NORTHWESTERN UNIVERSITY

Efficient Estimation with Smooth Penalization

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Economics

By

Sergey Gitlin

EVANSTON, ILLINOIS

September 2019

ABSTRACT

Efficient Estimation with Smooth Penalization

Sergey Gitlin

This dissertation proposes an oracle efficient estimator in the context of a sparse linear model. Chapter 1 introduces the penalty and the estimator that optimizes a penalized least squares objective. Unlike existing methods, the penalty is differentiable – once, and hence the estimator does not engage in model selection. This feature allows the estimator to reduce bias relative to a popular oracle efficient method (SCAD) when small, but not insignificant, coefficients are present. Consequently the estimator delivers a lower realized squared error of coefficients of interest. Furthermore, the objective function with the proposed penalty is shown to be convex; paired with differentiability, this ensures good computational properties of the estimator. Simulation evidence illustrates increased robustness of the estimator with the smooth penalty in the presence of small, but nonzero, coefficients.

Chapter 2 focuses on better understanding asymptotic properties of the proposed penalized estimator when the standard asymptotic approximation might be unsatisfactory,

and leveraging that understanding to improve inference. Conventional asymptotic analysis of efficient penalized estimators typically prohibits coefficients of a magnitude that lies in a certain range relative to the sampling error, and it is well understood that allowing for such coefficients can lead to slower rates of convergence of such estimators. I derive the asymptotic distribution for the penalized estimator with the once-differentiable penalty while allowing for coefficients in this range. The analysis is conducted under standard conditions on the tuning parameters, as well as under an alternative asymptotic framework that preserves nonlocal properties of these estimators. Inference by a modified bootstrap procedure is shown to be consistent both under the standard assumptions on tuning parameters that ensure oracle efficiency and under an alternative asymptotic view that excludes intermediate-magnitude coefficients but allows for nonnormal asymptotic distributions arising from penalization. Simulation evidence is presented that shows that the proposed smooth penalty paired with bootstrap inference provides good coverage together with smaller confidence intervals even under violations of exact sparsity that lead to poor performance by model-selection-based estimators.

Finally, Chapter 3 applies the proposed approach to reevaluate the effect of location-specific human capital on agricultural output using the data and framework of Bazzi et al. (2016). Authors consider a relocation program carried out in Indonesia that created exogenous variation in where migrants were settled. A key estimate in that work is that a one-standard-deviation increase in a measure of agricultural similarity between migrants' origins and destinations produced a 20% increase in rice productivity. The estimate comes from a regression with a relatively small set of controls, and I find that a more plausible estimate of the effect is 11%. Authors' original result appears to be driven by

omitted variable bias due to not including a small number of important controls, most notably education level of the local population, as measured by the average years of schooling of the locals. This finding fits well into the human capital transition mechanism envisioned in the original paper, in which interactions with more experienced locals would improve migrants' productivity. Since the agricultural similarity happens to correlate with local education levels in the dataset under study, omitting schooling from the regression increases the estimate on agricultural similarity.

Acknowledgements

I thank Joel Horowitz and Ivan Canay for the generous guidance, insightful advice and endless patience that made this work possible. Any and all errors are my own.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Efficient Estimation with Smooth Penalization

1.1. Introduction

Advances in traditional survey data collection and increasing availability of detailed data generated by tech companies have resulted in economic datasets that are 'big' in many ways: in particular both in the number of observations and in the number of covariates available for each observation. Consequently, improvements in data have enabled researchers to answer old questions with more precision and detail and ask new ones that would have previously been impossible to answer.

Conventional econometric tools in common use by practitioners are well suited to take advantage of the first aspect of Big Data – large number of observations – since their properties are typically derived for exactly such asymptotics. However, the second aspect – large number of available covariates, known as high dimensionality in the statistics literature and popularized by Varian (2014) as "fat data" – is less well covered in standard econometric toolset used in applied work.[1]

While many covariates may be available and may conceivably be related to the chosen outcome of economic interest, it is reasonable to suppose that some of those will not have a significant impact on that outcome of interest. This is known as *sparsity*: in the

[1]The idea that modeling grows in complexity as more observations become available is not new and predates the advent of Big Data. In a meta-analysis of wage equation literature Koenker (1988) decisively rejects the idea of fixed model dimensionality, in particular finding that in his sample model dimensionality grows as the fourth-root of sample size.

linear model context, the vector of coefficients is thought to be sparse, that is, has many elements that are zero.[2] Since including unnecessary covariates in a regression typically increases the variance of the estimates, and omitting important ones leads to omitted variable bias and ensuing inconsistency, a central question facing any applied researcher with a high-dimensional dataset is how to proceed with estimation without suffering from either predicament.[3]

I propose a rigorous approach to this problem with an estimator that exhibits optimal behavior in the standard sparse model settings and delivers improved performance under the challenging setting of deviations from the standard sparsity assumptions. I provide conditions on tuning parameters that allow the researcher to ensure convexity of the objective function that the estimator optimizes; paired with differentiability of the objective function this ensures that the estimator is computationally unburdensome. Such computational tractability is especially appealing in the practical settings if the researcher decides to use crossvalidation or bootstrap.

The approach uses penalized estimation with a novel penalty function. Key difference of the proposed penalty that sets it apart from existing methods is differentiability, specifically, the penalty is once differentiable at zero, with an infinite second derivative. I show that under conventional assumptions the proposed estimator achieves the *oracle*

[2]More generally, one can consider *approximate sparsity*, where some of the coeffients are very small, even if not exactly zero.

[3]A common approach to this problem involves researcher's judgement on what handful of covariates ought to be sufficiently relevant to warrant inclusion in the regression. To guard against omitted variable bias in particular, multiple regressions of varying dimensionality may be carried out (possibly with corresponding tests and levels of significance), so as to argue that the results aren't driven by the vagaries of the modeling choices. Unfortunately, such an approach is rarely carried out as a formally well-defined econometric procedure, and as such little can be said about the properties of the estimators, particularly when the final 'headline number' is chosen ad hoc from multiple reported specifications.

property: that is, it has the same asymptotic distribution as the *oracle estimator* that knows which covariates can be dropped from estimation. This result is similar to many other methods in the literature; nonetheless, in contrast to the existing literature this is the first penalized estimator with a differentiable objective function to do so.

However, the proposed estimator substantially improves upon a popular existing method (SCAD by Fan and Li (2001)) under very general conditions when conventional sparsity assumptions underpinning oracle efficiency results are violated. Under such violations my method reduces the worst squared error in the estimates of coefficients of interest. The magnitude of improvement is larger than the sampling error, and is driven by a reduction in omitted variable bias.

Conventional sparsity assumptions typically involve two restrictions: on the one hand, some coefficients are sufficiently large (larger than the sampling error); on the other hand, all the other coefficients are sufficiently small: either exactly zero (*exact sparsity*), or sufficiently smaller than the sampling error (*approximate sparsity*). This creates a *gap* in allowed coefficient magnitudes, and the common approach assumes no coefficients fall in that gap.

As has been highlighted by Leeb and Pötscher (2008) and related research, model-selection-consistent estimators like SCAD suffer from bad bias properties when we allow for coefficients to be in the gap, essentially due to dropping covariates that should not have been dropped. While the proposed estimator also suffers from a similar bias predicament, it suffers from it to a lesser extent. This is achieved due to a reduction in omitted variable bias inherent in model-selection-based estimators like SCAD, since the proposed estimator does not drop any covariates from the regression, and as such does not engage in model

selection. Curiously, the reduction in squared error is larger than the terms due to random sampling error (in particular the term that gives variance in expansion for squared error).

It is worth discussing the broader context of penalized estimation and its evolution as a tool of improving estimator precision. Many econometric estimators can be defined as maximizers or minimizers of certain objective functions (such as the sum of squared residuals in the case of OLS), and penalized estimation methods amend the objective by adding a *penalty* term that affects the properties of the estimator in a certain way that depends on the choice of the penalty. For example, information criteria, such as AIC (Akaike (1974)) and BIC (Schwarz (1978)), penalize the number of included covariates, i.e. the l_0 norm of the coefficient vector. We can imagine the researcher considering all possible subsets of covariates and using such information criteria to select the 'best' model, which would seem to be not too far from the heuristic procedure described earlier. However, even though a procedure based on l_0 penalization can be assured of good statistical properties in theory, searching over all possible subsets of covariates is a combinatorial problem, and hence such an approach is impossible in practice with any nontrivial number of potential regressors.

Since penalizing the l_0 norm is computationally unappealing, we can consider using a continuous penalty function to simplify optimization while possibly achieving the same effect. In a seminal contribution to this line of inquiry, Tibshirani (1996) proposed using l_1 penalty on estimated coefficients added to the least-squares objective as a way to simultaneously estimate coefficients on the 'relevant' covariates and eliminate some of the 'irrelevant' ones. The procedure, known as LASSO, ameliorates the variance problem of having too many covariates. LASSO achieves model size reduction by having a cusp

at zero with nonzero directional derivative on either side, which means that when the smooth main objective function is nearly flat in a given covariate coefficient at zero, Kuhn-Tucker conditions ensure that zero value for that coefficient is part of the solution to the optimization problem.

There are two potential issues with LASSO. One is that, except under very restrictive conditions, LASSO will select a model that is larger than necessary. That is, it will not eliminate all of the unnecessary regressors, and so the estimation will not be as efficient as possible. Another issue with LASSO is the bias it introduces. The fact that the coefficients are penalized in the same way (in terms of derivative of the penalty) regardless of their magnitude forces a tradeoff between bias for large coefficients and ability to drive small ones to zero.[4]

Various approaches have since been proposed to improve the properties of penalized estimation, notably bridge estimation, smoothly-clipped absolute deviation penalty (SCAD), and adaptive LASSO. They all aim to penalize coefficients that are close to zero more heavily than those that are further away, thereby reducing the bias for large coefficients without sacrificing model selection prowess.

SCAD, proposed by Fan and Li (2001), 'flattens' the penalty above a certain threshold, and keeps LASSO l_1 penalty below that threshold. If the threshold converges to zero at a slow enough rate, then all the large coefficients will eventually fall in the flat part of the penalty and hence won't suffer any penalization bias, while all the small ones will still be inside the penalized range and will be driven to zero. Fan and Li (2001) establish the asymptotic oracle property for SCAD: the resulting estimator has the same asymptotic

[4]This bias can be eliminated by using a post-selection estimator that keeps the covariates LASSO chose but does not penalize.

distribution as the oracle estimator: one that knows exactly which covariates should be included in the model from the outset. While theoretically attractive, SCAD suffers from finite-sample issues stemming from the reasons outlined above and the fact that it might not select the right model some of the time.

Huang, Horowitz, and Ma (2008) derive asymptotic properties (including oracle efficiency) for bridge estimators (see also Knight and Fu (2000)), defined as least-squares estimators with l_d penalty with $d \in (0, 1)$. This penalty function naturally has the desirable features of small derivative for large arguments and large (going to infinity) for small ones. However, unlike SCAD, penalized solutions are discontinuous in data: small variations in data may result in a jump in the value of the estimator from zero to something away from zero (or vice-versa).

Zou (2006) proposes weighting LASSO penaly with the inverse of a preliminary estimator, in effect penalizing smaller coefficients more. Adaptive LASSO retains attractive convexity property of LASSO while at the same time approximating the behavior of bridge estimators (see also Hastie, Tibshirani, and Friedman (2009)).

What all of these methods (and other similar approaches) share is the fact that they first solve the problem of *model selection*: they estimate some coefficients in the regression as exactly zero, effectively omitting corresponding covariates from estimation. Since under appropriate conditions these methods achieve *model selection consistency*, i.e. only keep the covariates needed and only drop those not needed, and since they don't incur too much bias for large coefficients, they achieve *oracle efficiency*: asymptotically, the estimates of large coefficients behave as if the researcher knew the right model from the start.

Nonetheless, as will be shown here, model selection (i.e. omitting the covariates deemed irrelevant) is not a necessary condition for oracle efficiency. In fact, it is enough for the coefficients on irrelevant covariates to converge to zero fast enough to avoid excess variance cost introduced by their estimation. In particular, we can use a differentiable penalty function to achieve the same asymptotic properties of the estimator. At the same time, first-order changes in the derivative of the usual least-squares objective function around sparse solutions ought to correspond to smaller changes in the estimate, and as such the second derivative of the objective at zero can't be finite. I therefore propose to use the l_q penalty around zero that is flattened away from zero as a solution that satisfies both of these requirements.

As the objective function will now be differentiable, none of the estimated coefficients will be exactly zero. As such, all of the regressors will be kept in the model, but those that would be estimated as zero by model-selection-consistent methods will have very small estimated coefficient values in my approach. While this is not, in and of itself, a benefit when the coefficients can be separated into large and small ones by the 'gap' assumption, it will allow us to reduce omitted variable bias in the worst cases when coefficients that are neither large nor small (I'll call them 'intermediate-magnitude') are present. As such, the estimates derived from my method will be more precise than those from SCAD in this setting; the practical significance of this being the improved robustness of the estimator against deviations from the standard sparsity assumptions that will be clarified in the main text.

While the proposed penalty is modelled overall on the SCAD penalty, the key distinction lies in the shape of the penalty around zero. To this end, the earliest use of l_q

regularization with $q \in (1, 2)$ among existing literature is Frank and Friedman (1993), who briefly consider it (as part of l_q with $q \in (0, \infty)$) as an extension of LASSO and ridge regression. Zou and Hastie (2005) also mention it as a possibility before settling on the elastic net as a different compromise between LASSO and ridge. Neither work evaluates its theoretical properties further. Knight and Fu (2000) consider bridge estimators with l_q penalties for $q > 0$ and derive some theoretical properties for them; however, due to the excessive bias incurred for large coefficients when $q > 1$, in their framework it is not possible to achieve efficient estimation in sparse models. None of the works mentioned recognizes the possibility of using such a method as an ingredient to achieve oracle efficiency without explicit model selection.

Theoretical results presented in this work are supported by evidence from simulations. In particular, when small nonzero coefficients are present, I find that smoothing out the penalty does reduce simulated mean squared error of coefficients of interest as expected. The optimization procedure is also extremely fast in practice, even with a general-purpose optimizer. The speed also seems to increase with smoothness, with smoother penalty resulting in faster estimation. Good numerical properties are aided by the fact that the objective function is convex; a simple condition ensuring convexity is provided.

The rest of the chapter is organized as follows. Section 1.2 presents the model and the motivation for seeking oracle efficiency. Section 1.3 introduces the proposed smooth penalty function and the corresponding penalized least squares estimator. Section 1.4 presents the three key properties of the method: objective function convexity, oracle efficiency under conventional assumptions and improvement in realized squared error over SCAD when insufficiently-small coefficients are allowed. Section 1.5 provides simulation

evidence illustrating increased robustness offered by the smooth penalty. Section 1.6 offers concluding remarks. All proofs are in the appendix.

1.2. Model and motivation

This section describes the key features of the model and provides motivation for using the proposed estimator. I will start by presenting the model in a notation that best illustrates the practical setting envisioned here. However, this notation is cumbersome for the theoretical arguments, and a more formal discussion is aided by a simplified notation. For this reason this second notation will be introduced, and the rest of the paper beyond this section will follow it.

1.2.1. The model

We have a sample of n observations from a linear model:

$$y_i = w_i'\theta + z_i'\psi + \varepsilon_i, \tag{1.1}$$

where w_i is a $k_0 \times 1$ vector of *covariates of* ex ante *interest* and z_i is a $(p_n - k_0) \times 1$ vector of potentially relevant *controls*.[5] All regressors are nonstochastic and standardized; details of this are addressed in Remark 1. ε_i is the error term with $E(\varepsilon_i) = 0$. p_n and k_0 are allowed to change with n; this feature will be addressed in Remark 2 at the end of the section.[6]

[5]We can also entertain a situation where there are no covariates of *ex ante* interest, and we want to estimate ψ, or all (or some) of its nonzero components. That is, the set of covariates of *ex ante* interest can be an empty set, i.e. $k_0 = 0$ is allowed. All theoretical results in this paper apply to this case.
[6]Dependence of θ and ψ on n is notationally suppressed for simplicity.

A key feature of the true model is that it possesses a *sparse* structure: that is, some coefficients on control covariates are large (and 'important'), and others are small (and 'unimportant'). 'Large' and 'small' is relative to the sample size; the exact details of this will be discussed in Section 1.4.[7] For example, any coefficient that does not vary with sample size and is not zero is 'large'. A fixed coefficient that is exactly zero is 'small'. For simplicity, I will assume *exact sparsity*: all small coefficients are exactly zero.[8] That is, some elements of ψ are equal to zero. Without loss of generality I order the control covariates in such a way that all the the controls with large coefficient values come first, followed by controls with zero coefficient values: $\psi = (\psi_1', \psi_2')'$, where $\psi_2 = 0$. I will use k_1 to denote the length of ψ_1, and let $k_n = k_0 + k_1$. Splitting z_i into $(z_{1,i}', z_{2,i}')'$ accordingly we can write Equation 1.1 as

$$y_i = w_i'\theta + z_{1,i}'\psi_1 + z_{2,i}'\psi_2 + \varepsilon_i.$$

Taking into account $\psi_2 = 0$ we can also write

$$y_i = w_i'\theta + z_{1,i}'\psi_1 + \varepsilon_i.$$

The researcher knows which covariates go into w_i versus z_i, but does not know which components of z_i are in $z_{1,i}$ and which are in $z_{2,i}$.

Since the role played by *ex ante* important covariates and controls with large coefficients will be similar, I adopt a notation that unifies them. Let $\beta_{10} = (\theta', \psi_1')'$ and

[7] Coefficients on covatiates of *ex ante* interest can be large, small or anything in between; due to their special status we want them in the regression regardless of the true value of their coefficients.

[8] An alternative approach would be to allow for coefficients to be 'small' but nonzero (approximate sparsity) similar to Horowitz and Huang (2013). For our purposes this costs more in clarity of presentation than it offers in insight, and so will be avoided here, but further addressed in subsection 1.4.3.

$\beta_{20} = \psi_2$, so that $\beta_0 = (\beta_{10}', \beta_{20}')' = (\theta', \psi')'$. Similarly, let $x_{1,i} = (w_i', z_{1,i}')'$, $x_{2,i} = z_{2,i}$ and $x_i = (x_{1,i}', x_{2,i}')'$. Then Equation 1.1 can be rewritten as

$$y_i = x_{1,i}'\beta_{10} + x_{2,i}'\beta_{20} + \varepsilon_i. \tag{1.2}$$

Taking into account $\beta_{20} = \psi_2 = 0$ we can also write

$$y_i = x_{1,i}'\beta_{10} + \varepsilon_i. \tag{1.3}$$

I will refer to the model in Equation 1.2 (equivalently Equation 1.1) as the *full model*. It will be assumed throughout that it is feasible to estimate this model by OLS, and thus I will refer to the OLS estimator of this model as the *full-model OLS estimator*.

I will call the model in Equation 1.3 the *oracle model*: this is the smallest parsimonious model that includes the covariates of interest. Since the researcher does not know which controls need to go into $x_{1,i}$, the *oracle OLS estimator* is infeasible.

Remark 1. Some details on regressor properties need to be addressed. First, following a common approach in the literature (see e.g. Huang, Horowitz, and Ma (2008)) I take the $n \times p_n$ matrix of covariates X_n to be nonstochastic. Equivalently, the results can be thought of as conditional on stochastic regressors. Second, I do not include a constant in the model; if needed, it can be accomodated by de-meaning the regressors and the outcome variable. Therefore, while not required for any results, it can be implicitly understood that each regressor has sample average zero. Third, the regressors are standardized. Let $X_n = (X_{1n}, X_{2n})$, where X_{1n} is the $n \times k_n$ matrix of *ex ante* important regressors and important controls. Let $\Sigma_n = \frac{1}{n} X_n' X_n$ be the regressor sample covariance matrix, which

is partitioned according to partitioning in Equation 1.2:

$$\Sigma_n = \frac{1}{n} X_n' X_n = \begin{pmatrix} \frac{1}{n} X_{1n}' X_{1n} & \frac{1}{n} X_{1n}' X_{2n} \\ \frac{1}{n} X_{2n}' X_{1n} & \frac{1}{n} X_{2n}' X_{2n} \end{pmatrix} = \begin{pmatrix} \Sigma_{1n} & \tilde{\Sigma}_n \\ \tilde{\Sigma}_n' & \Sigma_{2n} \end{pmatrix}. \tag{1.4}$$

I assume that all diagonal elements of Σ_n are equal to 1: the regressors are normalized to have unit sample variance. This is without loss of generality for the theoretical results. The same normalization approach should be followed in practice, with an appropriate rescaling of coefficients after estimation to correspond to the original scale of regressors.

Remark 2. As is implied by the notation, total number of covariates is allowed to change with n, and in particular it can grow with n. As a practical matter, growing p_n should not be interpreted literally, and is rather a theoretical tool to capture some features and limitations of estimators when the total number of covariates is large relative to sample size. One can imagine two obvious scenarios: more observations than covariates, and more covariates than observations. This work is focused on the first case: $p_n < n$. While there is a lot of research on the case of $p_n > n$ (see e.g. Van de Geer et al. (2014) and references therein), the case of $p_n < n$ is more prominent in applied economic research, where the number of covariates researchers commonly entertain for regression analysis can be large but is still usually smaller than the sample size. Focusing on this case allows us to avoid more restrictive conditions required in the $p_n > n$ case. On the other hand, the question of how to estimate linear models in the case of unknown 'optimal' specification is clearly present even for $p_n < n$, evidenced by the common practice of estimating a linear model with varying sets of controls and reporting the results of multiple specifications, as is done in Bazzi et al. (2016), chosen as the empirical application here.

The two main asymptotic results will differ in their stance on p_n. While the main result under conventional assumptions (Theorem 2) will allow a growing p_n, the result under violation of the 'gap' assumption (Theorem 3) will maintain that p_n is fixed. For the first result, growing p_n is required to highlight the cost of using a smooth penalty: namely, a slower-growing number of covariates can be accomodated with a smoother penalty. For the second result, even fixed p_n is sufficient to show that the proposed estimator improves worst squared error relative to SCAD. In exchange for giving up model complexity flexibility afforded by growing p_n, fixed p_n allows the proofs to be constructed in such a way so as to accomodate a broad class of problems under very weak assumptions on the error term.

1.2.2. Motivation: efficiency

In general terms, this work (and many other approaches in the literature) seeks to come up with a feasible procedure that replicates some of the properties of the oracle estimator. Since the full-model OLS estimator is feasible and the oracle estimator is not, it is worth asking whether there is any benefit to trying to mimic the latter.

To this end, consider a simplified model where the number of regressors is fixed and all coefficients are fixed. Moreover, $\Sigma_n \xrightarrow[n\to\infty]{} \Sigma > 0$, errors ε_i are i.i.d. with mean zero and variance σ^2, and regularity conditions on regressors are satisfied so that $\frac{1}{\sqrt{n}} \sum_{i=1}^n x_i \varepsilon_i \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \sigma^2 \Sigma)$.

Let

$$\hat{\beta}_F = \arg\min_b \sum_{i=1}^n (y_i - x_i' b)^2$$

be the OLS estimator of β_0 in the full model.

Then in the case of estimating the full model by OLS we have

$$\sqrt{n}(\hat{\beta}_F - \beta_0) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \sigma^2\Sigma^{-1}).$$

In particular, consider the asymptotic distribution of $\hat{\beta}_{F,1}$, the first k components of $\hat{\beta}_F$ that correspond to covariates of interest and controls with nonzero coefficient values. The expression above implies that

$$\sqrt{n}(\hat{\beta}_{F,1} - \beta_{10}) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \sigma^2 \left[\Sigma_1 - \tilde{\Sigma}\Sigma_2^{-1}\tilde{\Sigma}'\right]^{-1}),$$

where Σ_1, Σ_2 and $\tilde{\Sigma}$ are the corresponding limits of the parts of Σ_n defined in Equation 1.4.

On the other hand, for the infeasible oracle estimator $\hat{\beta}_O$ of β_{10} we have

$$\sqrt{n}(\hat{\beta}_O - \beta_{10}) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \sigma^2\Sigma_1^{-1}).$$

It is easy to verify that $\Sigma_1^{-1} - \left[\Sigma_1 - \tilde{\Sigma}\Sigma_2^{-1}\tilde{\Sigma}'\right]^{-1} \leq 0$ (in the sense of being negative semidefinite).[9] That is, for any linear combination of coefficients from β_{10}, oracle estimator is at least as efficient as, and potentially more efficient than, the full-model OLS estimator: the asymptotic variance and MSE are weakly lower. In the same vein, we can show that confidence intervals for coefficients of interest constructed using the oracle model will be smaller than those constructed using the full model.

1.3. The estimator

As has been discussed in the introduction, penalized estimation is a popular approach to achieving oracle efficiency, usually as a result of model selection. I also follow penalized

[9] In general it will be negative definite if $\tilde{\Sigma}\Sigma_2^{-1}\tilde{\Sigma}'$ is positive definite.

estimation approach, but one that does not engage in model selection, as will be clear from the discussion of the penalty function.

I estimate β_0 by the following penalized least squares estimator:

$$\hat{\beta}_n = \arg\min_b Q_n(b), \tag{1.5}$$

$$Q_n(b) = \sum_{i=1}^{n}(y_i - x_i'b)^2 + \lambda_n \sum_{j=k_0+1}^{p_n} \text{Pen}(b_j), \tag{1.6}$$

where the penalty function with smoothness parameter $\gamma \in (0,1)$ is

$$\text{Pen}(b) = \tau_n^{1+\gamma} \begin{cases} \frac{2}{1+\gamma}\left(\frac{|b|}{\tau_n}\right)^{1+\gamma}, & \text{when } \frac{|b|}{\tau_n} < 1; \\ \frac{1-\gamma}{1+\gamma} + a - \frac{1}{a-1}\left(a - \frac{|b|}{\tau_n}\right)^2, & \text{when } \frac{|b|}{\tau_n} \in [1,a]; \\ \frac{1-\gamma}{1+\gamma} + a, & \text{when } \frac{|b|}{\tau_n} > a. \end{cases} \tag{1.7}$$

Equivalently, the derivative of the penalty function is

$$\text{Pen}'(b) = 2\tau_n^{\gamma}\,\text{sgn}(b) \begin{cases} \left(\frac{|b|}{\tau_n}\right)^{\gamma}, & \text{when } \frac{|b|}{\tau_n} < 1; \\ \frac{a - \frac{|b|}{\tau_n}}{a-1}, & \text{when } \frac{|b|}{\tau_n} \in [1,a]; \\ 0, & \text{when } \frac{|b|}{\tau_n} > a. \end{cases} \tag{1.8}$$

Note that the objective function does not penalize coefficients on covariates of interest (i.e. the first k_0 components of β, or θ in the notation of Equation 1.1).

Observe that this one-dimensional penalty function is the $l_{1+\gamma}$ penalty[10] around zero paired with SCAD-like flattening away from zero. As such, the penalty, unlike SCAD,

[10]Notice that we do not raise the sum of $|b_j|^{1+\gamma}$ to the inverse power in order to make it the $l_{1+\gamma}$ norm of the coefficient vector.

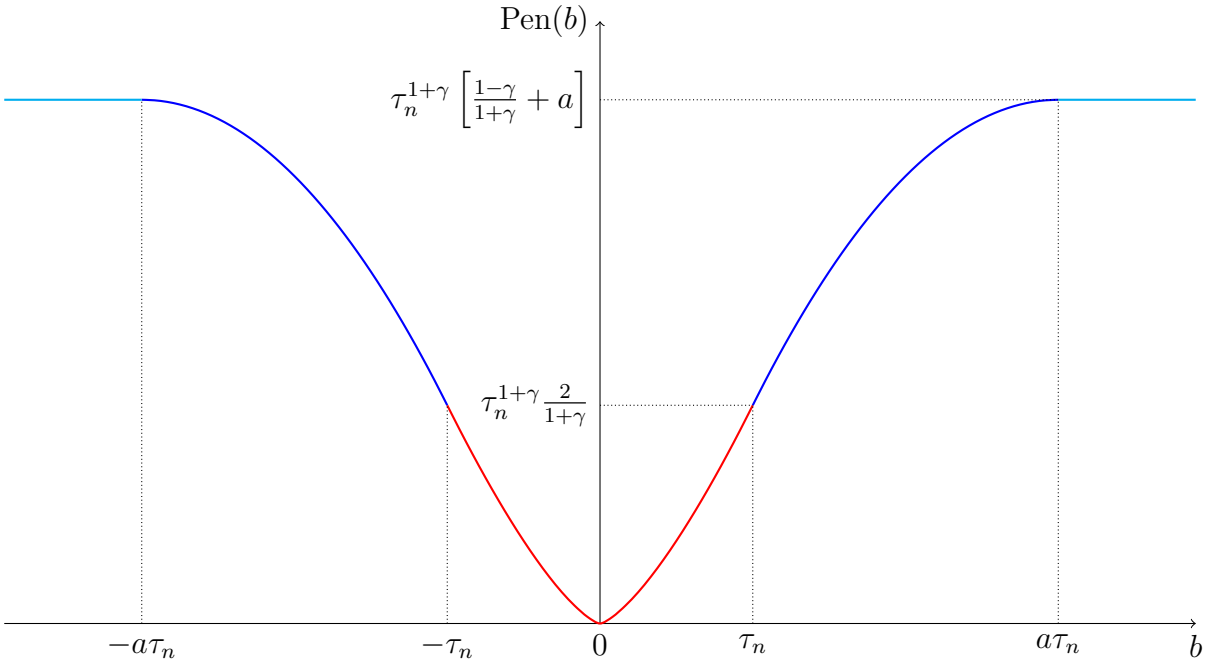

Figure 1.1. Plot of the penalty function for $\gamma = 0.4$, $\tau_n = 1$, $a = 3$.

LASSO or bridge, does have the first derivative at zero, but not the second (more specifically, the second derivative at zero is infinity: that is, the limit of it from both sides is $+\infty$). Notice that at one end of the spectrum $\gamma = 0$ corresponds to SCAD penalty, while at the other $\gamma = 1$ would be ridge penalty that is flattened away from zero similar to SCAD.

There are three tuning parameters: λ_n, τ_n and a, although a does not depend on the sample size. Moreover, one can think of γ as a choice variable as well, although any choice consistent with the relevant assumptions is acceptable. As will be seen in the discussion to follow, we can also make λ_n a function of τ_n, as is done in Fan and Li (2001), which would notionally reduce the number of tuning parameters. However, the roles λ_n and τ_n play are subtly different, and in particular pushing the boundaries on the allowed magnitude

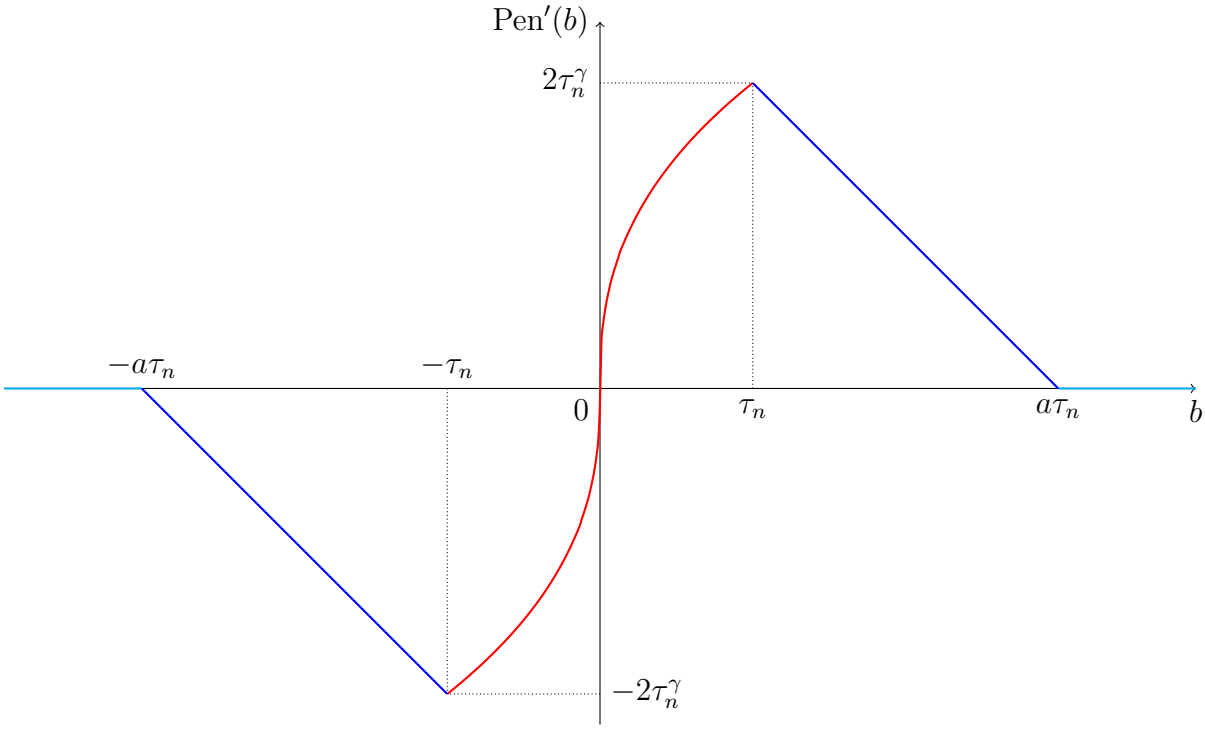

Figure 1.2. Plot of the derivative of the penalty function for $\gamma = 0.4$, $\tau_n = 1$, $a = 3$.

of λ_n can allow us to capture some finite-sample features of penalized estimators that are absent from the conventional asymptotic framework common in the literature.[11] Since we can still achieve the desired results with more flexibility in the choice of λ_n and τ_n than what would be dictated by the functional linkage similar to that of Fan and Li (2001), I will keep λ_n and τ_n separate for the sake of generality.

