On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators

Noureddine El Karoui

Abstract We study ridge-regularized generalized robust regression estimators, i.e.

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \rho_i(Y_i - X_i'\beta) + \frac{r}{2}\|\beta\|^2, \quad \text{where} \quad Y_i = \epsilon_i + X_i'\beta_0,$$

in the situation where $p/n$ tends to a finite non-zero limit. Our study here focuses on the situation where the errors $\epsilon_i$'s are heavy-tailed and $X_i$'s have an “elliptical-like” distribution. Our assumptions are quite general and we do not require homoskedasticity of $\epsilon_i$'s for instance. We obtain a characterization of the limit of $\|\hat{\beta} - \beta_0\|$, as well as several other results, including central limit theorems for the entries of $\hat{\beta}$.

Keywords High-dimensional inference · Random matrix theory · Concentration of measure · Proximal mapping · Regression M-estimates · Robust regression

Mathematics Subject Classification Primary 60F99; Secondary 62E20

I would like to thank Peter Bickel for many interesting discussions on this and related topics. I am grateful to the Center of Mathematical Sciences and Applications at Harvard University for hospitality while some of this work was being completed. I would like to thank two anonymous referees and an associate editor for interesting and constructive comments and questions from which the paper benefited. Support from NSF Grants DMS-0847647 (CAREER) and DMS-1510172 is gratefully acknowledged.
1 Introduction

Robust regression estimators are a standard and important tool in the toolbox of modern statisticians. They were introduced in the late sixties [36] and important early results appeared shortly thereafter [22, 23]. We recall that these estimators are defined as

\[ \hat{\beta}_\rho = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \rho \left( Y_i - X_i' \beta \right), \]

for \( \rho \) a function chosen by the user. Here \( Y_i \) is a scalar response and \( X_i \) is a vector of predictors in \( \mathbb{R}^p \). In the context we consider here, \( \rho \) will be a convex function. Naturally, one of the main reasons to use these estimators instead of the standard least-squares estimator is to increase the robustness of \( \hat{\beta}_\rho \) to outliers in e.g. \( Y_i \)'s. Formally, this robustness result can be seen through results of Huber (see [24]), in the low-dimensional case where \( p \) is fixed. Huber showed that when \( Y_i = X_i' \beta_0 + \epsilon_i \), and when \( \epsilon_i \)'s are i.i.d, under some mild regularity conditions, \( \hat{\beta}_\rho \) is asymptotically normal with mean \( \beta_0 \) and (asymptotic) covariance

\[ (X'X)^{-1} \mathbb{E} \left( \psi^2(\epsilon) \right) \left[ \mathbb{E} (\psi'(\epsilon)) \right]^2, \quad \text{where } \psi = \rho'. \]

The question of understanding the behavior of these estimators in the high-dimensional setting where \( p \) is allowed to grow with \( n \) was raised very early on in [23, p. 802, questions b–f]. These questions started being answered in the mid to late eighties in work of Portnoy and Mammen (e.g. [28, 31–34]). However, these papers covered the case where \( p/n \to 0 \) while \( p \to \infty \).

In the papers [16, 17], we explained (mixing, as in [23], rigorous arguments, simulations and heuristic arguments) that the case \( p/n \to \kappa \in (0, 1) \) yielded a qualitatively completely different picture for this class of problems. For instance, under various technical assumptions, we explained that the risk \( \| \hat{\beta}_\rho - \beta_0 \|_2 \) could be characterized through a system of two non-linear equations (sharing some characteristics with the one below), the distribution of the residuals could be found and was completely different of that of the \( \epsilon_i \)'s, by contrast with the low dimensional case. Furthermore, we showed in [4] that maximum likelihood estimators were in general inefficient in high-dimension and found dimension-adaptive loss functions \( \rho \) that yielded better estimators than the ones we would have gotten by using the standard maximum likelihood estimator, i.e. using \( \rho = -\log f_\epsilon \), where \( f_\epsilon \) is the density of the i.i.d errors \( \epsilon_i \)'s. (We subsequently showed in [15]—which is an initial version of the current paper—that the techniques we had proposed in [16] could be made mathematically rigorous under various assumptions. See also the paper [9] that handles only the case of i.i.d Gaussian predictors, whereas El Karoui [15] can deal with more general assumptions on the predictors. Donoho and Montanari [9] also make interesting connection with the Scherbina–Tirrozi model in statistical physics—see [38, 40]. For other interesting results using rigorous approximate message passing techniques, see also e.g. [2].)
In the current paper, we study a generic extension of the robust regression problem involving ridge regularization. In other words, we study the statistical properties of

\[ \hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \rho_i(Y_i - X_i'\beta) + \frac{\tau}{2} \|\beta\|^2, \text{ where } Y_i = \epsilon_i + X_i'\beta. \]

We will focus in particular on the case where there is no moment restriction on \( \epsilon_i \)'s. Furthermore, a key element of the study will be to show that the performance of \( \hat{\beta} \) is driven by the Euclidean geometry of the set of predictors \( \{X_i\}_{i=1}^n \). To do so, we will study “elliptical” models for \( X_i \)'s, i.e. \( X_i = \lambda_i X_i \), where \( X_i \) has for instance independent entries. We note that when \( \lambda_i \) is independent of \( X_i \) and \( \mathbb{E}(\lambda_i^2) = 1 \), \( \text{cov}(X_i) = \text{cov}(\lambda_i) \). Hence these families of distributions for \( X_i \)'s have the same covariance, but as we will see they yield estimators whose performance vary quite substantially with the distribution of \( \lambda_i \)'s. As we explain below, the role of \( \lambda_i \)'s is to induce a “non-spherical geometry” on the predictors; understanding the impact of \( \lambda_i \)'s on the performance of \( \hat{\beta} \) is hence a way to understand how the geometry of the predictors affects the performance of the estimator. We note that in the low-dimensional case, when \( X_i \)'s are i.i.d, \( X'X/n \to \text{cov}(X_1) \) in probability under mild assumptions, and hence the result of Huber mentioned in Eq. (2) shows that the limiting behavior of \( \hat{\beta}_\rho \) defined in Eq. (1) is the same under “elliptical” and non-elliptical models.

Our interest in elliptical distributions stems from the fact that, as we intuited for a related problem in [16], the behavior of quantities of the type \( X_i'QX_i \) for \( Q \) deterministic is at the heart of the performance of \( \hat{\beta} \). Hence, studying elliptical distribution settings both shed light on the impact of the geometry of predictors on the performance of the estimator and allow us to put to rest potential claims of “universality” of results obtained in the Gaussian (or geometrically similar) case. We note that in statistics there is a growing body of work showing the importance of predictor geometry on various high-dimensional problems (see e.g. [8,13,14,18,20]).

One main motivation for allowing \( \rho_i \) to change with \( i \) is that it might be natural to use different loss functions for different observations if we happen to have information about distributional inhomogeneities in \( \{X_i, Y_i\}_{i=1}^n \). For instance, one group of observations could have errors coming from one distribution and a second group might have errors with a different distribution. Another reason is to gain information on the case of weighted regression, in which case \( \rho_i = w_i \rho \). Also, this analysis can be used to justify rigorously some of the claims made in [16]. Finally, it may prove useful in some bootstrap studies (see e.g. [19] for example).

In the current paper, we consider the situation where \( \beta_0 \) is “diffuse”, i.e. all of its coordinates are small and it cannot be well approximated by a sparse vector. In this situation, use of ridge/\( \ell_2 \) penalization is natural. The paper also answers the question, raised by other researchers in statistics, of knowing whether the techniques of the initial version [15] could be used in the situation we are considering here. Finally, the paper shows that some of the heuristics of [3] can be rigorously justified.

When \( \rho_i = \rho \) for all \( i \), a natural question is to know whether we can find an optimal \( \rho \), in terms of prediction error for instance, as a function of the law of \( \epsilon_i \)'s—in effect asking similar questions to the ones answered by Huber [24] in low-dimension and
in [4] in high-dimension. However, the constraints we impose in the current paper on both the errors (i.e. we do not want them to have moments) and the functions $\rho_i$’s make part of the argument in [4] not usable and might require new ideas. So we will consider this “optimization over $\rho_i$’s and $\tau$” in future work, given that the current proof is already long.

The problem and setup considered in this paper are more natural in the context of robust regression than the ones studied in the initial version [15], where the chosen setup was targeted towards problems related to suboptimality of maximum likelihood methods. However, the strategy for the proof of the results here is similar to the strategy we devised in the initial [15]. There are three main conceptual novelties, that create important new problems: handling ellipticity and the fact that $\beta_0 \neq 0$ requires new ideas in the second part of the proof (i.e. “Appendix 4”). Dealing with heavy tails and appropriate loss functions impacts the whole proof and requires many changes compared to the proof of [15]. Conceptually, this latter part is also the most important, as it shows that all the approximations made in earlier heuristic papers are valid, even in the presence of heavy-tailed errors. This situation is of course the one where these approximations, while having clearly shown their usefulness in giving conceptual and heuristic understanding of the statistical problem, were the most mathematically “suspicious”. So it is interesting to see that they can be made to work rigorously, especially since the probabilistic heuristics developed in these earlier papers allow researchers to shed light quickly on non-trivial statistical problems.

We now state our results. We believe our notations are standard but refer the reader to section Notations (immediately before Eq. 9 below) in case clarification is needed.

## 2 Results

The main focus of the paper is in understanding the properties of

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \rho_i(Y_i - X_i'\beta) + \frac{\tau}{2}\|eta\|^2,$$

where $Y_i = X_i'\beta_0 + \epsilon_i$, \(3\)

and $\tau > 0$. For all $1 \leq i \leq n$, we have $\epsilon_i \in \mathbb{R}$ and $X_i \in \mathbb{R}^p$.

