}}/(2n)}\big)^{\complement} \,\leq\, 2p_{\boldsymbol{z}}^{1-c_{\mathrm{ep}}}\,,$$

where c_{ep} is as defined in Lemma 4.4.2 and $\sigma_{\boldsymbol{v}} = \max_{j\in[p_{\boldsymbol{x}}]}\sigma_{v^j}$.

Proof of Lemma 4.4.4. Note that, for a given $\lambda > 0$,

$$\mathcal{T}_{\boldsymbol{V}}(\lambda) \,=\, \bigcap_{j\in[p_{\boldsymbol{x}}]} \big\{\|\boldsymbol{Z}^\top \boldsymbol{v}^j/n\|_\infty \leq \lambda\big\}\,.$$

Hence

$$\begin{aligned}
&\mathrm{P}\,\mathcal{T}_{\boldsymbol{V}}\big(c\sigma_{\boldsymbol{v}}\sqrt{\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty \log p_{\boldsymbol{z}}/(2n)}\big)^{\complement} \\
&\qquad = \mathrm{P}\bigcup_{j\in[p_{\boldsymbol{x}}]} \big\{\|\boldsymbol{Z}^\top \boldsymbol{v}^j/n\|_\infty > c\sigma_{\boldsymbol{v}}\sqrt{\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty \log p_{\boldsymbol{z}}/(2n)}\big\} \\
&\qquad \leq \sum_{j\in[p_{\boldsymbol{x}}]} \mathrm{P}\big\{\|\boldsymbol{Z}^\top \boldsymbol{v}^j/n\|_\infty > c\sigma_{v^j}\sqrt{\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty \log p_{\boldsymbol{z}}/(2n)}\big\} \\
&\qquad \leq\, 2p_{\boldsymbol{x}}p_{\boldsymbol{z}}^{-c_{\mathrm{ep}}} \,\leq\, 2p_{\boldsymbol{z}}^{1-c_{\mathrm{ep}}}\,,
\end{aligned}$$

as claimed. □

Lemma 4.4.5 (Simultaneous control of $\|\boldsymbol{Z}^\top \boldsymbol{v}^j/n\|_\infty$, Gaussian noise)**.** *Suppose that* $\boldsymbol{V}$ *satisfies Assumption 2.1.3. For each* $j \in [p_{\boldsymbol{x}}]$*, define the set*

$$\mathcal{T}_{\boldsymbol{v}^j}(\lambda_j) \,:=\, \left\{4\|\boldsymbol{Z}^\top \boldsymbol{v}^j/n\|_\infty \leq \lambda_j\right\}. \tag{4.14}$$

Define the set

$$\mathcal{T}_{\mathbf{A}}(\lambda_j)_{j\in[p_{\boldsymbol{x}}]} \,:=\, \bigcap_{j\in[p_{\boldsymbol{x}}]} \mathcal{T}_{\boldsymbol{v}^j}(\lambda_j) \tag{4.15}$$

Then,

$$\mathrm{P}\big\{\mathcal{T}_{\mathbf{A}}(\sigma_{v^j} c\sqrt{\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty \log p_{\boldsymbol{z}}/n})_{j\in[p_{\boldsymbol{x}}]}\big\}^{\complement} \,\leq\, 2p_{\boldsymbol{z}}^{1-c_{\mathrm{ep}}},$$

where c_{ep} *is as defined in Lemma 4.4.2.*

Proof of Lemma 4.4.5. Note that

$$\mathrm{P}\Big\{\bigcap_{j\in[p_{\boldsymbol{x}}]} \mathcal{T}_{\boldsymbol{v}^j}(\sigma_{v^j} c\sqrt{\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty \log p_{\boldsymbol{z}}/n})\Big\}^{\complement}$$
$$= \mathrm{P}\Big\{\bigcup_{j\in[p_{\boldsymbol{x}}]} \left\{4\|\boldsymbol{Z}^\top \boldsymbol{v}^j/n\|_\infty > \sigma_{v^j} c\sqrt{\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty \log p_{\boldsymbol{z}}/n}\right\}\Big\}.$$

By Lemma 4.4.2, the union bound for the probability of the right-hand event above is $2p_{\boldsymbol{x}}p_{\boldsymbol{z}}^{-c_{\mathrm{ep}}}$. The result follows from the assumption under the presently studied regime that $p_{\boldsymbol{x}} \leq p_{\boldsymbol{z}}$. □

4.4.2 Materials required for Section 4.2.3

The following generic bound for $\|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1}$ can be combined with concentration results for specific distributions of the first-stage noise elements. We present it separately for the sake of modularity with respect to such assumptions.

Lemma 4.4.6 (Generic bound for $\|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1}$)**.** *Suppose that Assumption 4.2.2 holds. For each* $j \in [p_{\boldsymbol{x}}]$*, let* $r_j > 0$ *be arbitrary. Then, on the set* $\mathcal{T}_{\mathbf{A}}(r_j)_{j\in[p_{\boldsymbol{x}}]} = \bigcap_{j\in[p_{\boldsymbol{x}}]} \mathcal{T}_{\boldsymbol{v}^j}(r_j)$*,*

$$\|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1} \,\leq\, 4s_{\mathbf{A}} r_{\mathbf{A}}/\phi_{\mathbf{A}}^2 \,=\, 4\frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2} r_{\mathbf{A}},$$

where $r_{\mathbf{A}} = \max_{j\in[p_{\boldsymbol{x}}]} r_j$.

Proof of Lemma 4.4.6. For each $j \in [p_{\boldsymbol{x}}]$, [17, Theorem 6.1, Lemma 6.2] entails that

$$\begin{aligned}\mathcal{T}_{\boldsymbol{v}^j}(r_j) \;=\; \left\{4\|\boldsymbol{Z}^\top \boldsymbol{v}^j/n\|_\infty \le r_j\right\} \;&\subseteq\; \left\{\|\hat{\boldsymbol{\alpha}}^j - \boldsymbol{\alpha}^j\|_1 \le 4 s_{\boldsymbol{\alpha}^j} r_j/\phi_j^2\right\} \\ &\subseteq\; \left\{\|\hat{\boldsymbol{\alpha}}^j - \boldsymbol{\alpha}^j\|_1 \le 4 s_{\mathbf{A}} r_{\mathbf{A}}/\phi_{\mathbf{A}}^2\right\},\end{aligned}$$

where the latter containment follows from our specification of $s_{\mathbf{A}}, r_{\mathbf{A}}$, and $\phi_{\mathbf{A}}$. Take intersections over $j \in [p_{\boldsymbol{x}}]$ on both sides of the above display to conclude. □

Proof of Lemma 4.2.4. Lemma 4.4.6 entails that

$$\mathrm{P}\{\|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1} > 4 s_{\mathbf{A}} r_{\mathbf{A}}/\phi_{\mathbf{A}}^2\} \;\le\; \mathrm{P}\Big\{\bigcap_{j\in[p_{\boldsymbol{x}}]} \mathcal{T}_{\boldsymbol{v}^j}(r_j)\Big\}^{\complement} \;\le\; \sum_{j\in[p_{\boldsymbol{x}}]} \mathrm{P}\,\mathcal{T}_{\boldsymbol{v}^j}(r_j)^{\complement}.$$

Apply the estimate of Lemma 4.4.5 for the right-hand side to conclude. □

4.4.3 Materials required for Section 4.2.4

We require the following bound for Lemma 4.2.8

Lemma 4.4.7 (Control of $\|\widehat{\boldsymbol{D}}^\top \tilde{\boldsymbol{u}}/n\|_\infty$)**.** *Let* $\tilde{\boldsymbol{u}} = \boldsymbol{u} + [(\boldsymbol{D} - \widehat{\boldsymbol{D}}) + \boldsymbol{V}]\boldsymbol{\beta}$*. Then,*

$$\begin{aligned}\|\widehat{\boldsymbol{D}}^\top \tilde{\boldsymbol{u}}/n\|_\infty \;\le\; &\|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1} \|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty (\|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1}\|\boldsymbol{\beta}\|_1 + \|\mathbf{A}\|_{L_1}) \\ &+ (\|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1} + \|\mathbf{A}\|_{L_1})(\|\boldsymbol{Z}^\top \boldsymbol{V}/n\|_\infty \|\boldsymbol{\beta}\|_1 + \|\boldsymbol{Z}^\top \boldsymbol{u}/n\|_\infty).\end{aligned}$$

Proof of Lemma 4.4.7. Write $\widehat{\boldsymbol{D}}^\top = (\widehat{\boldsymbol{D}} - \boldsymbol{D})^\top + \boldsymbol{D}^\top$ to find that

$$\begin{aligned}\|\widehat{\boldsymbol{D}}^\top \tilde{\boldsymbol{u}}/n\|_\infty \;&=\; \big\|\big[(\widehat{\boldsymbol{D}} - \boldsymbol{D})^\top + \boldsymbol{D}^\top\big]\big[(\boldsymbol{D} - \widehat{\boldsymbol{D}})\boldsymbol{\beta} + (\boldsymbol{V}\boldsymbol{\beta} + \boldsymbol{u})\big]/n\big\|_\infty \\ &\le\; \|(\widehat{\boldsymbol{D}} - \boldsymbol{D})^\top (\widehat{\boldsymbol{D}} - \boldsymbol{D})\boldsymbol{\beta}/n\|_\infty + \|(\widehat{\boldsymbol{D}} - \boldsymbol{D})^\top (\boldsymbol{V}\boldsymbol{\beta} + \boldsymbol{u})/n\|_\infty \\ &\quad + \|\boldsymbol{D}^\top (\widehat{\boldsymbol{D}} - \boldsymbol{D})\boldsymbol{\beta}/n\|_\infty + \|\boldsymbol{D}^\top (\boldsymbol{V}\boldsymbol{\beta} + \boldsymbol{u})/n\|_\infty \\ &:=\; \mathrm{I}_1 + \mathrm{I}_2 + \mathrm{I}_3 + \mathrm{I}_4 \qquad (4.16)\end{aligned}$$

We treat each quantity in the right-hand side above in turn.