1.4. Properties of the estimator with smooth penalty

This section presents three theoretical properties of the estimator. The first one is convexity of the objective function, which is desirable from the computational point of view. This is a property shared by adaptive LASSO and by SCAD (under restrictions on

[11] This is the subject of Chapter 2 that focuses on inference with penalized estimators.

tuning parameter choices), but not by bridge estimators as defined in Huang, Horowitz, and Ma (2008). The second result is asymptotic normality and oracle property in a homoskedastic model with exact sparsity with a growing number of regressors. Even though oracle property is not unique to the proposed estimator (it is shared by SCAD, adaptive LASSO and bridge estimators under similar conditions), it is perceived as a necessary requirement in a sparse model. What is unique about the estimator with the smooth penalty is that it achieves oracle efficiency without model selection. The third result shows the benefits of it. It allows for approximate sparsity with small coefficients 'insufficiently small'; the result establishes that a sufficiently smooth penalty delivers a lower squared error than SCAD with probability approaching one. In other words, the proposed estimator is more robust to such scenarios, and this feature is a direct result of not engaging in model selection.

1.4.1. Convexity of the objective function

This part provides conditions under which the objective function is convex. While some penalized objective functions, such as least-squares with LASSO (and by extension adaptive LASSO) are obviously convex, the proposed penalty is not by itself convex, and so establishing convexity of the objective is not immediate. The following assumption provides conditions on tuning parameter choices that ensure convexity of the objective.

Assumption A1 (Objective function convexity and continuity of solutions)**.** *(a) Let* I_{p_n,k_0} *be a diagonal* $p_n \times p_n$ *matrix with the first* k_0 *diagonal elements equal to zero and*

the remainder equal to one. We have

$$\Sigma_n - \frac{\lambda_n \tau_n^{\gamma-1}}{n} \frac{1}{a-1} I_{p_n,k_0} > 0,$$

in the sense of the matrix being positive definite.

(b) (sufficient condition for A1(a)) Let $\rho_n = \text{Eig}_{\min}(\Sigma_n)$. $\rho_n > 0$ *and*

$$\frac{\lambda_n \tau_n^{\gamma-1}}{\rho_n n} < a - 1.$$

Assumption A1(b) provides a simple sufficient condition for A1(a) and is equivalent to it in the case $k_0 = 0$, that is, when all coefficients are penalized.

Lemma 1. *Suppose assumption A1(a) holds. Then the objective function* $Q_n(b)$ *is strictly convex.*

This result is important for the practical implications it carries. Convex optimization is a well-studied subject; a multitude of approaches exist to solving large-scale convex problems, and even conventional hill-descent algorithms can be guaranteed to converge to the unique global minimum in case of strict convexity. Contrast that with minimizing a non-convex (more specifically, non-quasi-concave) function, where there is a possibility of multiple local minima and no general way of knowing whether the estimator has converged to the global minimum.

Adding that the objective function is also differentiable (due to differentiable penalty) means that analytical gradient can be supplied to optimization algorithms, speeding up optimization in practice.

Remark 3. Since the proposed penalty converges uniformly to SCAD penalty as $\gamma \to 0$, a direct corollary to Lemma 1 is that the objective function for SCAD-penalized least squares is convex if the choice of tuning parameters for SCAD satisfies A1(a). In particular, since what I have denoted as λ_n and τ_n are linked in the original formulation of SCAD in Fan and Li (2001), a sufficient condition for convexity in that case is $\rho_n^{-1} < a-1$. The same condition has been previously derived in Zhang (2010).

1.4.2. Asymptotic properties: conventional framework

The results here establish consistency and asymptotic normality of the proposed penalized estimator, following the strategy adopted by Huang, Horowitz, and Ma (2008). Main focus here is on Theorem 2, and the consistency and superefficiency results are ingredients for it. I will state the assumptions and results first and then discuss the significance of results and some implications of the assumptions.

It is useful to introduce some notation first. For a real symmetric matrix A, let $\text{Eig}_{\max}(A)$ and $\text{Eig}_{\min}(A)$ denote the largest and smallest eigenvalues of A. For a vector v of length l, let $\|v\|_q = \left(\sum_{j=1}^{l} |v_j|^q\right)^{1/q}$, and $\|\cdot\| = \|\cdot\|_2$. The $\|\cdot\|_\infty$ norm is defined accordingly.

Assumption A2 (Errors). *(a) $\varepsilon_i, i = 1, \ldots, n$ are independent with mean zero.*

(b) $\varepsilon_i, i = 1, \ldots, n$ have nonzero finite variance σ^2.

Assumption A3 (Design). *(a) Let $\rho_n = \text{Eig}_{\min}(\Sigma_n)$. $\rho_n > 0$ for all n.*

(b)

$$\frac{1}{n} \max_{i=1,\ldots,n} x_{1,i}' \Sigma_{1n}^{-1} x_{1,i} \xrightarrow[n\to\infty]{} 0.$$

(c)

$$\frac{1}{n}\max_{i=1,\ldots,n} x_i'\Sigma_n^{-1}x_i \xrightarrow[n\to\infty]{} 0.$$

Assumption A4 (Parameters)**.** *(a) Let* $b_n = \min\{|\beta_{10j}|, k_0+1 \le j \le k_n\}$ *(*$b_n > 0$ *for all* n*). We have*

$$\tau_n + \left[\frac{p_n + \lambda_n k_n \tau_n^{1+\gamma}}{n\rho_n}\right]^{1/2} = o(b_n).$$

(b)

$$\tau_n \left(\rho_n^{-1}\left(\frac{p_n}{n}\right)^{1/2}\right)^{-1} \xrightarrow[n\to\infty]{} \infty.$$

(c) Let $\kappa_{1n} = \mathrm{Eig}_{\max}(\Sigma_{1n})$ *and* $\kappa_{2n} = \mathrm{Eig}_{\max}(\Sigma_{2n})$*. We have*

$$\left[\frac{p_n\kappa_{1n}}{\rho_n^2(\lambda_n/n^{(1+\gamma)/2})}\right]^{\frac{1}{1+\gamma}} \kappa_{2n}^{1/2} \xrightarrow[n\to\infty]{} 0.$$

The following theorem shows that under the maintained assumptions the proposed estimator is consistent at the same rate as the full-model OLS estimator.

Theorem 1 (Consistency)**.** *Suppose assumptions A2(a), A2(b) and A3(a) hold. Then*

$$\|\hat{\beta}_n - \beta_0\| = O_p\left(\frac{p_n(1+\lambda_n\tau_n^{1+\gamma})}{n\rho_n}\right)^{1/2}.$$

Suppose, moreover, that we have exact sparsity ($\beta_{20} = 0$*) and that A4(a) holds. Then*

$$\|\hat{\beta}_n - \beta_0\| = O_p\left(\rho_n^{-1}\left(\frac{p_n}{n}\right)^{1/2}\right).$$

The following lemma provides a bound on the magnitude of estimates of β_{20}, and will be used to establish conditions under which they go to zero sufficiently fast so as not to affect estimates of β_{10}. Consistent with the notation adopted for the true coefficient

vector, I partition the estimator as $\hat{\beta}_n = (\hat{\beta}'_{1n}, \hat{\beta}'_{2n})'$, where the first component has length k_n.

Lemma 2. *Under exact sparsity ($\beta_{20} = 0$) and assumptions A2(a), A2(b), A3(a), A4(a) and A4(b) we have*

$$\|\hat{\beta}_{2n}\|_{1+\gamma}^{1+\gamma} = O_p\left(\frac{p_n \kappa_{1n}}{\rho_n^2 \lambda_n}\right).$$

Equivalently

$$\|\sqrt{n}\hat{\beta}_{2n}\|_{1+\gamma}^{1+\gamma} = O_p\left(\frac{p_n \kappa_{1n}}{\rho_n^2 (\lambda_n / n^{(1+\gamma)/2})}\right).$$

Remark 4. I use the word 'superefficiency' to denote convergence of l_2 (and hence l_∞) norm of $\sqrt{n}\hat{\beta}_{2n}$ to zero in probability. Notice that with fixed p_n and ρ_n bounded away from zero $\frac{\lambda_n}{n^{(1+\gamma)/2}} \to \infty$ gives us superefficiency.[12] This is a key component to achieving oracle efficiency in the next result.

The following theorem established the asymptotic distribution of the penalized estimator with the smooth penalty.

Theorem 2 (Asymptotic distribution)**.** *Suppose conditions A2(a), A2(b), A3(a), A3(b), A3(c), A4(a), A4(b) and A4(c) are satisfied, and the model satisfies exact sparsity ($\beta_{20} = 0$). Let $\alpha_n = (\alpha'_{1n}, \alpha'_{2n})'$ be a sequence of $p_n \times 1$ vectors, where α_{1n} contains the first k_n components of α_n, and α_{2n} the rest. Let*

$$s_n^2 = \sigma^2 \left\{ \alpha'_{1n} \Sigma_{1n}^{-1} \alpha_{1n} + \alpha'_{2n} \left[\Sigma_{2n} - \tilde{\Sigma}'_n \Sigma_{1n}^{-1} \tilde{\Sigma}_n \right] \alpha_{2n} \right\}.$$

[12] We can bound the l_2 norm by the $l_{1+\gamma}$ norm since the space is of finite dimension $p_n - k_n$ for any n.

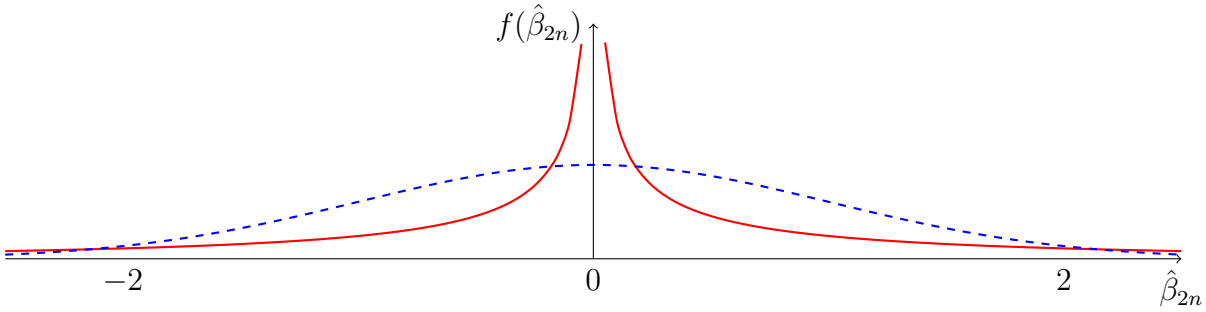

Figure 1.3. Plot of the distribution of zeros for $\gamma = 0.4$. Standard normal density in dashed.

Then

$$\frac{1}{s_n}\alpha_n' \begin{pmatrix} n^{1/2}(\hat{\beta}_{1n} - \beta_{10}) \\ \frac{\lambda_n}{n^{1/2}} \operatorname{sgn}(\hat{\beta}_{2n})|\hat{\beta}_{2n}|^{\gamma} \end{pmatrix} = n^{-1/2} \sum_{i=1}^{n} \varepsilon_i \frac{1}{s_n} \left\{ \alpha_{1n}' \Sigma_{1n}^{-1} x_{1,i} + \alpha_{2n}' \left(x_{2,i} - \tilde{\Sigma}_n' \Sigma_{1n}^{-1} x_{1,i} \right) \right\}$$

$$+ R_n$$

$$\xrightarrow[n\to\infty]{d} \mathcal{N}(0,1),$$

where

$$R_n = O_p\left(\left[\frac{p_n \kappa_{1n}}{\rho_n^2 (\lambda_n / n^{(1+\gamma)/2})} \right]^{\frac{1}{1+\gamma}} \kappa_{2n}^{1/2} \right) = o_p(1)$$

uniformly in α_n.

Theorem 2 establishes that the 'nonzeros' have the same asymptotic distribution as the oracle OLS estimator, and as such no efficiency is lost, asymptotically, from having unnecessary covariates in the penalized regression.

Moreover, we get a nonstandard distribution for estimates of 'zeros', depicted in Figure 1.3. The interesting feature of it is that for smaller values of γ most of the mass will be concentrated closer to zero, but at the same time tails will be fatter. So, for small γ,

most errors in estimates of 'zeros' will be small, but a few will be very large, relative to those coming from larger γ.

Curiously, the estimates of zeros and nonzeros are asymptotically uncorrelated, as follows from the variance formula in Theorem 2. The following corollary illustrates the result in the case where the total number of covariates is fixed:

Corollary 1 (Asymptotically uncorrelated estimates of 'zeros' and 'nonzeros')**.** *Suppose assumptions of Theorem 2 hold; moreover, suppose that the total number of covariates is fixed, i.e.* $p_n = p$*, and* $\Sigma_n \xrightarrow[n\to\infty]{} \Sigma > 0$*. Then*

$$\begin{pmatrix} n^{1/2}(\hat{\beta}_{1n} - \beta_{10}) \\ \frac{\lambda_n}{n^{1/2}} \operatorname{sgn}(\hat{\beta}_{2n})|\hat{\beta}_{2n}|^{\gamma} \end{pmatrix} \xrightarrow[n\to\infty]{d} \mathcal{N}\left(0, \sigma^2 \begin{pmatrix} \Sigma_1^{-1} & 0 \\ 0 & \Sigma_2 - \tilde{\Sigma}'\Sigma_1^{-1}\tilde{\Sigma} \end{pmatrix}\right),$$

where Σ_1*,* Σ_2 *and* $\tilde{\Sigma}$ *are the corresponding limits of the parts of* Σ_n *defined in Equation 1.4.*

The proof of Corollary 1 is a direct application of the Cramér-Wold device to the result of Theorem 2.

1.4.2.1. Regressor classification and inference. Theorem 2 by itself is not directly applicable to the task of inference since the researcher does not know a priori which coefficients are zeros and which are nonzeros. This means that even if the researcher is only interested in the inference on the *ex ante* important coefficients, she does not know a priori how to construct the matrix Σ_{1n} in order to do inference according to Theorem 2. The following result rectifies that.

Lemma 3. *Consider coefficients β_j with $j \in k_0 + 1, \ldots, p_n$. Let $S_0 = \{j : \beta_{0j} \neq 0\}$ and $\hat{S}_n = \left\{j : \left|\hat{\beta}_{nj}\right| \geq \kappa\tau_n\right\}$ for some $\kappa > 0$. Suppose assumptions of Theorem 2 hold. Then $\hat{S}_n = S_0$ with probability approaching 1 as $n \to \infty$.*

Any fixed κ works, and in particular choosing $\kappa \in [1, a]$ is natural: all covariates with estimated penalized coefficients in the inner zone of the penalty function are treated as components of X_{2n}, all covariates with estimated penalized coefficients in the flat outer zone of the penalty function are treated as components of X_{1n}, and κ governs the assignment of covariates whose coefficients fall in between. As lower κ will lead to weakly more covariates being included for the purposes of inference, confidence intervals obtained with $\kappa = 1$ will include as subsets those obtained with larger values of κ under the assumptions of Theorem 2.

A straightforward argument shows that Lemma 3 extends results of Theorem 2 so as to allow conducting inference using $\hat{S}_n$ as if it were S_0, in particular treating 'selected' covariates (together with coefficients of interest that were not penalized) as those corresponding to nonzero coefficients β_{10} for the purposes of Theorem 2. So, to construct confidence intervals for coefficients of interest we would simply compute standard errors as if we ran a regression with only controls being those with coefficients larger than $\kappa\tau_n$, and then construct confidence intervals as usual, only centered at the values estimated by the penalized estimator rather than OLS.[13]

A curious reader might wonder whether it is a good idea to run a second regression with only selected covariates following the thresholding procedure outlined above, and

[13]Note that even though coefficients of interest do not get thresholded, we still need thresholding on controls to establish which ones are deemed nonzeros, so as to construct Σ_{1n}.

report estimates and conduct inference based on this second regression. The short answer is no. As will be shown in subsection 1.4.3, a key benefit of utilizing the smooth penalty proposed in this work is the bias reduction it delivers in cases where the 'gap' assumption of no intermediate-magnitude coefficients is violated. Post-selection reestimation would be similar to utilizing a model-selection-based estimator like SCAD, and would suffer from the same omitted variable bias problem that is ameliorated by using the proposed estimator. Simulation evidence supports this recommendation: results in Section 1.5 illustrate the robustness of the proposed estimator in terms of mean-squared error, while results in Chapter 2 , in particular those in Figure 2.5, illustrate the corresponding inference benefits.

1.4.2.2. Discussion of assumptions. Assumptions A2, A3 and A4 are used for conventional theoretical results described above. Since the results seek to allow for a wide array of potential issues that can be encountered in high-dimensional estimation, the assumptions might appear complicated. To clarify the restrictions they place I will address each assumption in turn, and provide a simplified sufficient set of assumptions that illustrates these restrictions.

Assumption A2(a) is maintained to simplify proofs and allow the use of Laws of Large Numbers (LLNs) and Lindeberg-Feller Central Limit Theorem (CLT) that leverage independence. Assumption A2(b) is similarly chosen for simplicity and in line with conventional assumptions in the literature. It is conceivable that independence can be supplanted with a sufficiently-weakly-dependent requirement on errors and heteroskedasticity can be allowed provided some CLT can be found that applies to the relevant series in the proof of Theorem 2. This will likely involve extra restrictions on covariates, which will result in design assumptions becoming less tractable.

Assumption A3(a) is natural: without it we have perfect multicollinearity and a problem with identification. Evidently, this assumption precludes the $p_n > n$ case.

Assumption A3(c) is needed in the proof of the Lindeberg-Feller CLT for contrasts that include zeros in Theorem 2, and A3(b) is used for the same purpose when only *ex ante* important coefficients and large coefficients (i.e. the first k_n) are of interest. In particular, A3(c) can be dropped if we only need the asymptotic distribution of nonzeros, or just the coefficients of interest. These assumptions can be satisfied with a lower bound on the eigenvalues of Σ_n and a more detailed description of where the regressors come from. The approach taken here is agnostic about the origin of the regressors, and so I present the assumptions in this form to preserve generality.

Assumption A4 contains a lot of moving interdependent components, and thus might appear hard to interpret. For this reason I will provide simple sufficient conditions that ensure that it is satisfied. Moreover, the sufficient conditions will be designed so as to satisfy assumption A1(a) as well. As shown in subsection 1.4.1, assumption A1(a) ensures convexity of the objective function and continuity of solutions in the data. It is not needed for other results present in this work, in particular, for Theorem 2. However, I view these properties as highly desirable, continuity ensuring a certain stability of the estimator and convexity assuring convergence to the global minimum and fast optimization in practice. For these reasons I will include A1(b) (a simple sufficient condition for A1(a) which is equivalent to it in the case of $k_0 = 0$) in the discussion of constraints imposed by the assumptions. It is worth pointing out that A1(a) (and A1(b)) can be easily verified in an application: it only requires the knowledge of the regressor sample covariance matrix and of which regressors the researcher has deemed *ex ante* important (and thus excluded

from penalization). In particular, A1(b) can serve as a useful guide for the joint choice of λ_n and τ_n if the researcher deems convexity and continuity desirable in her application.

The sufficient conditions for satisfying A1(a) and A4 are given in the following lemma[14]:

Lemma 4. *Suppose the following conditions hold:*

(1) $\rho_n > 0$ *is bounded away from zero and* $\text{Eig}_{\max}(\Sigma_n)$ *is bounded above for all* n *large enough;*

(2) k_n *is bounded above for all* n *large enough;*

(3) *the smallest of coefficients* $\{\beta_{0,k_0+1}, \ldots, \beta_{0,k_n}\}$ *is bounded away from zero for all* n *large enough;*

(4) $p_n = o(n^{\frac{1-\gamma}{2}})$.

Then there exist sequences of real numbers λ_n, τ_n *such that assumptions A1(b) and A4 are satisfied. Specifically, these sequences are given by*

$$\begin{aligned} \lambda_n &= n^{\frac{1+\gamma}{2}} p_n f_n, \\ \tau_n &= n^{-1/2} p_n^{\frac{1}{1-\gamma}} \left(\frac{2}{\rho_n (a-1)} f_n \right)^{\frac{1}{1-\gamma}}, \end{aligned}$$

where f_n *is any sequence of real numbers such that* $f_n \to \infty$ *and* $f_n = o\left(n^{\frac{1-\gamma}{2}} p_n^{-1}\right)$.

Lemma 4 can be verified by directly verifying each part of assumptions A4 and A1(b), or derived from them by straightforward analysis. It is worth emphasizing that sequences f_n that we use to construct λ_n and τ_n always exist: these are simply sequences that grow slowly enough such that the last restriction is satisfied.

[14] Without the convexity requirement (A1) the following lemma would have $p_n = o\left(n^{\frac{2}{3+\gamma}}\right)$ as the bound on the number of covariates.

Most notable implication here is the rate of growth of p_n. For γ arbitrarily close to 0, i.e. method being close to SCAD, the rate can be arbitrarily close to $n^{1/2}$. Higher values of γ restrict p_n to growing slower, such that for γ close to 1, i.e. the method being close to ridge regression with clipped outer area of the penalty, p_n must either not grow or grow very slowly. In other words, there is a tradeoff between how fast we can allow the number of regressors to increase and how smooth we can let our penalty be. Faster growth of p_n necessitates harsher penalization (both less smooth function and higher λ_n).

Note that even with fixed p_n, Lemma 4 does not allow for $\gamma = 1$, as that would require that the sequence f_n goes both to infinity and to zero. In other words, the proposed estimator does not allow for the use of ridge penalty in the inner area. This highlights the importance of the core feature of the penalty: infinite second derivative at zero. With a finite second derivative, first-order variation in sampling error would lead to a similarly first-order variation in the estimates, precluding us from achieving superefficiency in Lemma 2 that is the key ingredient of Theorem 2.

In totality, the most restrictive assumption at the moment is that of a gap between large and small coefficients[15], or, in the notation of the simplified example in the previous part, the lower bound on large coefficients (b_n) given by A4(a) coupled with a restriction that all the rest are zero. While universally employed in the literature, it arguably hides the problems that penalized methods experience in practice by excluding the possibility that small-but-not-small-enough coefficients might be present (penalizing which introduces bias in estimates of other coefficients), and in doing so doesn't allow for theoretical

[15]In a way, this is what defines them as large and small. The important question is what happens when some coefficients are neither.

comparison between the proposed method and existing procedures under such challenging circumstances.

Of course real world coefficients do not change with sample size, and as such the assumption appears innocuous. The role of this assumption in the proofs is to keep estimates of 'nonzeros' far away from the penalized region, and to exclude any coefficients that might be erroneously penalized (asymptotically). Therefore with moderate sample sizes this assumption might be 'violated': we might have sampling error that is comparable to the magnitudes of some coefficients, and as such no reasonable choice of τ_n will allow us to avoid a nontrivial chance of some nonzero coefficients being in the non-flat area of the penalty, and hence treated as zeros.

The solution to this is to explicitly consider asymptotics that allow for coefficients to stay in the penalized region while not being small enough to be disregarded. This will be addressed in subsection 1.4.3.

1.4.2.3. Confidence interval length. Given that Theorem 2 promises oracle efficiency and the same asymptotic distribution as the oracle estimator, it is reasonable to expect narrower valid confidence intervals than those provided by the full-model OLS. This can be explored theoretically and numerically; I will cover the theoretical part here.

Given a consistent estimator of error variance (I will consider the case where the same estimator of error variance is used, e.g. the full-model OLS homoskedastic error variance estimator), and assuming homoskedasticity for the construction of confidence intervals, the length of the $1-\alpha$ confidence interval for coefficient j in the full-model OLS estimator is

$$2z_{1-\alpha/2}\sqrt{\left(\Sigma_n^{-1}\right)_{j,j}}\hat{\sigma}_n n^{-1/2}.$$

Constructing confidence intervals according to Theorem 2 after penalized estimation gives us confidence interval length

$$2z_{1-\alpha/2}\sqrt{\left(\Sigma_{1n}^{-1}\right)_{j,j}}\hat{\sigma}_n n^{-1/2}.$$

So, in this case "efficient" confidence intervals for nonzeros are narrower by a factor of

$$\sqrt{\left(\Sigma_n^{-1}\right)_{j,j} / \left(\Sigma_{1n}^{-1}\right)_{j,j}}.$$

How low can this be? If Σ_n is diagonal, then clearly this ratio is 1, and both confidence intervals have the same length. This is unrealistic, however. Even if regressors come from a distribution with a diagonal variance-covariance matrix, sample covariance matrix is unlikely to be diagonal. For example, I simulated 100 observations from a 50-component multivariate normal distribution with identity covariance, and normalized each component to have sample variance 1. I chose 8 components to be "nonzeros". Over 1000 MC replications, the average value of the CI length ratio for the first element on the diagonal (and similarly for others) is 1.37: that is, we would get on average an 1.37 shorter confidence intervals using the efficient penalized method (or oracle OLS) when we use the same error variance estimates.

The difference becomes even more pronounced when we allow for correlation in the true distribution of regressors, although then it matters more which 8 components are chosen as "nonzeros".

Note that this analysis applies equally to SCAD and other oracle efficient methods under appropriate conditions, and so all of them would obtain the same length of a confidence interval asymptotically.

1.4.3. Asymptotic properties: higher-order analysis under adverse conditions

We have shown that under assumptions common in the literature and with an appropriate choice of tuning parameters the proposed estimation procedure can achieve oracle efficiency, just as other methods in the literature. Furthermore, in Lemma 4 we found that smoother penalty functions (higher γ) can only be used with slower-growing number of covariates relative to those with lower γ. To this end, it is worth asking whether anything is gained by having smooth penalization, relative to nonsmooth methods like SCAD and others. We will address this question here.

So far we have assumed that there are two kinds of coefficients: small and large ones. Small ones in this work are defined as exactly zero (this is the assumption of exact sparsity: $\beta_{20} = 0$). Large ones, characterized by assumption A4(a), are kept further-than-sampling-error away from zero. This is a common approach in the literature. The assumption on small coefficients can be relaxed to allow for them to be small but nonzero, as is done in Horowitz and Huang (2013), where in the most favorable case the l_1 norm of β_{20} must be $o(n^{-1/2})$. Even with this relaxation there is still a *gap* between small and large coefficients which allows penalized estimation methods to distinguish between what is to be kept in the model and what is to be excluded.

It is well known, due to results by Leeb and Pötscher (2008) and related papers, that model-selection consistent estimators behave badly when we allow for some coefficients to be in this gap, i.e. when they are neither sufficiently large nor sufficiently small. It seems reasonable (and is shown formally in the proof of Theorem 3) that similar problems will plague the estimator proposed here. However, the smoothly penalized estimator delivers

a strict improvement in the worst squared error of the estimates of *ex ante* important coefficients through a reduction in omitted variable bias, as is shown in Theorem 3.

For this section, I will adjust the maintained assumptions, to strengthen them in some ways and significantly weaken in others.

Assumption A5 (Fixed dimension: regressors and errors)**.** *The total number of covariates is equal to p and is fixed (does not vary with n). Moreover:*

(a)

$$\Sigma_n \xrightarrow[n\to\infty]{} \Sigma > 0;$$

(b)

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i x_i = O_p(1);$$

$$E\left[\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i x_i\right)\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i x_i\right)'\right] = O(1).$$

Assumption A6 (Fixed dimension: coefficients)**.** *β_0 is given by $(\beta_{10}', h_n c')'$ with the length of the first component k and the second $p-k$ for fixed k, where*

(a) components $1, \ldots, k_0$ (for fixed k_0) of β_{10} are coefficients on covariates of ex ante *interest (which are not penalized), and*

$$n^{-1/2} = o\left(\min_{k_0+1\le j\le k} |\beta_{10j}|\right);$$

(b) $c \in \mathbb{R}^{p-k}$, $\|c\| \le 1$, $(\sqrt{n}\tau_n)^{1-\gamma} = o\left(\sqrt{n}h_n\right)$ and $h_n = o(\tau_n)$.

Assumption A7 (Fixed dimension: parameter choices)**.** *$a > 1$ and*

$$\tau_n = n^{-1/2} g_n,$$

$$\lambda_n = n^{(1+\gamma)/2} g_n^{1-\gamma} m,$$

where $m > 0$, $g_n \xrightarrow[n\to\infty]{} \infty$, $n^{-1/2} g_n = o(1)$ *and* $n^{-1/2} g_n = o\left(\min_{k_0+1 \le j \le k} |\beta_{10j}|\right)$.

The main strengthening of assumptions comes from maintaining that $p_n = p$ is fixed (does not vary with sample size). While it might be possible to replicate the results with a growing p_n similar to that of Theorem 2, it would require strengthening other assumptions. The setting considered here will be sufficient to discern important differences between methods that achieve exact model selection, mainly SCAD, and the proposed procedure under different choices of γ.

The major relaxation of assumptions comes from abandoning error independence and homoskedasticity. In fact, just about any sort of error process that would have led the researcher to believe that the full-model OLS estimator is asymptotically normal at a root-n rate would satisfy assumption A5(b).[16] In particular, this can accomodate heteroskedasticity, stationary time series settings, and more general error structures such as spatial dependence.

Assumption A7 corresponds to the most general form of tuning parameter choice for SCAD, and in particular is either identical to or less restrictive than the assumptions on the tuning parameter choice in Fan and Li (2001) and Horowitz and Huang (2013). The case of $\gamma > 0$ is tuned to correspond closest to chosen values for SCAD: the thresholding τ_n is in the same position and the rate of the magnitude of the largest value of the penalty

[16]The first component of A5(b) can be satisfied with (but does not require) $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i x_i \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \Sigma_N)$ for some covariance matrix Σ_N. The second component of A5(b) is a technical condition that is satisfied by common uniform integrability conditions.

is the same.[17] It is worth noting that A7 satisfies tuning parameter requirements of assumptions of Theorem 2, and in particular matches those in Lemma 4 for fixed number of covariates (for a specific value of m). It is also worth noting that the choice of m for SCAD is inconsequential for Theorem 3, and the choice of m for $\gamma > 0$ only affects the magnitude of the C_γ term, not its sign nor the rate on the difference between squared errors of SCAD and estimator with smooth penalty.

Assumption A6(a) (together with the restriction on τ_n in A7) is exactly assumption A4(a) in a fixed-dimension model with regressors satisfying A5(a), and simply characterized what it means for the coefficients to be 'large'. Assumption A6(b) is the main driving force behind the result: it specifies that while those intermediate coefficients are $o(\tau_n)$, they are still relatively large. So they will be penalized, leading to bias in their estimates and hence in the estimates of the 'important' coefficients. It is through reducing this version of omitted variable bias that the proposed estimator obtains the improvement in Theorem 3. The following lemma illustrates the root of the problem:

Lemma 5 (SCAD model selection)**.** *Let $\gamma = 0$. Suppose assumptions A5, A6(a) and A7 hold. Moreover, suppose that $\|\beta_{20}\|_\infty = o(\tau_n)$. Then $\hat{\beta}_{2n} = 0$ with probability approaching one as $n \to \infty$.*

What Lemma 5 says is that all coefficients that are $o(\tau_n)$ will be estimated as exactly zero by SCAD. In particular, this includes coefficients of the same magnitude as the sampling error ($n^{-1/2}$), as well as coefficients even larger than that. This will be the cause

[17] More importantly, it is exactly the relative rates of λ_n and τ_n given by A7 that give the estimator non-trivial asymptotic behavior when coefficients exactly proportional to τ_n are allowed. Detailed treatment of this case goes beyond the scope of the analysis here and is addressed in Chapter 2.

of omitted variable bias in estimates of β_{10}, and it is this bias that will be reduced by using the proposed smooth penalty. This is formally shown in the following theorem.

Theorem 3 (Asymptotic improvement in Realized Squared Error)**.** *Suppose assumptions A5 and A6 hold. Consider estimating $\alpha_1'\beta_{10} + \alpha_2'\beta_{20}$ for some chosen $\alpha = (\alpha_1', \alpha_2')'$ with two estimators: one with $\gamma = 0$ (SCAD) and one with $\gamma \in (0, 1]$ (estimator with smooth penalty), both tuned according to assumption A7 with the same sequence τ_n. Let $\bar{\alpha}_b = \left(\tilde{\Sigma}'\Sigma_1^{-1}\alpha_1 - \alpha_2\right)$ and $\bar{V} = \Sigma_2 - \tilde{\Sigma}'\Sigma_1^{-1}\tilde{\Sigma}$. Suppose that $\alpha_b \neq 0$.[18] Let I denote an identity matrix of the same dimension as $\bar{V}$. Let*

$$RSE_{SCAD} = \left[\alpha'\left(\hat{\beta}_{n,SCAD} - \beta_0\right)\right]^2$$

and

$$RSE_{\gamma} = \left[\alpha'\left(\hat{\beta}_{n} - \beta_0\right)\right]^2.$$

In case of $\gamma \in (0, 1)$ we have

$$\left(\frac{h_n}{\tau_n}\right)^{-\frac{1-\gamma}{\gamma}} h_n^{-2}\left[\sup_{c\in\mathbb{R}^{p-k},\|c\|\leq 1} RSE_{SCAD} - \sup_{c\in\mathbb{R}^{p-k},\|c\|\leq 1} RSE_{\gamma}\right] \geq C_{\gamma} + o_p(1),$$

where

$$C_{\gamma} = 2\bar{\alpha}_b'\left\{\frac{1}{m}\bar{V}\bar{\alpha}_b\right\}^{\frac{1}{\gamma}} \|\bar{\alpha}_b\|^{-\frac{1-\gamma}{\gamma}},$$

with the power understood as the sign-preserving element-wise (Hadamard) power. Moreover, $C_{\gamma} > 0$ for all γ high enough.