We prove four main results in the paper:

1. we characterize the $\ell_2$-risk of our estimator, i.e. $\|\hat{\beta} - \beta_0\|_2$;
2. we describe the behavior of the residuals $R_i = Y_i - X_i'\hat{\beta}$ and relate them to the leave-one-out prediction error $\tilde{r}_{i,(i)} = Y_i - X_i'\hat{\beta}_{(i)}$;
3. we obtain an approximate update formula for $\hat{\beta}$ when adding an observation (and show it is very accurate);
4. we provide central limit theorems for the individual coordinates of $\hat{\beta}$.

For the sake of clarity, we provide in the main text a series of assumptions that guarantee that our results hold. However, a more detailed and less restrictive statement of our assumptions is provided in the “Appendix”.

© Springer
2.1 Preliminaries and overview of technical assumptions

We use the notation \( \text{prox}(\rho) \) to denote the proximal mapping of the function \( \rho \), which is assumed to be convex throughout the paper. This notion was introduced in [29]. We recall that

\[
\text{prox}(c\rho)(x) = \arg\min_{y \in \mathbb{R}} \left( c\rho(y) + \frac{1}{2}(x - y)^2 \right), \quad \text{or equivalently,} \\
\text{prox}(c\rho)(x) = (\text{Id} + c\psi)^{-1}(x), \quad \text{where } \psi = \rho'.
\]

We refer the reader to [5,29], or [37, Sect. 7.3], for more details on this operation. Note that the previous definitions imply that

\[
\forall x, \quad \text{prox}(c\rho)(x) + c\psi(\text{prox}(c\rho)(x)) = x.
\]

We give examples of proximal mappings in the “Appendix 6”.

We now state some sufficient assumptions that guarantee that all the results stated below are correct. The main proofs are in the “Appendix”. The proofs done in the “Appendix” are done at a much greater level of generality than we are about to state and various aspects of those proofs require much weaker assumptions than those we present here. We start by giving an example where all of our conditions are met.

**Example** Our conditions are met when

- \( p/n \to \kappa \in (0, \infty) \).
- \( \epsilon_i \)'s are i.i.d Cauchy (with median at 0).
- \( X_i = \lambda_i X'_i \), where \( \lambda_i \in \mathbb{R} \) and \( X'_i \in \mathbb{R}^p \) are independent. \( \lambda_i \)'s are i.i.d with bounded support; \( X'_i \)'s are i.i.d with i.i.d \( \mathcal{N}(0,1) \) entries, or i.i.d entries with bounded support and mean 0 as well as variance 1. \( \{X'_i\}_{i=1}^n, \{\lambda_i\}_{i=1}^n \) and \( \{\epsilon_i\}_{i=1}^n \) are independent.
- \( \beta_0 \) is a “diffuse” vector with \( \beta_0(i) = u_{i,p}/\sqrt{p}, 0 \leq |u_{i,p}| \leq C \) and \( \sum_{i=1}^p u_{i,p}^2 = p \), i.e. \( \|\beta_0\|_2 = 1 \).
- \( \rho_i = \rho \) for all \( i \)'s and \( \rho \) is convex. \( \psi = \rho' \) is bounded and \( \psi' \) is Lipschitz and bounded. \( \text{sign}(\psi(x)) = \text{sign}(x) \) and \( \rho(x) \geq \rho(0) = 0 \).

We note that this last condition is satisfied for smoothed approximation of the Huber function, where the discontinuity in \( \psi' \) at say 1 is replaced by a linear interpolation; see below for more details. Note however that the Huber function has a priori no statistical optimality properties in the context we consider.

**Sufficient conditions for our results to hold**

- \( p/n \) has a finite non-zero limit.
- \( \rho_i \)'s are chosen from finitely many possible convex functions. If \( \psi_i = \rho'_i \), \( \sup_i \|\psi_i\|_\infty \leq K, \sup_i \|\psi'_i\|_\infty \leq K \), for some \( K \). \( \psi'_i \) is also assumed to be Lipschitz-continuous. Also, for all \( x \in \mathbb{R} \), \( \text{sign}(\psi_i(x)) = \text{sign}(x) \) and \( \rho_i(x) \geq \rho_i(0) = 0 \).
Theorem 2.1

As p, n tend to infinity while p = O(n log n), with p = O(n log n) being a median of G(\(X_i\)). We require the same assumption to hold when considering the columns of the n \times p design matrix \(X\). \(X_i\)'s have mean 0 and \(\text{cov}(\hat{X}_i) = \text{Id}_p\). We also assume that the coordinates of \(\hat{X}_i\) have moments of all order. Furthermore, for any given \(k\), the \(k\)th moment of the entries of \(\hat{X}_i\) is assumed to be bounded independently of \(n\) and \(p\).

- \(\mathbf{E}(\hat{\lambda}_i^2) = 1\), \(\mathbf{E}(\hat{\lambda}_i^4)\) is bounded and \(\sup_{1 \leq i \leq n} |\lambda_i|\) grows a most like \(C(\log n)^k\) for some \(k\). \(\hat{\lambda}_i\)'s may have different distributions, but the number of such possible distributions is finite.
- \(\epsilon_i\)'s are independent. They may have different distributions, but the number of such possible distributions is finite. Those distributions are assumed to have densities that are differentiable, symmetric and unimodal. Furthermore, we assume that if \(f_i\) is the density of one such distribution, \(\lim_{x \to \infty} xf_i(x) = 0\). \(\{\hat{X}_i\}_{i=1}^n, \{\lambda_i\}_{i=1}^n\) and \(\{\epsilon_i\}_{i=1}^n\) are independent.
- \(\|\beta_0\|_2\) remains bounded. Furthermore, \(\|\beta_0\|_\infty = O(n^{-e})\), where \(1/4 < e\).
- The fraction of time each possible combination of functions and distributions for \((\rho_1, \mathcal{L}(\epsilon_i), \mathcal{L}(\hat{\lambda}_i))\) appears in our problem has a limit as \(n \to \infty\). \(\mathcal{L}(\epsilon_i)\) and \(\mathcal{L}(\hat{\lambda}_i)\) are the laws of \(\epsilon_i\) and \(\hat{\lambda}_i\).

We now state our most important results (several others are in the “Appendix”, where we give the proof) and our proof strategy; naturally, the two go together to provide a sketch of proof. We postpone our discussion of both the assumptions and our results to Sect. 2.3.

2.2 Results and proof strategy

2.2.1 Characterization of the risk of \(\hat{\beta}\)

Consider \(\hat{\beta}\) defined in Eq. (3) and assume that \(\tau > 0\) is given, i.e. does not change with \(p\) and \(n\). Under the technical assumptions detailed in Sect. 2.1, we have:

\textbf{Theorem 2.1} As \(p, n\) tend to infinity while \(p/n \to \kappa \in (0, \infty)\), \(\text{var}(\|\hat{\beta} - \beta_0\|^2) \to 0\). Furthermore, \(\|\hat{\beta} - \beta_0\| \to r_p(\kappa)\) in probability, for \(r_p(\kappa)\) a deterministic scalar. Call \(W_i = \epsilon_i + r_p(\kappa)\lambda_i Z_i\), where \(Z_i\) is a \(\mathcal{N}(0, 1)\) random variable independent of \(\epsilon_i\) and \(\lambda_i\). Then there exists a constant \(c_p(\kappa) \geq 0\) such that

\[
\begin{aligned}
\left[ \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^n \mathbf{E}\left( (\text{prox}(c_p(\kappa)\hat{\lambda}_i^2 \rho_i))(W_i) \right) \right] & = 1 - \kappa + \tau c_p(\kappa), \\
\kappa \left[ \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^n \mathbf{E}\left( \frac{(W_i - \text{prox}(c_p(\kappa)\hat{\lambda}_i^2 \rho_i))(W_i)^2}{\hat{\lambda}_i^2} \right) \right] + \tau^2 \|\beta_0\|^2 c_p^2(\kappa) & = \kappa^2 r_p^2(\kappa).
\end{aligned}
\]

(4)
We note that
\[
\frac{(x - \text{prox}(c_{\rho}(\kappa)\lambda_i^2\rho_i)[x])^2}{\lambda_i^2} = c_{\rho}^2(\kappa)\lambda_i^2\psi_i^2(\text{prox}(c_{\rho}(\kappa)\lambda_i^2\rho_i)[x]),
\]
so in case \(\lambda_i\) takes the value 0, we can replace the expression on the left hand side by that on the right hand side, which does not involve dividing by \(\lambda_i^2\). This alternative expression also shows that there is no problem taking expectations in our equations.

The previous system can be reformulated in terms of \(\text{prox}((c_{\rho}(\kappa)\lambda_i^2\rho_i)*)(x)\), where \(f^*\) represents the Fenchel–Legendre dual of \(f\). Indeed, Moreau’s prox identity \([29]\) gives
\[
\text{prox}((c_{\rho})^*)(x) = x - \text{prox}(c_{\rho})(x).
\]
This is partly why we chose to write the system as we did, since it can be rephrased purely in terms of \(\text{prox}((c_{\rho}(\kappa)\lambda_i^2\rho_i)^*)\), a formulation that has proven useful in previous related problems (see \([4]\)).

We note that \(r_{\rho}(\kappa)\) and \(c_{\rho}(\kappa)\) will in general depend on \(\tau\), but we do not index those quantities by \(\tau\) to avoid cumbersome notations.

2.2.2 Organization of the proof and strategy

The proof is quite long so we now explain the main ideas and organization of the argument. Recall that if
\[
F(\beta) = \frac{1}{n} \sum^n_{i=1} \rho_i (Y_i - X'_i \beta) + \frac{\tau}{2} \|\beta\|^2,
\]
we have
\[
\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} F(\beta).
\]

The proof is broadly divided into three steps.