For I_1, write

$$\mathrm{I}_1 \;=\; \|(\widehat{\boldsymbol{D}} - \boldsymbol{D})^\top (\widehat{\boldsymbol{D}} - \boldsymbol{D})\boldsymbol{\beta}/n\|_\infty \;\le\; \|(\widehat{\boldsymbol{D}} - \boldsymbol{D})^\top (\widehat{\boldsymbol{D}} - \boldsymbol{D})/n\|_\infty \|\boldsymbol{\beta}\|_1.$$

Recall that $\widehat{\boldsymbol{D}} - \boldsymbol{D} = \boldsymbol{Z}(\widehat{\mathbf{A}} - \mathbf{A})$ and write

$$\begin{aligned}\|(\widehat{\boldsymbol{D}} - \boldsymbol{D})^\top(\widehat{\boldsymbol{D}} - \boldsymbol{D})/n\|_\infty &= \|(\widehat{\mathbf{A}} - \mathbf{A})^\top \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}(\widehat{\mathbf{A}} - \mathbf{A})\|_\infty \\ &\leq \|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1}^2 \|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty = \|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1}^2 \|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty ,\end{aligned}$$

where the second line follows from repeated application of Hölder's inequality Combine the two previous displays to conclude that

$$\mathrm{I}_1 \leq \|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1}^2 \|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty \|\boldsymbol{\beta}\|_1 . \tag{4.17}$$

For I_2, write

$$\begin{aligned}\mathrm{I}_2 &= \|(\widehat{\boldsymbol{D}} - \boldsymbol{D})^\top(\boldsymbol{V}\boldsymbol{\beta} + \boldsymbol{u})/n\|_\infty \\ &= \|(\widehat{\mathbf{A}} - \mathbf{A})^\top \boldsymbol{Z}^\top(\boldsymbol{V}\boldsymbol{\beta} + \boldsymbol{u})/n\|_\infty \\ &\leq \underbrace{\|(\widehat{\mathbf{A}} - \mathbf{A})^\top \boldsymbol{Z}^\top \boldsymbol{V}\boldsymbol{\beta}/n\|_\infty}_{\mathrm{I}_{2,\mathrm{a}}} + \underbrace{\|(\widehat{\mathbf{A}} - \mathbf{A})^\top \boldsymbol{Z}^\top \boldsymbol{u}/n\|_\infty}_{\mathrm{I}_{2,\mathrm{b}}} .\end{aligned}$$

Applications of Hölder's inequality yield

$$\mathrm{I}_{2,\mathrm{a}} \leq \|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1} \|\boldsymbol{Z}^\top \boldsymbol{V}/n\|_\infty \|\boldsymbol{\beta}\|_1 = \|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1} \|\boldsymbol{Z}^\top \boldsymbol{V}/n\|_\infty \|\boldsymbol{\beta}\|_1$$

and

$$\mathrm{I}_{2,\mathrm{b}} \leq \|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1} \|\boldsymbol{Z}^\top \boldsymbol{u}/n\|_\infty = \|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1} \|\boldsymbol{Z}^\top \boldsymbol{u}/n\|_\infty .$$

Combine the previous three displays to conclude that

$$\mathrm{I}_2 \leq \|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1} (\|\boldsymbol{Z}^\top \boldsymbol{V}/n\|_\infty \|\boldsymbol{\beta}\|_1 + \|\boldsymbol{Z}^\top \boldsymbol{u}/n\|_\infty) , \tag{4.18}$$

For the quantity I_3, write Write

$$\mathrm{I}_3 = \|\boldsymbol{D}^\top(\widehat{\boldsymbol{D}} - \boldsymbol{D})\boldsymbol{\beta}/n\|_\infty \leq \|\boldsymbol{D}^\top(\widehat{\boldsymbol{D}} - \boldsymbol{D})/n\|_\infty \|\boldsymbol{\beta}\|_1$$

and observe that

$$\boldsymbol{D}^\top(\widehat{\boldsymbol{D}} - \boldsymbol{D})/n = \mathbf{A}\boldsymbol{Z}^\top \boldsymbol{Z}(\widehat{\mathbf{A}} - \mathbf{A})/n = \mathbf{A}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}(\widehat{\mathbf{A}} - \mathbf{A}) ,$$

which yields

$$\begin{aligned}\|\boldsymbol{D}^\top(\widehat{\boldsymbol{D}}-\boldsymbol{D})/n\|_\infty &= \|\mathbf{A}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}(\widehat{\mathbf{A}}-\mathbf{A})\|_\infty \\ &\leq \|\mathbf{A}\|_{L_1}\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty\|(\widehat{\mathbf{A}}-\mathbf{A})\|_{L_1}\end{aligned}$$

after repeated application of Hölder's inequality. Conclude from the previous three displays that

$$\mathrm{I}_3 \leq \|\widehat{\mathbf{A}}-\mathbf{A}\|_{L_1}\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty\|\mathbf{A}\|_{L_1}\,. \tag{4.19}$$

For the quantity I_4, write

$$\begin{aligned}\mathrm{I}_4 &= \|\boldsymbol{D}^\top(\boldsymbol{V}\boldsymbol{\beta}+\boldsymbol{u})/n\|_\infty \\ &= \|\mathbf{A}^\top\boldsymbol{Z}^\top(\boldsymbol{V}\boldsymbol{\beta}+\boldsymbol{u})/n\|_\infty \\ &\leq \|\mathbf{A}\|_{L_1}\big(\|\boldsymbol{Z}^\top\boldsymbol{V}\boldsymbol{\beta}/n\|_\infty+\|\boldsymbol{Z}^\top\boldsymbol{u}/n\|_\infty\big) \\ &\leq \big(\|\boldsymbol{Z}^\top\boldsymbol{V}/n\|_\infty\|\boldsymbol{\beta}\|_1+\|\boldsymbol{Z}^\top\boldsymbol{u}/n\|_\infty\big)\|\mathbf{A}\|_{L_1}\,. \end{aligned}\tag{4.20}$$

The original claim follows from line (4.16) and lines (4.17)-(4.20). □

The following generic bound for $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_1$ can be combined with concentration results for specific distributions of the first- and second-stage noise elements. We present it separately for the sake of modularity with respect to such assumptions.

Lemma 4.4.8 (Generic bound for $\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_1$)**.** *Suppose that Assumption 4.2.5 holds. Let $\lambda_{\boldsymbol{V}},\lambda_{\boldsymbol{u}}>0$ be arbitrary. Set $r_{\boldsymbol{\beta}}$ according to Definition 4.2.6. Then, on the set $\mathcal{T}_{\boldsymbol{V}}(\lambda_{\boldsymbol{V}})\cap\mathcal{T}_{\boldsymbol{u}}(\lambda_{\boldsymbol{u}})$,*

$$\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_1 \leq 4\frac{s_{\boldsymbol{\beta}}}{\phi^2_{\boldsymbol{\beta}}}r_{\boldsymbol{\beta}}\,.$$

Proof of Lemma 4.4.8. By Theorem 4.4.1, if

$$\mathcal{T}_{\tilde{\boldsymbol{u}}}=\big\{4\|\widehat{\boldsymbol{D}}^\top\tilde{\boldsymbol{u}}/n\|_\infty\leq r\big\}\subseteq\big\{\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_1\leq 4s_{\boldsymbol{\beta}}r^2/\phi^2\big\}\,.$$

It therefore suffices to show that $\mathcal{T}_{\boldsymbol{V}}(\lambda_{\boldsymbol{V}}) \cap \mathcal{T}_{\boldsymbol{u}}(\lambda_{\boldsymbol{u}}) \subseteq \mathcal{T}_{\tilde{\boldsymbol{u}}}(r_{\boldsymbol{\beta}})$ for the present choice of $r_{\boldsymbol{\beta}}$. Lemma 4.4.7 gives the bound

$$\|\widehat{\boldsymbol{D}}^\top \tilde{\boldsymbol{u}}/n\|_\infty \leq \underbrace{\|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1} \|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty (m_{\boldsymbol{\beta}} \|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1} + m_{\mathbf{A}})}_{\mathrm{I}_1} + \underbrace{(\|\widehat{\mathbf{A}} - \mathbf{A}\|_{L_1} + m_{\mathbf{A}})(\|\boldsymbol{Z}^\top \boldsymbol{V}/n\|_\infty \|\boldsymbol{\beta}\|_1 + \|\boldsymbol{Z}^\top \boldsymbol{u}/n\|_\infty)}_{\mathrm{I}_2}.$$

Cite Lemma 4.4.6 to conclude that, on the set $\mathcal{T}_{\boldsymbol{V}}(\lambda_{\boldsymbol{V}})$,

$$\mathrm{I}_1 \leq 4\frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2} r_{\mathbf{A}} \|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty (4 m_{\boldsymbol{\beta}} \frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2} r_{\mathbf{A}} + m_{\mathbf{A}}),$$

Note that, on the set $\mathcal{T}_{\boldsymbol{V}}(\lambda_{\boldsymbol{V}}) \cap \mathcal{T}_{\boldsymbol{u}}(\lambda_{\boldsymbol{u}})$,

$$\mathrm{I}_2 \leq \frac{1}{4}\big(4\frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2} r_{\mathbf{A}} + m_{\mathbf{A}}\big)\big(m_{\boldsymbol{\beta}} \lambda_{\boldsymbol{V}} + \lambda_{\boldsymbol{u}}\big)$$

by specification. Multiply the two previous displays by 4 and combine with the third-previous display to conclude that $\mathcal{T}_{\boldsymbol{V}}(\lambda_{\boldsymbol{V}}) \cap \mathcal{T}_{\boldsymbol{u}}(\lambda_{\boldsymbol{u}}) \subseteq \mathcal{T}_{\tilde{\boldsymbol{u}}}(r_{\boldsymbol{\beta}})$ for the present choice of $r_{\boldsymbol{\beta}}$, as required. □

Proof of Lemma 4.2.8. Lemma 4.4.8 entails that

$$\big\{\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 > 4\frac{s_{\boldsymbol{\beta}}}{\phi_{\boldsymbol{\beta}}^2} r_{\boldsymbol{\beta}}\big\} \subseteq (\mathcal{T}_{\boldsymbol{V}}(\lambda_{\boldsymbol{V}}) \cap \mathcal{T}_{\boldsymbol{u}}(\lambda_{\boldsymbol{u}}))^{\complement} = \mathcal{T}_{\boldsymbol{V}}(\lambda_{\boldsymbol{V}})^{\complement} \cup \mathcal{T}_{\boldsymbol{u}}(\lambda_{\boldsymbol{u}})^{\complement}.$$