[18] This assumption excludes the case where asymptotically there would be no omitted variable bias in the desired linear combination of coefficients regardless of the value of β_{20} if we were to drop the 'controls' from the regression completely. Such a fortuitous occurence is a knife-edge case that is unlikely to be of practical significance.

In case of $\gamma = 1$ *we have*

$$h_n^{-2}\left[\sup_{c\in\mathbb{R}^{p-k},\|c\|\leq 1} RSE_{SCAD} - \sup_{c\in\mathbb{R}^{p-k},\|c\|\leq 1} RSE_{\gamma}\right] \;=\; \bar{C}_1 + o_p(1),$$

where

$$\bar{C}_1 = \bar{\alpha}_b'\left(\frac{1}{m}\bar{V} + I\right)^{-1}\left[\frac{1}{m^2}\bar{V}^2 + 2\frac{1}{m}\bar{V}\right]\left(\frac{1}{m}\bar{V} + I\right)^{-1}\bar{\alpha}_b > 0.$$

This result can be interpreted in two ways. First, where both α_1 and α_2 are potentially nonzero, Theorem 3 illustrates that using a smooth penalty yields lower worst prediction error. Second, the result can be interpreted as improving precision of estimation of particular contrasts, e.g. the ones that are focused on coefficients of *ex ante* interest. In this case, what Theorem 3 tells us is that when some of the coefficients are smaller than τ_n but nonetheless may not be 'small enough', their penalization affects estimates of the important coefficients in a way that is less severe for smoother penalties.

It is notable that this result is in terms of the *Realized Squared Error*, as opposed to the Mean Squared Error; that is, it considers the actual squared error in the contrast of interest as a random variable rather than taking its expectation. This is due to two facts. First, under the conditions of Theorem 3, specifically A6(b), the leading term in the expansion of the squared error is nonrandom and the same across different γ and for model selection methods. Second, and more notably, reduction in bias due to smooth penalty is larger than the terms that are due to random sampling error, making it possible to write the result in terms of RSE.[19]

[19]Note that the fact that the result is in terms of the realized squared error also implies that Theorem 3 extends to show that for any symmetric loss of the form $f\left(\left|\alpha'\left(\hat{\beta}_n - \beta_0\right)\right|\right)$ with f strictly increasing smoother penalty will produce lower worst loss with probability approaching one.

What is the magnitude of improvement delivered in Theorem 3? The answer to this question is driven by the multiplier $\left(\frac{h_n}{\tau_n}\right)^{-\frac{1-\gamma}{\gamma}} h_n^{-2}$ in the statement of the theorem. Note that in the case of $\gamma = 1$ this multiplier is exactly of the magnitude of the (squared) omitted variable bias, since h_n is the upper bound on how large the 'intermediate' coefficients are allowed to be. In particular, in case of $\gamma = 1$, the improvement in squared error is actually of the first order: as is demonstrated in the proof of Theorem 3, the appropriate rate on RSE for both SCAD and the smoothly-penalized estimator is h_n^2, exactly because of omitted variable bias. In the case of $\gamma < 1$ the result in Theorem 3 is a higher-order asymptotic refinement, and the rate of the magnitude of the improvement becomes less than the full omitted variable bias by the factor of $\left(\frac{h_n}{\tau_n}\right)^{\frac{1-\gamma}{\gamma}}$. Evidently the closer γ is to 1, the slower this factor converges to zero, and the closer the estimator is to achieving an improvement of the magnitude of the omitted variable bias.

Since assumptions A6 and A7 are intertwined through the sequence τ_n, it is worth illustrating the statement of Theorem 3 with an example. Suppose the researcher is satisfied that assumption A5 holds, and she is interested in the coefficient on the first regressor, which is not penalized (since it is of *ex ante* interest). The researcher maintains that 'large' coefficients on other covariates are bounded away from zero. She decided to use SCAD, and has chosen a sequence of tuning parameters that would ensure oracle efficiency according to the conditions of, and under the assumptions of, Fan and Li (2001) (or Horowitz and Huang (2013)). The theorem says that when some coefficients are smaller than the chosen thresholding parameter, but still relatively large, with probability approaching one the researcher would obtain a strictly lower squared error by using the proposed smooth penalty with the same tuning parameters and with a high enough value

of γ. While it is true that for any given h_n (with a fixed $\|c\|$) the researcher could pick a different sequence τ_n (specifically, smaller than h_n) to avoid omitted variable bias (with SCAD and with the smooth penalty), the researcher is unlikely to know the exact rate h_n; more importantly, no matter the choice of τ_n, there will always be sequences $h_n = o(\tau_n)$ that satisfy assumptions of Theorem 3.

This last point can also be illustrated by a game in which the researcher first decides on the tuning parameter values (and specifically on the sequence τ_n, representing the boundary at which the penalty will treat the coefficients as 'large') from all possible values that would deliver oracle efficiency in SCAD under standard sparsity assumptions; moreover, the researcher also has a choice of whether to use SCAD or a correspondingly-tuned smooth penalty. Then Nature moves and chooses the coefficients on controls with the goal of making the resulting estimate as far away from the truth as possible (in the sense of the squared error), with the only restriction that the coefficients Nature chooses be no larger in magnitude than a certain sequence $h_n = o(\tau_n)$. Theorem 3 shows that, provided h_n is not too restrictive, the researcher can strictly improve her payoff in this game with probability approaching 1 by choosing a sufficiently smooth penalty instead of SCAD. Moreover, as demonstrated above, this choice of smooth penalty and tuning parameters will still deliver oracle efficiency under the standard sparsity assumptions.

Two important qualifications need to be addressed. One is that the theorem is stated for "γ high enough". Due to nonlinear nature of the expression for C_γ for $\gamma \neq 1$, it is hard to characterize its sign, and so the argument in Theorem 3 works by continuity in the neighbourhood of $\gamma = 1$. While this might not sound useful for practitioners, in an application we can evaluate the sign of the sample analog of C_γ, since it only depends

on γ, Σ_n and α_1, the chosen contrast of interest. This sample analog converges to C_γ. However, due to nonstochastic nature of the regressors and no bounds on $\Sigma_n - \Sigma$, we can not test for the sign of C_γ.

The second qualification is that the theorem is written explicitly as a comparison to SCAD, and as such does not cover other oracle efficient model-selection-consistent estimators, such as those with adaptive LASSO, bridge or minimax concave penalties. Unfortunately it is less clear what a 'fair' comparison is in terms of tuning parameters between those penalties and the one proposed here; however, some analysis could be feasible. In particular, Theorem 3 can be straightforwardly adapted to comparison with minimax concave penalties as long as they satisfy two conditions: the derivative of the penalty above a certain threshold is zero, and for given tuning parameter sequences we can find a threshold τ_n such that $n^{-1/2} = o(\tau_n)$ and coefficients smaller than τ_n will be estimated as exactly zero with probability approaching one. Accomodating bridge penalties will require accounting for the bias they introduce due to nonzero derivative of the penalty even for large coefficients, and accomodating adaptive LASSO will involve similar concerns. Nonetheless, SCAD has emerged as a popular option in model-selection-consistent penalization literature (both due to prominence of Fan and Li (2001) and due to much subsequent work on it), and as such is a good benchmark.

Overall, what this section shows is that the proposed estimator can serve as a partial 'insurance policy' against the worst mistakes of the classic oracle efficient methods, achieving efficient estimation when the standard assumptions are satisfied but improving the worst-case performance under very general assumptions when small-but-not-small-enough coefficients are present.

1.5. Simulations

Theoretical results under conventional assumptions (Theorem 2) indicate that the proposed estimator is just as efficient, in the first-order asymptotic sense, as the oracle estimator, and hence as efficient as model selection based estimators that achieve oracle efficiency. The result on higher-order refinement in squared error tells us that when intermediate-magnitude coefficients are present, smoother penalties should have an advantage over model selection based estimators, at least in the worst cases. I will carry out numerical simulations to illustrate both points.

I will simulate the linear model considered throughout this chapter. Regressors will be drawn (once for each model) from a multivariate normal with covariance matrix with diagonal 1 and off-diagonal elements equal to 0.7.[20] This covariance structure ensures that the ordering of regressors is irrelevant, that is, permuting coefficients would not materially change the results. Regresssors will be standardized to have sample mean zero and variance one. I will treat regressor 1 as the "important" regressor and will not penalize the corresponding coefficient. Other regressors are treated as controls.

There will be $n = 250$ observations and $p = 20$ covariates in each simulation run. Errors are i.i.d. $\mathcal{N}(0, \sigma^2)$, where σ^2 is chosen so that the finite-sample MSE of the full-model OLS estimator of coefficient 1 would be 1.

I evaluate six simulation designs meant to illustrate settings of Theorem 2 and Theorem 3. In all of them the first coefficient is set to 1 and its estimate is not penalized. The second coefficient is set to 20, which is large enough to be in the outer area of the penalty with the choice of τ_n outlined above (which is about 1.8 in the simulations). So

[20] It is easy to verify that this produces a valid variance covariance matrix.

covariate 2 is thus a control covariate with a 'large' coefficient (unknown from the point of view of simulations), and even though its coefficient will be penalized, it will be in the flat area of the penalty. The simulation designs differ in the coefficients on the other 18 covariates. In the 'zero' design, those 18 coefficients are exactly zero, representing exact sparsity. The other 5 designs introduce progressively larger coefficients:

(1) Coefficients 2-20 are drawn from $U[0, 0.1]$.

(2) Coefficients 2-20 are drawn from $U[0, 0.2]$.

(3) Coefficients 2-20 are drawn from $U[0, 0.3]$.

(4) Coefficients 2-20 are drawn from $U[0, 0.4]$.

(5) Coefficients 2-20 are drawn from $U[0, 0.5]$.

For comparability of results across designs, all coefficients are drawn once from $U[0, 1]$ and then scaled appropriately for each design.

The 'zero' design captures the key feature of the ideal case: coefficients on controls are either zero or large, making it easy for model selection or penalization methods to distinguish between important and unimportant covariates.

The other designs are challenging in a sense that some of the coefficients on controls are smaller than τ_n but not sufficiently small to be irrelevant, leading to omitted variable bias due to penalization. Scaling coefficients up as we move from design 1 to design 5 should increase this bias. On the other hand, shrinking small coefficients also carries a benefit through a reduction in variance.

For all simulations, I will need to choose tuning parameters. I will conduct all simulations for various values of γ, so as to illustrate its effects, including $\gamma = 0$, i.e. SCAD. I will use $a = 3.7$ (as recommended by Fan and Li (2001) for SCAD). τ_n will be chosen in

each simulation run as the half-width of the widest 95% confidence interval that we would construct by running full-model OLS (and using homoskedastic OLS estimator of error variance). This seeks to mimic the idea that τ_n should be 'larger than sampling error'. Finally, λ_n will be chosen by crossvalidation in each simulation.

While in practice τ_n can, and probably should, be chosen by crossvalidation, I will keep it fixed here and report results for different values of γ to isolate its effects. γ can also be chosen by crossvalidation, as will be done in the empirical application, but here the goal of simulations is to highlight differences across γ and with SCAD, and so a grid of values of γ is reported.

I will focus on simulated mean squared error of estimates of coefficient 1. Note that I do not address coverage in the simulation results reported here; Chapter 2 focuses more specifically on inference and so I leave it to the simulations section in that chapter to place inference results in better context.

Results for SCAD, full-model OLS and smooth penalty with $\gamma \in \{0.1, 0.5, 0.9\}$ are plotted in Figure 1.4. Numerical results (including those for other values of γ) are reported in Table A.1 in the appendix, along with Figure A.1 showing simulated bias of estimates of coefficient 1.

In 'zero' design, all choices of γ yield lower MSE than OLS, but the lowest choices appear best, with the minimum at SCAD (i.e. $\gamma = 0$). This is to be expected, since we prefer the coefficients that are truly zero to be estimated as close to zero as possible, and low values of γ do that (the residual term in Theorem 2 is lower). SCAD estimates these coefficients as exactly zero with high probability, and so is practically the same as the oracle estimator.

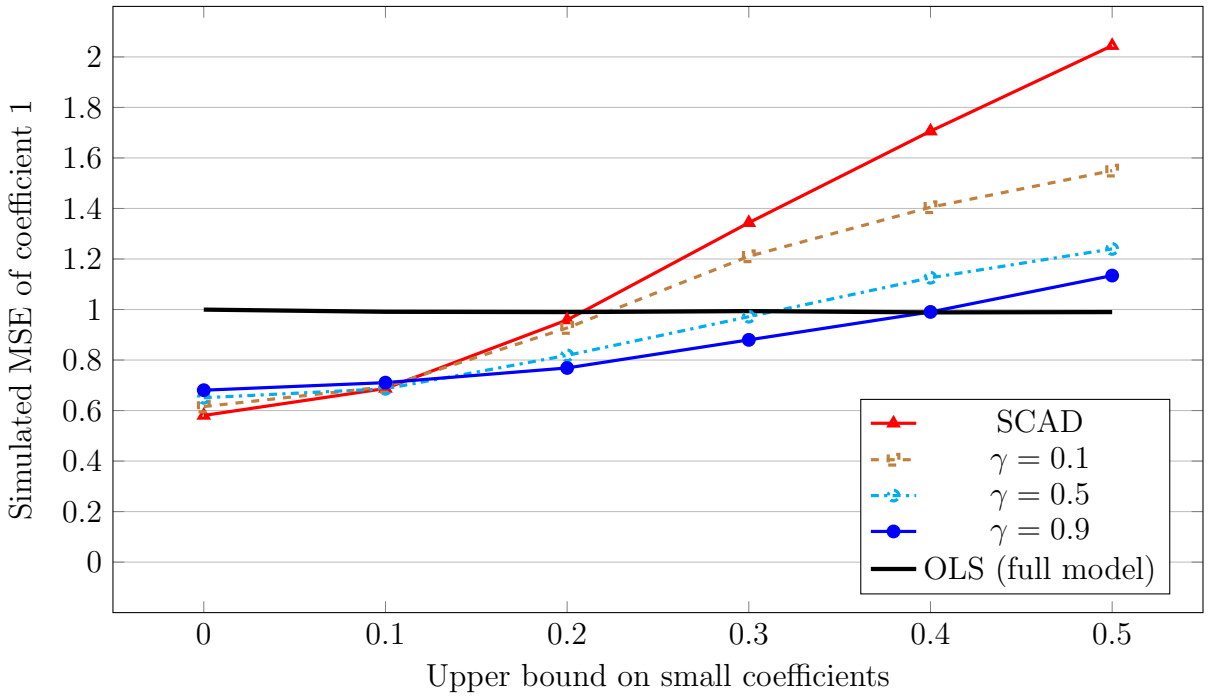

Figure 1.4. Simulated MSE of estimates of coefficient 1. All results are from 10 000 simulations (per model and estimator).

As we increase the magnitude of small coefficients, performance of SCAD estimator starts to worsen dramatically, to the point that in the last design the increase in MSE is equivalent to cutting sample size in half. Even moving to $\gamma = 0.1$ notably improves estimator performance, and $\gamma = 0.9$ proves the most robust. Even though it does somewhat worse than SCAD or lower values of γ in the 'zero' case, it still improves on OLS in that case. On the other hand, for upper bound 0.1, $\gamma = 0.9$ is nearly the same as other methods (and still improves on OLS), and clearly outperforms them for larger magnitudes of small coefficients. Note also that SCAD and $\gamma = 0.1$ break even with OLS around upper bound 0.2, and changing to $\gamma = 0.9$ moves the break-even point to 0.4. That is, high smoothness doubles the range of coefficient magnitudes where penalization is beneficial.

However, even $\gamma = 0.9$ does worse than OLS in the last case with the highest magnitude of small coefficients. This reflects the fact that omitted variable bias dominates variance reduction in this case.

Results support the intuition from Theorem 3 that smoother penalty performs better in the case when not-sufficiently-small coefficients are present, and also the intuition from Theorem 2 about behavior under exact sparsity.

In addition to the results presented above, average execution time with SCAD is almost twice as long as with smooth penalty with $\gamma = 0.1$, and higher γ increase execution speed even further. This suggest that smooth penalization might offer appealing computational advantages over nonsmooth SCAD. However, since faster algorithms can undoubtedly be created for both SCAD and smooth penalization estimators, the results on computational speed are tentative and as such I do not report the exact numbers.

Overall, simulation results suggest that smoother penalties might present better overall precision of estimation of coefficient of interest, improving upon full model OLS under ideal conditions and performing better than SCAD under challenging circumstances.

1.6. Conclusion

I have proposed a penalized estimation method that achieves oracle efficiency despite utilizing a differentiable penalty function. The proposed method is computationally easy and is particularly suitable to circumstances where model selection is not in itself required, such as when estimating a linear model with a large number of controls but small number of covariates of interest. For an applied researcher faced with a high-dimensional dataset,

the method provides a rigorous approach to carrying out efficient estimation of coefficients of interest without the need for *ad hoc* trimming of controls.

Importantly, the estimator improves on existing model selection based estimators when the 'gap' assumptions that preclude presence of coefficients comparable to the sampling error are violated. The proposed estimator achieves asymptotically lower worst realized squared error in the estimates of coefficients of interest, driven by a higher-order reduction in bias. This serves as a partial 'insurance policy' against undesirable properties when such coefficients are present.

The proposed method works well in simulations, reducing mean squared error of coefficients of interest relative to OLS across various settings. In particular, smoother versions of the penalty (higher γ) result in more robust performance when small-but-nonnegligible coefficients are present. The method is computationally easy, with estimation taking fractions of a second on a modern computer, and higher values of γ working faster than lower ones. Fast computational speed results in crossvalidation being unburdensome.

CHAPTER 2

Alternative Asymptotic Analysis of Once-Differentiable Penalty Estimator

2.1. Introduction

Modern economic analysis often deals with datasets with a large number of potentially relevant covariates. It is often not unreasonable to suppose that only a fraction of the covariates actually affect the outcome, but at the same time it is not always clear *ex ante* what those important covariates are. The question then is how to carry out estimation and inference in such settings so that efficiency gains of smaller models might be realized and yet model misspecification is also avoided.

A number of approaches have emerged to carry out estimation in such settings. Penalized estimation with a continuous penalty function has emerged as perhaps the most popular, with LASSO by Tibshirani (1996) and SCAD by Fan and Li (2001) being the most well known. Supposing that there is a true (and fixed) sparse data generating process underlying the data, SCAD can achieve *model selection consistency*, that is, keep all the covariates that are relevant to the outcome and drop the rest. Combined with the asymptotic lack of bias for nonzero coefficients, model selection delivers what is termed an *oracle property*: the estimator of nonzeros has the same asymptotic distribution as if the true, *oracle* model were known and estimated from the start. This immediately leads to straightforward inference that takes only estimated-as-nonzero coefficients into account.

Good properties of such estimators crucially depend on the absense of coefficients roughly comparable in magnitude to the sampling error. In particular, the oracle property is not uniform in coefficient size, and estimator performance can be greatly compromised by omitted variable bias when such intermediate-magnitude coefficients are present.[1] Chapter 1 proposed a penalized estimator with a once differentiable penalty function, showing that such an estimator can achieve oracle efficiency in the standard framework that prohibits coefficients of intermediate magnitude. At the same time, departing from model selection provides a reduction in estimator bias when such small coefficients are present, leading to a reduction in quadratic loss that is of a larger magnitude than even that of the sampling error. As shown in Chapter 1, inference based on the oracle distribution remains valid in the standard model, just as with model-selection-based estimators.

However, it is well understood in the literature (see e.g. Leeb and Pötscher (2008), Belloni, Chernozhukov, and Hansen (2014) and references therein) that the oracle distribution may be a less-than-ideal approximation to the finite sample behavior of penalized estimators. In finite samples we may experience both mistakes in classifying covariates as small or large, and local bias and variance distortions in estimator distributions coming from the penalty term. It is therefore desirable to consider alternative asymptotic approximations that might be better suited to capturing finite sample estimator behavior.

This paper explores such alternative asymptotic approaches in the context of the penalized least squares estimator proposed in Chapter 1. Three alternatives are considered,

[1]The range of what I call "intermediate-magnitude coefficients" is also known as the "gap" in the literature, after a common feature of assumptions required for estimators to achieve oracle efficiency: that of either having zero (or very small) or relatively large coefficient values, but nothing in between. A more detailed discussion of this assumption and its violations can be found in Chapter 1.

differing in which of the finite sample issues they capture and consequently differing in their usefulness to practitioners.

The first approach is the local asymptotic approximation that allows us to capture the full richness of finite sample behavior. Both classification mistakes and nontrivial covariance between large and small coefficients are captured, and consequently the approximation is the most realistic one. However, it is not useful for practical inference, as the asymptotic distribution depends on impossible-to-estimate drift parameters. Such an approximation is still interesting as a way of thinking about finite-sample performance of the estimator, and an example is evaluated that illustrates that the intuition from results obtained in Chapter 1 carries over to this more complex approximation. In particular, smoother penalty delivers slightly worse performance than when the classic sparse model is a correct representation of the DGP, but offers a much larger improvement under a wide range of violations of sparsity.

The second approach is what I call a "semi-local" asymptotic approximation. This approximation consideres tuning parameter choices that deliver correct classification and oracle efficiency in the standard sparse model. However, instead of assuming away intermediate coefficients, I explicitly consider coefficients proportional to the thresholding parameter. This allows us to capture bias and variance effects of having intermediate coefficients. The resulting asymptotic distribution is normal and incorporates bias and variance adjustments typical of delta method approaches. It is possible to estimate both the bias and the variance. However, I show that estimating the bias introduces the variance cost such that the confidence intervals obtained using such an approach are exactly

the same width as those from using the full-model OLS estimation and inference procedure. This result runs counter to the suggestion in Fan and Peng (2004), whereby bias is estimated but extra variance is not accounted for in the suggested inference procedure.

The third, and final, approach is to allow for nontrivial covariance between estimates of small and large coefficients but not for the classification errors and bias due to intermediate-magnitude coefficients. This covariance arises from the fact that while in the standard asymptotic framework of Theorem 2 estimates of 'zeros' converge to zero at a rate that is sufficiently fast to ignore them in the asymptotic distribution of 'nonzeros', in finite samples variation in these coefficients will introduce variation in estimates of large coefficients.[2] I show that the resulting nonstandard asymptotic distribution can be approximated by a modified bootstrap procedure similar to that used in Chatterjee and Lahiri (2011) for LASSO estimators. Moreover, the same modified bootstrap procedure is valid under the standard assumptions on coefficients and tuning parameters, and the researcher does not need to know which is the 'correct' regime to obtain valid inference via bootstrap. As such, inference by this modified bootstrap can be seen as a more robust approach that guards against underpenalization.

Simulation evidence supports theoretical findings and illustrates that the bootstrap behaves well even when the penalty level is not so high so as to render estimates of zero coefficients negligible, improving on the coverage of the confidence intervals constructed using the oracle distribution and still delivering narrower confidence intervals than those from the full-model OLS. Furthermore, building on the results of Chapter 1, simulation evidence suggests that even when exact sparsity is violated, a combination of a smoother

[2]In a model-selection-based estimator, a similar feature will be observed where the probability of selecting precisely the right model is less than one in a finite sample.

penalty and bootstrap inference continues to provide close-to-nominal coverage, capitalizing on reduced estimator bias that is a key feature of the smoothly-penalized estimator.

The remainder of the paper is structured as follows. Section 2.2 introduces the model and the notation, as well as the penalized least squares estimator with the smooth penalty proposed in Chapter 1. Section 2.3 discusses possible approaches to asymptotics as they relate to the choices of tuning parameters. Section 2.4 presents the local asymptotic approximation and illustrates the analogy to the results in Chapter 1. Section 2.5 considers the semi-local approximation and inference with estimated bias. Section 2.6 considers the asymptotic approximation that allows for nontrivial covariance between estimates of small and large coefficients but not for classification errors nor intermediate-magnitude coefficients, and shows that the modified bootstrap can be used to approximate the resulting asymptotic distribution, as well as the asymptotic distribution under standard assumptions on coefficients and tuning parameters. Section 2.7 presents simulation evidence on the comparative merits of bootstrap and oracle distribution. Section 2.8 concludes.

2.2. The model and the estimator

This section describes the model and the notation used in this paper, and reintroduces the penalized estimator proposed in Chapter 1.

The researcher observes a sample of n observations from the following model:

$$y_i = x_i'\beta_0 + \varepsilon_i,$$

where x_i is a $p \times 1$ vector of covariates, which includes $k_0 \geq 0$ *covariates of* ex ante *interest* and $p - k_0$ *controls*. Unlike Chapter 1, I treat p, the total number of covariates,

as fixed in this paper, as it simplifies the arguments while still allowing to focus on the interesting features of the estimator. The regressors are nonstochastic in line with the standard approach in the literature, and the error term ε_i has mean zero.

Some of the coefficients on controls are potentially zero: that is, the model can be written as

$$y_i = x'_{1,i}\beta_{10} + x'_{2,i}\beta_{20} + \varepsilon_i. \tag{2.1}$$

where β_{10} is of length $k \geq k_0$ and includes coefficients on covariates of *ex ante* interest and nonzero coefficients on controls. Correspondingly $\beta_{20} = 0$, and so consists of coefficients on those controls that have no mean effect on y. Taking the above into account we can write

$$y_i = x'_{1,i}\beta_{10} + \varepsilon_i. \tag{2.2}$$

The researcher's goal is to estimate β_0 and conduct inference on it; specifically, the researcher might focus on the coefficients on covariates of *ex ante* interest. One avenue is thus to estimate (and conduct inference based on) the *full model* in Equation 2.1 by OLS. Had the correct partitioning been known, the researcher would have wanted to estimate the *oracle model* in Equation 2.2, which could lead to lower estimator variance and narrower confidence intervals. Since the correct partitioning is unknown, the *oracle estimator* is infeasible.

I will consider here the estimator proposed in Chapter 1 and defined as follows:

$$\hat{\beta}_n = \arg\min_b Q_n(b), \tag{2.3}$$

$$Q_n(b) = \sum_{i=1}^{n}(y_i - x_i'b)^2 + \lambda_n \sum_{j=k_0+1}^{p} \text{Pen}(b_j), \tag{2.4}$$

where the penalty function is everywhere differentiable with the derivative

$$\text{Pen}'(b) = 2\tau_n^{\gamma}\,\text{sgn}(b) \begin{cases} \left(\frac{|b|}{\tau_n}\right)^{\gamma}, & \text{when } \frac{|b|}{\tau_n} < 1; \\ \frac{a-\frac{|b|}{\tau_n}}{a-1}, & \text{when } \frac{|b|}{\tau_n} \in [1,a]; \\ 0, & \text{when } \frac{|b|}{\tau_n} > a. \end{cases} \tag{2.5}$$

Figure 2.1 reproduces Figure 1.2 from Chapter 1 and illustrates the derivative of the penalty function.

The fact that the penalty is everywhere differentiable means that none of the estimated coefficients are exactly zero. Chapter 1 shows that this reduces worst-case estimator bias and through it estimation and prediction squared error when intermediate-magnitude coefficients are present.

It is worth highlighting that asymptotic approximations in this paper revolve around two tuning parameters in the expressions above: the *thresholding parameter* τ_n and the *scale parameter* λ_n.

Observe that the thresholding parameter determines what magnitude coefficients fall into what area of the penalty. In particular, the penalty has three areas: inner area (coefficients smaller than τ_n in absolute value), outer area (coefficients larger than $a\tau_n$ in absolute value), and an intermediate area that provides a smooth transformation from

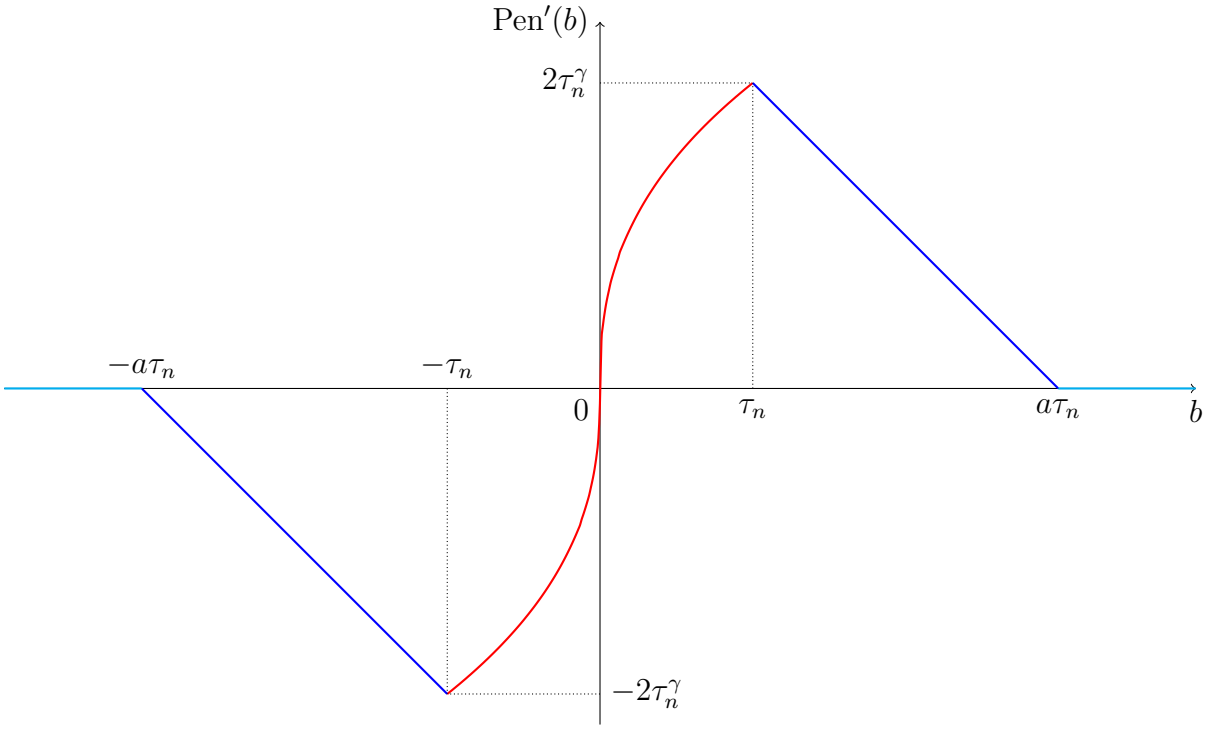

Figure 2.1. Plot of the derivative of the penalty function for $\gamma = 0.4$, $\tau_n = 1$, $a = 3$.

the inner area to the flat outer area. In this way the role of τ_n is to distinguish between small and large coefficients.

The scale parameter λ_n primarily serves to drive coefficients that are in the inner area of the penalty towards zero. How fast they converge to zero depends on how large λ_n is.

2.3. Possible approaches to asymptotic approximations

The main standard theoretical result in Chapter 1 – Theorem 2 – postulates what is known as the oracle property, popularized by Fan and Li (2001). This is a feature common to modern penalized estimators such as SCAD, bridge estimators and adaptive LASSO, and essentially establishes that under the right conditions the asymptotic distribution of

coefficients of interest (or otherwise "large" coefficients) is the same as if we knew and estimated the smallest parsimonious model by OLS from the start.

It is well understood in the literature (see e.g. Leeb and Pötscher (2008), Belloni, Chernozhukov, and Hansen (2014) and references therein) that the oracle distribution may be a less-than-ideal approximation to the finite sample behavior of penalized estimators. It is therefore worth looking into how the oracle property is achieved and what can be done differently to obtain a more realistic approximation to finite-sample behavior.

The conclusion of Theorem 2 is achieved with the following conditions on the two tuning parameters:

(1) τ_n should be large enough to exceed sampling error (Assumption A4(b)) but smaller than large coefficients (Assumption A4(a)).

(2) λ_n should be high enough so that true zero coefficients are driven to zero at a sufficiently fast rate (Assumption A4(c)).

The two conditions achieve subtly different effects. The first one ensures that, asymptotically, we can classify coefficients into "zeros" and "nonzeros" (or "small" and "large" ones). In particular, it is intertwined with the assumption that there are no coefficients of an intermediate magnitude, that is, ones which are large enough to be important yet not large enough to be easily distinguishable from zero. As such, this part ensures that there are no classification errors in Theorem 2, and so none of the coefficients that need to be kept in the model are biased, and all of the coefficients that are small are judged as such.