First step. The first idea is to relate \(\hat{\beta}\) and \(\hat{\beta}_{(i)}\), the solution of our optimization problem when the pair \((X_i, Y_i)\) is excluded from the problem. It is reasonable to expect that adding \((X_i, Y_i)\) will not change too much \(\tilde{r}_{j,(i)} = Y_j - X'_j \hat{\beta}_{(i)}\) when \(j \neq i\), and hence that \(\tilde{r}_{j,(i)} \simeq R_j = Y_j - X'_j \hat{\beta}\) when \(j \neq i\). Armed with this intuition, we can try to use a first-order Taylor expansion of \(\hat{\beta}\) around \(\hat{\beta}_{(i)}\) in the equation \(\nabla F(\hat{\beta}) = 0\) to relate the two vectors. This is what the first part of the proof does, by surmising an approximation \(\eta_i\) for \(\hat{\beta} - \hat{\beta}_{(i)}\) — following along the intuitive lines above but non-trivial to come up with at the level of precision we need. Much work is devoted to proving that this very informed guess is sufficiently accurate for our purposes. Since “the only thing we know” about \(\hat{\beta}\) is that \(\nabla F(\hat{\beta}) = 0\), we work on \(\nabla F(\hat{\beta}) - \nabla F(\hat{\beta}_{(i)} + \eta_i)\) to do so, and show in our preliminaries (see “Appendix 2”)

\[\circ\] Springer
that controlling this latter quantity is enough to control $\| \hat{\beta} - \hat{\beta}(i) - \eta_i \|$. Once our bound for $\| \hat{\beta} - \hat{\beta}(i) - \eta_i \|$ is established, we use it to bound $E \left( \| \hat{\beta} - \beta_0 \|^2 - \| \hat{\beta}(i) - \beta_0 \|^2 \right)$ and use a martingale inequality to deduce a bound on $\text{var} \left( \| \hat{\beta} - \beta_0 \|^2 \right)$, which we show goes to zero. The corresponding results are presented in Sect. 2.2.3 and the detailed mathematical analysis is in “Appendix 3”.

**Second step.** The second step of the proof is to relate $\hat{\beta}$ to another quantity $\hat{\gamma}$, which is the solution of our optimization problem when the last column of the matrix $X$ is excluded from the problem—see Sect. 2.2.4 below and “Appendix 4” for detailed mathematical analysis. Call $V$ the corresponding design matrix. In our setting, it is reasonable to expect that $r_{i, [p]} = Y_i - X_i(p)\beta_0(p) - V_i'\hat{\gamma} \simeq Y_i - X_i'\hat{\beta}$. A first order Taylor expansion of $\nabla F(\hat{\beta})$ around $(\hat{\gamma}', \beta_0(p))'$ and further manipulations yields an informed “guess”, denoted $\hat{b}$ below, for $\hat{\beta}$, and in particular for $\hat{\beta}_p$, the last coordinate of $\hat{\beta}$. A large amount of work is devoted to proving that the quantity we surmised—denoted $b_p$ below—approximates $\hat{\beta}_p$ sufficiently well for our purposes—once again by doing delicate computations on the corresponding gradients. Since $b_p$ has a reasonably nice probabilistic representation, it is possible to write $E \left( b_p^2 \right)$ is terms of other quantities appearing in the problem, such as $\psi_i(r_{i, [p]})$ (where $\psi_i = \rho_i'$) and a quantity $c_{\tau, p}$ that is the trace of the inverse of a certain random matrix. Because $b_p$ approximates $\hat{\beta}_p$ sufficiently well, our approximation of $E \left( b_p^2 \right)$ can be used to yield a good approximation of $E \left( \| \hat{\beta} - \beta_0 \|^2 \right)$. However, we want the approximation of $E \left( \| \hat{\beta} - \beta_0 \|^2 \right)$ to not depend on quantities that depend on $p$, such as $r_{i, [p]}$ and $c_{\tau, p}$. Further work is needed to show that the approximation of $E \left( \| \hat{\beta} - \beta_0 \|^2 \right)$ can be made in terms of $\hat{r}_{i, (i)}$’s—which we used in the first part of the proof—and a new quantity $c_{\tau}$, which is the trace of the inverse of a certain random matrix, as was $c_{\tau, p}$. The resulting approximation for $E \left( \| \hat{\beta} - \beta_0 \|^2 \right)$ is essentially the second equation of our system—see Proposition (2.4) for instance.

**Third step.** The last part of the proof—see Sect. 2.2.5 and “Appendix 5” for detailed mathematical analysis - is devoted to first showing that $\hat{r}_{i, (i)} = Y_i - X_i'\hat{\beta}(i)$ behaves asymptotically like $\epsilon_i + \lambda_i \sqrt{E \left( \| \hat{\beta} - \beta_0 \|^2 \right)} Z_i$, where $Z_i \sim \mathcal{N}(0, 1)$. The work done previously in the proof is extremely useful for that. Finally, we show that $c_{\tau}$ is asymptotically deterministic. The characterization of $c_{\tau}$ is essentially the first equation of our system—see Theorem 2.6 below. After all this is established, we can state for instance central limit theorems for $\hat{\beta}_p$ and interesting quantities that appear in our proof.

The following few subsubsections make all our intermediate results precise. Armed with the above explanation for our approach, they provide the reader with a clear overview of the arc of our proof. The detailed mathematical analysis is given in the “Appendix”.

### 2.2.3 Leave-one-observation out approximations

We call the residuals

$$R_i = Y_i - X_i'\hat{\beta} = \epsilon_i - X_i'\left( \hat{\beta} - \beta_0 \right).$$
We consider the situation where we leave the $i$th observation, $(X_i, Y_i)$, out. We call 
\[
\hat{\beta}(i) = \arg\min_{\beta \in \mathbb{R}^p} F_i(\beta), \quad \text{where } F_i(\beta) = \frac{1}{n} \sum_{j \neq i} \rho_j \left( \epsilon_j + X'_j \beta_0 - X'_j \beta \right) + \frac{\tau}{2} \| \beta \|^2.
\]

We use the notations
\[
\tilde{r}_{j,(i)} = \epsilon_j - X'_j (\hat{\beta}(i) - \beta_0) \quad \text{and} \quad S_i = \frac{1}{n} \sum_{j \neq i} \psi'_j (\tilde{r}_{j,(i)}) X_j X'_j.
\]

Note that $\tilde{r}_{j,(i)}$'s are simply the leave-one-out residuals (for $j \neq i$) and the leave-one-out prediction error (for $j = i$).

Let us consider 
\[
\tilde{\beta}_i = \hat{\beta}(i) + \frac{1}{n} (S_i + \tau I)^{-1} X_i \psi_i (\text{prox}(c_i \rho_i)(\tilde{r}_{i,(i)})) \triangleq \hat{\beta}(i) + \eta_i,
\]

where
\[
c_i = \frac{1}{n} X'_i (S_i + \tau I)^{-1} X_i, \quad \text{and} \quad \eta_i = \frac{1}{n} (S_i + \tau I)^{-1} X_i \psi_i (\text{prox}(c_i \rho_i)(\tilde{r}_{i,(i)})).
\]

We have the following theorem.

**Theorem 2.2** Under our technical assumptions, we have, for any fixed $k$, when $\tau$ is held fixed,
\[
\sup_{1 \leq i \leq n} \| \hat{\beta} - \tilde{\beta}_i \| = O_{L_k} \left( \frac{\text{polyLog}(n)}{n} \right).
\]

Also,
\[
\sup_{1 \leq i \leq n} \sup_{j \neq i} | \tilde{r}_{j,(i)} - R_j | = O_{L_k} \left( \frac{\text{polyLog}(n)}{n^{1/2}} \right),
\]
\[
\sup_{1 \leq i \leq n} | R_i - \text{prox}(c_i \rho_i)(\tilde{r}_{i,(i)}) | = O_{L_k} \left( \frac{\text{polyLog}(n)}{n^{1/2}} \right).
\]

Finally,
\[
\text{var} \left( \| \hat{\beta} - \beta_0 \|^2 \right) = O \left( \frac{\text{polyLog}(n)}{n} \right).
\]

A stronger version of this theorem is available in the “Appendix”. (We say that a sequence of random variables $W_n = O_{L_k} (1)$ if $\mathbb{E} \left( |W_n|^k \right)^{1/k} = O(1)$.)

There are two main reasons this theorem is interesting: it provides online-update formulas for $\hat{\beta}$ through $\tilde{\beta}_i$, with guaranteed approximation errors. Second, it relates the full residuals, whose statistical and probabilistic properties are quite complicated to the
much-simpler-to-understand “leave-one-out” prediction error, \( \tilde{r}_{i,(i)} \). Indeed, because \( X_i \) is independent of \( \hat{\beta}_{(i)} \) under our assumptions, the statistical properties of \( \hat{\beta}_{(i)}' X_i \) are much simpler to understand than those of \( \hat{\beta}' X_i \).

### 2.2.4 Leave-one-predictor out approximations

Let \( V \) be the \( n \times (p - 1) \) matrix corresponding to the first \( p - 1 \) columns of the design matrix \( X \). We call \( V \) in \( \mathbb{R}^{p-1} \) the vector corresponding to the first \( p - 1 \) entries of \( X_i \), i.e. \( V_i = (X_i(1), \ldots, X_i(p-1)) \). We call \( X(p) \) the vector in \( \mathbb{R}^n \) with \( j \)-th entry \( X_j(p) \), i.e. the \( p \)-th entry of the vector \( X_j \). When this does not create problems, we also use the standard notation \( X_j(p) \).

We use the notation \( \beta_0 = (\gamma_0' \beta_0(p))' \), i.e. \( \gamma_0 \) is the vector corresponding to the first \( p - 1 \) coordinates of \( \beta_0 \).