Thus,

$$\mathrm{P}\{\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 > 4\frac{s_{\boldsymbol{\beta}}}{\phi_{\boldsymbol{\beta}}^2} r_{\boldsymbol{\beta}}\} \leq \mathrm{P}\mathcal{T}_{\boldsymbol{V}}(\lambda_{\boldsymbol{V}})^{\complement} + \mathrm{P}\mathcal{T}_{\boldsymbol{u}}(\lambda_{\boldsymbol{u}})^{\complement}$$

Now cite the estimates of Lemmas 4.4.5 and 4.4.3. □

4.4.4 Materials required for Section 4.2.5

Proof of Lemma 4.2.9. Let S, s be as in the statement of Lemma 4.2.9, and let $\boldsymbol{\delta} \in \mathbb{R}^{p_{\boldsymbol{x}}} \setminus \{\mathbf{0}\}$ satisfying $\|\boldsymbol{\delta}_{S^{\complement}}\|_1 \leq 3\|\boldsymbol{\delta}_S\|_1$ be arbitrary. Write $\widehat{\boldsymbol{D}} = \boldsymbol{D} + (\widehat{\boldsymbol{D}} - \boldsymbol{D})$, so that

$$\begin{aligned}
\|\widehat{\boldsymbol{D}}\boldsymbol{\delta}\|_2^2 &= \|[\boldsymbol{D} + (\widehat{\boldsymbol{D}} - \boldsymbol{D})]\boldsymbol{\delta}\|_2^2 \\
&= \big\langle [\boldsymbol{D} + (\widehat{\boldsymbol{D}} - \boldsymbol{D})]\boldsymbol{\delta}, [\boldsymbol{D} + (\widehat{\boldsymbol{D}} - \boldsymbol{D})]\boldsymbol{\delta} \big\rangle \\
&= \|\boldsymbol{D}\boldsymbol{\delta}\|_2^2 + 2\langle \boldsymbol{D}^\top(\widehat{\boldsymbol{D}} - \boldsymbol{D})\boldsymbol{\delta},\, \boldsymbol{\delta}\rangle + \langle (\widehat{\boldsymbol{D}} - \boldsymbol{D})^\top(\widehat{\boldsymbol{D}} - \boldsymbol{D})\boldsymbol{\delta}, \boldsymbol{\delta}\rangle.
\end{aligned}$$

Thus,

$$\|\widehat{\boldsymbol{D}}\boldsymbol{\delta}\|_2^2/n \;\geq\; \|\boldsymbol{D}\boldsymbol{\delta}\|_2^2/n - \underbrace{2|\langle \boldsymbol{D}^\top(\widehat{\boldsymbol{D}}-\boldsymbol{D})\boldsymbol{\delta}/n, \boldsymbol{\delta}\rangle|}_{\mathrm{I}_1} \\ - \underbrace{|\langle (\widehat{\boldsymbol{D}}-\boldsymbol{D})^\top(\widehat{\boldsymbol{D}}-\boldsymbol{D})\boldsymbol{\delta}/n, \boldsymbol{\delta}\rangle|}_{\mathrm{I}_2}\,. \tag{4.21}$$

We now obtain bounds for the quantities $\mathrm{I}_1, \mathrm{I}_2$. From repeated applications of Hölder's inequality, write

$$\begin{aligned}
\mathrm{I}_1 \;\lesssim\; |\langle \boldsymbol{D}^\top(\widehat{\boldsymbol{D}}-\boldsymbol{D})\boldsymbol{\delta}/n, \boldsymbol{\delta}\rangle| \;&\leq\; \|\boldsymbol{D}^\top(\widehat{\boldsymbol{D}}-\boldsymbol{D})/n\|_\infty \|\boldsymbol{\delta}\|_1^2 \\
&=\; \|\mathbf{A}^\top \boldsymbol{Z}^\top \boldsymbol{Z}(\widehat{\mathbf{A}}-\mathbf{A})/n\|_\infty \|\boldsymbol{\delta}\|_1^2 \\
&\leq\; \|\mathbf{A}\|_{L_1} \|\boldsymbol{Z}^\top\boldsymbol{Z}/n\|_\infty \|\widehat{\mathbf{A}}-\mathbf{A}\|_{L_1} \|\boldsymbol{\delta}\|_1^2 \\
&\leq\; m_{\mathbf{A}} \|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty \|\widehat{\mathbf{A}}-\mathbf{A}\|_{L_1} \|\boldsymbol{\delta}\|_1^2
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{I}_2 \;&=\; |\langle (\widehat{\boldsymbol{D}}-\boldsymbol{D})^\top(\widehat{\boldsymbol{D}}-\boldsymbol{D})\boldsymbol{\delta}/n, \boldsymbol{\delta}\rangle| \\
&\leq\; \|(\widehat{\boldsymbol{D}}-\boldsymbol{D})^\top(\widehat{\boldsymbol{D}}-\boldsymbol{D})/n\|_\infty \|\boldsymbol{\delta}\|_1^2 \\
&=\; \|(\widehat{\mathbf{A}}-\mathbf{A})^\top \boldsymbol{Z}^\top \boldsymbol{Z}(\widehat{\mathbf{A}}-\mathbf{A})/n\|_\infty \|\boldsymbol{\delta}\|_1^2 \\
&\leq\; \|\widehat{\mathbf{A}}-\mathbf{A}\|_{L_1} \|\boldsymbol{Z}^\top\boldsymbol{Z}/n\|_\infty \|\widehat{\mathbf{A}}-\mathbf{A}\|_{L_1} \|\boldsymbol{\delta}\|_1^2 \\
&\leq\; \|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty \|\widehat{\mathbf{A}}-\mathbf{A}\|_{L_1}^2 \|\boldsymbol{\delta}\|_1^2\,.
\end{aligned}$$

Combine the previous two displays with (4.21) to find that

$$\|\widehat{\boldsymbol{D}}\boldsymbol{\delta}\|_2^2/n \;\geq\; \|\boldsymbol{D}\boldsymbol{\delta}\|_2^2/n - (2m_{\mathbf{A}}\|\widehat{\mathbf{A}}-\mathbf{A}\|_{L_1} + \|\widehat{\mathbf{A}}-\mathbf{A}\|_{L_1}^2)\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty \|\boldsymbol{\delta}\|_1^2\,.$$

By assumption, we have $\|\boldsymbol{\delta}\|_1 \leq 4\|\boldsymbol{\delta}_S\|_1$. Substitute this expression in the right-hand side above and multiply through by $s/\|\boldsymbol{\delta}_S\|_1^2$ to obtain

$$\frac{s\|\widehat{\boldsymbol{D}}\boldsymbol{\delta}\|_2^2}{n\|\boldsymbol{\delta}_S\|_1^2} \;\geq\; \frac{s\|\boldsymbol{D}\boldsymbol{\delta}\|_2^2}{n\|\boldsymbol{\delta}_S\|_1^2} - 16s\big(2m_{\mathbf{A}}\|\widehat{\mathbf{A}}-\mathbf{A}\|_{L_1} + \|\widehat{\mathbf{A}}-\mathbf{A}\|_{L_1}^2\big)\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty\,.$$

Thus, on the set $\{16s\big(2m_{\mathbf{A}}\|\widehat{\mathbf{A}}-\mathbf{A}\|_{L_1}+\|\widehat{\mathbf{A}}-\mathbf{A}\|_{L_1}^2\big)\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty \le \epsilon_1\}$, we have

$$\begin{aligned}
\frac{s\|\widehat{\boldsymbol{D}}\boldsymbol{\delta}\|_2^2}{n\|\boldsymbol{\delta}_S\|_1^2} &\ge \frac{s\|\boldsymbol{D}\boldsymbol{\delta}\|_2^2}{n\|\boldsymbol{\delta}_S\|_1^2} - \epsilon_1 \\
&= \left(\frac{s\boldsymbol{\delta}^\top\boldsymbol{\Sigma}_{\boldsymbol{d}}\boldsymbol{\delta}}{\|\boldsymbol{\delta}_S\|_1^2} - \frac{s\boldsymbol{\delta}^\top(\overline{\boldsymbol{\Sigma}}_{\boldsymbol{d}}-\boldsymbol{\Sigma}_{\boldsymbol{d}})\boldsymbol{\delta}}{\|\boldsymbol{\delta}_S\|_1^2}\right) - \epsilon_1 \\
&\ge \frac{s\boldsymbol{\delta}^\top\boldsymbol{\Sigma}_{\boldsymbol{d}}\boldsymbol{\delta}}{\|\boldsymbol{\delta}_S\|_1^2} - \frac{s\|\overline{\boldsymbol{\Sigma}}_{\boldsymbol{d}}-\boldsymbol{\Sigma}_{\boldsymbol{d}}\|_\infty\|\boldsymbol{\delta}\|_1^2}{\|\boldsymbol{\delta}_S\|_1^2} - \epsilon_1 \\
&\ge \frac{s\boldsymbol{\delta}^\top\boldsymbol{\Sigma}_{\boldsymbol{d}}\boldsymbol{\delta}}{\|\boldsymbol{\delta}_S\|_1^2} - 16s\|\overline{\boldsymbol{\Sigma}}_{\boldsymbol{d}}-\boldsymbol{\Sigma}_{\boldsymbol{d}}\|_\infty - \epsilon_1\,.
\end{aligned}$$