The second assumption ensures that the estimates of coefficients that are truly zero are so small as to not affect the nonzeros. It essentially allows us to treat the component

of variance in estimates of large coefficients that is due to variance in the estimates of small coefficients as negligible.

Among the two conditions on the tuning parameters, it is arguably easier to ensure that τ_n is larger than the sampling error of the estimator, for exampe by using the rule of thumb recommended in Chapter 1 (see e.g. the approach in Section 1.5), whereby we can set τ_n as the half-width of the widest confidence interval constructed using full-model OLS, where the nominal coverage is chosen suitably large.

It is less obvious what the appropriate finite-sample way of choosing λ_n should be. Simulations in Chapter 1 support the idea that crossvalidation works well both when the standard no-intermediate-coefficients assumption is replicated and even when some intermediate-magnitude coefficients are present. However, Wang, Li, and Tsai (2007) show in the context of SCAD that crossvalidation may lead to underpenalization, and so it is worth considering how we can carry out inference that remains valid under such conditions. Moreover, even if a data-driven way to choose λ_n can be devised that does indeed deliver λ_n sufficiently high (as BIC does in Wang, Li, and Tsai (2007)), it might be argued that a lower level of λ_n provides a better safeguard against penalization bias, and can be desirable in settings where the researcher is unsure of the validity of the standard sparsity assumptions.

Given the context of choices of tuning parameters, there are four possible combinations of views on whether we want to capture classification errors (and ensuing bias) and covariance between estimates of large and small coefficients.

Capturing neither effect is the standard asymptotic approach of Theorem 2 discussed above, achieved with sufficiently high τ_n, sufficiently high λ_n, and assuming away intermediate coefficients.

Capturing both effects can be achieved by using the local asymptotic approximation in Section 2.4, where neither τ_n nor λ_n are "large enough", and coefficients proportional to τ_n are allowed. As will be seen, while this might be seen as the best approximation theoretically, it is not useful for inference as the asymptotic distribution is dependent on impossible-to-estimate nuisance parameters. We must therefore turn to more restrictive approximations.

Capturing classification mistakes and penalization bias, but not covariance between coefficients estimated as "zero" and other ones, can be achieved under standard assumptions on tuning parameters but allowing for coefficients proportional to τ_n. This is considered in Section 2.5. While the resulting asymptotic distribution differs nontrivially from both the full-model OLS distribution and the oracle distribution derived in Theorem 2, estimating bias in order to carry out inference results in normal distribution with the variance that is exactly the same as that of full-model OLS, and so while realistic, this approach does not improve on simply estimating the full model by OLS as far as the width of confidence intervals is concerned.

Finally, we can capture covariance effects but not classification mistakes and penalization bias in a framework where there are no intermediate coefficients, τ_n satisfies standard assumptions but λ_n is not "large enough". This is the approach in Section 2.6. In contrast to the two alternative approaches above, the asymptotic distribution does not contain unknown parameters, and can be approximated by a modified bootstrap procedure, yielding

an inference tool that both improves the quality of the asymptotic approximation relative to the standard framework, and still delivers narrower confidence intervals than those offered by full-model OLS. Notably, this bootstrap procedure is also valid under the standard assumptions on τ_n and λ_n, and so can be recommended as a robust approach to inference for practical applications.

2.4. Local asymptotics

As described above, it is well understood that oracle efficiency in the standard sparse setting comes at a cost of poor performance when some coefficients fall into the no man's land of being neither small enough to warrant exclusion from the model nor large enough to be clearly distinguishable from zero. One clear way to evaluate performance in this case is to consider local asymptotics with model parameters approaching zero at just the right rate:

$$\beta_0 = \alpha_0 n^{-1/2},$$

where α_0 is a fixed p-vector.

For the asymptotics to be sensitive to specific choices of penalty function and tuning parameters, we need to capture both the possibility of mistakes in classifying coefficients based on location of the estimates, and the fact that for any given penalty multiplier there is a chance that the estimated coefficient will be large even if the true value is zero. To capture both effects we need to model τ_n and λ_n as slightly smaller than in the standard

asymptotic approach, specifically

$$\begin{aligned} \tau_n &= tn^{-1/2}, \\ \lambda_n &= ln^{\frac{1+\gamma}{2}} \end{aligned}$$

for some fixed t and l.

Note that now τ_n shinks at the the same rate as sampling error, and λ_n is just large enough to make the penalty term converge to a stable limit in local asymptotics.

Given the above choice of τ_n the magnitude of components of the Pittman drift α_0 relative to t and *at* determines whether a given coefficient is primarily "large" or "small", essentially by what area of the penalty function it falls into.

Notice also that the two tuning parameters are just smaller (by factors growing at an arbitrarily slow rate) than what is called for by Theorem 2 in Chapter 1, which would allow us to capture features of the estimator that are explicitly suppressed in Theorem 2 in Chapter 1. As such, the estimator will no longer exhibit oracle efficiency and the asymptotic distribution will be complicated, but the goal is for it to be a more realistic approximation to finite sample behavior.

I establish the following result for the case outlined above:

Theorem 4 (Local asymptotics)**.** *Suppose the total number of regressors and the number of* ex-ante *important regressors are fixed at p and k_0, respectively, regressor covariance matrix Σ_n converges to a positive definite limit Σ, true regression coefficients are given by $\beta_0 = \alpha_0 n^{-1/2}$ and the two tuning parameters are $\tau_n = tn^{-1/2}$ and $\lambda_n = ln^{\frac{1+\gamma}{2}}$. Assume that the errors are i.i.d. with mean 0 and variance σ^2. Let* $\text{Pen}(b;\tau)$ *denote the penalty*

function with the thresholding parameter set to τ*. Then*

$$\sqrt{n}\left(\hat{\beta}_n - \beta_0\right) \xrightarrow[n\to\infty]{d} \arg\min_u V_L(u),$$

where

$$V_L(u) = u'\Sigma u - 2u'W + l\sum_{j=k_0+1}^{p} \text{Pen}(\alpha_0 + u; t)$$

and $W \sim \mathcal{N}(0, \sigma^2\Sigma)$*.*

The proof of this result follows immediately from the arguments in Knight and Fu (2000).

As is evident from the result of the theorem above, the asymptotic distribution is nonstandard and, more importantly, depends on impossible-to-estimate Pittman drift parameters α_0. As such, this approximation can not be used for inference directly. However, it is still of interest in evaluating estimator performance in a theoretical setting.

While there does not appear to be a closed-form solution for the asymptotic distribution in the general case, it can be easily simulated for specific circumstances. In particular, we might want to explore how the mean squared error of the estimator varies with the Pittman drift. To this end, consider a two-covariate setup with the first coefficient large ($\alpha_{01} >> at$) and the other one small-to-moderate ($|\alpha_{02}| \in [0, at]$). We will compare the asymptotic mean squared error of the estimator of β_1 for different values of γ, i.e. for different shapes of the penalty function.

Since the penalty requires two tuning parameters (plus a as the third) for a given value of γ, we need a way to choose tuning parameters that would make results comparable across different γ. The only difference in the shape of the penalty lies in the inner area

of the penalty, and so we will set t and a the same across different γ, and then choose l (or λ_n) so that the shape of the penalty above t (or τ_n) is exactly the same for all γ. This amounts to choosing l so that the largest value of the derivative of the penalty is the same across γ. Suppose for $\gamma = 0$ we choose $l = l_0$. Then for a different γ the equivalent value of l is

$$l_\gamma = l_0 t^{-\gamma}.$$

We will consider asymptotic covariance matrix for regressors with correlation coefficient $\rho = 0.5$. The results depend on it in the obvious way: higher ρ will let estimator of coefficient 2 affect that of coefficient 1 more, in particular increasing the bias when α_{02} is nonzero.

Other parameters will be set as follows: $\sigma^2 = 1$, $a = 3.7$ per the recommendation in Fan and Li (2001), and $\alpha_{01} = 2at$, different for different values of t but always far above at.[3]

Note that in this model two-covariate OLS estimator of β_1 has asymptotic MSE of $\frac{1}{1-\rho^2} = 1.33$ and oracle asymptotic MSE of 1.

We will consider two choices of t and l_0:

(1) $t = 1$, $l_0 = 2$. This corresponds to the case when we would both make many classification mistakes (due to low value of t relative to σ), and we are penalizing a lot, so that, in case of second coefficient being really zero, we would estimate it as very close to zero most of the time and hence would get almost perfectly to oracle efficiency.

[3]We can consider a different setup where the first coefficient is not penalized, and then it would not matter where it is located. As it is, the main point in choosing a drift value for β_1 is to keep its estimates out of the penalized region.

(2) $t = 3$, $l_0 = 1$. This corresponds to the case where we make very few mistakes in classifying, and we are also penalizing less. So we would expect to be further from oracle efficiency in case $\beta_2 = 0$ but perhaps achieve better performance for all methods in the intermediate range where we incur most bias.

Results are summarized in plots in Figure 2.2 and Figure 2.3. In each figure, the first plot shows asymptotic MSE of $\hat{\beta}_1$ and the second shows the ratio MSE(γ)/MSE($\gamma = 0.1$) as a way of comparison.

I find that for all values of γ there is a large range of intermediate values of drift where the estimator has higher MSE than full-model OLS. This mirrors the intuition in Leeb and Pötscher (2008), in that the estimator (much like other penalized estimators) suffers from increased bias in the intermediate range of coefficient values as a cost of improving efficiency at zero.

Moreover, we find that higher values of γ lead us to two results: we lose some efficiency when the second coefficient is very close to zero, and gain some in a wide range where that coefficient is further away from zero but still in the penalized region. That is, it would be more reasonable to use low γ (or even SCAD) if we were very confident that the model is sparse with a large number of zeros and no intermediate parameters; however, if we suspect that some intermediate coefficients might be present, then larger values of γ provide some insurance againt very large bias and MSE due to those intermediate coefficients. This is in line with Theorem 3 in Chapter 1, illustrating that the reduction in bias due to smoother penalty dominates the whole of the variance term, implying that smoother penalties deliver lower squared error with intermediate-magnitude coefficients.

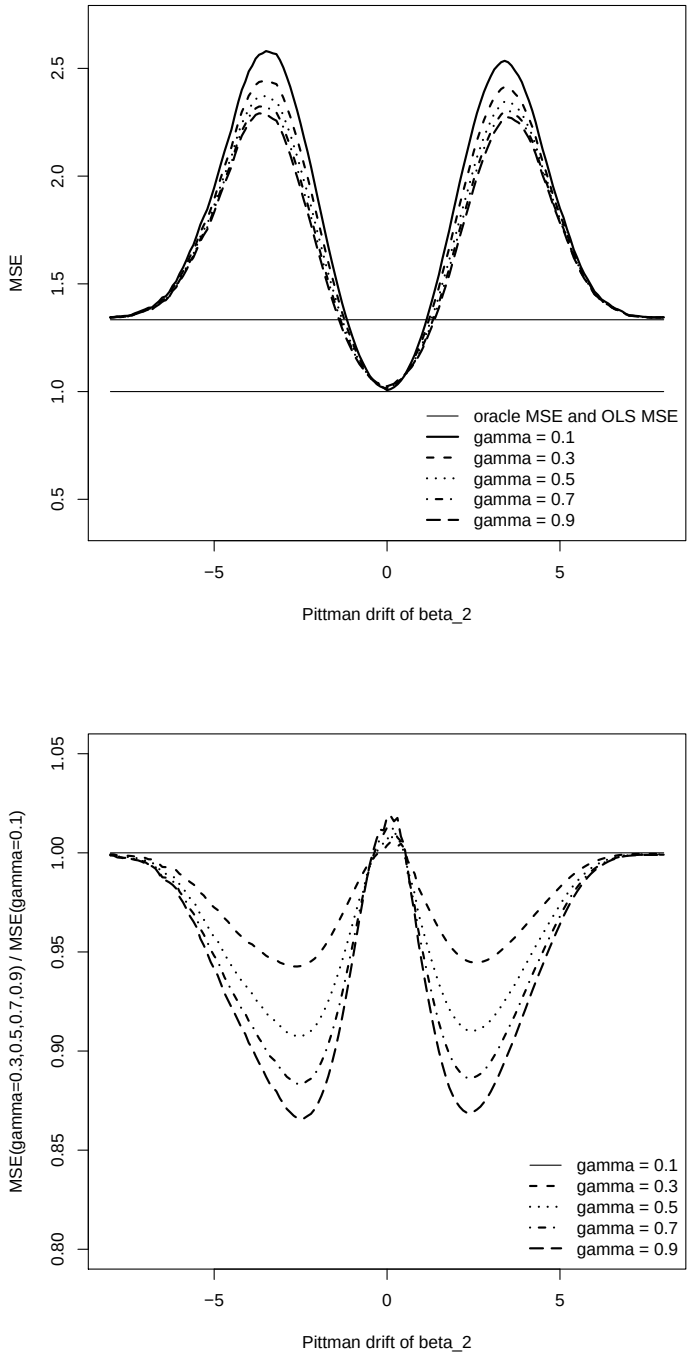

Figure 2.2. MSE and relative efficiency as a function of γ and β_2 – Case 1

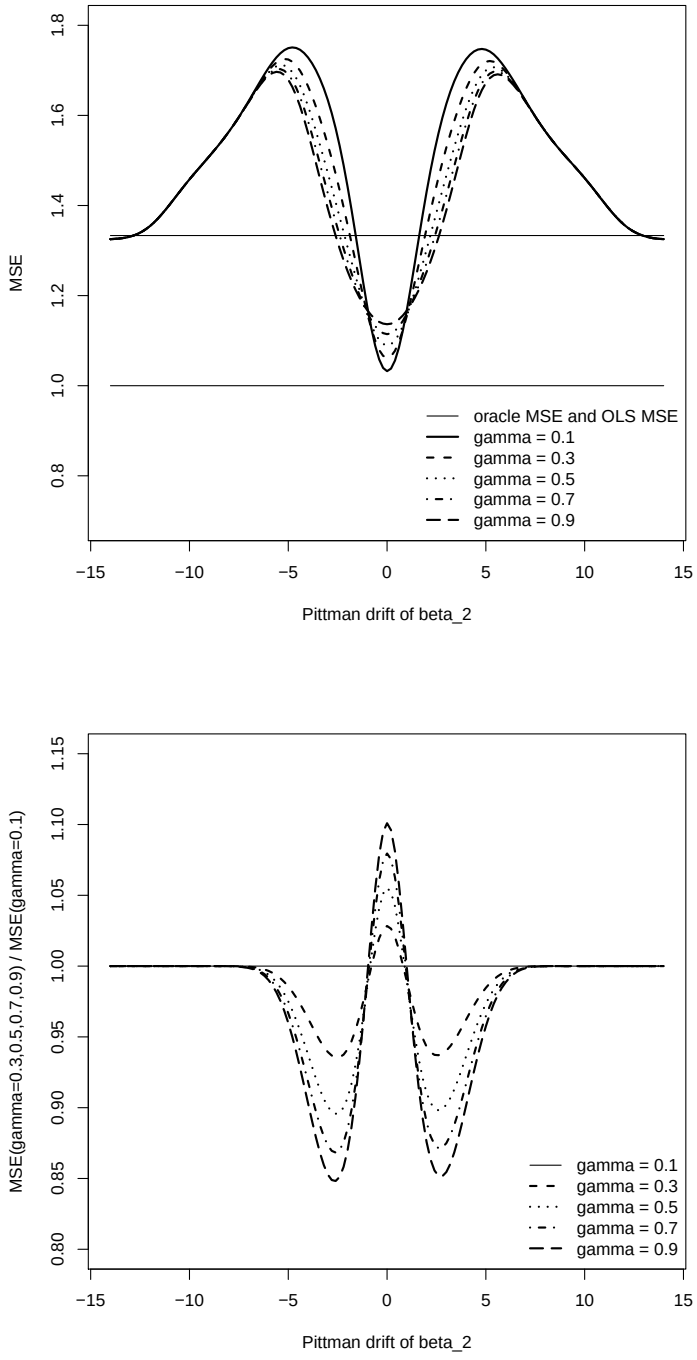

Figure 2.3. MSE and relative efficiency as a function of γ and β_2 – Case 2

As such, choosing γ amounts to resolving a tradeoff between excellent behavior when traditional assumptions on coefficients are a good approximation to the DGP, achieved by lower values of γ, and less-terribly-bad behavior when those assumptions do not capture key features of the DGP, and the bias problems are present, in which case higher values of γ provide more robust performance.

2.5. Semi-local asymptotics

As illustrated in the previous section, local asymptotics of the type considered there are unsuitable for inference due to dependence of asymptotic distribution on impossible-to-estimate drift parameters. We will consider an alternative approach to local asymptotics in which the coefficients are proportional to τ_n, but unlike in the previous section, $n^{1/2}\tau_n \to \infty$, consistent with the assumption on τ_n in the standard asymptotics. λ_n will also be chosen in a way that satisfies the standard assumptions, that is, high enough to suppress the effects of coefficients estimated as zero on the other ones.

To simplify the results, it helps to redefine the intermediate range of the penalty. The original definition of the penalty function specifies a transitional region from the inner part of the penalty to the flat outer part as a parabola, in the same way as in Fan and Li (2001). This means that the penalty function is not twice differentiable at the boundaries of the intermediate region. While this is irrelevant in asymptotic experiments that exclude coefficients comparable to the thresholding parameter, it becomes less convenient when such coefficients are allowed, potentially leading to non-normality in those coefficients that happen to center at the boundary of the transitional region. For this reason it is more convenient to redefine the penalty so that it is twice differentiable everywhere except for

zero, making the description of asymptotic behavior easier. Therefore consider a class of penalty functions defined by their derivative satisfying the following description:

Assumption A8 (Penalty function)**.** *Penalty function* $\text{Pen}(b;\tau_n)$ *(as a function of* b *with thresholding parameter* τ_n*) is continuously differentiable, with a derivative that satisfies the following conditions:* $\text{Pen}'(b;\tau_n) = 2\,\text{sgn}(b)|b|^\gamma$ *for* $\frac{|b|}{\tau_n} < 1$*;* $\text{Pen}'(b;\tau_n) = 0$ *for* $\frac{|b|}{\tau_n} > a$*;* $\text{Pen}'(b;\tau_n)$ *is continuously differentiable on* $b \neq 0$ *with finite derivative.*

I will rely on (asymptotic) objective function convexity to establish results in this section. To this end, Lemma 1 from Chapter 1 that establishes convexity of the objective function needs to be modified. The new assumption and convexity lemma are as follows:

Assumption A9 (Objective function convexity)**.** *(a) Let* I_{p,k_0} *be a diagonal* $p \times p$ *matrix with the first* k_0 *diagonal elements equal to zero and the remainder equal to one. We have*

$$\Sigma_n + \frac{1}{2}\frac{\lambda_n}{n}\min_{b\neq 0}\text{Pen}''(b;\tau_n)I_{p,k_0} > 0$$

in the sense of the matrix being positive definite.

(b) (sufficient condition for A9(a)) Let ρ_n *be the smallest eigenvalue of* Σ_n*. We have*

$$-\frac{1}{2}\frac{\lambda_n}{n}\min_{b\neq 0}\text{Pen}''(b;\tau_n) < \rho_n.$$

The modified convexity lemma thus reads as

Lemma 6. *Suppose the penalty function satisfies assumption A8, and assumption A9(a) holds. Then the objective function* $Q_n(b)$ *is strictly convex.*

The proof of Lemma 6 is a straightforward modification of the corresponding proof in Chapter 1.

We will make one weak assumption on errors together with regressors:

Assumption A10 (Asymptotic normality in OLS)**.**

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i x_i \xrightarrow[n\to\infty]{d} \mathcal{N}(0,\Sigma_N)$$

for some positive definite matrix Σ_N*.*

Assumption A10 allows for a very generic form of the error term, in particular allowing heteroskedasticity and cross-dependence as in time series settings. Essentially, whenever the researcher would be prepared to assume asymptotic normality at root-n rate with a full-model OLS regression (and have a way to consistently estimate the 'meat' of the asymptotic covariance matrix), she should be willing to entertain Assumption A10.

Regressor covariance matrix is assumed to converge to a positive definite limit:

Assumption A11 (Regressor covariance matrix)**.** *(a)* $\Sigma_n \to \Sigma$ *as* $n \to \infty$ *for some positive definite matrix* Σ*.*

(b) Assumption A11(a) holds; moreover

$$\|\Sigma_n - \Sigma\| = o\left(\frac{1}{\sqrt{n}\tau_n}\right).$$

Note that Assumption A11(b) is not a restriction on Σ_n, but rather an upper bound on how large τ_n can be, since whenever Assumption A11(a) holds, we can always let $\sqrt{n}\tau_n$ grow slowly enough such that A11(b) is satisfied.

Finally, we will require asymptotic convexity, which is achieved with a direct analogue of Assumption A9(*a*):

Assumption A12 (Asymptotic convexity)**.** *For the matrix* Σ *in assumption A11(a) and tuning parameters* λ_n, τ_n *such that* $\frac{\lambda_n}{n} = m\tau_n^{1-\gamma}$ *the following holds:*

$$\Sigma + \frac{1}{2} m \min_{b \neq 0} \text{Pen}''(b; 1) I_{p,k_0} > 0$$

in the sense of the matrix being positive definite.

Assumption A12 ensures that the asymptotic objective function is convex. Without it the estimator might not be continuous, which would preclude us from establishing asymptotic normality of the estimator.

Theorem 5 (Asymptotic distribution under semi-local asymptotics)**.** *Suppose conditions A8, A10, A11(b) and A12 are satisfied,* $a > 0$ *and* $\gamma \in (0, 1)$*. Let* $\tau_n = n^{-1/2} g_n$ *for some* $g_n \to \infty$ *and* $\frac{\lambda_n}{n} = m\tau_n^{1-\gamma}$ *for some fixed* $m > 0$*. Let* $\beta_0 = \alpha_0 \tau_n$*. Let*

$$b = \arg\min_v v'\Sigma v + m \sum_{j=k_0+1}^{p} \text{Pen}(\alpha_{0j} + v_j; 1).$$

Let β_{10} *capture the elements of* β_0 *such that either the corresponding coefficient is not penalized, or* $\alpha_{0j} + b_j \neq 0$*. Let* β_{20} *capture the rest, and let* $\beta_0 = (\beta_{10}', \beta_{20}')'$*. Partition* $\hat{\beta}_n$*,* b*,* α_0 *and* Σ_n *accordingly. Then*

$$\begin{pmatrix} \sqrt{n}\left(\hat{\beta}_{1n} - \beta_{10}\right) - \sqrt{n}\tau_n b_1 \\ \frac{\lambda_n}{\sqrt{n}} \hat{\beta}_{2n}^{\gamma} \end{pmatrix} = \begin{pmatrix} V_1^{-1} & 0 \\ -\tilde{\Sigma}_n' V_1^{-1} & I \end{pmatrix} W + o_p(1),$$

where $V_1 = \Sigma_{1n} + \frac{m}{2} \bigtriangledown^2 \text{Pen}(\alpha_{10} + b_1; 1)$ *and* $W \sim \mathcal{N}(0, \Sigma_N)$.

What Theorem 5 shows is that estimates of coefficients that are centered at zero converge faster than root-n, and that the asymptotic distribution of coefficients of interest (and others in group 1) is only affected by those in group 2 through the bias. A natural question is then whether we can estimate the bias and conduct inference. Observe that due to convexity, the minimization problem defining the bias has a unique solution characterized by the first-order condition:

$$b = -\frac{m}{2}\Sigma^{-1} \bigtriangledown \text{Pen}(\alpha_0 + b; 1),$$

or, utilizing the fact that $\tau_n^{-1}\left(\hat{\beta}_n - \beta_0\right) = b + o_p(1)$, we can rewrite the above as

$$b = -\frac{1}{2}\frac{\lambda_n}{n\tau_n}\Sigma^{-1} \bigtriangledown \text{Pen}\left(\hat{\beta}_n + o_p(1); \tau_n\right).$$

Naturally, we can use

$$\hat{b}_n = -\frac{1}{2}\frac{\lambda_n}{n\tau_n}\Sigma_n^{-1} \bigtriangledown \text{Pen}\left(\hat{\beta}_n; \tau_n\right)$$

as an estimator of the bias. However, replacing the unknown true bias in Theorem 5 with this estimate adds to the variance, which is captured in the following result:

Lemma 7 (Asymptotic distribution with estimated bias)**.** *Suppose assumptions A8, A10 and A11(a) are satisfied. Then*

$$\sqrt{n}\left(\hat{\beta}_n - \beta_0 - \tau_n\hat{b}_n\right) \xrightarrow[n\to\infty]{d} \mathcal{N}\left(0, \Sigma^{-1}\Sigma_N\Sigma^{-1}\right).$$

The result of Lemma 7 says that the confidence intervals constructed with the above estimate of the bias have the same width as those constructed from the full-model OLS. It

is not surprising: we cannot get narrower confidence intervals from penalized estimation without any assumptions on the possible magnitudes of coefficients.

2.6. Bootstrap

While the asymptotic distribution in Theorem 2 in Chapter 1 is normal (or a transformation of it in case of zeros), we have seen that under more realistic local asymptotics the asymptotic distribution is less tractable. What's worse is that it depends on Pittman drifts of small coefficients that can't be estimated consistently. Due to this fact it appears that bootstrap wouldn't work if we allow for arbitrary values of Pittman drift (the proof would go along the lines of Chatterjee and Lahiri (2010)), and it is not clear how to do (non-conservative) inference in this case.

As such, we will not seek to conduct inference in the presence of intermediate-value parameters, and will assume again that all parameters are either zero or 'large'. The question then is whether we must revert to the asymptotics in Theorem 2 in Chapter 1, or whether there is some middle ground that would be more realistic than oracle asymptotic distribution but where we could still construct a valid (and nonconservative) inference procedure.

One abstraction of Theorem 2 in Chapter 1 is that the estimators of zeros do not affect the estimators of nonzeros due to faster-than-root-n convergence. However, it ought to be clear that in a finite sample variability in the estimators of zeros will affect nonzeros (as can be seen under local asymptotics with zero drift). I will reintroduce the ability to distinguish zeros from nonzeros but will abandon superefficiency in order to keep interdependence of zeros and nonzeros in the asymptotic distribution. Theorem 6 shows that the

asymptotic distribution thus derived is nonstandard. I will then show that the modified bootstrap proposed in Chatterjee and Lahiri (2011) provides valid inference both in this case as well as the standard case where the estimates of zeros are negligibly small.

Assumption A13 describes details of the data-generating process. Assumptions A14(a) and A14(b) describe two alternative approaches to parameter choices that embody the different asymptotic approximations.

Assumption A13 (Errors and design)**.** *(a) The number of regressors is fixed and is equal to p.*

(b) Regressor covariance matrix Σ_n is nonsingular for all n and converges to a positive definite limit Σ.

(c) $\frac{1}{n}\sum_{i=1}^{n}\|x_i\|^3 = O(1)$.

(d) True regression coefficients are fixed (among which coefficients 1 to k are nonzero and $k+1$ to p are zero).

(e) The errors are i.i.d. with mean 0 and variance $\sigma^2 \in (0,\infty)$.

Assumption A14 (Parameter choices)**.** *The two tuning parameters are given by $\tau_n = n^{-1/2}g_n$ and $\lambda_n = f_n n^{\frac{1+\gamma}{2}}$.*

(a) (Oracle efficiency) $f_n \xrightarrow[n\to\infty]{} \infty$, $f_n g_n^{\gamma} = O(\log n)$ and $\frac{g_n}{\log n} \xrightarrow[n\to\infty]{} \infty$.

(b) (No oracle efficiency) $f_n = l \in (0,\infty)$ for all n, $g_n = O\left([\log n]^{1/\gamma}\right)$ and $\frac{g_n}{\log n} \xrightarrow[n\to\infty]{} \infty$.

Note that Assumption A13 is a simplified version of assumptions in Theorem 2, with the main difference being that we consider the number of regressors as fixed and the

coefficients as fixed. As such, whether we achieve oracle efficiency in this setting depends on the choice of tuning parameters.

The following two results highlight the different asymptotic approximations implied by the different assumptions A14(a) and A14(b) on parameter choices; the first result is a straightforward corollary of Theorem 2.

Corollary 2 (Asymptotic distribution with oracle efficiency)**.** *Suppose assumptions A13 and A14(a) hold. Then Theorem 2 applies. In particular,*

$$\begin{pmatrix} \sqrt{n}\left(\hat{\beta}_{1n} - \beta_{10}\right) \\ \frac{\lambda_n}{\sqrt{n}} \operatorname{sgn}(\hat{\beta}_{2n})|\hat{\beta}_{2n}|^{\gamma} \end{pmatrix} \xrightarrow[n\to\infty]{d} \mathcal{N}\left(0, \sigma^2 \begin{pmatrix} \Sigma_1^{-1} & 0 \\ 0 & \Sigma_2 - \tilde{\Sigma}'\Sigma_1^{-1}\tilde{\Sigma} \end{pmatrix}\right).$$

Theorem 6 (Asymptotic distribution without oracle efficiency)**.** *Suppose assumptions A13 and A14(b) hold. Then*

$$\sqrt{n}\left(\hat{\beta}_n - \beta_0\right) \xrightarrow[n\to\infty]{d} \arg\min_u V_F(u),$$

where

$$V_F(u) = u'\Sigma u - 2u'W + \frac{2}{1+\gamma} l \sum_{j=k+1}^{p} |u_j|^{1+\gamma}$$

and $W \sim \mathcal{N}(0, \sigma^2\Sigma)$.

The proof of Theorem 6 follows the argument in Theorem 4 and Knight and Fu (2000), with the only added detail that under the chosen tuning parameters estimates of zeros will be in the inner area of the penalty almost surely and estimates of nonzeros will be in the outer area almost surely (see Lemma 13 in the appendix).

Note that under Theorem 6 the variability in estimators of zeros will affect nonzeros, and so the limiting distribution is nonstandard. It can be easily simulated, and due to Lemma 13 we can correctly classify coefficients into zeros and nonzeros with probability 1, asymptotically, and hence we can use the asymptotic distribution to construct asymptotically valid confidence intervals for all coefficients.

However, it is both arguably easier and more robust to use bootstrap here. I now show that the modified bootstrap proposed in Chatterjee and Lahiri (2011) is consistent for the limiting distribution both under the conditions of Theorem 6 and under the conditions of Corollary 2. The modified bootstrap suitable for the estimator considered here is given by the following procedure:

(1) Compute the proposed penalized estimator $\hat{\beta}_n$.

(2) Construct the modified estimator $\tilde{\beta}_n$ as $\tilde{\beta}_{nj} = \hat{\beta}_{nj}\mathbb{I}\left\{\left|\hat{\beta}_{nj}\right| > \tau_n\right\}$ for $k_0+1 \leq j \leq p$ and $\tilde{\beta}_{nj} = \hat{\beta}_{nj}$ for $j \leq k_0$.

(3) Obtain residuals $r_i = y_i - x_i'\tilde{\beta}_n$ and recenter them as $r_i^* = r_i - \frac{1}{n}\sum_{i=1}^n r_i$.

(4) For a given number of bootstrap simulation iterations, obtain new residuals ε_i^* by resampling r_i^* with replacement, construct $y_i^* = x_i'\tilde{\beta}_n + \varepsilon_i^*$ and compute the penalized estimator β_n^* from $(y_i^*, x_i)_{i=1..n}$ using the same values of tuning parameters.

(5) Use the simulated distribution of $\sqrt{n}\left(\beta^* - \tilde{\beta}_n\right)$ as an approximation of the distribution of $\sqrt{n}\left(\hat{\beta}_n - \beta_0\right)$.

Theorem 7 (Modified bootstrap validity)**.** *Suppose assumption A13 holds. Moreover suppose either assumption A14(a) or assumption A14(b) holds. Then*

$$\varrho(\tilde{G}_n, G_n) \xrightarrow[n\to\infty]{} 0, a.s.,$$

where $\varrho(\cdot,\cdot)$ *denotes the Prohorov metric on the set of all probability measures on* $(\mathbb{R}^p, \mathcal{B}(\mathbb{R}^p))$*,* G_n *is the distribution of* $\sqrt{n}\left(\hat{\beta}_n - \beta_0\right)$ *and* $\tilde{G}_n$ *is the conditional distribution of* $\sqrt{n}\left(\beta^* - \tilde{\beta}_n\right)$ *given the errors* ε_i*.*

As a corollary of Theorem 7 we can use modified bootstrap to construct confidence intervals for individual coefficients.

Note the key implication of Theorem 7 for applied work: regardless of whether the researcher chose the penalty multiplier high enough to make estimates of 'zeros' irrelevantly small, inference by bootstrap is valid. As such, bootstrap serves as a robust way to conduct inference in practice.

2.7. Simulations

Since the main applied result in this work is the validity of the modified bootstrap procedure under lower-than-usual choices of scale parameter λ_n, I will seek to compare inference by the standard asymptotic approximation based on the oracle distribution (i.e. the one derived from Theorem 2) with inference by bootstrap across a range of settings and penalty smoothness parameter γ choices. The results will also be compared against inference based on the full-model OLS estimator, and that based on standard inference with SCAD.

It is worth considering what settings the bootstrap would be most valuable in, relative to the standard asymptotically normal approximation. Evidently, these are the settings where the sum total of extra variance from estimates of small coefficients is in some sense significant. That would be either because the scale parameter λ_n is relatively small, meaning that individual estimates of zeros are larger, or because there are a lot of zero (or small) coefficients, even if their individual variances are small.