Let us call \( \hat{\gamma} \) the solution of our optimization problem when we use the design matrix \( V \) instead of \( X \). In other words,

\[
\hat{\gamma} = \arg\min_{\gamma \in \mathbb{R}^{p-1}} \frac{1}{n} \sum_{i=1}^{n} \rho_i (\epsilon_i - V_i'(\gamma - \gamma_0)) + \frac{\tau}{2} \| \gamma \|^2. \tag{5}
\]

For stating the following results, we will rely heavily on the following definitions:

**Definition** We call the corresponding residuals \( \{r_i, [p]\}_{i=1}^{n} \), i.e. \( r_i, [p] = \epsilon_i + V_i' \gamma_0 - V_i' \hat{\gamma} \). Let

\[
u_p = \frac{1}{n} \sum_{i=1}^{n} \psi_i'(r_i, [p]) V_i X_i(p), \quad \mathcal{S}_p = \frac{1}{n} \sum_{i=1}^{n} \psi_i'(r_i, [p]) V_i V_i'.
\]

We have \( u_p \in \mathbb{R}^{p-1} \) and \( \mathcal{S}_p \) is \( (p - 1) \times (p - 1) \). We call

\[
\xi_n \triangleq \frac{1}{n} \sum_{i=1}^{n} X_i^2(p) \psi_i'(r_i, [p]) - u'_p (\mathcal{S}_p + \tau \text{Id})^{-1} u_p,
\]

\[
N_p \triangleq \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i(p) \psi_i(r_i, [p]).
\]

We call

\[
b_p \triangleq \beta_0(p) \frac{\xi_n}{\tau + \xi_n} + \frac{1}{\sqrt{n}} \frac{N_p}{\tau + \xi_n}, \tag{6}
\]

and

\[
\tilde{b} = \begin{bmatrix} \hat{\gamma} \\ \beta_0(p) \end{bmatrix} + [b_p - \beta_0(p)] \begin{bmatrix} -1 \\ (\mathcal{S}_p + \tau \text{Id})^{-1} \end{bmatrix}. \tag{7}
\]
Theorem 2.3  Under our Assumptions, we have, for any fixed $\tau > 0$,

$$\|\hat{\beta} - \tilde{b}\| \leq O_{L_k} \left( \frac{\text{polyLog}(n)}{n^{1/2} \wedge n^\theta} \right)^2.$$ 

In particular,

$$\sqrt{n}(\hat{\beta}_p - b_p) = O_{L_k} \left( \frac{\text{polyLog}(n)n^{1/2}}{n^{1/2} \wedge n^\theta} \right),$$

$$\sup_i |X'_i(\hat{\beta} - \tilde{b})| = O_{L_k} \left( \frac{\text{polyLog}(n)n^{1/2}}{\sqrt{n} \wedge n^\theta} \right),$$

$$\sup_i |R_i - r_i, [p]| = O_{L_k} \left( \frac{\text{polyLog}(n)}{\sqrt{n} \wedge n^\theta} \right) \vee \left[ \frac{\text{polyLog}(n)n^{1/2}}{n^{1/2} \wedge n^\theta} \right].$$

Let us call

$$c_{\tau} = \frac{1}{n} \text{trace} \left( (S + \tau \text{Id})^{-1} \right), \quad \text{where } S = \frac{1}{n} \sum_{i=1}^n \psi_i'(R_i)X_iX'_i.$$ 

We also have:

**Proposition 2.4**  Under our assumptions,

$$\left( \frac{p}{n} \right)^2 \mathbb{E} \left( \|\hat{\beta} - \beta_0\|^2 \right) = \frac{p}{n} \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left( [c_{\tau}\lambda_i \psi_i(\text{prox}(c_{\tau}\lambda_i^2 \rho_i)(\tilde{r}_{i,(i)}))]^2 \right)$$

$$+ \tau^2 \|\beta_0\|^2 \mathbb{E} \left( c_{\tau}^2 \right) + o(1).$$

Furthermore,

$$\sup_i |c_i - \lambda_i^2 c_{\tau}| = O_{L_k}(n^{-1/2} \text{polyLog}(n)).$$

2.2.5 Final steps and related results

**Lemma 2.5**  Under our assumptions, as $n$ and $p$ tend to infinity, $\tilde{r}_{i,(i)}$ behaves like $\epsilon_i + \lambda_i \sqrt{\mathbb{E} \left( \|\hat{\beta} - \beta_0\|^2 \right)} Z_i$, where $Z_i \sim \mathcal{N}(0, 1)$ is independent of $\epsilon_i$ and $\lambda_i$, in the sense of weak convergence.

Furthermore, if $i \neq j$, $\tilde{r}_{i,(i)}$ and $\tilde{r}_{j,(j)}$ are asymptotically (pairwise) independent. The same is true for the pairs $(\tilde{r}_{i,(i)}, \lambda_i)$ and $(\tilde{r}_{j,(j)}, \lambda_j)$.

**Theorem 2.6**  Under our assumptions, when $p/n \to \kappa \in (0, \infty)$, $\|\hat{\beta} - \beta_0\| \to r_\rho(\kappa)$, where $r_\rho(\kappa)$ is deterministic. Call $W_i = \epsilon_i + \lambda_i r_\rho(\kappa) Z_i$, where $Z_i \sim \mathcal{N}(0, 1)$ independent of $\epsilon_i$ and $\lambda_i$. Call

$$G_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{E} \left( \frac{1}{1 + x \lambda_i^2 \psi_i(\text{prox}(x \lambda_i^2 \rho_i)(W_i))} \right) \quad \text{and } G(x) = \lim_{n \to \infty} G_n(x),$$

\(\triangleright \) Springer
\[ H_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left( |x \lambda_i \psi_i (\text{prox}(x \lambda_i^2 \rho_i (W_i)))|^2 \right) \quad \text{and} \quad H(x) = \lim_{n \to \infty} H_n(x). \]

Under our assumptions, \( c_\tau \to c_\rho(\kappa) \) in probability, where \( c_\rho(\kappa) \) is the unique solution of the equation \( G(x) = 1 - \kappa + \tau x \). Furthermore, \( r_\rho(\kappa) \) solves

\[ \kappa^2 r_\rho^2(\kappa) = \kappa H(c_\rho(\kappa)) + \tau^2 \| \beta_0 \|^2 c_\rho(\kappa). \]

We note that the equation \( G(x) = 1 - \kappa + \tau x \) translates into the first equation of our system (4). This is a simple consequence of the properties of the derivative of Moreau’s proximal mapping—see Lemma 3.33.

The last equation of Theorem 2.6 is the second equation of our system (4). (The fact that the limits of \( G_n \) and \( H_n \) exist simply come from our assumptions that the proportion of times each possible triplet \((\rho_i, L(\epsilon_i), L(\lambda_i))\) appears has a limit as \( n \to \infty \).)

From this main theorem follows the following propositions.

**Proposition 2.7** \( \xi_n \to \xi \) in probability, where \( \xi = \kappa / c_\rho(\kappa) - \tau > 0 \).

\[ N_p \implies \mathcal{N}(0, v^2) \text{ where} \]

\[ v^2 = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left( \lambda_i^2 \psi_i^2 [\text{prox}(c_\rho(\kappa) \lambda_i^2 \rho_i (W_i))] \right). \]

Finally, when \( \beta_0(k) = O(n^{-1/2}) \),

\[ \sqrt{n}[(\tau + \xi) \hat{\beta}_k - \beta_0(k) \xi] \implies \mathcal{N}(0, v^2). \]

The previous result can be used with \( v^2 \) replaced by \( \hat{v}_n^2 = \frac{1}{n} \sum_{i=1}^{n} \lambda_i^2 \psi_i^2 [\text{prox}(c_\tau \lambda_i^2 \rho_i (\hat{r}_{i, i}))] \) and \( \xi \) replaced by \( \omega_n = p/(nc_\tau) - \tau \) in testing applications—see the discussion after Proposition 3.30 for justifications. Naturally, since for all \( x, \lambda_i^2 \psi_i^2 [\text{prox}(c_\tau \lambda_i^2 \rho_i (x))] = [x - \text{prox}(c_\tau \lambda_i^2 \rho_i (x))] / c_\tau \) when \( c_\tau > 0 \), \( \hat{v}_n^2 \) could also be written and computed using this alternative formulation.

We note that \( \omega_n \) is computable from the data. In our setup, \( \lambda_i \)'s are estimable using the scheme proposed in [14] and \( \hat{v}_n^2 \) can therefore also be estimated from the data. Hence, the previous proposition allows for testing the null hypothesis that \( \beta_0(k) = 0 \), for any \( 1 \leq k \leq p \).

We are also now in position to explain the behavior of the residuals.

**Proposition 2.8** When our assumptions are satisfied and we further assume that \( \lambda_i \)'s are uniformly bounded, we have

\[ \sup_{1 \leq i \leq n} |R_i - \text{prox}(\lambda_i^2 c_\rho(\kappa) \rho_i (\hat{r}_{i, i}))| = o_L(1). \]

The behavior of the residuals is therefore qualitatively very different in this high-dimensional setting than its counterpart in the low-dimensional setting.
2.3 Discussion of assumptions and results

2.3.1 Why consider elliptical-like predictors?

The study of elliptical distributions is quite classical in multivariate statistics (see [1]). As pointed out by various authors (see, in the context of statistics and random matrix theory [8,13,20]), the Gaussian distribution has a very peculiar geometry in high-dimension. It is therefore important to be able to study models that break away from these geometric restrictions, which are not particularly natural from the point of view of data analysts.

Under our assumptions, in light of Lemma 3.37, it is clear that

\[
\sup_{1 \leq i \leq n} \left| \frac{\|X_i\|^2}{p} - \lambda_i^2 \right| = o_P(1), \quad \text{and} \quad \sup_{i \neq j} \left| \frac{X_i'X_j}{p} \right| = o_P(1).
\]

In the Gaussian (or Gaussian-like case of i.i.d entries for \(X_i\), with e.g. bounded entries which satisfy the assumptions we stated above), \(\lambda_i = 1\). Hence, Gaussian or Gaussian-like assumptions imply that predictor vectors are situated near a sphere and are nearly orthogonal. (This simple geometry is of course closely tied to—or a manifestation of the—concentration of measure for convex 1-Lipschitz functions of those random variables.)

This is clearly not the case for elliptical predictors, though under our assumptions, \(\text{cov}(X_i) = \text{Id}_p\), even in the “elliptical” case we consider in the paper. So all the models we consider have the same covariance but the corresponding datasets may have different geometric properties.