From Cauchy-Schwartz we have $\|\boldsymbol{\delta}_S\|_1 \le \sqrt{s}\|\boldsymbol{\delta}_S\|_2$ and hence that $\|\boldsymbol{\delta}_S\|_1^2 \le s\|\boldsymbol{\delta}\|_2^2$. Substitute this bound into the first term on the right-hand side above to obtain

$$\begin{aligned}
\frac{s\|\widehat{\boldsymbol{D}}\boldsymbol{\delta}\|_2^2}{n\|\boldsymbol{\delta}_S\|_1^2} &\ge \frac{\boldsymbol{\delta}^\top\boldsymbol{\Sigma}_{\boldsymbol{d}}\boldsymbol{\delta}}{\|\boldsymbol{\delta}\|_2^2} - 16s\|\overline{\boldsymbol{\Sigma}}_{\boldsymbol{d}}-\boldsymbol{\Sigma}_{\boldsymbol{d}}\|_\infty - \epsilon_1 \\
&\ge \Lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{d}}) - 16s\|\overline{\boldsymbol{\Sigma}}_{\boldsymbol{d}}-\boldsymbol{\Sigma}_{\boldsymbol{d}}\|_\infty - \epsilon_1\,,
\end{aligned}$$

where $\Lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{d}})$ denotes the minimal eigenvalue of $\boldsymbol{\Sigma}_{\boldsymbol{d}}$. The right-hand side above does not depend on $\boldsymbol{\delta}$, so we may take the infimum of the left-hand side above over $\boldsymbol{\delta} \in \mathcal{C}(S)$ to write

$$\begin{aligned}
&\mathrm{P}\left\{\phi_\dagger^2(\widehat{\boldsymbol{D}},S) < \Lambda_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{d}}) - \epsilon_2 - \epsilon_1\right\} \\
&\qquad\le\ \mathrm{P}\big\{16s(2m_{\mathbf{A}}\|\widehat{\mathbf{A}}-\mathbf{A}\|_{L_1}+\|\widehat{\mathbf{A}}-\mathbf{A}\|_{L_1}^2)\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty > \epsilon_1\big\} \\
&\qquad\quad + \mathrm{P}\big\{16s\|\overline{\boldsymbol{\Sigma}}_{\boldsymbol{d}}-\boldsymbol{\Sigma}_{\boldsymbol{d}}\|_\infty > \epsilon_2\big\}\,,
\end{aligned}$$

as claimed. □

4.4.5 Materials required for Section 4.3

Lemmas 4.4.9, 4.4.11, 4.4.13, and 4.4.15 provide finite-sample bounds for the quantities $\|\boldsymbol{f}_\ell\|_\infty$ for $\ell \in \{1,2,3,4\}$ that are generic over various noise regimes. We present them separately for the sake of modularity with respect to such assumptions. Lemmas 4.4.10, 4.4.12, 4.4.14, and 4.4.16 in turn provide specific rates for the $\|\boldsymbol{f}_\ell\|_\infty$ under the Gaussian noise regime of Assumption 2.1.3.

Lemma 4.4.9 (Control of $\boldsymbol{f}_1$)**.** *Suppose that Assumption 4.2.2 and Conditions 3, 5 and that $\widehat{\boldsymbol{\Theta}}$ is chosen according to Assumption 4.3.2. Then, on the set $\mathcal{T}_{\boldsymbol{u}}(\lambda_{\boldsymbol{u}}) \cap \mathcal{T}_{\boldsymbol{\Theta}}(\mu)$, where $\lambda_{\boldsymbol{u}} > 0$ is arbitrary, the remainder term*

$$\boldsymbol{f}_1 \,=\, (\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})^\top \boldsymbol{D}^\top \boldsymbol{u}/\sqrt{n}$$

satisfies

$$\|\boldsymbol{f}_1\|_\infty \,\leq\, 2^{-q}\sqrt{n} m_{\mathbf{A}} c_q (m_{\boldsymbol{\Theta}}\mu)^{1-q} s_{\boldsymbol{\Theta}} \lambda_{\boldsymbol{u}}\,,$$

where c_q is as in Lemma 3.5.2.

Proof of Lemma 4.4.9. Lemma 3.5.2 entails that, on the set $\mathcal{T}_{\boldsymbol{\Theta}}(\mu)$,

$$\max_{j\in[p_{\boldsymbol{x}}]} \|\hat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j\|_1 \,\leq\, 2c_q(2m_{\boldsymbol{\Theta}}\mu)^{1-q} s_{\boldsymbol{\Theta}}\,.$$

We therefore find that

$$\begin{aligned}
\|(\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})^\top \boldsymbol{D}^\top \boldsymbol{u}/\sqrt{n}\|_\infty \,&\leq\, \sqrt{n}\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_{L_1} \|\boldsymbol{D}^\top \boldsymbol{u}/n\|_\infty \\
&\leq\, \sqrt{n}\|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_{L_1} \|\mathbf{A}\|_{L_1} \|\boldsymbol{Z}^\top \boldsymbol{u}/n\|_\infty \\
&\leq\, \sqrt{n} m_{\mathbf{A}} \|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_{L_1} \|\boldsymbol{Z}^\top \boldsymbol{u}/n\|_\infty \\
&\leq\, 2\sqrt{n} m_{\mathbf{A}} c_q (2m_{\boldsymbol{\Theta}}\mu)^{1-q} s_{\boldsymbol{\Theta}} \|\boldsymbol{Z}^\top \boldsymbol{u}/n\|_\infty\,.
\end{aligned}$$

On the set $\mathcal{T}_{\boldsymbol{u}}(\lambda_{\boldsymbol{u}})$ we have $\|\boldsymbol{Z}^\top \boldsymbol{u}/n\|_\infty \leq \lambda_{\boldsymbol{u}}/4$. From this bound and the previous display we conclude that, on the set $\mathcal{T}_{\boldsymbol{u}}(\lambda_{\boldsymbol{u}}) \cap \mathcal{T}_{\boldsymbol{\Theta}}(\mu)$,

$$\|(\boldsymbol{\Theta} - \widehat{\boldsymbol{\Theta}})^\top \boldsymbol{D}^\top \boldsymbol{u}/\sqrt{n}\|_\infty \,\leq\, 2^{-q}\sqrt{n} m_{\mathbf{A}} c_q (m_{\boldsymbol{\Theta}}\mu)^{1-q} s_{\boldsymbol{\Theta}} \lambda_{\boldsymbol{u}}\,,$$

as claimed. □

Lemma 4.4.10 (Control of $\boldsymbol{f}_1$, Gaussian noise)**.** *Suppose that (i) Assumption 4.2.2 and Conditions 3, 5 of Assumption 4.3.1 hold and (ii) Assumption 2.1.3 holds. Choose $\widehat{\boldsymbol{\Theta}}$ according to Assumption 4.3.2. Then,*

$$\begin{aligned}
&\mathrm{P}\left\{\|\boldsymbol{f}_1\|_\infty > 2^{-q} m_{\mathbf{A}} c_q \sigma_u c (m_{\boldsymbol{\Theta}}\mu)^{1-q} s_{\boldsymbol{\Theta}} \sqrt{m_{\boldsymbol{Z}} \log p_{\boldsymbol{z}}}\right\} \\
&\qquad\qquad \leq\, 2p_{\boldsymbol{z}}^{-c_{\mathrm{ep}}} + \mathrm{P}\big\{\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty > m_{\boldsymbol{Z}}\big\} + \mathrm{P}\,\mathcal{T}_{\boldsymbol{\Theta}}(\mu)^{\complement}\,,
\end{aligned}$$

where $c > 0$ and c_{ep} are as defined in Lemma 4.4.2. Consequently, if Conditions 1, 6, 7, and 8a of Assumption 4.3.1 also hold, then $\|\boldsymbol{f}_1\|_\infty = o_{\mathrm{P}}(1)$.

Proof of Lemma 4.4.10. Lemma 4.4.9 entails that

$$\mathrm{P}\left\{\|\boldsymbol{f}_1\|_\infty > 2^{-q}\sqrt{n}m_{\mathbf{A}}c_q(m_{\boldsymbol{\Theta}}\mu)^{1-q}s_{\boldsymbol{\Theta}}\lambda_{\boldsymbol{u}}\right\} \;\le\; \mathrm{P}\,\mathcal{T}_{\boldsymbol{u}}^{\complement} + \mathrm{P}\,\mathcal{T}_{\boldsymbol{\Theta}}(\mu)^{\complement}\,.$$

Substitute $\lambda_{\boldsymbol{u}}$ chosen according to Definition 4.2.7 into the display above and cite the estimate of Lemma 4.4.3 to deduce the original claim. □

Lemma 4.4.11 (Control of $\boldsymbol{f}_2$)**.** *Suppose that Assumption 4.2.2 and Condition 5 of Assumption 4.3.1 hold. Choose $\widehat{\boldsymbol{\Theta}}$ according to Assumption 4.3.2, set $r_{\mathbf{A}}$ according to (4.2) and let $\lambda_{\boldsymbol{u}} > 0$ be arbitrary. Then, on the set $\mathcal{T}_{\mathbf{A}}(\boldsymbol{r}) \cap \mathcal{T}_{\boldsymbol{u}}(\lambda_{\boldsymbol{u}}) \cap \mathcal{T}_{\boldsymbol{\Theta}}(\mu)$, the remainder term*

$$\boldsymbol{f}_2 \;=\; \widehat{\boldsymbol{\Theta}}^\top(\widehat{\boldsymbol{D}} - \boldsymbol{D})^\top\boldsymbol{u}/\sqrt{n}$$

satisfies

$$\|\boldsymbol{f}_2\|_\infty \;\le\; \sqrt{n}m_{\boldsymbol{\Theta}}\frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2}r_{\mathbf{A}}\lambda_{\boldsymbol{u}}$$

for n sufficiently large.

Proof of Lemma 4.4.11. Observe that

$$\widehat{\boldsymbol{\Theta}}^\top(\widehat{\boldsymbol{D}} - \boldsymbol{D})^\top\boldsymbol{u}/\sqrt{n} \;=\; \sqrt{n}\widehat{\boldsymbol{\Theta}}^\top(\widehat{\mathbf{A}} - \mathbf{A})^\top(\boldsymbol{Z}^\top\boldsymbol{u}/n)\,.$$

On the set $\mathcal{T}_{\boldsymbol{\Theta}}(\mu)$, each row $\boldsymbol{\theta}$ is feasible for Program 3.3.1. Then, $\|\hat{\boldsymbol{\theta}}_j\|_1 \le \|\boldsymbol{\theta}_j\|_1$ for each $j \in [p_{\boldsymbol{x}}]$ by specification. Lemmas 4.4.6 and 4.4.3 then entail that, on the set $\mathcal{T}_{\mathbf{A}}(\boldsymbol{r})\cap\mathcal{T}_{\boldsymbol{u}}(\lambda_{\boldsymbol{u}})\cap\mathcal{T}_{\boldsymbol{\Theta}}(\mu)$,

$$\begin{aligned}\|\widehat{\boldsymbol{\Theta}}^\top(\widehat{\boldsymbol{D}} - \boldsymbol{D})^\top\boldsymbol{u}/\sqrt{n}\|_\infty \;&\le\; \max_{j,k\in[p_{\boldsymbol{x}}]}\sqrt{n}\|\hat{\boldsymbol{\theta}}_j\|_1\|\hat{\boldsymbol{\alpha}}^k - \boldsymbol{\alpha}^k\|_1\|\boldsymbol{Z}^\top\boldsymbol{u}/n\|_\infty \\ &\le\; \sqrt{n}m_{\boldsymbol{\Theta}}\frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2}r_{\mathbf{A}}\lambda_{\boldsymbol{u}}\,,\end{aligned}$$

as claimed. □

Lemma 4.4.12 (Control of $\boldsymbol{f}_2$, Gaussian noise)**.** *Suppose that (i) Assumption 4.2.2 and Condition 5 of Assumption 4.3.1 hold and (ii) Assumption 2.1.3 holds. Choose $\widehat{\boldsymbol{\Theta}}$ according to Assumption 4.3.2; set $\boldsymbol{r} = (r_j)_{j=1}^{p_{\boldsymbol{x}}}$ according to Definition 4.2.3; and set $r_{\mathbf{A}}$ according to (4.2). Then,*

$$\mathrm{P}\Big\{\|\boldsymbol{f}_2\|_\infty > m_{\boldsymbol{\Theta}} c^2 \sigma_v \sigma_u m_{\boldsymbol{Z}} \frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2} \log p_{\boldsymbol{z}}/\sqrt{n}\Big\} \\ \leq 2p_{\boldsymbol{z}}^{1-c_{\mathrm{ep}}} + 2p_{\boldsymbol{z}}^{-c_{\mathrm{ep}}} + \mathrm{P}\big\{\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty > m_{\boldsymbol{Z}}\big\} + \mathrm{P}\,\mathcal{T}_{\boldsymbol{\Theta}}(\mu)^{\complement}\,.$$

Consequently, if Conditions 1, 6, 7, and 8b of Assumption 4.3.1 also hold, then $\|\boldsymbol{f}_2\|_\infty = o_{\mathrm{P}}(1)$.