One prime setting where λ_n can be restricted to be somewhat smaller is when convexity of the objective function is desirable. Recall that a sufficient condition for convexity given in Assumption A1(b) in Chapter 1 postulates that the largest value of λ_n that still ensures convexity given other tuning parameters declines with decline in the smallest eigenvalue of the regressor covariance matrix. So requiring convexity imposes an upper bound on the choice of λ_n, and this upper bound becomes more restrictive whenever regressors are "more correlated" with each other. As such, simulations will be conducted with and without convexity restriction on the choice of λ_n to compare the impact of requiring convexity on inference.

Simulation design will be based on that in Chapter 1. There will be $n = 250$ observations with a variable number of regressors, with regressors drawn from a normal distribution with all covariances equal to 0.7. The first coefficient will be considered coefficient of interest, and will not be penalized. The second coefficient will be set in the outer area of the penalty to be considered "large".

In the first group of simulation designs, the rest of the coefficients are set to zero to replicate exact sparsity. There will be either 20, 50 or 100 regressors, and for each

number of regressors estimation and inference with and without imposing convexity will be considered.

In the second group of designs, I will focus on exploring inference in the presence of small-but-nonzero coefficients. Settings with 20 covariates and with coefficients 3-20 drawn from a uniform distribution on $[0, 0.2]$, $[0, 0.4]$ and $[0, 0.6]$ will be simulated. Note that this design is almost identical to the one in Chapter 1, and the results can thus be considered as an extension of, and in conjunction with, those in Chapter 1.[4]

Confidence intervals will be constructed at the nominal level of 95%, and I will focus on the confidence intervals for the first coefficient.

Finally, in all simulations errors are drawn from a normal distribution with error variance chosen so as to yield asymptotic mean squared error of 1 for the full-model OLS estimator of the first coefficient.

Tuning parameters are chosen as follows: $a = 3.7$ as in Chapter 1, γ will be set to one of $\{0, 0.1, 0.5, 0.9\}$, τ_n will be chosen as the half-width of the largest 95% confidence interval constructed from full-model OLS, and λ_n will be chosen by crossvalidation, with and without the convexity restriction outlined above.

Standard asymptotic inference after penalized estimation will be carried out with thresholding at the level τ_n. That is to say, estimated coefficients with absolute value below τ_n will not be included in the construction of the covariance matrix for regressors with "large" coefficients. SCAD inference will rely on implicit thresholding, that is to say, all covariates with coefficients that are not estimated as exactly zero will be included

[4]The difference in collections of designs is that simulations in Chapter 1 covered a finer grid of values for the upper end of the uniform distribution, but did not include the [0,0.6] design.

in the covariance matrix. In actual simulations, this will mean that fewer regressors are excluded, as it appears that the implicit threshold is below τ_n.

The results are presented in Figures 2.4 and 2.5.

First, looking at the exactly sparse setting, we can see that confidence interval widths produced by the standard asymptotic approach are practically indistinguishable from those constructed with the oracle OLS estimator with only the first two regressors, which is to be expected. Bootstrap confidence intervals, on the other hand, are somewhat wider, and get wider with γ. Again, this is consistent with the fact that smoother penalties are less aggressive in driving small coefficients to zero. Enforcing convexity also widens the confidence intervals. In contrast, SCAD confidence intervals are somewhat wider, especially when we restrict λ_n to ensure convexity of the objective function. This is consistent with fewer of the "zeros" being estimated as exactly zero.

Looking at coverage in the exactly sparse setting, we can see that the coverage of the standard asymptotic approximation is better for lower γ, again consistent with the idea that smoother penalties allow for more variance in the estimates of zeros. However, coverage falls further below the nominal level for larger number of covariates, even for $\gamma = 0$ with thresholding at τ_n. Enforcing convexity makes this effect more pronounced, as the scale parameter λ_n is restricted to be lower. Coverage with SCAD inference is essentially at the nominal level. On the other hand, bootstrap coverage probability remains close to the nominal level regardless of γ, number of covariates and convexity restrictions. Taken together with the larger width of the confidence intervals, it suggests that the bootstrap successfully captures extra variance in the estimates of the coefficient of interest that is

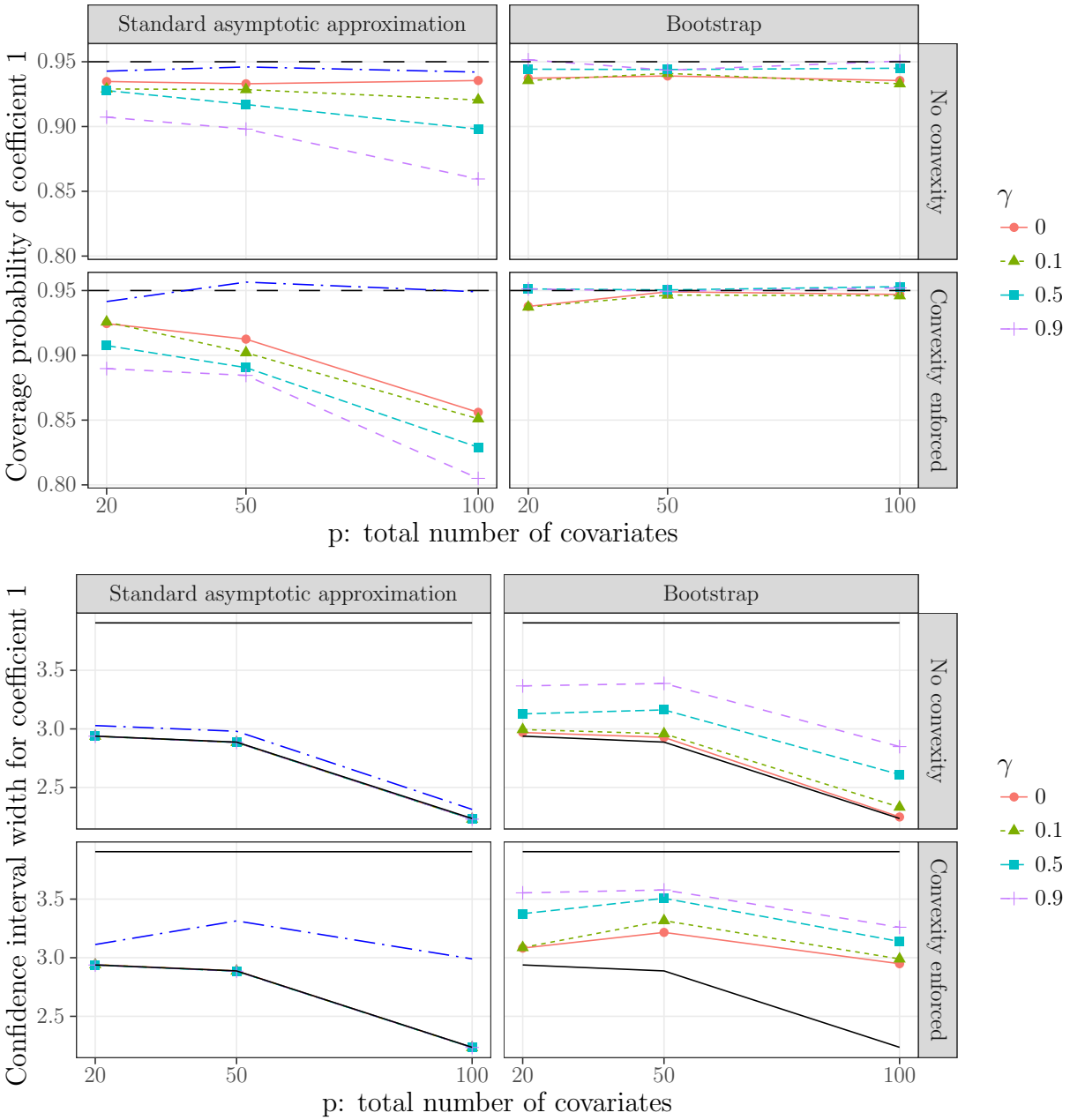

Figure 2.4. Coverage probability and width of the confidence interval of coefficient 1 under exact sparsity. Confidence interval widths by full-model OLS and OLS with only the first two covariates in solid black. Results for SCAD inference in dash-dot blue.

suppressed in the standard asymptotic approach for smooth penalties, supporting the results of Theorem 6 and Theorem 7.

Considering the case with small nonzero coefficients presented in Figure 2.5, we can see that the standard asymptotic approximation now performs relatively badly for all values of γ, although for different reasons. As shown in the simulations in Chapter 1 and additional results in Figure A.1 in the Appendix, and consistent with the theoretical results, smoother penalties reduce bias, so the contribution of bias to undercoverage should be less for higher γ. However, the extra variance effect of estimating small coefficients with smooth penalties is still present, just as in the exactly sparse case. Moreover, inference by SCAD still leads to noticeable undercoverage, even with much wider confidence intervals than that by smooth penalties with thresholding. This highlights the fact that bias, and the reduction in bias due to utilizing smoother penalty, dominates the variance effects, and suggests that the improvement in estimator precision derived in Theorem 3 carries over to an improvement in inference, even in the settings where the sparsity assumption might be a poor approximation of the data-generating process.

Interestingly, bootstrap achieves coverage rates close to nominal for $\gamma = 0.9$, which is consistent with it capturing the variance component. Bootstrap performs less well for low values of γ, due to the fact that it can not account for the larger bias in those settings. At the same time as achieving near-nominal coverage, inference by bootstrap with $\gamma = 0.9$ still noticeably reduces the width of the confidence interval relative to estimation by full-model OLS. Equally importantly, the "pseudo-oracle" OLS that only includes the first two covariates leads to coverage far below the nominal, illustrating the danger of estimating smaller potentially misspecified models.

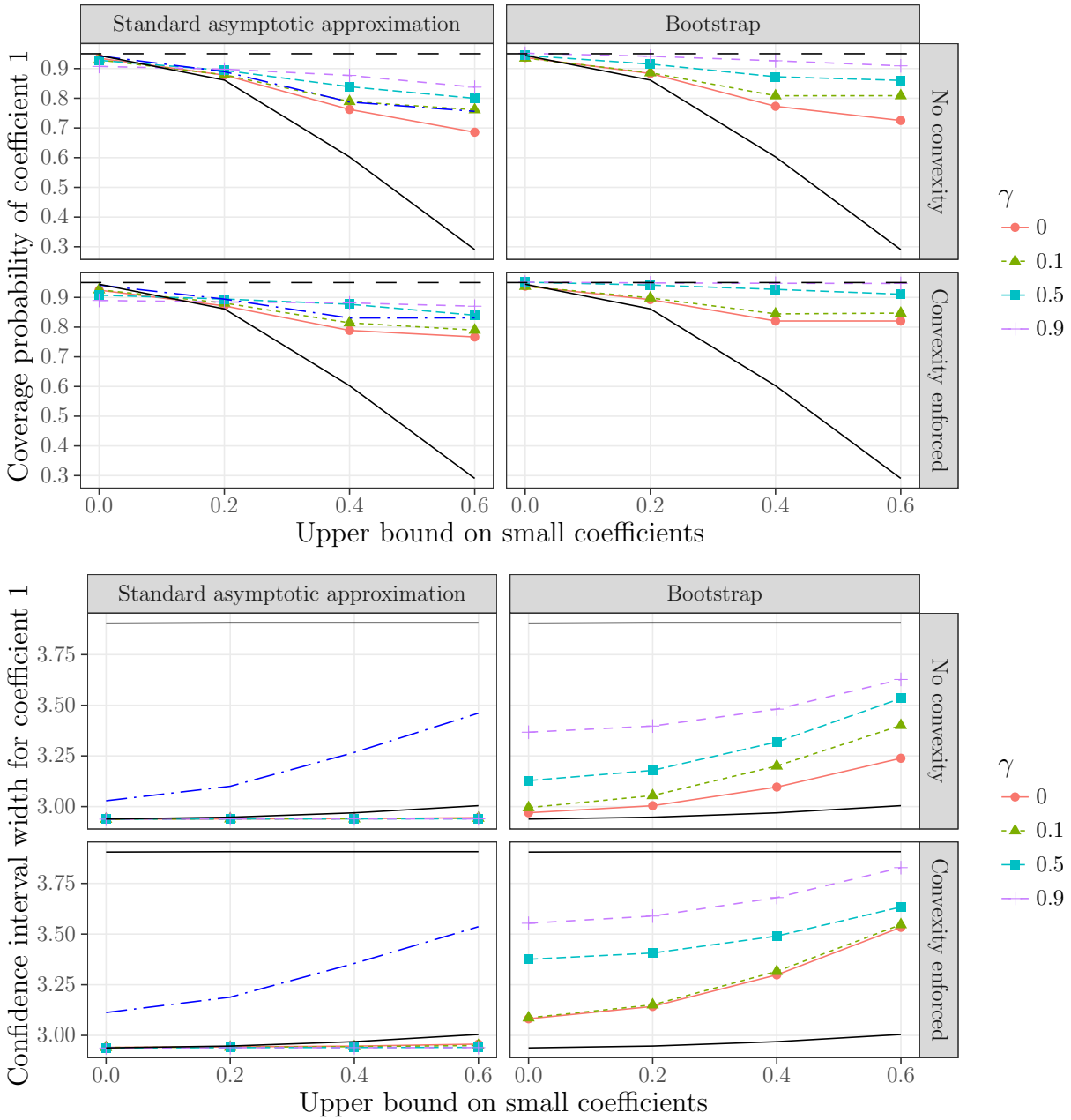

Figure 2.5. Coverage probability and width of the confidence interval of coefficient 1 under violations of exact sparsity. Coverage probability by OLS with only the first two covariates and confidence interval widths by full-model OLS and OLS with only the first two covariates in solid black. Results for SCAD inference in dash-dot blue.

Overall, simulation results strongly support the premise that excess variance due to estimates of zero coefficients needs to be explicitly accounted for when carrying out inference based on estimators with the smooth penalty. Moreover, inference by bootstrap with an estimator with a high value of γ appears to deliver an attractive combination of robust inference across a wide range of settings together with reduced length of confidence intervals relative to the full-model OLS estimator. Taken together with simulation results in Chapter 1, these results suggest that using a high value of gamma for estimation and subsequently using bootstrap for inference provides a good practical approach that delivers close-to-oracle performance under ideal circumstances and substantial robustness both to deviations from sparsity and to nontrivial covariance across estimates of zero and nonzero coefficients.

2.8. Conclusion

Asymptotic approximations to distributions of penalized estimators based on the concept of oracle property are appealing in their simplicity but are not always a good approximation to estimator properties in practice. I have considered alternative approaches to constructing asymptotic approximations to finite sample estimator behavior, including local asymptotics, semilocal asymptotics and asymptotics with no penalization bias but nontrivial variance effects of estimates of small coefficients on the estimates of the large ones. Among the three approaches, only the third one is amenable to a workable inference procedure that is materially different from simply not penalizing at all. This is achieved by a modified bootstrap procedure. Moreover, the same modified bootstrap procedure is also valid in the standard asymptotic framework that delivers oracle efficiency, and can

therefore be used without the researcher having to make a bet on whether her choice of tuning parameters achieves sufficient penalization or underpenalizes.

Simulation evidence strongly supports the need for this alternative asymptotic approximation even in situations that fit perfectly into the standard exactly sparse framework. Moreover, bootstrap paired with a smooth penalty with a high smoothness parameter continues to work well and deliver close-to-nominal coverage even in the settings where exact sparsity is violated, and penalization bias becomes an important consideration. This dovetails with the results in Chapter 1 that illustrate the reduction in bias achieved by smoother penalties. Such an approach also greatly improves on the coverage of confidence intervals from a small misspecified "pseudo-oracle" model, highlighting that estimating smaller *ad hoc* models under specification uncertainty should not be relied upon as an applied practice. At the same time, bootstrap confidence intervals remain smaller than those constructed from estimating the full model, providing a credible alternative to estimating an *ad hoc* model when some measure of efficiency is desirable.

Overall, the simulation evidence together with the theoretical results suggests that inference by bootstrap after estimation with a penalty with a high smoothness parameter is noticeably more robust to deviations from the standard sparse framework, and can provide an appropriate balance between increasing efficiency in ideal settings and delivering reliable estimation and inference in less-than-ideal circumstances.

CHAPTER 3

Empirical Application

3.1. Introduction and baseline results

This chapter illustrates the empirical applicability of the proposed estimation procedure. For this purpose, I will reevaluate some key results of Bazzi et al. (2016).

Bazzi et al. (2016) address an interesting question in development and broader human capital literature: to what extent is human capital location specific? The question is a natural one, especially in the context of agricultural production: after all, even experienced rice growers would find their skills out of place if, for example, tasked with deer herding in the tundra. While no one would advocate for such an extreme relocation programme, the question of the extent of skill specificity is pertinent in any migration context, and in particular is important for evaluating effects of climate change, which is expected to displace a large number of people from areas subject to flooding due to rising ocean levels. Higher skill specificity would imply larger adjustment costs in the case of population relocation, as well as heightened policy focus on matching immigrants to most-suitable destinations.

Quantifying skill transferability is nontrivial due to endogeneity concerns: voluntary migrant workers might settle in places that are more suitable to their skill set. Bazzi et al. (2016) address this problem by examining a large-scale resettlement program in Indonesia that, by and large, resettled migrants randomly to newly created villages, providing the

desired exogenous variation. To formally quantify skill specificity, the authors introduce *agroclimatic similarity* as a measure of similarity of one location to another, based on factors such as topography, hydrology, climate and soil properties. Aggregating this measure at a village level then characterizes how well matched the population in a given village is to their prior skill set. Since rice is the dominant agricultural output in the region in question, and since rice production is location-dependent and less standardized than other crops, village-level rice output (per hectare, log) is therefore a useful measure of skill-sensitive productivity.

In the terminology of this dissertation, agroclimatic similarity is the covariate of *ex ante* interest. In fact, coefficient on it is the only coefficient authors report in Table 3 that I am reanalyzing, implying that other covariates are only included to the extend that they could be confounding the estimate on agrosimilarity.

Bazzi et al. (2016) consider 83 additional regressors as controls. They are grouped in 7 categories: island fixed effects, predetermined village controls, origin province migrant shares, log weighted average distance to origins, weighted average predetermined controls at origins, predetermined controls at destinations, and demographics and schooling. In Table 3, Bazzi et al. (2016) report the results of 5 different linear regressions: there is the smallest one that only includes island fixed effects, and the largest one that includes all the controls; what I have called the *full model*. The other three specifications include island fixed effects and add various combinations of groups of other controls. All specifications are estimated by OLS, which is feasible since there are 600 observations (more than covariates).[1]

[1]I have replicated all the regressions reported in Table 3 and obtained the same numbers as Bazzi et al. (2016).

Bazzi et al. (2016) Table 3 – Effects of Agroclimatic Similarity on Rice Productivity

	(1)	(2)	(3)	(4)	(5)
Panel A. Rice productivity					
Agroclimatic similarity	0.204	0.182	0.210	0.151	0.166
	(0.064)	(0.045)	(0.075)	(0.057)	(0.068)
Number of Villages	600	600	600	600	600
R^2	0.149	0.032	0.178	0.281	0.318
Island fixed effects	Yes	Yes	Yes	Yes	Yes
Predetermined village controls (x_j)	Yes	No	Yes	Yes	Yes
Origin province migrant shares	No	No	Yes	No	Yes
log weighted avg. distance to origins	No	No	Yes	No	Yes
Weighted avg. predetermined controls, origins	No	No	Yes	No	Yes
Predetermined controls, destinations	No	No	No	Yes	Yes
Demographics and schooling	No	No	No	Yes	Yes

Note: This reproduction of Table 3 of Bazzi et al. (2016) omits *Panel B*, which carries out a placebo test by using cash crop productivity instead of rice productivity as the dependent variable.

The headline number in this analysis is the result of the regression with island fixed effects and only village controls, giving an estimate of 0.204 (i.e. 20%) increase in rice productivity for one standard deviation increase in agrosimilarity. While the estimate from the full model is smaller at 0.166, Bazzi et al. (2016) argue that it is not statististically significantly different, and so keep the 20% as the main result.

Note that if we consider island fixed effect as must-have controls (since they are included in all specifications the authors report), that leaves 80 controls that might or might not be included, with $2^{80} = 1.2 \times 10^{24}$ (1.2 septillion) potential combinations to explore. Even if we settled to only include and exclude the covariates in given groups, that would still leave $2^6 = 64$ specifications to analyse. While running 64 linear regressions is feasible,

the fact that only 5 specifications are reported, and that no further mention of specification search is made, leads me to believe no exhaustive search was performed and no formal analysis was carried out to choose which results to report.

3.2. Estimation

I reestimate the effect of agroclimatic similarity in this setting by using the proposed smooth penalization procedure. As described above, island fixed effects are included in all specifications the authors report, so I will treat them as must-have controls and therefore include them in the set of covariates that will not be penalized, in addition to agrosimilarity.

All covariates are demeaned and standardized to have sample covariance 1, and the outcome variable (log rice output per hectare) is demeaned. Consequently no constant is included.

The main question is the choice of tuning parameters: γ, λ, τ and a. I will set $a = 3.7$ consistent with the recommendation by Fan and Li (2001) for SCAD. The other parameters are chosen by leave-one-out crossvalidation. That is, for a given grid of potential choices of γ, λ and τ, crossvalidation criterion is computed for each point in the grid and the tuning parameters corresponding to the smallest value of crossvalidation criterion are chosen for the final estimations.

The choice of the range of τ is more complicated here than in the simulations since the smallest eigenvalue is low (3×10^{-4}), indicating multicollinearity, and the range of values for confidence interval widths for standardized covariates is wide, with the largest being 40 times the smallest. As such, I choose a log scale range between the half-widths of

the largest and the smallest 95% confidence intervals for coefficients (based on assuming homoskedastic errors).

Range of λ is chosen to include on the upper end values produced by linking λ to τ in the same manner as Fan and Li (2001), and going down about 50-fold, so that the lower end of the range isn't binding.

Crossvalidation yields a choice of $\gamma = 0.5$. It is worth asking whether Theorem 3 is useful here, i.e. whether this value of γ yields an asymptotic reduction in squared error promised by the theorem for high enough values of γ. Since we are only interested in the coefficient on agrosimilarity, we are using $\alpha_1 = (1, 0, .., 0)'$ as the contrast of interest. Using coefficients that are estimated larger than τ as 'nonzeros'[2] and plugging in sample covariance matrix for the dataset we get a positive value for the sample analog of constant C_γ in Theorem 3, i.e. smooth penalization should yield lower worst squared error than a model-selection based method like SCAD asymptotically under conditions of Theorem 3.

Estimating the model with parameter values chosen by crossvalidation yields an estimate of 0.11, i.e. 11% increase in rice productivity for one standard deviation increase in agrosimilarity, rather than 20% used as the headline number by Bazzi et al. (2016). The marked difference in the result appears largely due to omisssion of covariates in the group of predetermined (pre-resettlement) destination controls, such as destination literacy rates, schooling levels, technology penetration (TV, radio), proportion of population in trade and worker wages.[3] Note that this estimate lies both in the 95% confidence interval

[2]Changing the threshold to $a\tau$ instead of τ does not change the sign, so the argument still holds.

[3]In fact, adding this group of controls (destination controls minus potential crop yields) to island fixed effects and village controls (authors' preferred regression) and just running OLS gives an estimate of 0.08. However, since this work is explicitly advocating against model selection, I will rely on the 0.11 estimate that the penalized estimation method produces.

of the authors' preferred specification (specification 1) and of the full model (specification 5).

3.3. Inference

Carrying out inference here is complicated by the fact that assuming that the errors are i.i.d. is likely unreasonable. While the improvement in squared error result of Theorem 3 applies in this setting, the bootstrap validity result only applies to the i.i.d. context. As such, I will take three approaches to inference here, all of which are mostly illustrative.

The first approach is to reestimate standard errors via only considering "nonzero" covariates as in the standard asymptotic approach, but to construct the covariance matrix under the author's chosen assumptions on the error structure. As argued in subsubsection 1.4.2.1, this requires determining which controls to treat as "important" and which to discard. This is done by designating controls with coefficients larger than $\kappa\tau$ as "important", for $\kappa \in [1, a]$. I will consider values of 1 and a, 1 being a more conservative choice as it treats more covariates as "important", but will potentially deliver wider confidence intervals.

Specifically, to reestimate standard errors, I use authors' code for spatial HAC variance estimation with a set of controls with correspondingly large coefficients. To be precise, formal results on inference by standard asymptotic approximation do not cover inference in the model without error independence and homoskedasticity. Hence inference results here are tentative and are meant to illustrate the simplest way of conducting inference when conditions for oracle efficiency are satisfied. While I conjecture that oracle efficiency, and valid inference by the same standard asymptotic approximation via thresholding, can

be achieved under more general error structures (like the one assumed in Bazzi et al. (2016)), I offer no theoretical results on that here.

Using $\kappa = 1$ yields 26 controls, with the corresponding standard error on the coefficient for agrosimilarity of 0.057. Using $\kappa = a = 3.7$ yields 16 controls and standard error of 0.051. Corresponding 95% confidence intervals are $(0.003, 0.227)$ and $(0.015, 0.215)$.

It is notable that for both values of κ, standard errors are smaller than those in authors' preferred specification (0.064) and those from the full model (0.068), leading to narrower confidence intervals. Together with a lower point estimate the upper end of the confidence interval that I obtain here is noticeably lower than what could be constructed from Bazzi et al. (2016). Overall this approach leads to the conjecture that the effect of agroclimatic similarity on rice productivity is likely lower than that reported in Bazzi et al. (2016), but still different from zero at 95% confidence level.

The second and third approaches carry out inference under the i.i.d. assumption, by the standard asymptotic approximation and bootstrap, respectively. While this fits into the inference results in Chapter 1 and Chapter 2, it is unlikely to be a reasonable assumption for the given dataset. Instead, it is an illustration of relative merits of these approaches, in particular in comparison to homoskedastic inference from full-model OLS and smaller specifications.

Results are presented in Table 3.1. In addition to results for $\gamma = 0.5$ chosen by crossvalidation, I also present results for $\gamma = 0.9$ for comparison.

Note that confidence interval widths obtained by the standard asymptotic approximation are noticeably smaller than those by bootstrap and smaller than all but one OLS specification, highlighting the potential value of excluding some covariates in obtaining

Table 3.1. Inference under i.i.d. errors.

			OLS		
	(1)	(2)	(3)	(4)	(5)
Estimate	0.204	0.182	0.210	0.151	0.166
95% conf. int.	[0.077,0.331]	[0.080,0.284]	[0.071,0.350]	[0.000,0.301]	[-0.004,0.337]
Conf. int. width	0.254	0.204	0.279	0.301	0.340
			Penalized estimation		
	$\gamma = 0.5$			$\gamma = 0.9$	
Estimate	0.112			0.149	
			Standard asymptotic approximation		
95% conf. int.	[-0.004,0.227]			[0.030,0.269]	
Conf. int. width	0.231			0.239	
			Bootstrap		
95% conf. int.	[-0.040,0.251]			[0.007,0.308]	
Conf. int. width	0.291			0.300	

narrower confidence intervals. Bootstrap confidence intervals are wider than those under standard asymptotic approximation, but still narrower than those from the full-model OLS.

3.4. Conclusion

Reevaluation of importance of location-specific human capital in rice productivity post-resettlement in Indonesia addressed in Bazzi et al. (2016) highlights the dangers of *ad hoc* approach to model selection. I argue that the value of the key estimate obtained by the authors is driven by omitted variable bias due to a relatively small number of covariates, and obtain an estimate half the size. More detailed analysis of important controls suggests that they all belong to a single group of destination region covariates. However, even though a linear regression on a specification that included those covariates could have revealed a noticeable change in the coefficient of interest, choosing this specification would

have required either a good deal of luck or a significant effort to search over a large number of specifications, as well as choosing the best one after the search.

While theoretical results on inference in this dissertation do not admit the error structure deemed appropriate by Bazzi et al. (2016), a reevaluation of inference under i.i.d. error assumption highlights the potential reduction in confidence interval width achieved by using either the standard asymptotic approximation or bootstrap. Together with theoretical results on reduction in squared error under violations of exact sparsity and simulation evidence suggesting that bootstrap with smooth penalty is reasonably robust to departures from the standard framework, these results suggest that there are gains to using smoother penalties both for estimation and inference in the circumstances where the researcher expects some degree of sparsity and would like to leverage that, but can not be entirely sure that the standard assumptions reflect the DGP.

References

Akaike, H. (1974): "A new look at the statistical model identification," *IEEE transactions on automatic control*, 19, 716–723.

Bazzi, S., A. Gaduh, A. D. Rothenberg, and M. Wong (2016): "Skill Transferability, Migration, and Development: Evidence from Population Resettlement in Indonesia," *American Economic Review*, 106, 2658–2698.

Belloni, A., V. Chernozhukov, and C. Hansen (2014): "Inference on treatment effects after selection among high-dimensional controls," *The Review of Economic Studies*, 81, 608–650.

Chatterjee, A. and S. N. Lahiri (2010): "Asymptotic properties of the residual bootstrap for lasso estimators," *Proceedings of the American Mathematical Society*, 138, 4497–4509.

——— (2011): "Bootstrapping lasso estimators," *Journal of the American Statistical Association*, 106, 608–625.

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

Fan, J. and H. Peng (2004): "Nonconcave penalized likelihood with a diverging number of parameters," *The Annals of Statistics*, 32, 928–961.

Frank, L. E. and J. H. Friedman (1993): "A statistical view of some chemometrics regression tools," *Technometrics*, 35, 109–135.

Hastie, T., R. Tibshirani, and J. Friedman (2009): *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*, Springer Series in Statistics, Springer.

Horowitz, J. L. and J. Huang (2013): "Penalized estimation of high-dimensional models under a generalized sparsity condition," *Statistica Sinica*, 23, 725–748.

Huang, J., J. L. Horowitz, and S. Ma (2008): "Asymptotic properties of bridge estimators in sparse high-dimensional regression models," *The Annals of Statistics*, 36, 587–613.

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

Koenker, R. (1988): "Asymptotic theory and econometric practice," *Journal of Applied Econometrics*, 3, 139–147.

Leeb, H. and B. M. Pötscher (2008): "Sparse estimators and the oracle property, or the return of Hodges estimator," *Journal of Econometrics*, 142, 201–211.

Rockafellar, R. T. (2015): *Convex Analysis*, Princeton University Press.

Schwarz, G. (1978): "Estimating the dimension of a model," *The Annals of Statistics*, 6, 461–464.

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

Van de Geer, S., P. Bühlmann, Y. Ritov, R. Dezeure, et al. (2014): "On asymptotically optimal confidence regions and tests for high-dimensional models," *The Annals of Statistics*, 42, 1166–1202.

Varian, H. R. (2014): "Big Data: New Tricks for Econometrics," *Journal of Economic Perspectives*, 28, 3–28.

Wang, H., R. Li, and C.-L. Tsai (2007): "Tuning parameter selectors for the smoothly clipped absolute deviation method," *Biometrika*, 94, 553–568.

Zhang, C.-H. (2010): "Nearly unbiased variable selection under minimax concave penalty," *The Annals of Statistics*, 38, 894–942.

Zou, H. (2006): "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, 101, 1418–1429.

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

Appendices

CHAPTER A

Appendix to Chapter 1

A.1. Additional simulation results

	Upper bound on small coefficients					
	0	0.1	0.2	0.3	0.4	0.5
SCAD	0.58	0.69	0.96	1.34	1.71	2.04
$\gamma = 0.1$	0.62	0.69	0.93	1.21	1.40	1.55
$\gamma = 0.2$	0.62	0.70	0.90	1.10	1.28	1.39
$\gamma = 0.3$	0.64	0.69	0.86	1.06	1.22	1.34
$\gamma = 0.4$	0.64	0.69	0.83	1.01	1.16	1.27
$\gamma = 0.5$	0.65	0.69	0.82	0.97	1.13	1.24
$\gamma = 0.6$	0.66	0.68	0.81	0.96	1.11	1.22
$\gamma = 0.7$	0.67	0.69	0.78	0.91	1.05	1.18
$\gamma = 0.8$	0.68	0.70	0.78	0.90	1.03	1.13
$\gamma = 0.9$	0.68	0.71	0.77	0.88	0.99	1.13
OLS	1.00	0.99	0.99	0.99	0.99	0.99

Table A.1. Simulated MSE of estimates of coefficient 1. All results from 10 000 simulations.