We show in the paper that the role of the distribution of \(\lambda_i\)’s in the performance of the estimator depends on much more than its second moment, as Theorem 2.1 makes very clear. This is a situation that is similar to corresponding results in random matrix theory—see e.g. [13,18]. It is therefore clear here again that predictor geometry (as measured by \(\lambda_i\)) plays a key role in the performance of our estimators in high-dimension. This is in sharp contrast with the low-dimensional setting—see [24]—which shows that in low-dimensional robust regression, what matters is only \(\text{cov}(X_i)\).

These types of studies are also interesting and we think important as they clearly show that there is little hope of statistically meaningful “universality” results derived from Gaussian design results: moving from independent Gaussian assumptions for the entries of \(X_i\) to i.i.d assumptions does not change the geometry of the predictors, which appears to be key here as our proof’s reliance on concentration of quadratic forms in \(X_i\) makes clear. As such, while interesting on many counts, for instance to allow discrete predictors, moving from Gaussian to i.i.d assumptions is not a very significant perturbation of the model for statistical purposes. This is why we chose to work under elliptical assumptions. See also [8] for similar observations in a different statistical context.

In conclusion, the generalized elliptical models we study in this paper prove also that many models may be such that the predictors have the same covariance \(\text{cov}(X_i)\) but yield very different performance when it comes to \(\lim \|\hat{\beta} - \beta_0\|\). They therefore
provide a meaningful perturbation of the Gaussian assumption, give us insights into the impact of predictor geometry on the behavior of our estimators, and give us a rough idea of the subclass of models for which we can expect similar (or “universal”) performance for $\hat{\beta}$.

**Examples of distribution for $\mathcal{X}_i$ satisfying our concentration assumptions** Corollary 4.10 in [27] shows that our assumptions are satisfied if $\mathcal{X}_i$ has independent entries bounded by $1/(2\sqrt{c})$. Theorem 2.7 in [27] shows that our assumptions are satisfied if $\mathcal{X}_i$ has independent entries with density $f_k, 1 \leq k \leq p$ such that $f_k(x) = \exp(-u_k(x))$ and $u_k''(x)\geq \sqrt{c}$ for some $c > 0$. Then $c = c/2$. This is in particular the case for the case where $\mathcal{X}_i$ has i.i.d $\mathcal{N}(0, 1)$ entries: then $c = 1$ and $c = 1/2$. We discuss briefly after Lemma 3.35 in the “Appendix” the impact of choosing other types of concentration assumptions.

2.3.2 Non-sparse $\beta_0$: why consider $\ell_2$/ridge-regularization?

In this paper, we consider the case where $\beta_0$ cannot—in general—be approximated in $\ell_2$-norm by a sparse vector. This is a situation that is thought to not be uncommon in biology (see, in a slightly different context [12], and many similar references), where sparsity assumptions are often/sometimes in doubt.

In other words, if $s$ is a sparse vector (e.g. with support of size $o(p)$), we necessarily have when $\beta_0$ is diffuse (i.e. all of its entries are roughly of size $p^{-1/2}$) $\|\beta_0 - s\| \not\rightarrow 0$. In the situation we consider, it is in fact unclear whether any estimator can be consistent in $\ell_2$ for $\beta_0$. One interesting aspect of our study is that the System (4) might allow us to optimize (at least in certain circumstances) over the functions $\rho_i$’s we consider to get the best performing estimator in the class of ridge-regularized robust regression estimators for $\beta_0$ and hence potentially beat sparse estimators (in the same line of thought, there are of course numerous applied examples where ridge regression outperforms Lasso in terms of prediction error).

Finally, one benefit of our analysis is that we have a central limit theorem for the coordinates of $\hat{\beta}$ (see Proposition 2.7), which makes testing possible. In the situation where $\beta_0$ has some large entries (of size up to $n^{-1/4-\eta}, \eta > 0$) and many small ones [of size $o(n^{-1/2})$], this central limit theorem and its more refined version in Proposition 3.30 could help in designing better performing estimators by using scaled versions of $\{\hat{\beta}_k\}_{k=1}^p$, which we would threshold according to the result of our test. In other words, these central limit theorems for the coordinates of $\hat{\beta}$ are the gateway to the construction of Hodges-type estimators in the setup we consider.

2.3.3 A remark on the fixed design case

We have worked in this paper with a certain class of random designs. It is not unusual to do so in robust regression studies—see the classic papers by Portnoy [31,32,34]. In many areas of applications, it is also unclear why statisticians should limit themselves to the study of fixed designs, in particular when they do not have control over the choice of the values of the predictors, i.e. they cannot design their experiments.
However, it is also interesting to understand what remains valid of our analysis in the case of fixed design. We note that our analysis gives already a few results in this direction.

In fact, since we have shown that $\operatorname{var} (\| \hat{\beta} - \beta_0 \|^2) \to 0$, we have shown that

$$E \left( \operatorname{var} (\| \hat{\beta} - \beta_0 \|^2 | X) \right) \to 0$$

and

$$\operatorname{var} (E (\| \hat{\beta} - \beta_0 \|^2 | X)) \to 0,$$

because

$$\operatorname{var} (\| \hat{\beta} - \beta_0 \|^2) = E \left( \operatorname{var} (\| \hat{\beta} - \beta_0 \|^2 | X) \right) + \operatorname{var} (E (\| \hat{\beta} - \beta_0 \|^2 | X)).$$

Therefore, with probability (over the design $X$) going to 1,

$$\| \hat{\beta} - \beta_0 \|^2 - r_\rho(\kappa) \to 0 \text{ in } P_{\{\epsilon_i\}_{i=1}^n} \text{-probability.}$$

($P_{\{\epsilon_i\}_{i=1}^n}$-probability simply refers to probability statements with respect to the random $\epsilon_i$’s, the only source of randomness if the design matrix $X$ is assumed to be fixed.) In other words, if the design is fixed, but results from one random draw of a $n \times p$ matrix satisfying our distributional assumptions, Theorem 2.1 applies with probability (over the choice of design matrix) going to 1.

We note that $\| \hat{\beta} - \beta_0 \|$ is an especially important quantity in terms of prediction error in our context, which is why our short discussion above focused on this quantity: if we are given a new predictor vector $X_{\text{new}}$, we would naturally predict an unobserved response $Y_{\text{new}}$ by $X'_{\text{new}} \hat{\beta}$ and hence, if $Y_{\text{new}} = \epsilon_{\text{new}} + X'_{\text{new}} \beta_0$, our prediction error will be $PE_{\text{new}} = \epsilon_{\text{new}} + X'_{\text{new}} (\beta_0 - \hat{\beta})$. Of course, if $X_{\text{new}}$ has mean 0 and satisfies $\operatorname{cov} (X_{\text{new}}) = \operatorname{Id}_p$, $E_{\text{new}} [(X'_{\text{new}} (\beta_0 - \hat{\beta}))^2] = \| \beta_0 - \hat{\beta} \|^2_2$. Hence the expected squared prediction error will be $\operatorname{var} (\epsilon_{\text{new}}) + \| \beta_0 - \hat{\beta} \|^2_2$, provided $\epsilon_{\text{new}}$ is independent of $X_{\text{new}}$.

2.3.4 Optimization with respect to $\tau$ and $\rho$

Just as the classic work of Huber on robust regression started by establishing central limit theorems for the estimator of interest (as a function of $\rho$) and proceeded to find optimal methods in various contexts (see [24]), one objective of our work is to pave the way for answering optimality questions in the setting we consider. An important first step to do so is therefore to obtain results such as Theorem 2.1.

A natural question is therefore to ask what are the optimal $\rho_i$’s in the context we consider, where optimality might be defined in terms of minimizing $r_\rho(\kappa)$ in Theorem 2.1 or $\upsilon^2$ in Proposition 2.7. For an example of such a study for $r_\rho(\kappa)$ in a slightly different context, see [4]. Similarly, optimization over $\tau$ should be possible. We leave however these questions for future work, since they are of a more analytic nature. (We have had success in [4] in the situation where $\lambda_i = 1$ and the errors are log-concave and hence not heavy-tailed, but the technique we employed in that paper does not apply readily here.)
We also note that in our context the optimal \( \tau \) – say for prediction – is in general not going to be close to 0, so the fact that our current study requires \( \tau > 0 \) is not a problem (see also [3]).

As can intuitively be seen from Proposition 2.7, the fact that \( \| \hat{\beta} - \beta_0 \| \) goes to a non-zero constant has two sources: bias induced by the ridge-regularization and the fact that each of the \( p \) coordinates has fluctuations of size \( n^{-1/2} \). Its asymptotically deterministic character comes on the other hand from our analysis in “Appendix 3”. This is in contrast with the low-dimensional case where \( p \) is fixed, where \( \| \hat{\beta} - \beta_0 \| \) goes to a non-zero constant simply because of bias issues and the law of large numbers.

In light of Proposition 2.7, it is clear that \( \tau \) plays in this problem a fairly similar role to the one it plays in low-dimension: it trades bias for variability in each individual coordinate. By contrast with the low-dimensional case however, even if we consider the case \( p < n \), the fact that the number of coordinates is of the same order of magnitude as \( n \) means that even for small \( \tau \) (and hence low-bias), \( \| \hat{\beta} - \beta_0 \| \) has a non-zero limit. The fact that this limit can be somewhat large is what suggests using values of \( \tau \) that are not close to 0. Interestingly, this is of course the situation encountered in practice in many real-world problems where \( p \) and \( n \) are large. A case in point is the situation where \( \beta_0 = 0 \), in which case the optimal value of \( \tau \) is clearly \( \infty \); using \( \tau = 0 \) would put us back in the situation of [17], which would result in much worse performance for \( \| \hat{\beta} - \beta_0 \| \).