Proof of Lemma 4.4.12. Lemma 4.4.11 entails that

$$\mathrm{P}\left\{\|\boldsymbol{f}_2\|_\infty > \frac{1}{4}\sqrt{n} m_{\boldsymbol{\Theta}} \frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2} r_{\mathbf{A}} \lambda_{\boldsymbol{u}}\right\} \leq \mathrm{P}\,\mathcal{T}_{\mathbf{A}}(\boldsymbol{r})^{\complement} + \mathrm{P}\,\mathcal{T}_{\boldsymbol{u}}(\lambda_{\boldsymbol{u}})^{\complement} + \mathrm{P}\,\mathcal{T}_{\boldsymbol{\Theta}}(\mu)^{\complement}\,.$$

The present choice of r_j for $j \in [p_{\boldsymbol{x}}]$ entails that $r_{\mathbf{A}} \geq \sigma_v c\sqrt{\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty \log p_{\boldsymbol{z}}/n}$. Substitute the latter quantity and the choice of $\lambda_{\boldsymbol{u}} = \sigma_u c\sqrt{\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty \log p_{\boldsymbol{z}}/n}$ into the previous display and cite the estimates of Corollaries 4.4.5, 4.4.3 to deduce the original claim. □

Lemma 4.4.13 (Control of $\boldsymbol{f}_3$)**.** *Suppose that Assumptions 4.2.2 and 4.2.5 and Condition 5 of Assumption 4.3.1 hold. Choose $\widehat{\boldsymbol{\Theta}}$ according to Assumption 4.3.2; let $\boldsymbol{r} = (r_j)_{j\in[p_{\boldsymbol{x}}]} > \mathbf{0}$ and $\lambda_{\boldsymbol{u}} > 0$ be arbitrary. Then, on the set*

$$\mathcal{T}_{\mathbf{A}}(\boldsymbol{r}) \cap \mathcal{T}_{\boldsymbol{u}}(\lambda_{\boldsymbol{u}}) \cap \mathcal{T}_{\boldsymbol{V}}(\lambda_{\boldsymbol{V}}) \cap \mathcal{T}_{\boldsymbol{\Theta}}(\mu)\,,$$

the remainder term

$$\boldsymbol{f}_3 = \widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{D}}^\top(\boldsymbol{X} - \widehat{\boldsymbol{D}})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})/\sqrt{n}$$

satisfies

$$\|\boldsymbol{f}_3\|_\infty \leq 8 m_{\boldsymbol{\Theta}} m_{\mathbf{A}} \sqrt{n}\Big(4\frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2} r_{\mathbf{A}} + \lambda_{\boldsymbol{u}}\Big)\frac{s_{\boldsymbol{\beta}}}{\phi_{\boldsymbol{\beta}}^2} r_{\boldsymbol{\beta}}$$

Proof of Lemma 4.4.13. We first observe that

$$\|\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{D}}^\top(\boldsymbol{X}-\widehat{\boldsymbol{D}})(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})/\sqrt{n}\|_\infty \;\leq\; \sqrt{n}\|\widehat{\boldsymbol{\Theta}}\|_{L_1}\|\widehat{\boldsymbol{D}}^\top(\boldsymbol{X}-\widehat{\boldsymbol{D}})/n\|_\infty\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\|_1\,.$$

Now,

$$\begin{aligned}\widehat{\boldsymbol{D}}^\top(\boldsymbol{X}-\widehat{\boldsymbol{D}})/n &= \widehat{\boldsymbol{D}}^\top(\boldsymbol{D}+\boldsymbol{u}-\widehat{\boldsymbol{D}})/n\\ &= \widehat{\boldsymbol{D}}^\top(\boldsymbol{D}-\widehat{\boldsymbol{D}})/n+\widehat{\boldsymbol{D}}^\top\boldsymbol{u}/n\\ &= \widehat{\mathbf{A}}^\top(\boldsymbol{Z}^\top\boldsymbol{Z}/n)(\mathbf{A}-\widehat{\mathbf{A}})+\widehat{\boldsymbol{D}}^\top\boldsymbol{u}/n\,.\end{aligned}$$

For the first term on the right-hand side above, write

$$\begin{aligned}\|\widehat{\mathbf{A}}^\top(\boldsymbol{Z}^\top\boldsymbol{Z}/n)(\mathbf{A}-\widehat{\mathbf{A}})\|_\infty &\leq \|\mathbf{A}^\top(\boldsymbol{Z}^\top\boldsymbol{Z}/n)(\mathbf{A}-\widehat{\mathbf{A}})\|_\infty\\ &\quad+\|(\widehat{\mathbf{A}}-\mathbf{A})^\top(\boldsymbol{Z}^\top\boldsymbol{Z}/n)(\mathbf{A}-\widehat{\mathbf{A}})\|_\infty\\ &\leq \|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty[m_{\mathbf{A}}\|\widehat{\mathbf{A}}-\mathbf{A}\|_{L_1}+\|\widehat{\mathbf{A}}-\mathbf{A}\|_{L_1}^2]\,.\end{aligned}$$

On the set $\mathcal{T}_{\mathbf{A}}(\boldsymbol{r})$, the right-hand side above is less than or equal to $2m_{\mathbf{A}}\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty\|\widehat{\mathbf{A}}-\mathbf{A}\|_{L_1}$ for n sufficiently large by Lemma 4.4.6 and the hypotheses of the present lemma. For the second term on the right-hand side of two displays previous, write

$$\begin{aligned}\|\widehat{\boldsymbol{D}}^\top\boldsymbol{u}/n\|_\infty &= \|\widehat{\mathbf{A}}^\top\boldsymbol{Z}^\top\boldsymbol{u}/n\|_\infty\\ &= \|(\mathbf{A}+[\widehat{\mathbf{A}}-\mathbf{A}])^\top(\boldsymbol{Z}^\top\boldsymbol{u}/n)\|_\infty\\ &\leq 2m_{\mathbf{A}}\|\boldsymbol{Z}^\top\boldsymbol{u}/n\|_\infty\,,\end{aligned}$$

where the final line holds on the set $\mathcal{T}_{\mathbf{A}}(\boldsymbol{r})$ for n sufficiently large by Lemma 4.4.6. Thus, on the set $\mathcal{T}_{\mathbf{A}}(\boldsymbol{r})\cap\mathcal{T}_{\boldsymbol{u}}(\lambda_{\boldsymbol{u}})$, we have

$$\begin{aligned}\|\widehat{\boldsymbol{D}}^\top(\boldsymbol{X}-\widehat{\boldsymbol{D}})/n\|_\infty &\leq 2m_{\mathbf{A}}(\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty\|\widehat{\mathbf{A}}-\mathbf{A}\|_{L_1}+\|\boldsymbol{Z}^\top\boldsymbol{u}/n\|_\infty)\\ &\leq 2m_{\mathbf{A}}(4\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty\frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2}r_{\mathbf{A}}+\lambda_{\boldsymbol{u}})\,,\end{aligned}$$

where the latter substitutions are justified by Lemma 4.4.6 and the definition of $\mathcal{T}_{\boldsymbol{u}}(\lambda_{\boldsymbol{u}})$.

On the set $\mathcal{T}_{\boldsymbol{\Theta}}(\mu)$, each row $\boldsymbol{\theta}$ is feasible for Program 3.3.1. Then, $\|\hat{\boldsymbol{\theta}}_j\|_1\leq\|\boldsymbol{\theta}_j\|_1$ for each $j\in[p_{\boldsymbol{x}}]$ by specification.

Finally, Lemma 4.4.8 entails that, for the present choice of $r_{\boldsymbol{\beta}}$,

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 \leq 4\frac{s_{\boldsymbol{\beta}}}{\phi^2_{\boldsymbol{\beta}}} r_{\boldsymbol{\beta}}$$

on the set $\mathcal{T}_{\boldsymbol{V}}(\lambda_{\boldsymbol{V}}) \cap \mathcal{T}_{\boldsymbol{u}}(\lambda_{\boldsymbol{u}})$.