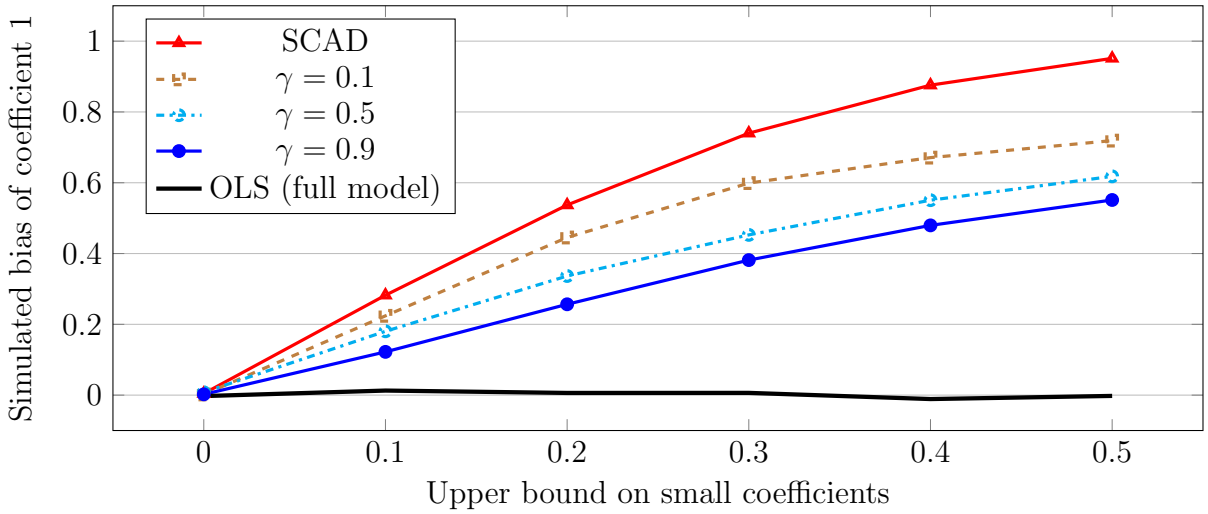

Figure A.1. Simulated bias of estimates of coefficient 1. All results from 10 000 simulations.

A.2. Proofs

Proof of Lemma 1. In this proof I will show that a weaker version of A1(a) that replaces the positive definite requirement with positive semidefinite is sufficient for convexity, and that A1(a) as stated ensures strict convexity.

To simplify notation I will show convexity of $Q_n(b)/n$. We have

$$\begin{aligned}\frac{Q_n(b)}{n} &= \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i'b)^2 + \frac{\lambda_n}{n}\sum_{j=k_0+1}^{p_n}\text{Pen}(b_j) \\ &= \frac{1}{n}\sum_{i=1}^{n}y_i^2 - \left[\frac{2}{n}\sum_{i=1}^{n}y_i x_i'\right]b + b'\left[\frac{1}{n}\sum_{i=1}^{n}x_i x_i'\right]b + \frac{\lambda_n}{n}\sum_{j=k_0+1}^{p_n}\text{Pen}(b_j).\end{aligned}$$

Observe that the first two terms on the last line form an affine function of b, so it is enough to show convexity of

$$f(b) = b'\Sigma_n b + \frac{\lambda_n}{n}\sum_{j=k_0+1}^{p_n}\text{Pen}(b_j).$$

Observe that f is everywhere differentiable and the gradient is given by

$$\bigtriangledown f = 2\Sigma_n b + \frac{\lambda_n}{n} \bigtriangledown \text{Pen}(b),$$

where $\bigtriangledown \text{Pen}(b)$ is a vector with the first k_0 components equal to zero and the rest equal to $\text{Pen}'(b_j)$ for corresponding j.

Consider two arbitrary points $b_1, b_2 \in \mathbb{R}^{p_n}$, $b_1 \neq b_2$, and let $\Delta b = b_2 - b_1$. Consider function

$$G(\alpha) = f\left((1-\alpha)b_1 + \alpha b_2\right)$$

on $\alpha \in [0,1]$. It is enough to show convexity of $G(\alpha)$.

By Theorem 24.2 of Rockafellar (2015), it is enough to show that the function

$$\begin{aligned} g(\alpha) &= \bigtriangledown f\left((1-\alpha)b_1 + \alpha b_2\right)'(b_2 - b_1) \\ &= 2b_1'\Sigma_n \Delta b + 2\Delta b'\Sigma_n \Delta b \alpha + \frac{\lambda_n}{n} \sum_{j=k_0+1}^{p_n} \text{Pen}'(b_{1j} + \alpha \Delta b_j)\Delta b_j \end{aligned}$$

is nondecreasing in α to prove convexity. A straightforward adjustment to the proof of Theorem 24.2 of Rockafellar (2015) shows that it is enough to show that $g(\alpha)$ is strictly increasing in α to show strict convexity.

Consider two arbitrary values $\alpha_1 < \alpha_2$, and let $\Delta\alpha = \alpha_2 - \alpha_1$. We have

$$\begin{aligned} g(\alpha_2) - g(\alpha_1) &= 2\Delta\alpha\Delta b'\Sigma_n \Delta b \\ &\quad + \frac{\lambda_n}{n} \sum_{j=k_0+1}^{p_n} \underbrace{\left[\text{Pen}'(b_{1j} + \alpha_2 \Delta b_j) - \text{Pen}'(b_{1j} + \alpha_1 \Delta b_j)\right]}_{\Delta \text{Pen}'_j} \Delta b_j. \end{aligned}$$

Now for each j we have

$$\Delta \operatorname{Pen}'_j \Delta b_j \geq -\frac{2}{(a-1)\tau_n^{1-\gamma}} \Delta\alpha (\Delta b_j)^2$$

(verify first for $\Delta b_j > 0$ and then argue the opposite case the same way).

Therefore we have

$$\begin{aligned} g(\alpha_2) - g(\alpha_1) &\geq 2\Delta\alpha \Delta b' \Sigma_n \Delta b - 2\Delta\alpha \frac{\lambda_n}{(a-1)n\tau_n^{1-\gamma}} \underbrace{\sum_{j=k_0+1}^{p_n} (\Delta b_j)^2}_{\Delta b' I_{p_n,k_0} \Delta b} \\ &= 2\Delta\alpha \Delta b' \left[\Sigma_n - \frac{\lambda_n}{(a-1)n\tau_n^{1-\gamma}} I_{p_n,k_0} \right] \Delta b. \end{aligned}$$

Since $\Delta\alpha > 0$, it is enough for the matrix inside square brackets to be positive semi-definite to ensure that $g(\alpha_2) - g(\alpha_1) \geq 0$. This is assured by the weak inequality version of assumption A1(a). Moreover, under the strict inequality version of A1(a) we have $g(\alpha_2) - g(\alpha_1) > 0$, and so the objective function is strictly convex. □

I will need the following lemma from Huang, Horowitz, and Ma (2008):

Lemma 8. *Let u be a $p_n \times 1$ vector. Under assumptions A2(a) and A2(b)*

$$E \sup_{\|u\|<\delta} \left| \sum_{i=1}^{n} \varepsilon_i x_i' u \right| \leq \delta\sigma n^{1/2} p_n^{1/2}.$$

I will prove Theorem 1 under more general conditions that allow for small coefficients that are not exactly zero but are nonetheless very close to zero. To this end assumption A15 defines how small these coefficients need to be. Assumption A15 is trivially satisfied

under exact sparsity, therefore Theorem 1 presented in the main text is a special case of the result proved here.

Assumption A15 (Approximate sparsity). *Let* $b_{2n} = \max\{|\beta_{0j}|, k_n + 1 \leq j \leq p_n\}$.

(a)

$$b_{2n} = O\left(\left[\frac{k_n}{p_n}\right]^{\frac{1}{1+\gamma}} \tau_n\right).$$

(b)

$$b_{2n}^{1+\gamma} \lambda_n \rho_n \frac{p_n - k_n}{p_n} = O(1).$$

Theorem 8 (Generalized version of Theorem 1). *Suppose* $\gamma \in [0, 1]$ *and assumptions A2(a), A2(b) and A3(a) hold. Then*

$$\|\hat{\beta}_n - \beta_0\| = O_p\left(\frac{p_n + \lambda_n\left[k_n \tau_n^{1+\gamma} + (p_n - k_n) b_{2n}^{1+\gamma}\right]}{n\rho_n}\right)^{1/2}.$$

Suppose moreover that A4(a), A15(a) and A15(b) hold. Then

$$\|\hat{\beta}_n - \beta_0\| = O_p\left(\rho_n^{-1}\left(\frac{p_n}{n}\right)^{1/2}\right).$$

Proof of Theorem 8 (and Theorem 1 as a special case of it). This proof follows closely the one for Theorem 1 from Huang, Horowitz, and Ma (2008).

By the definition of $\hat{\beta}_n$, we have

$$\sum_{i=1}^{n}(y_i - x_i'\hat{\beta}_n)^2 + \lambda_n \sum_{j=k_0+1}^{p_n} \text{Pen}(\hat{\beta}_{nj}) \leq \sum_{i=1}^{n}(y_i - x_i'\beta_0)^2 + \lambda_n \sum_{j=k_0+1}^{p_n} \text{Pen}(\beta_{0j}).$$

Dropping the penalty term on the LHS and letting $\eta_n = \lambda_n \sum_{j=k_0+1}^{p_n} \text{Pen}(\beta_{0j})$ we get

$$\begin{aligned}\eta_n &\geq \sum_{i=1}^{n}(y_i - x_i'\hat{\beta}_n)^2 - \sum_{i=1}^{n}(y_i - x_i'\beta_0)^2 \\ &= \sum_{i=1}^{n}[x_i'(\hat{\beta}_n - \beta_0)]^2 + 2\sum_{i=1}^{n}\varepsilon_i x_i'(\beta_0 - \hat{\beta}_n).\end{aligned}$$

Let $\Delta_n = n^{1/2}(\Sigma_n)^{1/2}(\hat{\beta}_n - \beta_0)$, $D_n = n^{-1/2}(\Sigma_n)^{-1/2}X'$ and $\varepsilon = (\varepsilon_1, .., \varepsilon_n)'$. Then

$$\sum_{i=1}^{n}[x_i'(\hat{\beta}_n - \beta_0)]^2 + 2\sum_{i=1}^{n}\varepsilon_i x_i'(\beta_0 - \hat{\beta}_n) = \Delta_n'\Delta_n - 2(D_n\varepsilon)'\Delta_n$$

and

$$\Delta_n'\Delta_n - 2(D_n\varepsilon)'\Delta_n - \eta_n \leq 0,$$

$$||\Delta_n - D_n\varepsilon||^2 - ||D_n\varepsilon||^2 - \eta_n \leq 0,$$

$$||\Delta_n - D_n\varepsilon|| \leq ||D_n\varepsilon|| + \eta_n^{1/2}.$$

By the triangle inequality

$$||\Delta_n|| \leq ||\Delta_n - D_n\varepsilon|| + ||D_n\varepsilon||.$$

So

$$||\Delta_n|| \leq 2||D_n\varepsilon|| + \eta_n^{1/2}.$$

Using the fact that $xy \leq \frac{1}{2}(x^2 + y^2)$ for two scalars x and y we have

$$||\Delta_n||^2 \leq 6||D_n\varepsilon||^2 + 3\eta_n.$$

Let d_i be the i'th column of D_n, so that $D_n\varepsilon = \sum_{i=1}^n d_i\varepsilon_i$. Since ε_i and ε_j are uncorrelated for $i \neq j$ by A2(a), and by A2(b), we have

$$E||D_n\varepsilon||^2 = \sum_{i=1}^n ||d_i||^2 E(\varepsilon_i^2) = \sigma^2 \operatorname{trace}(D_n'D_n) = \sigma^2 p_n, \tag{A.1}$$

$$E||\Delta_n||^2 \leq 6\sigma^2 p_n + 3\eta_n. \tag{A.2}$$

Splitting the penalty into that for 'nonzeros' and 'zeros' we have:

$$\begin{aligned}
\eta_n &= \lambda_n \sum_{j=k_0+1}^{p_n} \operatorname{Pen}(\beta_{0j}) \\
&\leq \lambda_n k_n \max_b \operatorname{Pen}(b) + \lambda_n(p_n - k_n)\operatorname{Pen}(b_{2n}) \\
&\leq \lambda_n k_n \operatorname{Pen}(a\tau_n) + \lambda_n(p_n - k_n)\frac{2}{1+\gamma} b_{2n}^{1+\gamma} \\
&= \lambda_n k_n \tau_n^{1+\gamma}\left[\frac{1-\gamma}{1+\gamma} + a\right] + \lambda_n(p_n - k_n)\frac{2}{1+\gamma} b_{2n}^{1+\gamma} \\
&\leq \lambda_n \left[k_n\tau_n^{1+\gamma} + (p_n - k_n)b_{2n}^{1+\gamma}\right](a+1).
\end{aligned}$$

So

$$nE\left[(\hat\beta_n - \beta_0)'\Sigma_n(\hat\beta_n - \beta_0)\right] \leq 6\sigma^2 p_n + 3(a+1)\lambda_n\left[k_n\tau_n^{1+\gamma} + (p_n - k_n)b_{2n}^{1+\gamma}\right].$$

By min-max theorem and A3(a) $||\hat\beta_n - \beta_0||^2 \leq \rho_n^{-1}(\hat\beta_n - \beta_0)'\Sigma_n(\hat\beta_n - \beta_0)$ so by Jensen's inequality

$$E||\hat\beta_n - \beta_0|| \leq \left[\frac{6\sigma^2 p_n + 3(a+1)\lambda_n\left[k_n\tau_n^{1+\gamma} + (p_n - k_n)b_{2n}^{1+\gamma}\right]}{n\rho_n}\right]^{1/2}.$$

Hence

$$||\hat{\beta}_n - \beta_0|| = O_p\left(\left[\frac{p_n + \lambda_n\left[k_n\tau_n^{1+\gamma} + (p_n - k_n)b_{2n}^{1+\gamma}\right]}{n\rho_n}\right]^{1/2}\right).$$

This completes the first part of the proof. For the second part, let $r_n = \rho_n\left(\frac{n}{p_n}\right)^{1/2}$. We want to show that

$$r_n||\hat{\beta}_n - \beta_0|| = O_p\left(1\right).$$

For each n partition the parameter space minus β_0 into 'shells'

$$S_{m,n} = \left\{\beta : r_n||\beta - \beta_0|| \in \left(2^{m-1}, 2^m\right]\right\},$$

where m is an integer. Since $\hat{\beta}_n$ minimizes sample objective, we have, for all $\epsilon_n > 0$,

$$P(r_n||\hat{\beta}_n - \beta_0|| > 2^{M-1}) \leq \sum_{m \geq M, 2^m \leq \epsilon_n r_n} P\left(\inf_{\beta \in S_{m,n}}\left(Q_n(\beta) - Q_n(\beta_0)\right) \leq 0\right) + P\left(2||\hat{\beta}_n - \beta_0|| > \epsilon_n\right).$$

We will pick a sequence $\epsilon_n = \left[\frac{p_n + \lambda_n k_n \tau_n^{1+\gamma}}{n\rho_n}\right]^{1/2}\zeta_n$, with ζ_n such that $\zeta_n \to \infty$ and $\epsilon_n = o(b_n)$. From the result in the first part of the theorem and from A4(a) and A15(a) we have that the last term converges to zero as $n \to \infty$. So we are left to show that the first term goes

to zero.

$$\begin{aligned}
Q_n(\beta) - Q_n(\beta_0) =& \sum_{i=1}^{n}(y_i - x_i'\beta)^2 - \sum_{i=1}^{n}(y_i - x_i'\beta_0)^2 \\
&+ \lambda_n \sum_{j=k_0+1}^{k_n} \text{Pen}(\beta_j) + \lambda_n \sum_{j=k_n+1}^{p_n} \text{Pen}(\beta_j) - \lambda_n \sum_{j=k_0+1}^{p_n} \text{Pen}(\beta_{0j}) \\
\geq& \sum_{i=1}^{n}[x_i'(\beta - \beta_0)]^2 - 2\sum_{i=1}^{n}\varepsilon_i x_i'(\beta - \beta_0) \\
&+ \lambda_n \sum_{j=k_0+1}^{k_n} [\text{Pen}(\beta_j) - \text{Pen}(\beta_{10j})] - \lambda_n \sum_{j=k_n+1}^{p_n} \text{Pen}(\beta_{0j}) \\
\equiv& I_{1n} + I_{2n} + I_{3n} + I_{4n}.
\end{aligned}$$

On $S_{m,n}$ we have

$$I_{1n} > n\rho_n 2^{2(m-1)} r_n^{-2}.$$

Let $\underline{I}_{4n} = \lambda_n(p_n - k_n)b_{2n}^{1+\gamma}\frac{2}{1+\gamma}$. We have

$$I_{4n} \geq -\underline{I}_{4n}.$$

Using assumption A4(a) we can bound I_{3n} at zero for n large enough. With $\tau_n = o(b_n)$ $\text{Pen}(\beta_{10j}) = \max_b \text{Pen}(b)$ for all $j : k_0 + 1 \leq j \leq k_n$, for all n such that $a\tau_n < b_n$. Also, for all $j : k_0 + 1 \leq j \leq k_n$,

$$\begin{aligned}
|\beta_j| &\geq b_n - ||\beta - \beta_0||_\infty \\
&\geq b_n - ||\beta - \beta_0|| \\
&\geq b_n - \epsilon_n,
\end{aligned}$$

where the last line holds on all shells considered above. Since $\epsilon_n = o(b_n)$ and $\tau_n = o(b_n)$, for all n large enough $|\beta_j| > a\tau_n$ and $\text{Pen}(\beta_j) = \max_b \text{Pen}(b)$ for all $j : k_0 + 1 \leq j \leq k_n$ on all shells considered above. Hence for all n large enough $I_{3n} = 0$ on all shells $S_{m,n}$ considered above.

So

$$Q_n(\beta) - Q_n(\beta_0) \geq -|I_{2n}| + n\rho_n 2^{2(m-1)} r_n^{-2} - \underline{\mathrm{I}}_{4n}.$$

Hence, by Markov inequality and Lemma 8,

$$\begin{aligned} P\left(\inf_{\beta \in S_{m,n}} (Q_n(\beta) - Q_n(\beta_0)) \leq 0\right) &\leq P\left(\sup_{\beta \in S_{m,n}} |I_{2n}| \geq n\rho_n 2^{2(m-1)} r_n^{-2} - \underline{\mathrm{I}}_{4n}\right) \\ &\leq \frac{2\sigma n^{1/2} p_n^{1/2} 2^m r_n^{-1}}{n\rho_n 2^{2(m-1)} r_n^{-2} - \underline{\mathrm{I}}_{4n}} \\ &= \frac{\sigma}{2^{m-3} - \underline{\mathrm{I}}_{4n} \rho_n p_n^{-1} 2^{-m-1}}. \end{aligned}$$

We have

$$\underline{\mathrm{I}}_{4n} \rho_n p_n^{-1} 2^{-m-1} \leq \lambda_n \frac{p_n - k_n}{p_n} b_{2n}^{1+\gamma} \rho_n 2^{-m},$$

which by A15(b) is smaller than $\frac{1}{4} 2^{m-3}$ for all M large enough, $m \geq M$.

Finally

$$\sum_{m \geq M, 2^m \leq \epsilon_n r_n} P\left(\inf_{\beta \in S_{m,n}} (Q_n(\beta) - Q_n(\beta_0)) \leq 0\right) \leq \sum_{m \geq M} \frac{\sigma}{2^{m-4}} = \frac{\sigma}{2^{M-5}}.$$

This goes to zero for every $M = M_n \to \infty$. So we've shown that $r_n ||\hat{\beta}_n - \beta_0|| = O_p(1)$, which completes the proof.

□

Lemma 9. *Suppose assumptions of Theorem 8 (A2(a), A2(b), A3(a), A4(a), A15(a) and A15(b)) hold, and moreover that $b_{2n} = o(\tau_n)$ and A4(b) hold. Then with probability approaching 1 estimates of 'zeros' will be in the inner area of the penalty. That is,*

$$\|\hat{\beta}_{2n}\|_\infty = o_p(\tau_n).$$

Proof of Lemma 9. By the triangle inequality and Theorem 8

$$\begin{aligned}\|\hat{\beta}_{2n}\|_\infty &\leq \|\hat{\beta}_{2n} - \beta_{20}\|_\infty + \|\beta_{20}\|_\infty \\ &\leq \rho_n^{-1}\left(\frac{p_n}{n}\right)^{1/2} + b_{2n}.\end{aligned}$$

The first term is $o_p(\tau_n)$ by A4(b) and the second term is $o(\tau_n)$ by assumption. □

Proof of Lemma 2. Let $h_n = \rho_n^{-1}\left(\frac{p_n}{n}\right)^{1/2}$. By Theorem 1 with A4(a), there exists $C < \infty$ such that $\|\hat{\beta}_n - \beta_0\| \leq h_n C$ with probability approaching 1. Let $\hat{\beta}_{1n} = \beta_{10} + h_n u_1$ and $\hat{\beta}_{2n} = \beta_{20} + h_n u_2 = h_n u_2$. We have $\|u_1\|^2 + \|u_2\|^2 \leq C^2$ by the argument above. The gradient of the objective function is

$$\frac{\partial Q_n(b)}{\partial b} = -2\sum_{i=1}^{n}(y_i - x_i'b)\cdot x_i + \lambda_n \bigtriangledown \mathrm{Pen}(b),$$

where $\bigtriangledown \mathrm{Pen}(b)$ is the gradient of the function $\sum_{j=k_0+1}^{p_n}\mathrm{Pen}(b_j)$.

Setting the gradient to zero as our first-order condition and using the notation introduced above we have

$$0 = -2\sum_{i=1}^{n}(\varepsilon_i - h_n x_{1,i}'u_1 - h_n x_{2,i}'u_2)\cdot x_i + \lambda_n \bigtriangledown \mathrm{Pen}(\hat{\beta}_n). \tag{A.3}$$

By assumption A4(b) $\frac{h_n}{\tau_n} \to 0$, so the estimators of zeros will be in the inner range of the penalty function. By assumption A4(a) and Theorem 1 the estimators of nonzeros will be in the outer (flat) range of the penalty function with probability approaching 1. We rewrite the first-order conditions for the part of the vector corresponding to zeros (i.e. the last $p_n - k_n$ components):

$$0 = -\sum_{i=1}^{n}(\varepsilon_i - h_n x'_{1,i}u_1 - h_n x'_{2,i}u_2)\cdot x_{2,i} + \lambda_n \operatorname{sgn}(u_2)|h_n u_2|^{\gamma}, \tag{A.4}$$

where the last term is understood as the vector with j'th component $\operatorname{sgn}(u_{2j})|h_n u_{2j}|^{\gamma}$.

Transpose the RHS of Equation A.4 and multiply by $h_n u_2$ to get a scalar condition:

$$h_n^2\sum_{i=1}^{n}(x'_{2,i}u_2)^2 + h_n^2\sum_{i=1}^{n}x'_{1,i}u_1\cdot x'_{2,i}u_2 - h_n\sum_{i=1}^{n}\varepsilon_i\cdot x'_{2,i}u_2 + \lambda_n\|h_n u_2\|_{1+\gamma}^{1+\gamma} = 0. \tag{A.5}$$

We can bound the first two terms from below as

$$\begin{aligned}
h_n^2\sum_{i=1}^{n}(x'_{2,i}u_2)^2 + h_n^2\sum_{i=1}^{n}x'_{1,i}u_1\cdot x'_{2,i}u_2 &\geq \frac{1}{2}h_n^2\sum_{i=1}^{n}(x'_{2,i}u_2)^2 \\
&\quad - \frac{1}{2}h_n^2\sum_{i=1}^{n}(x'_{2,i}u_2)^2 - \frac{1}{2}h_n^2\sum_{i=1}^{n}(x'_{1,i}u_1)^2 \\
&= -\frac{1}{2}h_n^2\sum_{i=1}^{n}(x'_{1,i}u_1)^2 \\
&\geq -\frac{1}{2}nh_n^2\operatorname{Eig}_{\max}(\Sigma_{1n})\|u_1\|^2 \\
&\geq -\frac{p_n}{\rho_n^2}\kappa_{1n}C^2
\end{aligned}$$

on the sets of probability approaching 1 where $\|u\| \leq C$.

We can bound the third term by bounding its expectation using A2(a) and A2(b):

$$\begin{aligned} E\left|\sum_{i=1}^{n} \varepsilon_i \cdot x'_{2,i} u_2\right| &\leq \left[E\left(\sum_{i=1}^{n} \varepsilon_i \cdot x'_{2,i} u_2\right)^2\right]^{1/2} \\ &= \sigma\left[\sum_{i=1}^{n} (x'_{2,i} u_2)^2\right]^{1/2} \\ &\leq \sigma n^{1/2} \operatorname{Eig}_{\max}(\Sigma_{2n})^{1/2} \|u_2\| \\ &\leq \sigma n^{1/2} p_n^{1/2} C \end{aligned}$$

on the sets of probability approaching 1 where $\|u\| \leq C$. So

$$h_n \sum_{i=1}^{n} \varepsilon_i \cdot x'_{2,i} u_2 = O_p\left(\frac{p_n}{\rho_n}\right).$$

Plugging these results back into Equation A.5 we get

$$\begin{aligned} \lambda_n \|h_n u_2\|_{1+\gamma}^{1+\gamma} &= O_p\left(\frac{p_n \kappa_{1n}}{\rho_n^2} + \frac{p_n}{\rho_n}\right), \\ \|\hat{\beta}_{2n}\|_{1+\gamma}^{1+\gamma} &= O_p\left(\frac{p_n \kappa_{1n}}{\rho_n^2 \lambda_n}\right). \end{aligned}$$

□

Proof of Theorem 2. Rewrite the first-order conditions from Equation A.3 separately for zeros and nonzeros as a system of two (vector) equations, where by results of Theorem 1 and assumptions A4(a) and A4(b) nonzeros are in the flat outer part of the

penalty and zeros are in the inner part:

$$\begin{cases} -\sum_{i=1}^{n} \varepsilon_i x_{1,i} + h_n \sum_{i=1}^{n} x_{1,i} x'_{1,i} u_1 + h_n \sum_{i=1}^{n} x_{1,i} x'_{2,i} u_2 = 0; \\ \\ -\sum_{i=1}^{n} \varepsilon_i x_{2,i} + h_n \sum_{i=1}^{n} x_{2,i} x'_{1,i} u_1 + h_n \sum_{i=1}^{n} x_{2,i} x'_{2,i} u_2 + \lambda_n \operatorname{sgn}(u_2)|h_n u_2|^{\gamma} = 0. \end{cases}$$

$$\begin{cases} \Sigma_{1n} h_n u_1 + \tilde{\Sigma}_n h_n u_2 = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i x_{1,i}; \\ \\ \frac{\lambda_n}{n} \operatorname{sgn}(u_2)|h_n u_2|^{\gamma} + \Sigma_{2n} h_n u_2 + \tilde{\Sigma}'_n h_n u_1 = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i x_{2,i}. \end{cases}$$

Since by A3(a) we assumed that $\rho_n > 0$, we can solve for $h_n u_1$ in the first equation and plug it into the second:

$$\begin{cases} h_n u_1 = \Sigma_{1n}^{-1} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i x_{1,i} - \Sigma_{1n}^{-1} \tilde{\Sigma}_n h_n u_2; \\ \\ \frac{\lambda_n}{n} \operatorname{sgn}(u_2)|h_n u_2|^{\gamma} + \Sigma_{2n} h_n u_2 - \tilde{\Sigma}'_n \Sigma_{1n}^{-1} \tilde{\Sigma}_n h_n u_2 = \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i x_{2,i} - \tilde{\Sigma}'_n \Sigma_{1n}^{-1} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i x_{1,i}. \end{cases}$$

Multiplying by $\sqrt{n}$ and rearranging we get the following form of the first-order conditions:

(A.6)

$$\begin{cases} n^{1/2} h_n u_1 = \dfrac{1}{n^{1/2}} \displaystyle\sum_{i=1}^{n} \varepsilon_i \left\{ \Sigma_{1n}^{-1} x_{1,i} \right\} - \Sigma_{1n}^{-1} \tilde{\Sigma}_n n^{1/2} h_n u_2; \\ \dfrac{\lambda_n}{n^{1/2}} \operatorname{sgn}(u_2)|h_n u_2|^{\gamma} = \dfrac{1}{n^{1/2}} \displaystyle\sum_{i=1}^{n} \varepsilon_i \left\{ x_{2,i} - \tilde{\Sigma}'_n \Sigma_{1n}^{-1} x_{1,i} \right\} - \left[\Sigma_{2n} - \tilde{\Sigma}'_n \Sigma_{1n}^{-1} \tilde{\Sigma}_n \right] n^{1/2} h_n u_2. \end{cases}$$

For a sequence of p_n-vectors α_n, partition $\alpha_n = (\alpha_{1n}', \alpha_{2n}')'$, where α_{1n} contains the first k_n components of α_n, and α_{2n} the rest. Let

$$s_n^2 = \sigma^2 \left\{ \alpha_{1n}' \Sigma_{1n}^{-1} \alpha_{1n} + \alpha_{2n}' \left[\Sigma_{2n} - \tilde{\Sigma}_n' \Sigma_{1n}^{-1} \tilde{\Sigma}_n \right] \alpha_{2n} \right\}.$$

Then

$$\begin{aligned} \frac{1}{s_n} \alpha_n' \begin{pmatrix} n^{1/2} h_n u_1 \\ \frac{\lambda_n}{n^{1/2}} \operatorname{sgn}(u_2) |h_n u_2|^\gamma \end{pmatrix} =& \frac{1}{n^{1/2}} \sum_{i=1}^n \varepsilon_i \frac{1}{s_n} \left\{ \alpha_{1n}' \Sigma_{1n}^{-1} x_{1,i} + \alpha_{2n}' \left(x_{2,i} - \tilde{\Sigma}_n' \Sigma_{1n}^{-1} x_{1,i} \right) \right\} \\ & - \frac{1}{s_n} \alpha_{1n}' \Sigma_{1n}^{-1} \tilde{\Sigma}_n n^{1/2} h_n u_2 \\ & - \frac{1}{s_n} \alpha_{2n}' \left[\Sigma_{2n} - \tilde{\Sigma}_n' \Sigma_{1n}^{-1} \tilde{\Sigma}_n \right] n^{1/2} h_n u_2. \end{aligned}$$

Observe that $\Sigma_{2n} - \tilde{\Sigma}_n' \Sigma_{1n}^{-1} \tilde{\Sigma}_n = \frac{1}{n} Z' M_W Z$ and hence it is symmetric positive semi-definite. Since it is the Schur complement of Σ_{1n} in Σ_n and by A3(a) we assumed that $\rho_n > 0$, it is positive definite. Consider the last two terms in the expression above. Let

$$R_n = \frac{1}{s_n} \alpha_{1n}' \Sigma_{1n}^{-1} \tilde{\Sigma}_n n^{1/2} h_n u_2 + \frac{1}{s_n} \alpha_{2n}' \left[\Sigma_{2n} - \tilde{\Sigma}_n' \Sigma_{1n}^{-1} \tilde{\Sigma}_n \right] n^{1/2} h_n u_2.$$

Then by repeated application of Cauchy-Schwarz inequality and by min-max theorem

$$\begin{aligned} R_n^2 \leq & 2 \frac{\alpha_{1n}' \Sigma_{1n}^{-1} \alpha_{1n}}{s_n^2} \left(n^{1/2} h_n u_2 \right)' \tilde{\Sigma}_n' \Sigma_{1n}^{-1} \tilde{\Sigma}_n \left(n^{1/2} h_n u_2 \right) \\ & + 2 \frac{\alpha_{2n}' \left[\Sigma_{2n} - \tilde{\Sigma}_n' \Sigma_{1n}^{-1} \tilde{\Sigma}_n \right] \alpha_{2n}}{s_n^2} \left(n^{1/2} h_n u_2 \right)' \left[\Sigma_{2n} - \tilde{\Sigma}_n' \Sigma_{1n}^{-1} \tilde{\Sigma}_n \right] \left(n^{1/2} h_n u_2 \right) \\ \leq & 2 \sigma^{-2} \left(n^{1/2} h_n u_2 \right)' \Sigma_{2n} \left(n^{1/2} h_n u_2 \right) \\ \leq & 2 \sigma^{-2} \operatorname{Eig}_{\max}(\Sigma_{2n}) \| n^{1/2} h_n u_2 \|^2. \end{aligned}$$

Hence by Lemma 2 we have

$$R_n = O_p\left(\left[\frac{p_n\kappa_{1n}}{\rho_n^2(\lambda_n/n^{(1+\gamma)/2})}\right]^{\frac{1}{1+\gamma}}\kappa_{2n}^{1/2}\right)$$

uniformly in α_n. By A4(c) this is also $o_p(1)$ uniformly in α_n.