2.3.5 Possible extensions

Less smooth \( \rho \)’s and \( \psi \)’s While our approach is quite general and allows us to handle designs that are far from being Gaussian, the proof presented in this paper still requires some smoothness concerning \( \rho_i \)’s and \( \psi_i \)’s. On the one hand, results such as the ones obtained in [4] suggest that it is often the case that optimal loss functions in high-dimension are smoother than in low dimension. So the fact that we require \( \psi_i \)’s to be smooth is a source of less concern that it would be in low dimension. (Note also that the classic papers [28,32] also require smoothness properties on \( \psi \).)

Though it is unclear whether the Huber function is optimal in any sense for the problems we are looking at, and hence whether it warrants a special focus, let us discuss this function in some detail. For the sake of simplicity let us focus on the situation where the transition from quadratic to linear happens at \( x = \pm 1 \). Then

\[
\psi(x) = \begin{cases} 
    x & \text{if } |x| \leq 1 \\
    \text{sign}(x) & \text{if } |x| \geq 1 
\end{cases}
\]

So \( \psi \) is not differentiable at 1. However, it is easy to approximate this function by a function whose derivative is Lipschitz. As a matter of fact, if \( 0 < \eta < 1 \), \( \psi_\eta \) such that

\[
\psi_\eta'(x) = \begin{cases} 
    1 & \text{if } |x| \leq 1 - \eta \\
    \frac{1-|x|}{\eta} & \text{if } |x| \in (1 - \eta, 1) \\
    0 & \text{if } |x| \geq 1 
\end{cases}
\]
is $1/\eta$-Lipschitz. Furthermore, $\psi_\eta$, the corresponding antisymmetric function with, when $x \geq 0$,

$$
\psi_\eta(x) = \begin{cases}
  x & \text{if } 0 \leq x \leq 1 - \eta, \\
  1 - \frac{\eta}{2} - \frac{(1-x)^2}{2\eta} & \text{if } x \in (1 - \eta, 1), \\
  1 - \frac{\eta}{2} & \text{if } x \geq 1,
\end{cases}
$$

can be made to be arbitrarily close to $\psi$ and similarly for the corresponding $\rho_\eta$, picked such that $\rho_\eta(0) = 0$.

Our results apply to $\rho_\eta$, for any $\eta > 0$. It seems quite likely that with a bit (and possibly quite a bit) of further approximation theoretic work, it should be possible to establish results similar to Theorem 2.1 for the Huber function by taking the limit of corresponding results for $\rho_\eta$ with $\eta$ arbitrarily small.

We note that most of our proof (in particular “Appendices 3 and 4”) is actually valid with functions $\rho_i$’s that can change with $n$. In particular, many results hold when $\psi_i$ are $L_i(n)$-Lipschitz with $L_i(n) \leq C n^\alpha$. So one strategy to handle the case of the Huber function could be to use $\psi_n$ with $\eta_n = 1/\log(n)$ for instance and strengthen the arguments of “Appendix 5” in the Appendix—in this very specific case where $\psi_n$ has a limit—to get the Huber case as a limiting result. Because our proof is already long, we leave the details to the interested reader and might consider this problem in detail in future work.

Weighted robust regression One motivation for working on the problem at the level of generality we dealt with is that our results should allow us to tackle among other things weighted robust regression. For instance if $\epsilon_i$’s or $\lambda_i$’s in our model had different distributions, it would be natural to pick the corresponding $\rho_i$’s either as completely different functions, or maybe as $\rho_i = w_i \rho$, with $w_i$ deterministic but possibly depending on the distribution of $\epsilon_i$’s or $\lambda_i$’s. In the case where $\epsilon_i$’s and $\lambda_i$’s come from finitely many possible distributions, our results handle this situation.

Most of our results—i.e. those of “Appendices 3 and 4”—are true even when $w_i$’s are allowed to take a possibly infinite set of different values. If $\epsilon_i$’s are i.i.d, $\lambda_i$’s are i.i.d and $w_i$’s are i.i.d and these three groups of random variables are independent of each other, our arguments can be made to go through without much extra difficulties. The main potential problem is in “Appendix 5”, but then distributional symmetry between the $R_i$’s on one hand and the $\tilde{r}_{i,(i)}$ on the other hand becomes helpful, as it had in [15]. So it is very likely that our results could be extended to cover this case at relatively little technical cost.

3 Conclusion

We have studied ridge-regularized robust regression estimators in the high-dimensional context where $p/n$ has a finite non-zero limit. Our study has highlighted the importance of the geometry of the predictors in this problem: two models with similar covariance but different predictor geometry will in general yield estimators with very different performance. We have shown this result by studying the random design case

\[ \text{Springer} \]
in the context of elliptical predictors and looking at the influence of the “ellipticity parameter” $\lambda_i$ on our results. Importantly, this shows that no statistically meaningful “universality” results can be derived from the study of Gaussian or i.i.d-designs, since their geometry is so peculiar (i.e. they are limited to the case $\lambda_i = 1$ for all $i$’s). The technique used in the paper seems versatile enough to be useful for several other high-dimensional M-estimation problems.

We have also obtained central limit theorems for the coordinates of $\hat{\beta}$ that can be used for testing whether $\beta_0(k) = 0$ for any $1 \leq k \leq p$. However, our focus was mostly on the case where $\beta_0$ is diffuse, with all coordinates small but contributing to $Y_i = \epsilon_i + X_i' \beta_0$. Our results also provide a very detailed understanding of the properties of the residuals $R_i$.

All these results were obtained without moment requirements on the errors $\epsilon_i$’s.

Finally, our characterization of the risk of these estimators raises interesting analytic questions related to finding optimal loss functions $\rho_i$’s in the context we consider. We plan to study these questions in the future.

Open Access  This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

Appendix 1: Assumptions and technical elements

Recall that the focus of the paper is on understanding the properties of

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \rho_i \left( \epsilon_i - X_i' (\beta - \beta_0) \right) + \frac{\tau}{2} \| \beta \|^2$$

(8)

where $\tau > 0$. For all $1 \leq i \leq n$, we have $\epsilon_i \in \mathbb{R}$ and $X_i \in \mathbb{R}^p$.

Different parts of the proof require different assumptions. So we label them accordingly.

Most of our proof (“Appendices 3 and 4”) is carried out for functions $\rho_{i,n}$ that may vary with $n$, so our assumptions reflect this and we carry out most of our work at this level of generality. However, we do not make the dependence of $\rho_{i,n}$ on $n$ explicit to avoid cumbersome notations. Having these results available should make future work on weighted regression or work of a more approximation-theoretic nature (for instance using a sequence $\rho_{i,n}$ to approximate a function $\rho_i$ that is not smooth) easier. This is one of the prime motivations for working at this level of generality.

Naturally, our assumptions are more and more restrictive as the proof progresses, so the summary of assumptions we provided in the main text is obtained by going through the assumptions and simply tallying the more restrictive ones. A sketch of proof is provided in Sect. 2.2.2, which should be helpful in navigating the detailed proof we provide in this “Appendix”.

Before we delve into the details of the assumptions needed for each part of the proof to work, we summarize for the convenience of the reader the assumptions we need for the whole proof to go through.
Assumptions under which the whole proof goes through

- **A1** $p/n$ has a finite non-zero limit
- **A2** $\|\beta_0\|$ remains bounded. Furthermore, $\|\beta_0\|_\infty = O(n^{-\theta})$, for $\theta > 1/4$
- **A3** $\rho_i$’s are twice differentiable, and convex. If $\psi_i = \rho_i^\prime$, we assume that $\text{sign}(\psi_i(x)) = \text{sign}(x)$ and $\rho_i \geq 0 = \rho_i(0)$. Furthermore, there exists $C$ such that $\|\psi_i\|_\infty \leq C$, $\|\psi_i\|_\infty \leq C$ and $\psi_i$ is assumed to be Lipschitz. The functions $\rho_i$’s can be chosen among finitely many possible functions (over all $n$).
- **A4** $X_i = \lambda_i \chi_i$, $\lambda_i$’s are random variables with $\lambda_i \in \mathbb{R}$. $\chi_i$'s are independent and identically distributed. Their distribution is allowed to change with $p$ and $n$. The entries of $\chi_i$ are independent. Furthermore, for any 1-Lipschitz (with respect to Euclidean norm) convex function $G$, if $m_{G(\chi_i)}$ is a median of $G(\chi_i)$, for any $t > 0$, $P(|G(\chi_i) - m_{G(\chi_i)}| > t) \leq C_n \exp(-c_n t^2)$, $C_n$ and $c_n$ can vary with $n$. For simplicity, we assume that $1/c_n = O(\text{polyLog}(n))$ and $C_n$ is bounded in $n$. $\chi_i$’s have mean 0 and variance 1. (See [27, Corollary 4.10].)
- **A5** $\lambda_i$’s are independent of each other and $\{\chi_i\}_{i=1}^n$. $\mathbb{E}(\lambda_i^2) = 1$ and $\mathbb{E}(\lambda_i^4) \leq C$ and $\sup_{1 \leq i \leq n} |\lambda_i| = O_Lk(\text{polyLog}(n))$.
- **A6** $\epsilon_i$’s are independent of $\{\chi_i\}_{i=1}^n$ and $\{\lambda_i\}_{i=1}^n$ and of each other. Furthermore, for any $r \in \mathbb{R}$, if $Z \sim \mathcal{N}(0, 1)$, independent of $\epsilon_i$, $\epsilon_i + rZ$ has a (differentiable) density $f_{r, \epsilon}$ which is increasing on $(-\infty, 0)$ and decreasing on $(0, \infty)$. Finally, $\lim_{|r| \to \infty} tf_{r, \epsilon}(t) = 0$.
- **A7** $\lambda_i$’s can have different distributions. Similarly, $\epsilon_i$’s can have different distributions. However, the number of choices for the triplet $(\rho_i, \mathcal{L}(\lambda_i), \mathcal{L}(\epsilon_i))$ is finite (over all $n$). Furthermore, the fraction of times each such triplet appears in our problem—see Eq. (8) has a limit. $\mathcal{L}(\epsilon_i)$ just means the law of $\epsilon_i$.