Combining the foregoing results, we see that, on the set

$$\mathcal{T}_{\mathbf{A}}(\boldsymbol{r}) \cap \mathcal{T}_{\boldsymbol{u}}(\lambda_{\boldsymbol{u}}) \cap \mathcal{T}_{\boldsymbol{V}}(\lambda_{\boldsymbol{V}}) \cap \mathcal{T}_{\boldsymbol{\Theta}}(\mu)\,,$$

it holds that

$$\|\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{D}}^\top(\boldsymbol{X} - \widehat{\boldsymbol{D}})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})/\sqrt{n}\|_\infty \leq 8m_{\boldsymbol{\Theta}} m_{\mathbf{A}} \sqrt{n}(4\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty \frac{s_{\mathbf{A}}}{\phi^2_{\mathbf{A}}} r_{\mathbf{A}} + \lambda_{\boldsymbol{u}}) \frac{s_{\boldsymbol{\beta}}}{\phi^2_{\boldsymbol{\beta}}} r_{\boldsymbol{\beta}}\,,$$

as claimed. □

Lemma 4.4.14 (Control of $\boldsymbol{f}_3$, Gaussian noise)**.** *Suppose that (i) Assumptions 4.2.2 and 4.2.5 and Condition 5 of Assumption 4.3.1 hold and (ii) Assumption 2.1.3 holds. Choose $\widehat{\boldsymbol{\Theta}}$ according to Assumption 4.3.2; set $\boldsymbol{r} = (r_j)_{j=1}^{p_{\boldsymbol{x}}}$ according to Definition 4.2.3; set $r_{\mathbf{A}}$ according to (4.2); set $\lambda_{\boldsymbol{u}}$ and $\lambda_{\boldsymbol{V}}$ according to Definition 4.2.7; and set $r_{\boldsymbol{\beta}}$ according to Definition 4.2.6. Then,*

$$\begin{aligned}\mathrm{P}\big\{\|\boldsymbol{f}_3\|_\infty > 8m_{\boldsymbol{\Theta}} m_{\mathbf{A}} \sqrt{n}(4m_{\boldsymbol{Z}} \frac{s_{\mathbf{A}}}{\phi^2_{\mathbf{A}}} r_{\mathbf{A}} + \lambda_{\boldsymbol{u}}) \frac{s_{\boldsymbol{\beta}}}{\phi^2_{\boldsymbol{\beta}}} r_{\boldsymbol{\beta}}\big\} \\ \leq 2p_{\boldsymbol{z}}^{1-c_{\mathrm{ep}}} + 2p_{\boldsymbol{z}}^{-c_{\mathrm{ep}}} + \mathrm{P}\big\{\|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\|_\infty > m_{\boldsymbol{Z}}\big\} + \mathrm{P}\,\mathcal{T}_{\boldsymbol{\Theta}}(\mu)^{\complement}\,.\end{aligned}$$

Consequently, if Conditions 1, 2, 6, 7, and 8b of Assumption 4.3.1 also hold, then $\|\boldsymbol{f}_3\|_1 = o_{\mathrm{P}}(1)$.

Proof of Lemma 4.4.14. Lemma 4.4.13 entails that

$$\mathrm{P}\big\{\|\boldsymbol{f}_3\|_\infty > 8m_{\boldsymbol{\Theta}} m_{\mathbf{A}} \sqrt{n}(4m_{\boldsymbol{Z}} \frac{s_{\mathbf{A}}}{\phi^2_{\mathbf{A}}} r_{\mathbf{A}} + \lambda_{\boldsymbol{u}}) \frac{s_{\boldsymbol{\beta}}}{\phi^2_{\boldsymbol{\beta}}} r_{\boldsymbol{\beta}}\big\} \leq \mathrm{P}\,\mathcal{T}_{\boldsymbol{V}}(\lambda_{\boldsymbol{V}})^{\complement} + \mathrm{P}\,\mathcal{T}_{\boldsymbol{u}}(\lambda_{\boldsymbol{u}})^{\complement} + \mathrm{P}\,\mathcal{T}_{\boldsymbol{\Theta}}(\mu)^{\complement}\,.$$

Substitute the present choices of tuning parameters into the display above and cite the estimates of Corollaries 4.4.5, 4.4.3 to deduce the first claim. Expand the the present choices

of tuning parameters to find

$$\begin{aligned} 8m_{\boldsymbol{\Theta}}m_{\mathbf{A}}\sqrt{n}(4m_{\boldsymbol{Z}}\frac{s_{\mathbf{A}}}{\phi_{\mathbf{A}}^2}r_{\mathbf{A}}+\lambda_{\boldsymbol{u}})\frac{s_{\boldsymbol{\beta}}}{\phi_{\boldsymbol{\beta}}^2}r_{\boldsymbol{\beta}} \\ \lesssim\ & s_{\mathbf{A}}^3 s_{\boldsymbol{\beta}}(\log p_{\boldsymbol{z}})^{3/2}/n + s_{\mathbf{A}}^2 s_{\boldsymbol{\beta}}(\log p_{\boldsymbol{z}})/\sqrt{n} \\ & + s_{\mathbf{A}}^2 s_{\boldsymbol{\beta}}(\log p_{\boldsymbol{z}})^{3/2}/n + s_{\mathbf{A}} s_{\boldsymbol{\beta}}\log p_{\boldsymbol{z}}/\sqrt{n}\,, \end{aligned}$$

from which the latter claim follows. □

Lemma 4.4.15 (Control of $\boldsymbol{f}_4$)**.** *Suppose that Assumption 4.2.5 and Condition 5 of Assumption 4.3.1 hold. Choose $\widehat{\boldsymbol{\Theta}}$ according to Assumption 4.3.2; let $\lambda_{\boldsymbol{V}}, \lambda_{\boldsymbol{u}} > 0$ be arbitrary, and set $r_{\boldsymbol{\beta}}$ according to Definition 4.2.6. Then, on the set $\mathcal{T}_{\boldsymbol{V}}(\lambda_{\boldsymbol{V}}) \cap \mathcal{T}_{\boldsymbol{u}}(\lambda_{\boldsymbol{u}}) \cap \mathcal{T}_{\boldsymbol{\Theta}}(\mu)$, the remainder term*

$$\boldsymbol{f}_4 \ =\ \sqrt{n}(\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{I})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})$$

satisfies

$$\|\boldsymbol{f}_4\|_\infty \ \le\ 4\sqrt{n}\mu\frac{s_{\boldsymbol{\beta}}}{\phi_{\boldsymbol{\beta}}^2}r_{\boldsymbol{\beta}}\,.$$

Proof of Lemma 4.4.15. Note first that

$$\begin{aligned} \|\sqrt{n}(\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{I})(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\|_\infty \ &\le\ \sqrt{n}\|\widehat{\boldsymbol{\Theta}}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}} - \boldsymbol{I}\|_\infty\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|_1 \\ &\le\ \sqrt{n}\mu\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|_1\,, \end{aligned}$$

where the latter inequality follows from the specification of $\widehat{\boldsymbol{\Theta}}$ and the fact that Program 3.3.1 is feasible given a. By Lemma 4.4.8, on the set $\mathcal{T}_{\boldsymbol{V}}(\lambda_{\boldsymbol{V}}) \cap \mathcal{T}_{\boldsymbol{u}}(\lambda_{\boldsymbol{u}}) \cap \mathcal{T}_{\boldsymbol{\Theta}}(\mu)$,

$$\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|_1 \ \le\ 4\frac{s_{\boldsymbol{\beta}}}{\phi_{\boldsymbol{\beta}}^2}r_{\boldsymbol{\beta}}\,.$$

Combine the two previous displays to deduce the original claim. □

Lemma 4.4.16 (Control of $\boldsymbol{f}_4$, Gaussian noise)**.** *Suppose that (i) Assumption 4.2.5 and Condition 5 of Assumption 4.3.1 hold and (ii) Assumption 2.1.3 holds. Choose $\widehat{\boldsymbol{\Theta}}$ according*

to Assumption 4.3.2; set $\boldsymbol{r} = (r_j)_{j=1}^{p_{\boldsymbol{x}}}$ *according to Definition 4.2.3; set* $r_{\mathbf{A}}$ *according to (4.2); set* $\lambda_{\boldsymbol{u}}$ *and* $\lambda_{\boldsymbol{V}}$ *according to Definition 4.2.7; and set* $r_{\boldsymbol{\beta}}$ *according to Definition 4.2.6. Then,*

$$\mathrm{P}\big\{\|\boldsymbol{f}_4\|_\infty > 4\sqrt{n}\mu \frac{s_{\boldsymbol{\beta}}}{\phi_{\boldsymbol{\beta}}^2} r_{\boldsymbol{\beta}}\big\} \;\leq\; 2p_{\boldsymbol{z}}^{1-c_{\mathrm{ep}}} + 2p_{\boldsymbol{z}}^{-c_{\mathrm{ep}}} + \mathrm{P}\,\mathcal{T}_{\boldsymbol{\Theta}}(\mu)^{\complement}$$

Consequently, if Conditions 1, 6, 7, and 8c of Assumption 4.3.1 also hold, then $\|\boldsymbol{f}_4\|_1 = o_{\mathrm{P}}(1)$.

Proof of Lemma 4.4.16. Lemma 4.4.15 entails that

$$\mathrm{P}\big\{\|\boldsymbol{f}_4\|_\infty > 4\sqrt{n}\mu \frac{s_{\boldsymbol{\beta}}}{\phi_{\boldsymbol{\beta}}^2} r_{\boldsymbol{\beta}}\big\} \;\leq\; \mathrm{P}\,\mathcal{T}_{\boldsymbol{V}}(\lambda_{\boldsymbol{V}})^{\complement} + \mathrm{P}\,\mathcal{T}_{\boldsymbol{u}}(\lambda_{\boldsymbol{u}})^{\complement} + \mathrm{P}\,\mathcal{T}_{\boldsymbol{\Theta}}(\mu)\,.$$

Substitute the present choices of tuning parameters into the display above and cite the estimates of Corollaries 4.4.5, 4.4.3 to deduce the first claim. Expand the the present choices of tuning parameters to find

$$\sqrt{n}\mu \frac{s_{\boldsymbol{\beta}}}{\phi_{\boldsymbol{\beta}}^2} r_{\boldsymbol{\beta}} \;\lesssim\; \mu s_{\boldsymbol{\beta}}\big(s_{\mathbf{A}}^2 \log p_{\boldsymbol{z}}/\sqrt{n} + s_{\mathbf{A}}\sqrt{\log p_{\boldsymbol{z}}}\big)$$

from which the second claim follows. □

Proof of Lemma 4.3.3. The result follows from Lemmas 4.4.10, 4.4.12, 4.4.14, and 4.4.16. □

Chapter 5

NUMERICAL EXPERIMENTS

In this chapter, we present a Monte Carlo simulation study of the finite-sample properties of the inferential procedure developed in Chapter 3 using the two-stage Lasso studied in Chapter 4. Our objective is to test this method under a variety of parameter configurations chosen to reflect settings of practical interest. In Section 5.1, we describe the general scheme according to which the data for each trial are generated and the metrics gathered for each configuration. In Section 5.2, we enumerate the specific parameter configurations studied and discuss the results.