Let $v_i = \varepsilon_i \frac{1}{n^{1/2}s_n}\left\{\alpha_{1n}'\Sigma_{1n}^{-1}x_{1,i} + \alpha_{2n}'\left(x_{2,i} - \tilde{\Sigma}_n'\Sigma_{1n}^{-1}x_{1,i}\right)\right\}$. We want to apply the Lindeberg-Feller CLT to $\sum_{i=1}^n v_i$. Observe that

$$\begin{aligned}
\text{Var}\left(\sum_{i=1}^n v_i\right) =& \frac{\sigma^2}{ns_n^2}\sum_{i=1}^n\left\{\alpha_{1n}'\Sigma_{1n}^{-1}x_{1,i} + \alpha_{2n}'\left(x_{2,i} - \tilde{\Sigma}_n'\Sigma_{1n}^{-1}x_{1,i}\right)\right\}^2 \\
=& \frac{\sigma^2}{s_n^2}\Big\{\alpha_{1n}'\Sigma_{1n}^{-1}\alpha_{1n} + \alpha_{2n}'\left[\Sigma_{2n} - \tilde{\Sigma}_n'\Sigma_{1n}^{-1}\tilde{\Sigma}_n\right]\alpha_{2n} \\
& + 2\alpha_{1n}'\Sigma_{1n}^{-1}\left(\tilde{\Sigma}_n - \Sigma_{1n}\Sigma_{1n}^{-1}\tilde{\Sigma}_n\right)\alpha_{2n}\Big\} \\
=& \frac{\sigma^2}{s_n^2}\left\{\alpha_{1n}'\Sigma_{1n}^{-1}\alpha_{1n} + \alpha_{2n}'\left[\Sigma_{2n} - \tilde{\Sigma}_n'\Sigma_{1n}^{-1}\tilde{\Sigma}_n\right]\alpha_{2n}\right\} \\
=& 1.
\end{aligned}$$

Consider the term inside the limit in the Lindeberg condition:

$$\begin{aligned}
\sum_{i=1}^n E\left[v_i^2; |v_i| > \epsilon\right] =& \frac{1}{ns_n^2}\sum_{i=1}^n\left\{\alpha_{1n}'\Sigma_{1n}^{-1}x_{1,i} + \alpha_{2n}'\left(x_{2,i} - \tilde{\Sigma}_n'\Sigma_{1n}^{-1}x_{1,i}\right)\right\}^2 \\
& \cdot E\left[\varepsilon_i^2; \left|\varepsilon_i\frac{\alpha_{1n}'\Sigma_{1n}^{-1}x_{1,i} + \alpha_{2n}'\left(x_{2,i} - \tilde{\Sigma}_n'\Sigma_{1n}^{-1}x_{1,i}\right)}{n^{1/2}s_n}\right| > \epsilon\right].
\end{aligned}$$

From the way we defined s_n^2 we have $\frac{1}{ns_n^2}\sum_{i=1}^n \left\{\alpha_{1n}'\Sigma_{1n}^{-1}x_{1,i} + \alpha_{2n}'\left(x_{2,i} - \tilde{\Sigma}_n'\Sigma_{1n}^{-1}x_{1,i}\right)\right\}^2 = \frac{1}{\sigma^2}$. Hence a sufficient condition for the Lindeberg condition to hold is

$$\frac{\max_{i=1,\ldots,n}\left\{\alpha_{1n}'\Sigma_{1n}^{-1}x_{1,i} + \alpha_{2n}'\left(x_{2,i} - \tilde{\Sigma}_n'\Sigma_{1n}^{-1}x_{1,i}\right)\right\}^2}{\sigma^{-2}ns_n^2} \xrightarrow[n\to\infty]{} 0.$$

The above condition is satisfied by satisfying the following two:

$$\frac{\max_{i=1,\ldots,n}\left\{\alpha_{1n}'\Sigma_{1n}^{-1}x_{1,i}\right\}^2}{n\left\{\alpha_{1n}'\Sigma_{1n}^{-1}\alpha_{1n}\right\}} \xrightarrow[n\to\infty]{} 0, \tag{A.7}$$

$$\frac{\max_{i=1,\ldots,n}\left\{\alpha_{2n}'\left(x_{2,i} - \tilde{\Sigma}_n'\Sigma_{1n}^{-1}x_{1,i}\right)\right\}^2}{n\left\{\alpha_{2n}'\left[\Sigma_{2n} - \tilde{\Sigma}_n'\Sigma_{1n}^{-1}\tilde{\Sigma}_n\right]\alpha_{2n}\right\}} \xrightarrow[n\to\infty]{} 0, \tag{A.8}$$

when the respective contrast coefficients α_{1n} and α_{2n} are nonzero. If one of them is zero, the corresponding condition can be dropped, and only the condition for the nonzero contrast needs to be satisfied.

For Equation A.7 we have

$$\begin{aligned}\frac{\max_{i=1,\ldots,n}\left\{\alpha_{1n}'\Sigma_{1n}^{-1}x_{1,i}\right\}^2}{n\left\{\alpha_{1n}'\Sigma_{1n}^{-1}\alpha_{1n}\right\}} &\leq \frac{\left\{\alpha_{1n}'\Sigma_{1n}^{-1}\alpha_{1n}\right\}\max_{i=1,\ldots,n}\left\{x_{1,i}'\Sigma_{1n}^{-1}x_{1,i}\right\}}{n\left\{\alpha_{1n}'\Sigma_{1n}^{-1}\alpha_{1n}\right\}} \\ &= \frac{1}{n}\max_{i=1,\ldots,n} x_{1,i}'\Sigma_{1n}^{-1}x_{1,i} \xrightarrow[n\to\infty]{} 0,\end{aligned}$$

with the last line holding by assumption A3(b).

Let $S_c = \left[\Sigma_{2n} - \tilde{\Sigma}_n' \Sigma_{1n}^{-1} \tilde{\Sigma}_n\right]$. Observe that

$$\begin{aligned}\alpha_{2n}' \left(x_{2,i} - \tilde{\Sigma}_n' \Sigma_{1n}^{-1} x_{1,i}\right) &= \alpha_{2n}' \left(-\tilde{\Sigma}_n' \Sigma_{1n}^{-1}, I_{p_n - k_n}\right) \begin{pmatrix} x_{1,i} \\ x_{2,i} \end{pmatrix} \\ &= \alpha_{2n}' S_c^{1/2} S_c^{-1/2} \left(-\tilde{\Sigma}_n' \Sigma_{1n}^{-1}, I_{p_n - k_n}\right) x_i.\end{aligned}$$

Then by Cauchy-Schwarz inequality

$$\begin{aligned}\frac{\left\{\alpha_{2n}' \left(x_{2,i} - \tilde{\Sigma}_n' \Sigma_{1n}^{-1} x_{1,i}\right)\right\}^2}{n \left\{\alpha_{2n}' S_c \alpha_{2n}\right\}} &\leq \frac{1}{n} x_i' \begin{pmatrix} -\Sigma_{1n}^{-1} \tilde{\Sigma}_n \\ I_{p_n - k_n} \end{pmatrix} S_c^{-1} \left(-\tilde{\Sigma}_n' \Sigma_{1n}^{-1}, I_{p_n - k_n}\right) x_i \\ &= \frac{1}{n} x_i' \begin{pmatrix} \Sigma_{1n}^{-1} \tilde{\Sigma}_n S_c^{-1} \tilde{\Sigma}_n' \Sigma_{1n}^{-1} & -\Sigma_{1n}^{-1} \tilde{\Sigma}_n S_c^{-1} \\ -S_c^{-1} \tilde{\Sigma}_n' \Sigma_{1n}^{-1} & S_c^{-1} \end{pmatrix} x_i \\ &= \frac{1}{n} x_i' \left[\Sigma_n^{-1} - \begin{pmatrix} \Sigma_{1n}^{-1} & 0 \\ 0 & 0 \end{pmatrix}\right] x_i,\end{aligned}$$

where the last line follows from the formula for block matrix inversion by inverting Σ_n via the inverses of Σ_{1n} and its Schur complement.

Therefore Equation A.8 is satisfied if

$$\frac{1}{n} \max_{i=1,\ldots,n} x_i' \Sigma_n^{-1} x_i \xrightarrow[n\to\infty]{} 0,$$

which holds by assumption A3(c).

Hence by the Lindeberg-Feller CLT we have $\sum_{i=1}^n v_i \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1)$. □

Lemma 10 (Bound on the estimator error)**.** *Let* $\gamma \in [0, 1]$. *Suppose assumptions A5, A6(a) and A7 hold. Moreover, suppose that* $\|\beta_{20}\|_\infty = o(\tau_n)$. *Then*

$$\|\hat{\beta}_n - \beta_0\|_\infty = o_p(\tau_n).$$

Proof of Lemma 10. Let $\delta_n = \hat{\beta}_n - \beta_0$ and $N = \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i x_i$. Replicating the first part of the proof of Theorem 8 with a straightforward adjustment in Equation A.1 for A5(b) instead of A2 we get

$$\|\delta_n\| = O_p(\tau_n), \tag{A.9}$$

and so the same applies to $\|\delta_n\|_\infty$. We will now show that $\|\delta_n\| = o_p(\tau_n)$, which will imply the same for $\|\delta_n\|_\infty$.

Fix arbitrary $\Delta > 0$ and $\epsilon > 0$. We want to show that

$$P\left(\|\tau_n^{-1}\delta_n\| > \Delta\right) < \epsilon$$

for all n large enough.

By Equation A.9 we can find $C > 0$ such that

$$P\left(\|\tau_n^{-1}\delta_n\| > C\right) < \frac{\epsilon}{2}$$

for all n large enough. So consider $\delta_{n,C}$ defines as

$$\tau_n^{-1}\delta_{n,C} = \arg\min_{u \in B(0,C)} V_n(u),$$

$$V_n(u) = u'\Sigma_n u - 2\frac{1}{\sqrt{n}\tau_n}u'N + \frac{\lambda_n}{n\tau_n^2}\sum_{j=k_0+1}^{p} \mathrm{Pen}(\beta_{0j} + \tau_n u_j),$$

where $B(0, C)$ is the closed (and hence compact) l_2 ball of radius C around zero. Observe that

$$P\left(\delta_n = \delta_{n,C}\right) > 1 - \frac{\epsilon}{2}$$

for all n large enough.

By assumption A6(a)

$$\tau_n = o(\beta_{0j} + \tau_n u_j)$$

for $j = k_0 + 1, \ldots, k$. So for all n large enough

$$\tau_n^{-1}\delta_{n,C} = \arg\min_{u \in B(0,C)} V_n^*(u),$$

$$V_n^*(u) = u'\Sigma_n u - 2\frac{1}{\sqrt{n}\tau_n}u'N + \frac{\lambda_n}{n\tau_n^2}\sum_{j=k+1}^{p} \mathrm{Pen}(\beta_{0j} + \tau_n u_j),$$

where $V_n^*(\cdot)$ omits the penalty terms for coefficients $j = k_0 + 1, \ldots, k$ since they are a constant function of u on $B(0, C)$ for n large enough.

Let $\text{Pen}(b,t)$ denote the penalty function with the thresholding parameter set to t. That is, in the notation of Equation 1.7 $\text{Pen}(b) = \text{Pen}(b,\tau_n)$. Then, under A7,

$$\begin{aligned} V_n^*(u) &= u'\Sigma_n u - 2\frac{1}{\sqrt{n}\tau_n}u'N + \frac{\lambda_n\tau_n^{1+\gamma}}{n\tau_n^2}\sum_{j=k+1}^{p}\text{Pen}(\tau_n^{-1}\beta_{0j}+u_j,1) \\ &= u'\Sigma_n u - 2\frac{1}{\sqrt{n}\tau_n}u'N + m\sum_{j=k+1}^{p}\text{Pen}(\tau_n^{-1}\beta_{0j}+u_j,1). \end{aligned}$$

Now by $\|\beta_{20}\|_\infty = o(\tau_n)$ and A5(b) $V_n^*(u)$ converges in probability uniformly in $u \in B(0,C)$ to

$$V^*(u) = u'\Sigma u + m\sum_{j=k+1}^{p}\text{Pen}(u_j,1),$$

and hence $\tau_n^{-1}\delta_{n,C} \xrightarrow[n\to\infty]{p} 0$. Therefore for all n large enough

$$P\left(\|\tau_n^{-1}\delta_{n,C}\| > \Delta\right) < \frac{\epsilon}{2}.$$

Therefore for all n large enough

$$P\left(\|\tau_n^{-1}\delta_n\| > \Delta\right) < \epsilon.$$

□

Proof of Lemma 5. Suppose that for some $j = k+1,\ldots,p$ $\hat{\beta}_{n,j} \neq 0$. Without loss of generality let $\hat{\beta}_{n,j} > 0$. Let $N = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i x_i$. Then the objective function has the partial derivative with respect to its j'th argument at $\hat{\beta}_n$:

$$\begin{aligned} \frac{1}{n}\frac{\partial Q_n}{\partial b_j} &= 2\left[\left\{\Sigma_n\left(\hat{\beta}_n - \beta_0\right)\right\}_j - n^{-1/2}N_j + \frac{\lambda_n}{n}\right] \\ &= 2\left[o_p(\tau_n) + o_p(\tau_n) + m\tau_n\right], \end{aligned}$$

where I used Lemma 10, A5(b), $\|\beta_{20}\|_\infty = o(\tau_n)$ and A7 to get the last line. Hence this partial derivative is larger than zero with probability approaching 1, i.e. $\hat{\beta}_n$ is not the minimum of the objective function. Using the fact that $\hat{\beta}_{2n}$ is of bounded dimension we conclude that $\hat{\beta}_{2n} = 0$ with probability approaching 1. □

The following lemma seeks to clarify the use of the phrase "with probability approaching one" when proving convergence in probability and similar results (including deriving bounds on rates of convergence that hold in the o_p sense).

Lemma 11 (Convergence in probability and the "with probability approaching one" argument)**.** *Suppose that for event A we have $P(A) \xrightarrow[n\to\infty]{} 1$ (A holds with probability approaching one). Suppose moreover that for two random variables X_n and Y_n we have $X_n(\omega) = Y_n(\omega)\,\forall\,\omega \in A$. Finally, suppose $f(Y_n) \xrightarrow[n\to\infty]{p} 0$ for a given scalar-valued function f. Then*

$$f(X_n) \xrightarrow[n\to\infty]{p} 0.$$

Proof of Lemma 11. Let $\varepsilon > 0$.

$$\begin{aligned} P\left(|f(X_n)| > \varepsilon\right) &= P\left(|f(X_n)| > \varepsilon \cap A\right) + P\left(|f(X_n)| > \varepsilon \cap \bar{A}\right) \\ &= P\left(|f(Y_n)| > \varepsilon \cap A\right) + P\left(|f(X_n)| > \varepsilon \cap \bar{A}\right) \\ &\leq \underbrace{P\left(|f(Y_n)| > \varepsilon\right)}_{\xrightarrow[n\to\infty]{} 0\,\forall\,\varepsilon>0} + \underbrace{P\left(\bar{A}\right)}_{\xrightarrow[n\to\infty]{} 0}. \end{aligned}$$

Therefore $f(X_n) \xrightarrow[n\to\infty]{p} 0$. □

PROOF OF THEOREM 3. Consider the objective function that the penalized estimator minimizes:

$$Q_n(b) = \sum_{i=1}^{n} \varepsilon_i^2 + (b-\beta_0)' \sum_{i=1}^{n} x_i x_i'(b-\beta_0) - 2\sum_{i=1}^{n} \varepsilon_i x_i'(b-\beta_0) + \lambda_n \sum_{j=k_0+1}^{p} \text{Pen}(\beta_{0j}+b_j-\beta_{0j}).$$

Let $\delta_n = \hat{\beta}_n - \beta_0$ and $N = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i x_i$. Partition δ_n and N according to the partition of β_0. Then δ_n minimizes

$$V_n(u) = u'\Sigma_n u - 2\frac{1}{\sqrt{n}} u'N + \frac{\lambda_n}{n} \sum_{j=k_0+1}^{p} \text{Pen}(\beta_{0j}+u_j).$$

By Lemma 10 $\delta_n = o_p(\tau_n)$.

By Lemma 5 $\hat{\beta}_{2n} = 0$ with probability approaching one for $\gamma = 0$. So with probability approaching 1 $\delta_{2n,SCAD} = -h_n c$, and since $\delta_n = o_p(\tau_n)$ and $\tau_n = o(\min_{k_0+1 \le j \le k} |\beta_{10j}|)$ by A7 the first-order condition for $\delta_{1n,SCAD}$ gives

$$\delta_{1n,SCAD} = \frac{1}{\sqrt{n}} \Sigma_{1n}^{-1} N_1 - \Sigma_{1n}^{-1} \tilde{\Sigma}_n \delta_{2n},$$

and therefore

$$\begin{aligned} h_n^{-1}\left(\alpha_1'\delta_{1n,SCAD}+\alpha_2'\delta_{2n,SCAD}\right) &= \frac{1}{\sqrt{n}h_n}\alpha_1'\Sigma_{1n}^{-1}N_1+\alpha_1'\Sigma_{1n}^{-1}\tilde{\Sigma}_n c-\alpha_2' c, \\ \left(h_n^{-1}\alpha'\delta_{n,SCAD}\right)^2 &= \left(\alpha_1'\Sigma_{1n}^{-1}\tilde{\Sigma}_n-\alpha_2'\right)cc'\left(\tilde{\Sigma}_n'\Sigma_{1n}^{-1}\alpha_1-\alpha_2\right) \\ &\quad +\frac{1}{nh_n^2}\alpha_1'\Sigma_{1n}^{-1}N_1N_1'\Sigma_{1n}^{-1}\alpha_1 \\ &\quad +2\frac{1}{\sqrt{n}h_n}\left(\alpha_1'\Sigma_{1n}^{-1}\tilde{\Sigma}_n-\alpha_2'\right)cN_1'\Sigma_{1n}^{-1}\alpha_1 \\ &= \left(\alpha_1'\Sigma_{1n}^{-1}\tilde{\Sigma}_n-\alpha_2'\right)cc'\left(\tilde{\Sigma}_n'\Sigma_{1n}^{-1}\alpha_1-\alpha_2\right) \\ &\quad +O_p\left(\frac{1}{\sqrt{n}h_n}\right), \end{aligned}$$

with the O_p term uniform in c. Let

$$\alpha_b=\tilde{\Sigma}_n'\Sigma_{1n}^{-1}\alpha_1-\alpha_2.$$

We have

$$\begin{aligned} \sup_{c,\|c\|\le 1}\left(h_n^{-1}\alpha'\delta_{n,SCAD}\right)^2 &= \sup_{c,\|c\|\le 1}\left(\alpha_b'c\right)^2+O_p\left(\frac{1}{\sqrt{n}h_n}\right) && \text{(A.10)} \\ &= \left(\alpha_b'c\right)^2\Big|_{c=\|\alpha_b\|^{-1}\alpha_b}+O_p\left(\frac{1}{\sqrt{n}h_n}\right) && \text{(A.11)} \\ &= \|\alpha_b\|^2+O_p\left(\frac{1}{\sqrt{n}h_n}\right) && \text{(A.12)} \\ &= \|\alpha_b\|^2+o_p\left(\left[\frac{h_n}{\tau_n}\right]^{\frac{1-\gamma}{\gamma}}\right). && \text{(A.13)} \end{aligned}$$

The last line holds by A6.

We now consider $\gamma > 0$. Since V_n is differentiable for $\gamma > 0$, we have

$$\Sigma_n \delta_n - \frac{1}{\sqrt{n}} N + \frac{1}{2} \frac{\lambda_n}{n} \bigtriangledown \text{Pen}(\beta_0 + \delta_n) = 0,$$

where $\bigtriangledown \text{Pen}(b)$ is the gradient of the function $\sum_{j=k_0+1}^{p} \text{Pen}(b_j)$. Since $\|\delta_n\|_\infty = o_p(\tau_n)$ we have, with probability approaching one (here and in the derivations to follow)[1],

$$\begin{cases} \Sigma_{1n}\delta_{1n} + \tilde{\Sigma}_n \delta_{2n} = \frac{1}{\sqrt{n}} N_1; \\ \tilde{\Sigma}_n' \delta_{1n} + \Sigma_{2n}\delta_{2n} = \frac{1}{\sqrt{n}} N_2 - \frac{\lambda_n}{n} \left(\beta_{20} + \delta_{2n}\right)^\gamma ; \end{cases}$$

where the power is understood as the sign-preserving element-wise (Hadamard) power (here and in the derivations to follow).

Let $V = \Sigma_{2n} - \tilde{\Sigma}_n' \Sigma_{1n}^{-1} \tilde{\Sigma}_n$. Rearranging the above expression we get

$$\text{(A.14)} \qquad \begin{cases} \delta_{1n} = \frac{1}{\sqrt{n}} \Sigma_{1n}^{-1} N_1 - \Sigma_{1n}^{-1} \tilde{\Sigma}_n \delta_{2n}; \\ V\delta_{2n} = \frac{1}{\sqrt{n}} \left[N_2 - \tilde{\Sigma}_n' \Sigma_{1n}^{-1} N_1 \right] - \frac{\lambda_n}{n} \left(\beta_{20} + \delta_{2n}\right)^\gamma . \end{cases}$$

Let $W = N_2 - \tilde{\Sigma}_n' \Sigma_{1n}^{-1} N_1$. Rewrite the second part:

$$\text{(A.15)} \qquad V \left(\tau_n^{-1}\beta_{20} + \tau_n^{-1}\delta_{2n}\right) + \frac{\lambda_n}{n\tau_n^{1-\gamma}} \left(\tau_n^{-1}\beta_{20} + \tau_n^{-1}\delta_{2n}\right)^\gamma = \frac{h_n}{\tau_n} Vc + \frac{h_n}{\tau_n} \frac{1}{\sqrt{n} h_n} W.$$

We will first consider the case of $\gamma \in (0, 1)$, and afterwards come back to Equation A.15 to consider the case of $\gamma = 1$.

[1] See Lemma 11 for a formal clarification of this statement. In this context, the equations that follow hold exactly for δ_n on the event $\{\|\delta_n\|_\infty < \tau_n\}$, which happens with probability approaching one as $n \to \infty$. In the interests of clarity, I do not introduce separate notation for a new 'version' of δ_n (Y_n in the statement of Lemma 11) that always (rather than with probability approaching one) satisfies these equations. However, the reader can safely assume that we revert back to the original δ_n at any statement that involves o_p terms.

Since $\|\tau_n^{-1}\beta_{20} + \tau_n^{-1}\delta_{2n}\|_\infty = o_p(1)$, $\frac{\lambda_n}{n\tau_n^{1-\gamma}} = m$ and $\|W\|_\infty = O_p(1)$ we have

$$\|\tau_n^{-1}\beta_{20} + \tau_n^{-1}\delta_{2n}\|_\infty = O_p\left(\left[\frac{h_n}{\tau_n}\right]^{\frac{1}{\gamma}}\right),$$

or

$$\|c + h_n^{-1}\delta_{2n}\|_\infty = O_p\left(\left[\frac{h_n}{\tau_n}\right]^{\frac{1}{\gamma}-1}\right) = o_p(1).$$

So

$$\begin{aligned}
\left(\frac{h_n}{\tau_n}\right)^{\gamma}\left(c + h_n^{-1}\delta_{2n}\right)^{\gamma} &= \frac{h_n}{\tau_n}\frac{1}{m}Vc - \frac{h_n}{\tau_n}\frac{1}{m}V\left(c + h_n^{-1}\delta_{2n}\right) + \frac{h_n}{\tau_n}\frac{1}{\sqrt{n}h_n}\frac{1}{m}W, \\
\left(\frac{h_n}{\tau_n}\right)^{\gamma-1}\left(c + h_n^{-1}\delta_{2n}\right)^{\gamma} &= \frac{1}{m}Vc - \frac{1}{m}V\left(c + h_n^{-1}\delta_{2n}\right) + \frac{1}{\sqrt{n}h_n}\frac{1}{m}W.
\end{aligned}$$

Now since $\sqrt{n}h_n \to \infty$ we have

$$\begin{aligned}
\left(\frac{h_n}{\tau_n}\right)^{\gamma-1}\left(c + h_n^{-1}\delta_{2n}\right)^{\gamma} &= \frac{1}{m}Vc + o_p(1), \\
\left(\frac{h_n}{\tau_n}\right)^{\frac{\gamma-1}{\gamma}}\left(c + h_n^{-1}\delta_{2n}\right) &= \left(\frac{1}{m}Vc\right)^{\frac{1}{\gamma}} + o_p(1), \\
\delta_{2n} &= h_n\left[-c + \left(\frac{h_n}{\tau_n}\right)^{\frac{1-\gamma}{\gamma}}\left(\frac{1}{m}Vc\right)^{\frac{1}{\gamma}} + o_p\left(\left[\frac{h_n}{\tau_n}\right]^{\frac{1-\gamma}{\gamma}}\right)\right],
\end{aligned}$$

where the second line holds by CMT and the o_p term is uniform in c.

Going back to the first part of Equation A.14 and plugging in the above result:

$$\delta_{1n} = \frac{1}{\sqrt{n}}\Sigma_{1n}^{-1}N_1 + h_n\Sigma_{1n}^{-1}\tilde{\Sigma}_n c - h_n\left(\frac{h_n}{\tau_n}\right)^{\frac{1-\gamma}{\gamma}}\Sigma_{1n}^{-1}\tilde{\Sigma}_n\left(\frac{1}{m}Vc\right)^{\frac{1}{\gamma}} + o_p\left(h_n\left[\frac{h_n}{\tau_n}\right]^{\frac{1-\gamma}{\gamma}}\right).$$

Therefore for a given $(\alpha_1', \alpha_2')'$ we have

$$\begin{aligned} h_n^{-1}\left(\alpha_1'\delta_{1n} + \alpha_2'\delta_{2n}\right) &= \alpha_1'\Sigma_{1n}^{-1}\tilde{\Sigma}_n c - \alpha_2' c + \frac{1}{\sqrt{n}h_n}\alpha_1'\Sigma_{1n}^{-1}N_1 \\ &\quad - \left(\frac{h_n}{\tau_n}\right)^{\frac{1-\gamma}{\gamma}} \alpha_1'\Sigma_{1n}^{-1}\tilde{\Sigma}_n \left(\frac{1}{m}Vc\right)^{\frac{1}{\gamma}} + \left(\frac{h_n}{\tau_n}\right)^{\frac{1-\gamma}{\gamma}} \alpha_2' \left(\frac{1}{m}Vc\right)^{\frac{1}{\gamma}} \\ &\quad + o_p\left(\left[\frac{h_n}{\tau_n}\right]^{\frac{1-\gamma}{\gamma}}\right) \\ &= \alpha_b' c + \frac{1}{\sqrt{n}h_n}\alpha_1'\Sigma_{1n}^{-1}N_1 - \left(\frac{h_n}{\tau_n}\right)^{\frac{1-\gamma}{\gamma}} \alpha_b' \left(\frac{1}{m}Vc\right)^{\frac{1}{\gamma}} \\ &\quad + o_p\left(\left[\frac{h_n}{\tau_n}\right]^{\frac{1-\gamma}{\gamma}}\right). \end{aligned}$$

Hence the squared error is

$$\begin{aligned} \left(h_n^{-1}\alpha'\delta_n\right)^2 &= \alpha_b' c c' \alpha_b \\ &\quad + \frac{1}{nh_n^2}\alpha_1'\Sigma_{1n}^{-1}N_1 N_1'\Sigma_{1n}^{-1}\alpha_1 \\ &\quad + 2\frac{1}{\sqrt{n}h_n}\alpha_b' c N_1'\Sigma_{1n}^{-1}\alpha_1 \\ &\quad - 2\left(\frac{h_n}{\tau_n}\right)^{\frac{1-\gamma}{\gamma}} \alpha_b' \left(\frac{1}{m}Vc\right)^{\frac{1}{\gamma}} c'\alpha_b \\ &\quad - 2\frac{1}{\sqrt{n}h_n}\left(\frac{h_n}{\tau_n}\right)^{\frac{1-\gamma}{\gamma}} \alpha_b' \left(\frac{1}{m}Vc\right)^{\frac{1}{\gamma}} N_1'\Sigma_{1n}^{-1}\alpha_1 \\ &\quad + \left(\frac{h_n}{\tau_n}\right)^{2\frac{1-\gamma}{\gamma}} \alpha_b' \left(\frac{1}{m}Vc\right)^{\frac{1}{\gamma}} \left(\left(\frac{1}{m}Vc\right)^{\frac{1}{\gamma}}\right)' \alpha_b \\ &\quad + o_p\left(\left[\frac{h_n}{\tau_n}\right]^{\frac{1-\gamma}{\gamma}}\right), \end{aligned}$$

with the o_p term uniform in c. Note that the last two terms before o_p term can be absorbed into it as well, so

$$\begin{aligned}\left(h_n^{-1}\alpha'\delta_n\right)^2 &= \alpha_b' cc'\alpha_b \\ &\quad + \frac{1}{nh_n^2}\alpha_1'\Sigma_{1n}^{-1}N_1N_1'\Sigma_{1n}^{-1}\alpha_1 \\ &\quad + 2\frac{1}{\sqrt{n}h_n}\alpha_b' cN_1'\Sigma_{1n}^{-1}\alpha_1 \\ &\quad - 2\left(\frac{h_n}{\tau_n}\right)^{\frac{1-\gamma}{\gamma}}\alpha_b'\left(\frac{1}{m}Vc\right)^{\frac{1}{\gamma}}c'\alpha_b \\ &\quad + o_p\left(\left[\frac{h_n}{\tau_n}\right]^{\frac{1-\gamma}{\gamma}}\right),\end{aligned}$$

with the o_p term uniform in c. Now, with $\left(\frac{1}{\sqrt{n}\tau_n}\right)^{\gamma} = o\left(\frac{h_n}{\tau_n}\right)$ the second and the third terms on the right-hand side can also be absorbed into the o_p term:

$$\begin{aligned}\left(h_n^{-1}\alpha'\delta_n\right)^2 &= \alpha_b' cc'\alpha_b \\ &\quad - 2\left(\frac{h_n}{\tau_n}\right)^{\frac{1-\gamma}{\gamma}}\alpha_b'\left(\frac{1}{m}Vc\right)^{\frac{1}{\gamma}}c'\alpha_b \\ &\quad + o_p\left(\left[\frac{h_n}{\tau_n}\right]^{\frac{1-\gamma}{\gamma}}\right),\end{aligned}$$

with the o_p term uniform in c.

Now, taking supremum over $\|c\| \leq 1$ we have

$$\begin{aligned}\sup_{c,\|c\|\leq 1}\left(h_n^{-1}\alpha'\delta_n\right)^2 &\leq \sup_{c,\|c\|\leq 1}\left[\alpha_b' cc'\alpha_b - 2\left(\frac{h_n}{\tau_n}\right)^{\frac{1-\gamma}{\gamma}}\alpha_b'\left(\frac{1}{m}Vc\right)^{\frac{1}{\gamma}}c'\alpha_b\right] \\ &\quad + o_p\left(\left[\frac{h_n}{\tau_n}\right]^{\frac{1-\gamma}{\gamma}}\right).\end{aligned}$$

Let

$$S_n(c) = \alpha_b' c c' \alpha_b - 2\left(\frac{h_n}{\tau_n}\right)^{\frac{1-\gamma}{\gamma}} \alpha_b' \left(\frac{1}{m} V c\right)^{\frac{1}{\gamma}} c' \alpha_b.$$

Observe that S_n is even: $S_n(c) = S_n(-c)$. Since S_n is a continuous function on a compact support, it has a maximum and a maximizer, call it $\hat{c}_n$. Since S_n is even, $-\hat{c}_n$ is also a maximizer. We will restrict attention to maximizers belonging to some half-ball $S \subset \{c : \|c\| \leq 1\}$ such that $(c \in S \Leftrightarrow -c \notin S)$; let $\bar{S}$ be the closure of S. Moreover we will require that $\|\tilde{\Sigma}'\Sigma_1^{-1}\alpha_1 - \alpha_2\|^{-1}\left(\tilde{\Sigma}'\Sigma_1^{-1}\alpha_1 - \alpha_2\right)$ is in the interior of S relative to $\{c : \|c\| \leq 1\}$ (feasible since $\tilde{\Sigma}'\Sigma_1^{-1}\alpha_1 - \alpha_2 \neq 0$), e.g. $S \subset \bar{S} = \{c : \|c\| \leq 1, c'\alpha_b \geq 0\}$. So we will let $\hat{c}_n \in S$ without loss of generality. Then, since $\|\hat{c}_n\| \leq 1$,

$$\begin{aligned}
\sup_{c,\|c\|\leq 1} S_n(c) &= \alpha_b' \hat{c}_n \hat{c}_n' \alpha_b - 2\left(\frac{h_n}{\tau_n}\right)^{\frac{1-\gamma}{\gamma}} \alpha_b' \left(\frac{1}{m} V \hat{c}_n\right)^{\frac{1}{\gamma}} \hat{c}_n' \alpha_b \\
&\leq \|\alpha_b\|^2 - 2\left(\frac{h_n}{\tau_n}\right)^{\frac{1-\gamma}{\gamma}} \alpha_b' \left(\frac{1}{m} V \hat{c}_n\right)^{\frac{1}{\gamma}} \hat{c}_n' \alpha_b.
\end{aligned}$$

$S_n(c)$ converges uniformly to

$$S(c) = \left(\alpha_1'\Sigma_1^{-1}\tilde{\Sigma} - \alpha_2'\right) c c' \left(\tilde{\Sigma}'\Sigma_1^{-1}\alpha_1 - \alpha_2\right),$$

so $S(c)$ has a unique maximizer $\bar{c} \in \bar{S}$ and it is $\bar{c} = \|\tilde{\Sigma}'\Sigma_1^{-1}\alpha_1 - \alpha_2\|^{-1}\left(\tilde{\Sigma}'\Sigma_1^{-1}\alpha_1 - \alpha_2\right)$; it is in the interior of $\bar{S}$ relative to $\{c : \|c\| \leq 1\}$ by the choice of $\bar{S}$. Hence by the usual asymptotic theory $\hat{c}_n \xrightarrow[n\to\infty]{} \bar{c}$. Hence

$$\alpha_b' \left(\frac{1}{m} V \hat{c}_n\right)^{\frac{1}{\gamma}} \hat{c}_n' \alpha_b \xrightarrow[n\to\infty]{} \bar{\alpha}_b' \left(\frac{1}{m} \bar{V} \bar{\alpha}_b\right)^{\frac{1}{\gamma}} \|\bar{\alpha}_b\|^{-\frac{1-\gamma}{\gamma}},$$

where $\bar{V} = \lim_{n\to\infty} V$ and $\bar{\alpha}_b = \tilde{\Sigma}'\Sigma_1^{-1}\alpha_1 - \alpha_2$.