We note that the entries of $\chi_i$ do not need to have the same distribution. Our condition A4 is satisfied when $\chi_i$ have i.i.d $\mathcal{N}(0, 1)$ entries, or independent entries that are bounded by a constant and have mean 0 and variance 1. (See [27, Corollary 4.10].)

Condition A7 just means that if for instance $\rho_i = \rho$ and $\lambda_i$’s are i.i.d but $\epsilon_i$’s can come from 3 distributions, the fraction of $\epsilon_i$’s coming from each of these three distributions has a limit as $n \to \infty$. This last condition mostly plays a role in guaranteeing that we can take limits in various expressions. The simplest case is of course when $\rho_i = \rho$, $\lambda_i$’s are i.i.d and $\epsilon_i$’s are i.i.d, in which case there is only one possible choice for the triplet $(\rho_i, \mathcal{L}(\lambda_i), \mathcal{L}(\epsilon_i))$.

We now state the conditions under which we carry out the proof. We state them in one place for the convenience of the reader. A discussion follows immediately after the statement of all the conditions.
First part of the proof (“Appendix 3”)

For the first part of the proof (i.e. “leave-one-observation-out”), we work under the following assumptions:

- **O1**: \( p/n \) has a finite non-zero limit.
- **O2**: \( \rho_i \)'s are twice differentiable, convex and non-linear. \( \psi_i = \rho_i' \). Note that \( \psi_i' \geq 0 \) since \( \rho_i \) is convex. We assume that \( \text{sign}(\psi_i(x)) = \text{sign}(x) \) and \( \rho_i \geq 0 = \rho_i(0) \).
- **O3**: \( \sup_{x,i} |\psi_i(x)| \leq C \text{polyLog}(n) \) where \( C \) is constant. This is natural in the context of robust statistics, since it means that we allow \( \rho_i \)'s to grow at most linearly at infinity. This assumption is for instance verified for Huber functions. Furthermore, \( \psi_i' \) is assumed to be \( L_i(n) \)-Lipschitz with \( L_i(n) \leq C n^\alpha, \alpha \geq 0 \). (We here have in mind smoothed Huber functions.) We also assume that \( \sup_i \|\psi_i'\|_\infty \leq C \text{polyLog}(n) \). Finally, we assume that \( \frac{1}{n} \sum_{i=1}^n \|\psi_i^2\|_\infty \leq C \), where \( C \) is a constant independent of \( n \).
- **O4**: \( X_i = \lambda_i X_i \). \( \lambda_i \)'s are random variables with \( \lambda_i \in \mathbb{R} \). \( X_i \in \mathbb{R}^p \) are independent and identically distributed. Their distribution is allowed to change with \( p \) and \( n \). Furthermore, for any \( 1 \)-Lipschitz (with respect to Euclidean norm) convex function \( G \), if \( m_G(X_i) \) is a median of \( G(X_i) \), for any \( t > 0 \), \( P(|G(X_i) - m_G(X_i)| > t) \leq C_n \exp(-c_n t^2) \). \( C_n \) and \( c_n \) can vary with \( n \). For simplicity, we assume that \( 1/C_n = o(\text{polyLog}(n)) \) and \( C_n \) is bounded in \( n \). \( \lambda_i \)'s have mean \( 0 \) and \( \text{cov}(\lambda_i^t) = \text{Id}_p \). We also assume that the coordinates of \( X_i \) have moments of all order. Furthermore, for any given \( k \), the \( k \)th moment of the entries of \( X_i \) is assumed to be bounded independently of \( n \) and \( p \). \( \{X_i\}_{i=1}^n \) and \( \{\lambda_i\}_{i=1}^n \)'s are independent.
- **O5**: \( \{X_i\}_{i=1}^n \) and \( \{\lambda_i\}_{i=1}^n \) are independent of \( \{\epsilon_i\}_{i=1}^n \). \( \epsilon_i \)'s are independent of each other.
- **O6**: \( \sup_{1 \leq i \leq n} |\lambda_i| \overset{\Delta}{=} L_n = O_{L_k}(\text{polyLog}(n)) \) and \( \lambda_i \)'s are independent. Furthermore, \( E(\lambda_i^2) = 1 \). (Note that this implies that \( \text{cov}(X_i) = \text{cov}(\lambda_i^2) \).)
- **O7**: \( 1 - 2\alpha > 0 \) and \( \|\beta_0\| = O(\text{polyLog}(n)) \).

Note that we do not assume that \( \epsilon_i \)'s have identical distributions. Assumption **O4** is satisfied for instance when \( X_i \) are \( \mathcal{N}(0, \text{Id}_p) \) or have i.i.d entries bounded by \( \text{polyLog}(n) \)—see [27] (this reference guarantees the concentration result we require is satisfied; the moment conditions need to be checked by other methods, but this is generally much simpler, as the case of Gaussian random variables clearly shows). Importantly, note that **O4** does not require the entries of \( X_i \) to be independent; see [27] or [13] for examples of \( X_i \) satisfying **O4** with dependent entries. In other respects, the assumption \( E(\lambda_i^2) = 1 \) plays a very minor role mathematically and could be relaxed to \( E(\lambda_i^2) \) is uniformly bounded without problems. Statistically, it is however important as it guarantees that \( \text{cov}(X_i) = \text{Id}_p \) in all the models we consider.

Second part of the proof (“Appendix 4”)

For the second part of the proof (i.e. “leave-one-predictor-out”), we need all the previous assumptions and
• **P1**: \( X_i \)'s have independent entries. Furthermore, for \( 1 \leq k \leq p \), the vectors 
\( \Theta_k = (X_1(k), \ldots, X_n(k)) \) in \( \mathbb{R}^n \) satisfy: for any 1-Lipschitz (with respect to 
Euclidean norm) convex function \( G \), if \( m_{G(\Theta_k)} \) is a median of \( G(\Theta_k) \), for any 
\( t > 0 \), 
\[ P(|G(\Theta_k) - m_{G(\Theta_k)}| > t) \leq C_n \exp(-c_nt^2), \]
\( C_n \) and \( c_n \) can vary with \( n \). 
As above, we assume that \( 1/c_n = O(\text{polyLog}(n)) \).

• **P2**: 
\[ \frac{1}{n} \sum_{i=1}^{n} \|\psi_i\|_{\infty} = O(1). \]

• **P3**: 
\[ \|\beta_0\|_{\infty} = O(n^{-\epsilon}), \] where \( \epsilon > 0 \). Furthermore, 
\[ \|\beta_0\|_2 \leq C, \] where \( C \) is a 
constant independent of \( p \) and \( n \). \( \epsilon \) satisfies \( \alpha + 1/4 - \epsilon < 0 \).

• **P4**: 
\[ 1/2 - 2\alpha > 0 \text{ and } \min(1/2, \epsilon) - \alpha - 1/4 > 0. \] The latter implies that 
\[ \min(1/2, \epsilon) > 0. \]

We note that according to Corollary 4.10 and the discussion that follows in [27], 
Assumptions O4 and P1 are compatible. O4 and P1 are for instance satisfied if the 
entries of \( X_i \)'s are independent and bounded by polyLog \( (n) \). Another example is the 
case of \( X_i \sim \mathcal{N}(0, \text{Id}_p) \), in which case \( c_n \) is a constant independent of the dimension.

Note that we do not assume that the entries of \( X_i \) have the same distribution.

We note that if for instance \( \alpha = 1/12 \) and \( \epsilon = 5/12 \), all the conditions in P3–P4 
are satisfied. When \( \alpha = 0 \), they simply become \( \epsilon > 1/4 \).

**Last part of the proof (“Appendix 5”)**

For the last part of the proof, when we combine everything together, we will need the 
following assumptions on top of all the others:

• **F1**: the \( \epsilon_i \)'s may have different distributions; however, they may only come from 
finitely many distributions. Furthermore, for any \( r \in \mathbb{R} \), if \( Z \sim \mathcal{N}(0, 1) \), independent 
of \( \epsilon_i, \epsilon_i + rZ \) has a differentiable density \( f_{i,r} \) which is increasing on \((-\infty, 0)\) 
and decreasing on \((0, \infty)\). Finally, \( \lim_{|t| \to \infty} tf_{i,r}(t) = 0 \).

• **F2**: 
\[ \frac{1}{n} \sum_{i=1}^{n} \|\psi_i\|_{\infty} = O(1). \] \( \psi_i \) has Lipschitz constant \( L_i(n) \). Furthermore, 
\[ \frac{1}{n} \sum_{i=1}^{n} L_i(n) \|\psi_i\|_{\infty} = O(1). \]

• **F3**: \( \alpha < 1/6 \) and \( \alpha + 1/3 < 2 \min(1/2, \epsilon) \)

• **F4**: there exists \( C \) independent of \( n \) and \( p \) such that \( E(\lambda_i^2) \leq C \).

• **F5**: \( \lambda_i \)'s may have different distributions, but the set of possible distributions for \( \lambda_i \) 
is finite. Similarly, \( \rho_i \) may be different functions, but the set of possible functions 
\( \rho_i \) may be finite. Also, the number of distinct triplets \( (\rho_i, \mathcal{L}(\epsilon_i), \mathcal{L}(\lambda_i)) \) is finite 
(over all \( n \)). Furthermore, the proportion of each such distinct triplet has a limit as 
\( n \to \infty \).

Condition F3 is clearly satisfied in the case \( \alpha = 1/12 \) and \( \epsilon = 5/12 \) we mentioned 
above. On the other hand, condition F5 requires that \( \alpha = 0 \), since it prevents \( \rho_i \) from 
changing with \( n \). (We note that since F5 is required only at the very end of the proof, 
one could probably weaken its requirements considerably if another situation that the 
one we investigate really called for it.)