5.1 General experimental design

Each trial consists of a data-generation step and an estimation step. We specify the regression parameters $\boldsymbol{\beta}$ and $\mathbf{A}$ for the data-generation step as follows. For each configuration, we set the second-stage regression parameter $\boldsymbol{\beta}$ according to $\beta_j = 1$ for $j \in S_{\boldsymbol{\beta}}$ and $\beta_j = 0$ otherwise, where $S_{\boldsymbol{\beta}} \subset [p_{\boldsymbol{x}}]$ is a random set of $s_{\boldsymbol{\beta}}$ indices generated by uniformly random draws from $[p_{\boldsymbol{x}}]$ without replacement. Similarly, we set the first-stage regression parameters $\boldsymbol{\alpha}^j$ for $j \in [p_{\boldsymbol{x}}]$ according to $\alpha_k^j = 1$ for $k \in S_j$ and $\alpha^k = 0$ otherwise, where $S_j \subset [p_{\boldsymbol{z}}]$ is a random set of $s_{\mathbf{A}}$ indices generated by uniformly random draws from $[p_{\boldsymbol{z}}]$ without replacement. We let $s_{\boldsymbol{\beta}}, s_{\mathbf{A}}$ vary over configurations.

Having specified the regression parameters, we then draw n i.i.d. observations $(y_i, \boldsymbol{x}_i, \boldsymbol{z}_i)$

according to

$$\begin{aligned} \boldsymbol{z}_i &\sim \mathcal{N}_{p_{\boldsymbol{z}}}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{z}}) \,, \\ (u_i, \boldsymbol{v}_i) \,|\, \boldsymbol{z}_i &\sim \mathcal{N}_{1+p_{\boldsymbol{x}}}(\mathbf{0}, \boldsymbol{\Sigma}_{u\boldsymbol{v}}) \,, \\ x_{ij} &= \langle \boldsymbol{z}_i, \boldsymbol{\alpha}^j \rangle + v_{ij} \,, \\ y_i &= \langle \boldsymbol{x}_i, \boldsymbol{\beta} \rangle + u_i \,, \end{aligned}$$

where $n, p_{\boldsymbol{x}}, p_{\boldsymbol{z}}, \boldsymbol{\beta}, \{\boldsymbol{\alpha}^j\}_{j \in [p_{\boldsymbol{x}}]}$, and the structure of $\boldsymbol{\Sigma}_{\boldsymbol{z}}$ vary amongst configurations. For all configurations, we set

$$\boldsymbol{\Sigma}_{u\boldsymbol{v}} = \begin{pmatrix} \sigma_u & \boldsymbol{\sigma}_{u\boldsymbol{v}}^\top \\ \boldsymbol{\sigma}_{u\boldsymbol{v}} & \sigma_v \boldsymbol{I} \end{pmatrix} ,$$

where $\sigma_u, \sigma_v = \sqrt{.7}$ are held fixed and $\boldsymbol{\sigma}_{u\boldsymbol{v}} = (\sigma_{uv^1}, \ldots, \sigma_{uv^{p_{\boldsymbol{x}}}})$ is given as follows. For each configuration, we set one σ_{uv^j} chosen at random equal to .5, nine σ_{uv^j} chosen at random equal to .25, and the remaining σ_{uv^j} equal to .05. The present covariance structure for the noise reflects the constraint that $\boldsymbol{\Sigma}_{u\boldsymbol{v}}$ be positive-definite; our choices of σ_{uv^j} are an attempt to balance this requirement with the goal of studying non-trivial correlations between the first- and second-stage noise elements.

We consider two forms for the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{z}}$. The first is a Toeplitz (TZ) structure given by

$$\boldsymbol{\Sigma}_{\boldsymbol{z}}^{\mathrm{TZ}}\big|_{jk} = \rho^{|j-k|} \,, \qquad j, k \in [p_{\boldsymbol{z}}] \,, \qquad \rho = 0.8 \,.$$

The second is a circulant-symmetric (CS) structure given for $j \leq k$ by

$$\boldsymbol{\Sigma}_{\boldsymbol{z}}^{\mathrm{CS}}\big|_{jk} = \begin{cases} 1 & k = j \,, \\ 0.1 & k \in \{j+1, \ldots, j+5\} \cup \{j + p_{\boldsymbol{z}} - 5, \ldots, j + p_{\boldsymbol{z}} - 1\} \,, \\ 0 & \text{otherwise} \,. \end{cases}$$

Within a configuration study, the random quantities $\boldsymbol{z}_i, u_i$, and $\boldsymbol{v}_i$ are re-drawn for each trial; the quantities $\boldsymbol{\beta}, \{\boldsymbol{\alpha}^j\}_{j=1}^{p_{\boldsymbol{x}}}, \boldsymbol{\Sigma}_{\boldsymbol{z}}, \boldsymbol{\sigma}_{u\boldsymbol{v}}, n, p_{\boldsymbol{x}}, p_{\boldsymbol{z}}$ are held fixed.

For the estimation step of each trial, we compute the first- and second-stage Lasso as defined in Section 4.1 using the `glmnet` package [24]. Tuning parameters r for the Lasso

estimators are selected by 10-fold cross-validation over a grid $\{r_\ell\}_{\ell=1}^{L}$, where $L = 100$, $r_L = .01r_1$, and r_1 is the least quantity for which the respective Lasso estimator is identically 0. The rows $\hat{\boldsymbol{\theta}}_j$ of $\widehat{\boldsymbol{\Theta}}$ are obtained as solutions to the respective Program 3.3.1 with tuning paramter μ_j chosen as follows. For each $j \in [p_{\boldsymbol{x}}]$, we set

$$\mu_j \;:=\; \kappa \times \inf_{\boldsymbol{\theta} \in \mathbb{R}^{p_{\boldsymbol{x}}}} \|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}}\boldsymbol{\theta} - \boldsymbol{e}_j\|_\infty \,,$$

where $\boldsymbol{e}_j$ denotes the j^{th} canonical basis vector in $p_{\boldsymbol{x}}$ dimensions. To obtain the infimum, we cast minimize$_{\boldsymbol{\theta} \in \mathbb{R}^{p_{\boldsymbol{x}}}} \|\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}}\boldsymbol{\theta} - \boldsymbol{e}_j\|_\infty$ as a linear programming problem, which we solve using MOSEK optimization software [40]. Note that, under this choice of μ_j, the respective Program 3.3.1 is guaranteed feasible. The factor $\kappa > 1$ is chosen to balance the performance of $\widehat{\boldsymbol{\Theta}}$ as a surrogate inverse for $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{d}}$, for which a smaller κ is desirable, with the size of the objective function $\|\boldsymbol{\theta}\|_1$, for which a larger κ is desirable. The following results were obtained under $\kappa = 1.2$, which was chosen based on practical considerations.

In a given trial, we set $\alpha = 0.05$ and compute the respective $(1-\alpha)\%$ confidence interval

$$\hat{\mathcal{I}}_{\alpha,j} \;=\; \left[\tilde{\beta}_j - z_\alpha \widehat{\mathrm{SE}}(\tilde{\beta}_j),\; \tilde{\beta}_j + z_\alpha \widehat{\mathrm{SE}}(\tilde{\beta}_j)\right],$$

or each component $\tilde{\beta}_j$ of $\tilde{\boldsymbol{\beta}}$, where $z_\alpha = \Phi^{-1}(1-\alpha/2)$ and $\widehat{\mathrm{SE}}(\tilde{\beta}_j) = \sqrt{\mathbb{E}_n[(y_i - \hat{\boldsymbol{\beta}}\boldsymbol{X})^2 \langle \hat{\boldsymbol{\theta}}_j, \hat{\boldsymbol{d}}_i\rangle^2]}$. For each configuration of n, $p_{\boldsymbol{x}}$, $p_{\boldsymbol{z}}$, $s_{\boldsymbol{\beta}}$, $s_{\mathbf{A}}$, $\boldsymbol{\Sigma}_{\boldsymbol{z}}$, we generate $N = 100$ trials and calculate the average coverage $\widehat{\mathrm{cvg}}$ for the 95% confidence intervals $\hat{\mathcal{I}}_{\alpha,j}$ about components of $\tilde{\boldsymbol{\beta}}$ and the average interval length $\widehat{\mathrm{len}}$ given by

$$\widehat{\mathrm{cvg}} \;=\; \frac{1}{p_{\boldsymbol{x}}} \sum_{j=1}^{p_{\boldsymbol{x}}} \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{\beta_j \in \hat{\mathcal{I}}_{\alpha,j}\}\,, \qquad \widehat{\mathrm{len}} \;=\; \frac{1}{p_{\boldsymbol{x}}} \sum_{j=1}^{p_{\boldsymbol{x}}} \frac{1}{N} \sum_{i=1}^{N} \mathrm{len}(\hat{\mathcal{I}}_{\alpha,j})\,.$$

For each configuration, we also provide the average mean squared error of the second-stage Lasso estimator given by

$$\widehat{\mathrm{MSE}} \;=\; \frac{1}{N} \sum_{i=}^{N} \frac{1}{p_{\boldsymbol{x}}} \sum_{j=1}^{p_{\boldsymbol{x}}} (\hat{\beta}_j - \beta_j)^2\,.$$

We present the results for the study described above in Section 5.2.

5.2 Specifications and results

We conduct simulations according the design described in Section 5.1 for all configurations belonging to

$$\underbrace{\begin{pmatrix} (50, 75, 100) \\ (100, 125, 150) \\ (300, 400, 500) \end{pmatrix}}_{(n\,, p_{\boldsymbol{x}}\,, p_{\boldsymbol{z}})} \times \underbrace{\begin{pmatrix} (3, 5) \\ (10, 15) \\ (20, 25) \end{pmatrix}}_{(s_{\boldsymbol{\beta}}\,, s_{\mathbf{A}})} \times \underbrace{\begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{z}}^{\mathrm{CS}} \\ \boldsymbol{\Sigma}_{\boldsymbol{z}}^{\mathrm{TZ}} \end{pmatrix}}_{\boldsymbol{\Sigma}_{\boldsymbol{z}}}.$$

The results, which are presented in Tables 5.1 and 5.2, show that our estimator achieves close to nominal coverage under a variety of configurations. We also see that arguably the greatest determinant of coverage is the relative magnitude of $p_{\boldsymbol{x}}$ and $p_{\boldsymbol{z}}$ to the size of the active set $s_{\boldsymbol{\beta}}$. As the latter grows, coverage diverges from the nominal level. This phenomenon is expected, since the bounds for the estimation error of the Lasso is proportional to the size of the active set. Finally, we observe that the covariance structure of the instrumental variables $\boldsymbol{z}_i$ has a strong influence on coverage: the Toeplitz structure features greater correlation amongst the instrumental variables in general, and this is reflected in coverage that tends to be farther from the nominal level than in the case of the circulant-symmetric covariance structure. These results suggest that our proposed method of inference for the low-dimensional components of a high-dimensional regression vector is relevant to practical scenarios that may exhibit non-trivial degrees of correlation between the noise components and nontrivial correlation amongst the instrumental variables.