Finally, putting all of the results together we get

$$\sup_{c,\|c\|\leq 1}\left(h_n^{-1}\alpha'\delta_{n,SCAD}\right)^2-\sup_{c,\|c\|\leq 1}\left(h_n^{-1}\alpha'\delta_n\right)^2 \geq 2\left(\frac{h_n}{\tau_n}\right)^{\frac{1-\gamma}{\gamma}}\bar{\alpha}_b'\left(\frac{1}{m}\bar{V}\bar{\alpha}_b\right)^{\frac{1}{\gamma}}\|\bar{\alpha}_b\|^{-\frac{1-\gamma}{\gamma}} + o_p\left(\left[\frac{h_n}{\tau_n}\right]^{\frac{1-\gamma}{\gamma}}\right),$$

or

$$\left(\frac{h_n}{\tau_n}\right)^{-\frac{1-\gamma}{\gamma}}h_n^{-2}\left[\sup_{c,\|c\|\leq 1}\left(\alpha'\delta_{n,SCAD}\right)^2-\sup_{c,\|c\|\leq 1}\left(\alpha'\delta_n\right)^2\right] \geq 2\bar{\alpha}_b'\left(\frac{1}{m}\bar{V}\bar{\alpha}_b\right)^{\frac{1}{\gamma}}\|\bar{\alpha}_b\|^{-\frac{1-\gamma}{\gamma}} + o_p\left(1\right).$$

Since the main term on the right-hand side is a continuous function of γ for $\gamma > 0$ we will show that it is strictly positive for the limiting case $\gamma = 1$, which will imply that the same holds for all $\gamma < 1$ high enough.

Observe that $\bar{V} = \lim_{\to\infty}\left[\Sigma_{2n} - \tilde{\Sigma}_n'\Sigma_{1n}^{-1}\tilde{\Sigma}_n\right] = \Sigma_2 - \tilde{\Sigma}'\Sigma_1^{-1}\tilde{\Sigma}$ is positive definite since it is the Schur complement of Σ_1 in Σ, and Σ is positive definite. Therefore, for $\gamma = 1$

$$2\bar{\alpha}_b'\left(\frac{1}{m}\bar{V}\bar{\alpha}_b\right)^{\frac{1}{1}}\|\bar{\alpha}_b\|^{-\frac{1-1}{1}} = \frac{2}{m}\underbrace{\bar{\alpha}_b'\bar{V}\bar{\alpha}_b}_{\text{Quadratic form with } \bar{V}>0} > 0.$$

We have proved the theorem for $\gamma \in (0, 1)$. We now come back to Equation A.15 to complete the proof for $\gamma = 1$.

Rewrite Equation A.15 for $\gamma = 1$ considering that $\frac{\lambda_n}{n\tau_n^{1-\gamma}} = m$:

$$\begin{aligned} V\left(\tau_n^{-1}\beta_{20} + \tau_n^{-1}\delta_{2n}\right) + m\left(\tau_n^{-1}\beta_{20} + \tau_n^{-1}\delta_{2n}\right)^1 &= \frac{h_n}{\tau_n}Vc + \frac{h_n}{\tau_n}\frac{1}{\sqrt{n}h_n}W, \\ (V + mI)\left[h_n^{-1}\beta_{20} + h_n^{-1}\delta_{2n}\right] &= Vc + \frac{1}{\sqrt{n}h_n}W, \\ c + h_n^{-1}\delta_{2n} &= (V + mI)^{-1}Vc + (V + mI)^{-1}\frac{1}{\sqrt{n}h_n}W, \\ \delta_{2n} &= h_n\left\{\left[-I + (V + mI)^{-1}V\right]c + o_p(1)\right\}, \end{aligned}$$

with the o_p term uniform in c.

So

$$\delta_{2n} = h_n\left[-\left(\frac{1}{m}V + I\right)^{-1}c + o_p(1)\right],$$

with the o_p term uniform in c.

Going back to Equation A.14 we have

$$\delta_{1n} = \frac{1}{\sqrt{n}}\Sigma_{1n}^{-1}N_1 - \Sigma_{1n}^{-1}\tilde{\Sigma}_n\delta_{2n},$$

and therefore

$$\begin{aligned} h_n^{-1}\left(\alpha_1'\delta_{1n} + \alpha_2'\delta_{2n}\right) &= \alpha_1'\Sigma_{1n}^{-1}\tilde{\Sigma}_n\left(\frac{1}{m}V + I\right)^{-1}c - \alpha_2'\left(\frac{1}{m}V + I\right)^{-1}c + o_p(1), \\ h_n^{-1}\alpha'\delta_n &= \alpha_b'\left(\frac{1}{m}V + I\right)^{-1}c + o_p(1), \end{aligned}$$

with the o_p term uniform in c.

Now

$$\begin{aligned}
\sup_{c,\|c\|\leq 1}\left(h_n^{-1}\alpha'\delta_n\right)^2 &= \sup_{c,\|c\|\leq 1}\left[\alpha_b'\left(\frac{1}{m}V+I\right)^{-1}c\right]^2+o_p(1)\\
&= \left[\alpha_b'\left(\frac{1}{m}V+I\right)^{-1}c\right]^2\Bigg|_{c=\left\|\left(\frac{1}{m}V+I\right)^{-1}\alpha_b\right\|^{-1}\left(\frac{1}{m}V+I\right)^{-1}\alpha_b}+o_p(1)\\
&= \left\|\left(\frac{1}{m}V+I\right)^{-1}\alpha_b\right\|^2+o_p(1).
\end{aligned}$$

Combining the above with Equation A.12 we have

$$\begin{aligned}
&\sup_{c,\|c\|\leq 1}\left(h_n^{-1}\alpha'\delta_{n,SCAD}\right)^2-\sup_{c,\|c\|\leq 1}\left(h_n^{-1}\alpha'\delta_n\right)^2\\
&=\|\alpha_b\|^2-\left\|\left(\frac{1}{m}V+I\right)^{-1}\alpha_b\right\|^2+o_p(1)\\
&=\alpha_b'\left[I-\left(\frac{1}{m}V+I\right)^{-2}\right]\alpha_b+o_p(1)\\
&=\alpha_b'\left(\frac{1}{m}V+I\right)^{-1}\left[\frac{1}{m^2}V^2+2\frac{1}{m}V\right]\left(\frac{1}{m}V+I\right)^{-1}\alpha_b+o_p(1).
\end{aligned}$$

We can replace α_b and V with their respective limits $\bar{\alpha}_b$ and $\bar{V}$:

$$\begin{aligned}
&\sup_{c,\|c\|\leq 1}\left(h_n^{-1}\alpha'\delta_{n,SCAD}\right)^2-\sup_{c,\|c\|\leq 1}\left(h_n^{-1}\alpha'\delta_n\right)^2\\
&=\bar{\alpha}_b'\left(\frac{1}{m}\bar{V}+I\right)^{-1}\left[\frac{1}{m^2}\bar{V}^2+2\frac{1}{m}\bar{V}\right]\left(\frac{1}{m}\bar{V}+I\right)^{-1}\bar{\alpha}_b+o_p(1),
\end{aligned}$$

where the leading term is greater than zero since $\bar{V}>0$ (as shown in the $\gamma\in(0,1)$ part of this proof) and $\bar{\alpha}_b\neq 0$ by assumption.

□

CHAPTER B

Appendix to Chapter 2

B.1. Proofs

PROOF OF THEOREM 5. Consider the objective function that the penalized estimator minimizes:

$$Q_n(b) = \sum_{i=1}^{n} \varepsilon_i^2 + (b-\beta_0)' \sum_{i=1}^{n} x_i x_i' (b-\beta_0) - 2\sum_{i=1}^{n} \varepsilon_i x_i' (b-\beta_0) + \lambda_n \sum_{j=k_0+1}^{p} \text{Pen}(\beta_{0j} + b_j - \beta_{0j}).$$

Let $\delta_n = \hat{\beta}_n - \beta_0$ and $N = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varepsilon_i x_i$. Then δ_n minimizes

$$V_n(u) = u'\Sigma_n u - 2\frac{1}{\sqrt{n}} u'N + \frac{\lambda_n}{n} \sum_{j=k_0+1}^{p} \text{Pen}(\beta_{0j} + u_j). \tag{B.1}$$

Since this function is differentiable, we have

$$\Sigma_n \delta_n - \frac{1}{\sqrt{n}} N + \frac{1}{2} \frac{\lambda_n \tau_n^{\gamma}}{n} \bigtriangledown \text{Pen}(\tau_n^{-1}\beta_0 + \tau_n^{-1}\delta_n; 1) = 0,$$

or

$$\frac{n}{\lambda_n \tau_n^{\gamma}} \delta_n = -\frac{1}{2} \Sigma_n^{-1} \bigtriangledown \text{Pen}(\tau_n^{-1}\beta_0 + \tau_n^{-1}\delta_n; 1) + \frac{\sqrt{n}}{\lambda_n \tau_n^{\gamma}} \Sigma_n^{-1} N. \tag{B.2}$$

Using the notation $\tau_n = n^{-1/2} g_n$ and $\lambda_n = n^{\frac{1+\gamma}{2}} f_n$ with $g_n \to \infty$, $f_n \to \infty$, note that $\frac{\sqrt{n}}{\lambda_n \tau_n^{\gamma}} = \frac{1}{f_n g_n^{\gamma}} \to 0$, $\bigtriangledown \text{Pen}(\cdot; 1)$ is bounded and $N = O_p(1)$ by Assumption A10 and does

not depend on coefficients, so

$$\begin{aligned}\delta_n &= O_p\left(\frac{\lambda_n \tau_n^{\gamma}}{n}\right) \\ &= O_p\left(\frac{f_n g_n^{\gamma}}{\sqrt{n}}\right)\end{aligned}$$

uniformly in β_0.

This provides us with a bound on convergence rate of the estimator. For more precise results we need to consider sequences $\tau_n^{-1}\beta_0$ and $\tau_n^{-1}\delta_n$. We will consider semi-local sequences $\beta_0 = \tau_n \alpha_0$ so as to keep the first component meaningful. From the result above we have

$$\begin{aligned}\tau_n^{-1}\delta_n &= O_p\left(\frac{\lambda_n}{n\tau_n^{1-\gamma}}\right) \\ &= O_p\left(\frac{f_n}{g_n^{1-\gamma}}\right).\end{aligned}$$

Three distinct cases are possible here, depending on whether the ratio inside the O_p converges to zero, is equal to a constant, or diverges to infinity. In this theorem I focus on the second case.

Case 1: if $f_n = o\left(g_n^{1-\gamma}\right)$, then $\tau_n^{-1}\delta_n = o_p(1)$ and, from Equation B.2,

$$\frac{n}{\lambda_n \tau_n^{\gamma}}\delta_n \xrightarrow[n\to\infty]{p} -\frac{1}{2}\Sigma^{-1} \bigtriangledown \mathrm{Pen}(\alpha_0; 1).$$

Case 2: if $f_n = m g_n^{1-\gamma}$ (or $\frac{\lambda_n}{n} = m\tau_n^{1-\gamma}$), then, rewriting Equation B.1, we have

$$\tau_n^{-1}\delta_n = \arg\min_v v'\Sigma_n v - 2\frac{1}{\sqrt{n}\tau_n}v'N + m\sum_{j=k_0+1}^{p}\mathrm{Pen}(\tau_n^{-1}\beta_{0j} + v_j; 1).$$

So, by the usual asymptotic argument,

$$\tau_n^{-1}\delta_n \xrightarrow[n\to\infty]{p} \underbrace{\arg\min_v v'\Sigma v + m\sum_{j=k_0+1}^{p} \mathrm{Pen}(\alpha_{0j}+v_j;1)}_{\equiv b}.$$

Let $D_n = \tau_n^{-1}\delta_n - b$. Returning to the first-order condition for δ_n we have

$$\Sigma_n b + \Sigma_n D_n + \frac{m}{2} \bigtriangledown \mathrm{Pen}(\alpha_0 + b + D_n; 1) = \frac{1}{\sqrt{n}\tau_n} N.$$

Let D_{1n} capture the elements of D_n such that either the corresponding coefficient is not penalized, or $\alpha_{0j} + b_j \neq 0$. Let D_{2n} contain the rest. Then since $D_n \xrightarrow[n\to\infty]{p} 0$, we can expand the gradient of the penalty according to Taylor expansion for elements corresponding to D_{1n} and as a penalty at D_{2n} for the rest:

$$\bigtriangledown \mathrm{Pen}(\alpha_0 + b + D_n; 1) = \bigtriangledown \mathrm{Pen}(\alpha_0 + b; 1) + \begin{pmatrix} \bigtriangledown^2 \mathrm{Pen}(\alpha_{10} + b_1; 1) D_{1n} + o(D_{1n}) \\ 2D_{2n}^{\gamma} \end{pmatrix}$$

with probability approaching 1.

So

$$(\Sigma_n - \Sigma)\, b + \left[\underbrace{\Sigma b + \frac{m}{2} \bigtriangledown \mathrm{Pen}(\alpha_0 + b; 1)}_{=0 \text{ by definition of b}}\right] + \Sigma_n D_n + \frac{m}{2} \begin{pmatrix} \bigtriangledown^2 \mathrm{Pen}(\alpha_{10} + b_1; 1) D_{1n} + o(D_{1n}) \\ 2D_{2n}^{\gamma} \end{pmatrix} = \frac{1}{\sqrt{n}\tau_n} N.$$

Splitting Σ_n into Σ_{1n}, Σ_{2n} and $\tilde{\Sigma}_n$ according to the split into D_{1n} and D_{2n}, we have

$$\begin{cases} \Sigma_{1n} D_{1n} + \tilde{\Sigma}_n D_{2n} + \frac{m}{2} \bigtriangledown^2 \operatorname{Pen}(\alpha_{10} + b_1; 1) D_{1n} + o(D_{1n}) = \frac{1}{\sqrt{n}\tau_n} N_1 - [(\Sigma_n - \Sigma)b]_1 ; \\ \tilde{\Sigma}_n' D_{1n} + \Sigma_{2n} D_{2n} + m D_{2n}^{\gamma} = \frac{1}{\sqrt{n}\tau_n} N_2 - [(\Sigma_n - \Sigma)b]_2 . \end{cases}$$

Continuing with the first part of the FOC, and utilizing Assumption A11(b),

$$\underbrace{\left[\Sigma_{1n} + \frac{m}{2} \bigtriangledown^2 \operatorname{Pen}(\alpha_{10} + b_1; 1)\right]}_{\equiv V_1} D_{1n} + o(D_{1n}) = \frac{1}{\sqrt{n}\tau_n} N_1 - \tilde{\Sigma}_n D_{2n} + o\left(\frac{1}{\sqrt{n}\tau_n}\right),$$

$$D_{1n} = \frac{1}{\sqrt{n}\tau_n} V_1^{-1} N_1 - V_1^{-1} \tilde{\Sigma}_n D_{2n} + o_p\left(\frac{1}{\sqrt{n}\tau_n} N_1 + \tilde{\Sigma}_n D_{2n}\right),$$

where V_1 is invertible for n large enough as a corollary of assumptions A11(a) and A12.

Plugging the above result into the second part of the FOC we have

$$\left[\Sigma_{2n} - \tilde{\Sigma}_n' V_1^{-1} \tilde{\Sigma}_n\right] D_{2n} + o_p(D_{2n}) + m D_{2n}^{\gamma} = \frac{1}{\sqrt{n}\tau_n} \left[N_2 - \tilde{\Sigma}_n' V_1^{-1} N_1\right] - o_p\left(\frac{1}{\sqrt{n}\tau_n}\right).$$

It follows from assumption A12 that the first term on the left-hand side is $O(D_{2n})$, so

$$D_{2n}^{\gamma} = \frac{1}{\sqrt{n}\tau_n} \frac{1}{m} \left[N_2 - \tilde{\Sigma}_n' V_1^{-1} N_1\right] + o_p\left(\frac{1}{\sqrt{n}\tau_n}\right).$$

Returning to D_{1n} we have

$$D_{1n} = \frac{1}{\sqrt{n}\tau_n} V_1^{-1} N_1 + o_p\left(\frac{1}{\sqrt{n}\tau_n}\right),$$

and putting the two together

$$\sqrt{n}\tau_n \begin{pmatrix} D_{1n} \\ D_{2n}^{\gamma} \end{pmatrix} = \begin{pmatrix} V_1^{-1} & 0 \\ -\frac{1}{m}\tilde{\Sigma}_n' V_1^{-1} & \frac{1}{m} I \end{pmatrix} N + o_p(1).$$

The conclusion of the theorem follows from A10.

□

Proof of Lemma 7. Writing out the first-order condition (see Equation B.2 in the proof of Theorem 5) we have

$$\sqrt{n}\left(\hat{\beta}_n - \beta_0 + \underbrace{\frac{1}{2}\frac{\lambda_n}{n}\Sigma_n^{-1} \bigtriangledown \text{Pen}\left(\hat{\beta}_n; \tau_n\right)}_{=-\tau_n \hat{b}_n}\right) = \Sigma_n^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\varepsilon_i x_i,$$

with the result following from assumptions A10 and A11(a). □

Proof of Corollary 2. We will verify the assumptions of Theorem 2.

Assumptions A2(a) and A2(b) are satisfied by assumption A13(e).

Assumption A3(a) is satisfied by assumption A13(b).

Assumption A3(c) is satisfied by assumptions A13(b) and A13(c): observe that

$$x_i'\Sigma_n^{-1}x_i \leq \rho_n^{-1}\|x_i\|^2.$$

□

I will use Lemma 4.1 and a modification of Lemma 4.2 of Chatterjee and Lahiri (2010) for obtaining the asymptotic distribution and later for ensuring bootstrap consistency.

Lemma 12 (Lemma 4.1 in Chatterjee and Lahiri (2010)). *Suppose that Σ_n converges to a positive definite limit Σ, $\frac{1}{n}\sum_{i=1}^n \|x_i\|^3 = O(1)$ and the errors ε_i are i.i.d. with mean 0 and variance σ^2. Then*

$$\left\| \sum_{i=1}^n \varepsilon_i x_i \right\| = o(n^{1/2}\log n), a.s.$$

Lemma 13 (Modification of Lemma 4.2 in Chatterjee and Lahiri (2010)). *Suppose that Σ_n converges to a positive definite limit Σ, $\frac{1}{n}\sum_{i=1}^n \|x_i\|^3 = O(1)$ and the errors ε_i are i.i.d. with mean 0 and variance $\sigma^2 \in (0,\infty)$. Suppose moreover that $\lambda_n \tau_n^\gamma = O(n^{1/2}\log n)$. Then*

$$\left\| \sqrt{n}\left(\hat{\beta}_n - \beta_0\right) \right\| = O(\log n), a.s.$$

Proof of Lemma 13. This proof follows the arguments in the proof of Lemma 4.2 in Chatterjee and Lahiri (2010) and those in the proof of Theorem 1 in Chapter 1.

Let $T_n = \sqrt{n}\left(\hat{\beta}_n - \beta_0\right)$ and $N = \frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i x_i$. Then T_n minimizes

$$V_n(u) = u'\Sigma_n u - 2u'N + \lambda_n \sum_{j=k_0+1}^{p} \left[\mathrm{Pen}(\beta_{0j} + n^{-1/2}u_j) - \mathrm{Pen}(\beta_{0j})\right].$$

Let ρ_n be the smallest eigenvalue of Σ_n. Using the fact that the largest derivative of the penalty function is $2\tau_n^\gamma$ we have

$$\begin{aligned} V_n(u) &\geq \rho_n\|u\|^2 - 2\|u\|\|N\| - 2pn^{-1/2}\lambda_n\tau_n^\gamma\|u\| \\ &= \|u\|\left(\rho_n\|u\| - 2\|N\| - 2pn^{-1/2}\lambda_n\tau_n^\gamma\right) \equiv V_{1,n}(u). \end{aligned}$$

Now for $\|u\| \geq C\log n$ with C large enough we have

$$\rho_n\|u\| > 2\|N\| + 2pn^{-1/2}\lambda_n\tau_n^\gamma, a.s.$$

by Lemma 12 and assumption on λ_n, τ_n. Therefore $V_{1,n}(u) > 0$ and hence $V_n(u) > 0$ with probability 1 on the set $\{u : \|u\| \geq C \log n\}$. Since $V_n(0) = 0$, the minimizer of V_n lies in the set $\{u : \|u\| < C \log n\}$ with probability 1. □

Lemma 13 is of interest beyond its use in the proof of modified bootstrap consistency. In particular, note that the restrictions on tuning parameters admit values required for Theorem 2 in Chapter 1, and this lemma does not impose any conditions on coefficients (in particular, does not exclude 'intermediate' values). So, with a careful choice of λ_n and τ_n we can achieve oracle efficiency (when intermediate coefficients are excluded) while maintaining the same bound on almost sure convergence ($n^{-1/2} \log n$, allowing for any coefficients) as the OLS and methods that do not provide oracle efficiency. It is curios because we know from Leeb and Pötscher (2008) and relater works that we cannot estimate regression coefficients at root-n rate uniformly if we use a model-selection-consistent method like SCAD. It seems reasonable that the method proposed here would suffer the same fate, and this lemma provides a bound on how bad this slowdown in convergence is.

Corollary 3. *Suppose conditions of Lemma 13 hold. Moreover, suppose that all coefficients are fixed, and that $\tau_n = o(1)$. Then*

$$\left\| \sqrt{n} \left(\tilde{\beta}_n - \beta_0 \right) \right\| = O(\log n), a.s.$$

Proof of Corollary 3. Observe that with probability 1 there is $N > 0$ such that for $n > N$ only estimates of coefficients that are equal to zero may be thresholded. Therefore

$$\left| \sqrt{n} \left(\tilde{\beta}_n - \beta_0 \right) \right| \leq \left| \sqrt{n} \left(\hat{\beta}_n - \beta_0 \right) \right|$$

elementwise (a.s. for n large enough), and the conclusion of the corollary follows by Lemma 13. □

Lemma 14 (Modification of Lemma 4.3 in Chatterjee and Lahiri (2010)). *Let* $s_n^2 = \frac{1}{n}\sum_{i=1}^n (r_i - \bar{r}_n)^2$ *and* $\hat{\mu}_{3,n} = \frac{1}{n}\sum_{i=1}^n |r_i - \bar{r}_n|^3$. *Assume that* $\frac{1}{n}\sum_{i=1}^n \|x_i\|^2 = O(1)$ *and the errors* ε_i *are i.i.d. with mean* 0 *and variance* $\sigma^2 \in (0, \infty)$. *Moreover, assume that the conditions of Corollary 3 hold. Then*

$$\left|s_n^2 - \sigma^2\right| + n^{-1/2}\hat{\mu}_{3,n} \xrightarrow[n\to\infty]{a.s.} 0.$$

Proof of Lemma 14 follows the one in Chatterjee and Lahiri (2010) with the only difference that Corollary 3 for the thresholded estimator supplants the original Lemma 4.2 in the argument.

Lemma 15 (Lemma 4.4 in Chatterjee and Lahiri (2010)). *For each* $n \geq 1$, *let* $\{\eta_{i,n}\}_{i=1}^n$ *be a collection of random variables on* $(\Omega, \mathcal{F}, P)$ *such that given* $\mathcal{E}$ *(a sub-sigma-algebra of* $\mathcal{F}$*),* $\{\eta_{i,n}\}_{i=1}^n$ *are i.i.d. with* $E_*(\eta_{i,n}) = 0$, *and* $|E_*(\eta_{i,n})^2 - t^2| + n^{-1/2}E_*|\eta_{i,n}|^3 \to 0$ *as* $n \to \infty$ *with probability* 1 *for some* $t \in (0, \infty)$ *(where* $E_*(\cdot) = E(\cdot|\mathcal{E})$*). Also, suppose that* Σ_n *converges to a positive definite limit* Σ *and that* $\frac{1}{n}\sum_{i=1}^n \|x_i\|^3 = O(1)$ *as* $n \to \infty$. *Then*

$$\mathcal{L}\left(n^{-1/2}\sum_{i=1}^n x_i\eta_{i,n}|\mathcal{E}\right) \xrightarrow[n\to\infty]{d} \mathcal{N}\left(0, t^2\Sigma\right), a.s.,$$

where $\mathcal{L}(\cdot|\mathcal{E})$ *denotes the conditional distribution given* $\mathcal{E}$.

Proof of Theorem 7. The proof follows that of Theorem 1 of Chatterjee and Lahiri (2011) with adjustments for a different penalty function. It relies on Lemmas

4.2 - 4.4 from Chatterjee and Lahiri (2010) applying to the residuals from the modified estimator $\tilde{\beta}_n$. Lemma 4.2 holds under the conditions of, and as a corollary of, modified Lemma 4.2 for $\hat{\beta}_n$ described above. A modification of Lemma 4.3 can be established provided 4.2 holds, and resampled residuals ε_i^* satisfy the conditions of Lemma 4.4 given Lemma 4.3.

Let $(\Omega, \mathcal{F}, P)$ be the underlying probability space. Let $N^* = \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i^* x_i$. Then T_n^* minimizes

$$V_n^*(u) = u'\Sigma_n u - 2u'N^* + \lambda_n \sum_{j=k_0+1}^{p} \left[\text{Pen}(\tilde{\beta}_{n,j} + n^{-1/2}u_j) - \text{Pen}(\tilde{\beta}_{n,j}) \right].$$

Note that Lemma 15 applies to resampled residuals ε_i^* due to Lemma 14. Therefore

$$\mathcal{L}\left(N^* | \mathcal{E}\right) \xrightarrow[n\to\infty]{d} \mathcal{N}\left(0, \sigma^2\Sigma\right), a.s., \tag{B.3}$$

where $\mathcal{E}$ is the sigma-algebra generated by $\{\varepsilon_i\}_{i=1}^n$.

Let $A \in \mathcal{F}$ be a set of probability 1 such that for every $\omega \in A$

$$\left\| \sqrt{n}\left(\tilde{\beta}_n - \beta_0\right) \right\| = O(\log n)$$

(by Corollary 3) and

$$\mathcal{L}\left(N^* | \mathcal{E}\right)(\omega) \xrightarrow[n\to\infty]{d} \mathcal{N}\left(0, \sigma^2\Sigma\right).$$

Fix $\omega \in A$. There exists $N > 0$ such that for all $n > N$

$$\tilde{\beta}_{n,j} = \hat{\beta}_{n,j}, \forall\, j : 1 \le j \le k$$

and

$$\tilde{\beta}_{n,j} = 0, \forall\, j : k+1 \leq j \leq p.$$

Therefore for all $n > N$ we have

$$V_n^*(u) = u'\Sigma_n u - 2u'N^* + \lambda_n \sum_{j=k_0+1}^{k} \left[\text{Pen}(\tilde{\beta}_{n,j} + n^{-1/2}u_j) - \text{Pen}(\tilde{\beta}_{n,j})\right]$$
$$+ \lambda_n \sum_{j=k+1}^{p} \text{Pen}(n^{-1/2}u_j).$$

Similarly to Lemma 13 we can establish that

$$\left\|\sqrt{n}\left(\beta_n^* - \tilde{\beta}_n\right)\right\| = O(\log n), a.s.$$

Restricting attention to $B \subset A$ such that $P(B) = 1$ and $\left\|\sqrt{n}\left(\beta_n^* - \tilde{\beta}_n\right)\right\| = O(\log n)$ for every $\omega \in B$, and fixing ω, we have that for all n large enough $\beta_{n,j}^*$ is in the outer area of the penalty for $k_0 + 1 \leq j \leq k$ and in the inner area of the penalty for $k+1 \leq j \leq p$, so T_n^* minimizes

$$V_n^{**}(u) = u'\Sigma_n u - 2u'N^* + \frac{2}{1+\gamma} f_n \sum_{j=k+1}^{p} |u_j|^{1+\gamma}. \tag{B.4}$$

Here we will consider the cases under assumptions A14(a) and A14(b) separately; this will help highlight that while estimator behavior under the two choices of tuning parameters is different, modified bootstrap inference is valid under both of them.

First consider the case under A14(b). We have

$$V_n^{**}(u) = u'\Sigma_n u - 2u'N^* + \frac{2}{1+\gamma} l \sum_{j=k+1}^{p} |u_j|^{1+\gamma}.$$

Following the argument in Theorem 1 of Chatterjee and Lahiri (2011) and those in Knight and Fu (2000), we can establish that

$$\mathcal{L}\left(V_n^{**}(\cdot)|\mathcal{E}\right)(\omega) \xrightarrow[n\to\infty]{d} \mathcal{L}\left(V_F(\cdot)\right)$$

on the space of all functions on $\mathbb{R}^p$ that are uniformly bounded on compact subsets of $\mathbb{R}^p$ (where V_F is defined in Theorem 6). Therefore

$$\mathcal{L}\left(T_n^*|\mathcal{E}\right)(\omega) \xrightarrow[n\to\infty]{d} \mathcal{L}\left(T_\infty\right),$$

where T_∞ is distributed as in Theorem 6. Since this holds for all $\omega \in B$ and $P(B) = 1$ the conclusion of the theorem follows.

Now consider the case under A14(a). Maintaining the same split of p-vector u (and correspondingly N^* and Σ_n) into $(u_1', u_2')'$, where u_1 has k elements (consistent with assumption A13), we have, from Equation B.4,

$$\begin{cases} \frac{\partial V_n^{**}}{\partial u_1} = 2\Sigma_{1n}u_1 - 2N_1^* + 2\tilde{\Sigma}_n u_2; \\ \\ \frac{\partial V_n^{**}}{\partial u_2} = 2\tilde{\Sigma}_n' u_1 + 2\Sigma_{2n}u_2 - 2N_2^* + 2f_n u_2^{\gamma}; \end{cases}$$

where the power is understood as the sign-preserving element-wise (Hadamard) power (here and in the derivations to follow).

Hence by first-order conditions for $T_n^* = (T_{1n}^{*\prime}, T_{2n}^{*\prime})'$ we have

$$\begin{cases} T_{1n}^* = \Sigma_{1n}^{-1} N_1^* - \Sigma_{1n}^{-1} \tilde{\Sigma}_n T_{2n}^*; \\ \\ \tilde{\Sigma}_n' T_{1n}^* + \Sigma_{2n} T_{2n}^* - N_2^* + f_n T_{2n}^{*\,\gamma} = 0. \end{cases}$$

Substituting the first equation into the second and rearranging we get

$$\begin{cases} T_{1n}^* = \Sigma_{1n}^{-1} N_1^* - \Sigma_{1n}^{-1} \tilde{\Sigma}_n T_{2n}^*; \\ \\ \left(\Sigma_{2n} - \tilde{\Sigma}_n' \Sigma_{1n}^{-1} \tilde{\Sigma}_n\right) T_{2n}^* + f_n T_{2n}^{*\,\gamma} = N_2^* - \tilde{\Sigma}_n' \Sigma_{1n}^{-1} N_1^*. \end{cases}$$

Following the same argument as in Theorem 2 we have

$$f_n T_{2n}^{*\,\gamma} = N_2^* - \tilde{\Sigma}_n' \Sigma_{1n}^{-1} N_1^* + o_p(1),$$

and hence

$$\begin{cases} T_{1n}^* = \Sigma_{1n}^{-1} N_1^* + o_p(1); \\ \\ f_n T_{2n}^{*\,\gamma} = N_2^* - \tilde{\Sigma}_n' \Sigma_{1n}^{-1} N_1^* + o_p(1). \end{cases}$$

Rewriting further we get

$$\begin{pmatrix} T_{1n}^* \\ \\ f_n T_{2n}^{*\,\gamma} \end{pmatrix} = \begin{pmatrix} \Sigma_{1n}^{-1} & 0 \\ \\ -\tilde{\Sigma}_n' \Sigma_{1n}^{-1} & I_{p-k} \end{pmatrix} N^* + o_p(1).$$

Finally, utilizing B.3 we have

$$\mathcal{L}\left(\begin{pmatrix} T_{1n}^{*} \\ f_n T_{2n}^{*\,\gamma} \end{pmatrix} | \mathcal{E}\right) \xrightarrow[n\to\infty]{d} \mathcal{N}\left(0, \sigma^2\Omega\right), a.s.,$$

where

$$\begin{aligned} \Omega &= \begin{pmatrix} \Sigma_1^{-1} & 0 \\ -\tilde{\Sigma}'\Sigma_1^{-1} & I_{p-k} \end{pmatrix} \begin{pmatrix} \Sigma_1 & \tilde{\Sigma} \\ \tilde{\Sigma}' & \Sigma_2 \end{pmatrix} \begin{pmatrix} \Sigma_1^{-1} & -\Sigma_1^{-1}\tilde{\Sigma} \\ 0 & I_{p-k} \end{pmatrix} \\ &= \begin{pmatrix} \Sigma_1^{-1} & 0 \\ 0 & \Sigma_2 - \tilde{\Sigma}'\Sigma_1^{-1}\tilde{\Sigma} \end{pmatrix}, \end{aligned}$$

which completes the proof. □