Table 5.1: Simulation Results for Circulant-Symmetric $\boldsymbol{\Sigma}_{\boldsymbol{z}}$

$(n,\, p_{\boldsymbol{x}},\, p_{\boldsymbol{z}})$	$(s_{\boldsymbol{\beta}}, s_{\boldsymbol{\alpha}^j})$	$\widehat{\text{cvg}}$	$\widehat{\text{len}}$	$\text{MSE}(\hat{\boldsymbol{\beta}})$
	(3, 5)	0.942	0.225	0.004
(50, 75, 100)	(10, 15)	0.843	0.608	0.058
	(20, 25)	0.647	1.221	0.514
	(3, 5)	0.947	0.157	0.002
(100, 125, 150)	(10, 15)	0.930	0.190	0.003
	(20, 25)	0.752	0.471	0.070
	(3, 5)	0.947	0.094	0.001
(300, 400, 500)	(10, 15)	0.961	0.067	0.000
	(20, 25)	0.959	0.093	0.001

Table 5.2: Simulation Results for Toeplitz $\boldsymbol{\Sigma}_{\boldsymbol{z}}$

$(n,\, p_{\boldsymbol{x}},\, p_{\boldsymbol{z}})$	$(s_{\boldsymbol{\beta}}, s_{\boldsymbol{\alpha}^j})$	$\widehat{\text{cvg}}$	$\widehat{\text{len}}$	$\text{MSE}(\hat{\boldsymbol{\beta}})$
	(3, 5)	0.895	0.201	0.005
(50, 75, 100)	(10, 15)	0.546	0.516	0.171
	(20, 25)	0.333	0.917	0.914
	(3, 5)	0.942	0.140	0.001
(100, 125, 150)	(10, 15)	0.545	0.219	0.030
	(20, 25)	0.232	0.477	0.364
	(3, 5)	0.952	0.092	0.001
(300, 400, 500)	(10, 15)	0.927	0.064	0.000
	(20, 25)	0.729	0.108	0.005

BIBLIOGRAPHY

[1] T. Amemiya. The nonlinear two-stage least-squares estimator. *J. Econometrics*, 2(2):105–110, 1974.

[2] T. Amemiya. The maximum likelihood and the nonlinear three-stage least squares estimator in the general nonlinear simultaneous equation model. *Econometrica*, 45(4):955–968, 1977.

[3] T. Amemiya. *Advanced Econometrics*. Harvard University Press, 1985.

[4] T. Anderson and H. Rubin. The asymptotic properties of estimates of the parameters of a single equation in a complete system of stochastic equations. *Ann. Math. Stat.*, 21(4):570–582, 1950.

[5] J. Angrist and J-S. Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press, 2009.

[6] A. Antoniadis. Comments on: ℓ_1-penalization for mixture regression models. *Test*, 19(2):257–258, 2010.

[7] Y. Baraud, C. Giraud, and S. Huet. Estimator selection in the Gaussian setting. *Ann. Inst. H. Poincaré Probab. Statist.*, 50(3):1092–1119, 2014.

[8] R. Basmann. A generalized classical method of linear estimation of coefficients in a structural equation. *Econometrica*, 25(1):77–83, 1957.

[9] A. Belloni, D. Chen, V. Chernozhukov, and C. Hansen. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429, 2012.

[10] A. Belloni and V. Chernozhukov. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547, 2013.

[11] A. Belloni, V. Chernozhukov, and C. Hansen. Inference for high-dimensional sparse econometric models. *Advances in Economics and Econometrics*, 2011.

[12] A. Belloni, V. Chernozhukov, and L. Wang. Square-root Lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791–806, 2011.

[13] D. Benkeser and M. Van Der Laan. The highly adaptive Lasso estimator. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 689–696, Oct 2016.

[14] P. Bickel, C. Klaassen, Y. Ritov, and J. Wellner. *Efficient and adaptive inference in semiparametric models*. Spring-Verlag, New York, 1998.

[15] P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732, 2009.

[16] J. Bien, I. Gaynanova, J. Lederer, and C. L. Müller. Non-convex global minimization and false discovery rate control for the TREX. *J. Comput. Graph. Statist.*, 2017.

[17] P. Bühlmann and S. van de Geer. *Statistics for High-Dimensional Data*. Springer, 2011.

[18] F. Bunea, J. Lederer, and Y. She. The group square-root Lasso: Theoretical properties and fast algorithms. *IEEE Trans. Inf. Theor.*, 60(2):1313–1325, February 2014.

[19] T. Cai, W. Liu, and X. Luo. A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.*, 106(464):594–607, 2011.

[20] M. Chichignoud, J. Lederer, and M. Wainwright. A practical scheme and fast algorithm to tune the Lasso with optimality guarantees. *J. Mach. Learn. Res.*, 17:1–17, 2016.

[21] D. Chtelat, J. Lederer, and J. Salmon. Optimal two-step prediction in regression. *Electron. J. Statist.*, 11(1):2519–2546, 2017.

[22] A. Dalalyan, M. Hebiri, and J. Lederer. On the prediction performance of the Lasso. *Bernoulli*, 23(1):552–581, 02 2017.

[23] J. Fan and Y. Liao. Endogeneity in high dimensions. *Ann. Statist.*, 42(3):872–917, 2014.

[24] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, 33(1):1–22, 2010.

[25] A. Gallant and D. Jorgenson. Statistical inference for a system of simultaneous, nonlinear, implicit equations in the context of instrumental variable estimation. *J. Econometrics*, 11(2-3):275–302, 1979.

[26] E. Gautier and A. Tsybakov. High-dimensional instrumental variables regression and confidence sets. *ArXiv:1105.2454v4*, 2014.

[27] C. Giraud. *Introduction to High-Dimensional Statistics*. Monographs on statistics and applied probability (Series); 139. CRC Press, Taylor & Francis Group., 2014.

[28] C. Giraud, S. Huet, and N. Verzelen. High-dimensional regression with unknown variance. *Statist. Sci.*, 27(4):500–518, 2012.

[29] L. Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054, 1982.

[30] T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical learning with sparsity: The Lasso and generalizations*. Monographs on statistics and applied probability (Series); 143. Boca Raton: CRC Press, Taylor & Francis Group., 2015.

[31] G. Imbens, S. Donald, and W. Newey. Empirical likelihood estimation and consistent tests with conditional moment restrictions. *J. Econometrics*, 117(1):55–93, 2003.

[32] A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.*, 15:2869–2909, 2014.

[33] B.-Y. Jing, Q.-M. Shao, and Q. Wang. Self-normalized Cramér-type large deviations for independent random variables. *Ann. Probab.*, 31(4):2167–2215, 2003.

[34] H. Kelejian. Two-stage least squares and econometric systems linear in parameters but nonlinear in endogenous variables. *J. Amer. Statist. Assoc.*, 66:373–374, 1971.

[35] K. Knight and W. Fu. Asymptotics for lasso-type estimators. *Ann. Statist.*, 28(5):1356–1378, 10 2000.

[36] M. Van Der Laan and S. Dudoit. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples. *U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 130.*, 2003.

[37] J. Lederer and C. Müller. Don't Fall for Tuning Parameters: Tuning-Free Variable Selection in High Dimensions With the TREX. *arXiv:1404.0541*, 2014.

[38] H. Liu, L. Wang, and T. Zhao. Calibrated multivariate regression with application to neural semantic basis discovery. *J. Mach. Learn. Res.*, 16:1579–1606, 2015.

[39] N. Meinshausen and P. Bühlmann. Stability selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 72(4):417–473, 2010.

[40] MOSEK ApS. *MOSEK Rmosek Package 8.1.0.34*, 2017.

[41] A. Mousavi, A. Maleki, and R. Baraniuk. Consistent parameter estimation for lasso and approximate message passing. *To appear in Ann. Statist.*

[42] W. Newey. Efficient instrumental variables estimation of nonlinear models. *Econometrica*, 58(4):809–837, 1990.

[43] M. Neykov, Y. Ning, J. Liu, and H. Liu. A unified theory of confidence regions and testing for high dimensional estimating equations. *arXiv:1510.08986*, 2015.

[44] B. Pötscher and H. Leeb. On the distribution of penalized maximum likelihood estimators: The LASSO, SCAD, and thresholding. *J. Multivar. Anal.*, 100(9):2065–2082, 2009.

[45] M. Rudelson and S. Zhou. Reconstruction from anisotropic random measurements. *JMLR Workshop Conf. Proc.*, 23:10.1–10.28, 2012.

[46] J. Sargan. The estimation of economic relationships using instrumental variables. *Econometrica*, 26(3):393–415, 1958.

[47] R. Shah and R. Samworth. Variable selection with error control: Another look at stability selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 75(1):55–80, 2013.

[48] N. Städler, P. Bühlmann, and S. van de Geer. ℓ_1-penalization for mixture regression models. *Test*, 19(2):209–256, 2010.

[49] J. Stock and F. Trebbi. Retrospectives: Who invented instrumental variable regression? *J. Econ. Perspect.*, 17(3):177–194, 2003.

[50] T. Sun and C.-H. Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012.

[51] H. Theil. Repeated least-squares applied to complete equation systems. *Centraal Planbureau Memorandum*, 1953.

[52] S. van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.*, 3:1360–1392, 2009.

[53] S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.*, 42(3):1166–1202, 2014.

[54] S. van de Geer and A. Muro. On higher order isotropy conditions and lower bounds for sparse quadratic forms. *Electron. J. Stat.*, 8:3031–3061, 2014.

[55] M. van der Laan, S. Dudoit, and A. van der Vaart. The cross-validated adaptive epsilon-net estimator. *Statistics & Decisions*, 24(3):373–395, 2006.

[56] A. van der Vaart, S. Dudoit, and M. van der Laan. Oracle inequalities for multi-fold cross validation. *Statistics & Decisions*, 24(3):351–371, 2006.

[57] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Yonina C. Eldar and Gitta Kutyniok, editors, *Compressed Sensing: Theory and Applications*, chapter 5, pages 210–268. Cambridge University Press, 2012.

[58] C.-H. Zhang and S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 76(1):217–242, 2014.