We refer the reader to Lemma 3.39 and the discussion immediately following it for 
examples of densities for \( \epsilon_i \)'s satisfying F1. We note that smooth symmetric (around 
0) log-concave densities will for instance satisfy all the assumptions we made about 
the \( \epsilon_i \)'s. See [25,26] for instance. This is also the case for the Cauchy distribution (see 
Theorem 1.6 in [7]). The latter is the most relevant reference here since we care about 
heavy-tailed \( \epsilon_i \)'s.
Assumptions concerning the loss functions  We wanted to investigate in this paper the situation where $\epsilon_i$’s have no moment restrictions, as befits “classical” robust statistics studies. As such, it is natural to assume that $\psi_i$’s remain bounded, which is part of assumption A3. We think that interesting results can be found in “Appendices 3 and 4”: in particular for dealing with $\psi_i$ functions that are not smooth and require approximations by $\psi_{i,n}$ functions that are smooth, but also for bootstrap studies for instance. This is why the paper handles those cases, even though we do not use fully these results in the main statements of the paper. We note that in specific cases, one would simply need to modify various arguments in “Appendix 5” to handle limits of those $\psi_{i,n}$.

Assumptions concerning the predictors  Assumption O4 is a bit stronger than we will need. For instance, “Appendices 3 and 4” do not actually require the $X_i$’s to have identical distributions; “Appendix 5” would work if we assumed that $X_i$’s were coming from finitely many distributions, with the proportion of $X_i$’s picked from a particular distribution having a limit as $n \to \infty$. The functions $G$ we are working with will either be linear or square-root of quadratic forms, so we could limit our assumptions to those functions. However, as documented in [27] and discussed briefly in the introduction, a large number of natural or “reasonable” distributions satisfy the O4 assumptions. Our choice of having a potentially varying $c_n$ is motivated by the idea that we could, for instance, relax an assumption of boundedness of the entries of $X_i$’s—that guarantees that O4 is satisfied when $X_i$ has independent entries—and replace it by an assumption concerning the moments of the entries of $X_i$’s and a truncation of triangular arrays argument (see for instance [42]). We also refer the interested reader to [13] for a short list of distributions satisfying O4, compiled from various parts of [27]. Finally, we could replace the $\exp(-c_n t^2)$ upper bound in O4 by $\exp(-c_n t^\beta)$ for some fixed $\beta > 0$ and it seems that all our arguments would go through. We chose not to work under these more general assumptions because it would involve extra book-keeping and does not enlarge the set of distributions we can consider enough to justify this extra technical cost. Importantly, O4 allows the entries of $X_i$’s to be dependent.

To give a concrete example, let us consider the situation where the entries of $X_i$ are independent and symmetric with an exponential density chosen to have variance 1. Then it is clear that $\sup_{i,j} |X_i(j)| \leq K[\log(n)]^2$ almost surely as $n, p \to \infty$. Our analysis and assumptions then apply to the predictors $X_i = \lambda_i s_n \Gamma_{i,n}$, with $\Gamma_{i,n} = \Gamma_i \Gamma_i' \leq K[\log(n)]^2$, where $s_n$ is chosen so that $1/s_n^2 = \text{var}(\Gamma_{i,n})$ (the variance of the entries of $\Gamma_{i,n}$ is not 1, since it is a truncation of $X_i$ but it is easy to see that $\text{var}(\Gamma_{i,n}) \to 1$). Note that $X_i = X_i$ almost surely, and therefore our statistical problem is not affected. Very minor modifications to the arguments of “Appendix 5” are then needed to handle $s_n$ and show that our results go through. Naturally the same argument could be made for other (non-exponential) distributions as long as $\sup_{i,j} |X_i(j)| \leq K[\log(n)]^2$. We note that our method should also be able to handle $c_n$ such that $1/c_n$ grows faster that polyLog$(n)$ and hence deal with an even broader class of predictor...
distributions, but we chose not to do this in full details to limit the book-keeping burden in an already long proof.

**Notations** We will repeatedly use the following notations: \( Y_i = X_i' \beta_0 + \epsilon_i \); polyLog\((n)\) is used to replace a power of \( \log(n) \); \( \lambda_{\max}(M) \) denotes the largest eigenvalue of the matrix \( M \); \( \| M \|_2 \) denotes the largest singular value of \( M \). We call \( \hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} X_i X_i' \) the usual sample covariance matrix of the \( X_i \)'s when \( X_i \)'s are known to have mean 0.

We say that \( X \preceq Y \) in \( L_k \) if \( \mathbf{E} (|X|^k) \leq \mathbf{E} (|Y|^k) \). We write \( X \preccurlyeq Y \) to say that the random variables \( X \) and \( Y \) are equal in law. We use the usual notation \( \beta(i) \) to denote the regression vector we obtain when we do not use the pair \((X_i, Y_i)\) or \((X_i, \epsilon_i)\) in our optimization problem, a.k.a the leave-one-out estimate. We will also use the notation \( X(i) \) to denote \( \{X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n\} \). We use the notation \((a, b)\) for either the interval \((a, b)\) or the interval \([a, b)\): in several situations, we will have to localize quantities in intervals using two values \( a \) and \( b \) but we will not know whether \( a < b \) or \( b > a \). We denote by \( X \) the \( n \times p \) design matrix whose \( i \)th row is \( X_i \). We write \( a \wedge b \) for \( \min(a, b) \) and \( a \vee b \) for \( \max(a, b) \). If \( A \) and \( B \) are two symmetric matrices, \( A \succeq B \) means that \( A - B \) is positive semi-definite, i.e. \( A \) is greater than \( B \) in the positive-definite/Loewner order. The notations \( o_p \), \( O_p \) are used with their standard meanings, see e.g. [41, p. 12] for definitions. For the random variable \( W \), we use the definition \( \| W \|_{L_k} = [\mathbf{E} (|W|^k)]^{1/k} \). For sequences of random variables \( W_n, Z_n \), we use the notation \( W_n = o_{L_k}(Z_n) \) (resp \( W_n = o_L(Z_n) \)) when \( \| W_n \|_{L_k} = O(\| Z_n \|_{L_k}) \) (resp \( \| W_n \|_{L_k} = o(\| Z_n \|_{L_k}) \)). For a vector \( v \) in \( \mathbb{R}^p \), \( \|v\| \) is its Euclidean norm, whereas \( \|v\|_\infty = \max_{1 \leq k \leq p} |v(k)| \). For a function \( f \) from \( \mathbb{R} \) to \( \mathbb{R} \), \( \|f\|_\infty = \sup_{x \in \mathbb{R}} |f(x)| \).

**Remarks** We call

\[
F(\beta) = \frac{1}{n} \sum_{i=1}^{n} \rho_i(\epsilon_i + X'_i \beta_0 - X'_i \beta) + \frac{\tau}{2} \| \beta \|^2. \tag{9}
\]

Note that under our assumptions on \( \rho, \beta(i) \) is defined as the solution of

\[
f(\beta) = 0 \tag{10}
\]

\[
\nabla F = f(\beta) = \frac{1}{n} \sum_{i=1}^{n} -X_i \psi_i(\epsilon_i + X'_i \beta_0 - X'_i \beta) + \tau \beta. \tag{11}
\]

Recall the following important definitions.

**Definition** We call

\[
R_i = \epsilon_i + X'_i \beta_0 - X'_i \beta(i) \text{ (i.e. the residuals),} \tag{12}
\]

\[
S = \frac{1}{n} \sum_{i=1}^{n} \psi'_i(R_i)X_i X'_i, \tag{13}
\]

\[
c_\tau = \frac{1}{n} \text{trace} (S + \tau \text{Id})^{-1}. \tag{14}
\]
Appendix 2: Preliminaries

General remarks

**Proposition 3.1** Let $\beta_1$ and $\beta_2$ be two vectors in $\mathbb{R}^p$. Then, when $\rho_i$’s are convex and twice-differentiable,

\[ \|\beta_1 - \beta_2\| \leq \frac{1}{\tau} \|f(\beta_1) - f(\beta_2)\|. \]  

(15)

**Proof** Let $\beta_1$ and $\beta_2$ be two vectors in $\mathbb{R}^p$. We have by definition

\[ f(\beta_1) - f(\beta_2) = \tau(\beta_1 - \beta_2) + \frac{1}{n} \sum_{i=1}^{n} X_i \left[ \psi_i (\epsilon_i + X_i' \beta_0 - X_i' \beta_2) - \psi_i (\epsilon_i + X_i' \beta_0 - X_i' \beta_1) \right]. \]

We can use the mean value theorem to write

\[ \psi_i (\epsilon_i + X_i' \beta_0 - X_i' \beta_2) - \psi_i (\epsilon_i + X_i' \beta_0 - X_i' \beta_1) = \psi_i' (\gamma_{\epsilon_i + X_i' \beta_0, X_i' \beta_1, X_i' \beta_2}) X_i (\beta_1 - \beta_2), \]

where $\gamma_{\epsilon_i + X_i' \beta_0, X_i' \beta_1, X_i' \beta_2}$ is in the interval $(\epsilon_i + X_i' \beta_0 - X_i' \beta_1, \epsilon_i + X_i' \beta_0 - X_i' \beta_2)$—recall that we do not care about the order of the endpoints in our notation.

Hence,

\[ f(\beta_1) - f(\beta_2) = \tau(\beta_1 - \beta_2) + \frac{1}{n} \sum_{i=1}^{n} \psi_i' (\gamma_{\epsilon_i + X_i' \beta_0, X_i' \beta_1, X_i' \beta_2}) X_i X_i' (\beta_1 - \beta_2), \]

which we write

\[ f(\beta_1) - f(\beta_2) = (S_{\beta_1, \beta_2} + \tau \text{Id}_p)(\beta_1 - \beta_2), \]

(16)

where

\[ S_{\beta_1, \beta_2} = \frac{1}{n} \sum_{i=1}^{n} \psi_i' (\gamma_{\epsilon_i + X_i' \beta_0, X_i' \beta_1, X_i' \beta_2}) X_i X_i'. \]

(17)

This shows that

\[ \beta_1 - \beta_2 = (S_{\beta_1, \beta_2} + \tau \text{Id}_p)^{-1} (f(\beta_1) - f(\beta_2)). \]

Since $\rho_i$’s are convex, $\psi_i' = \rho_i''$ is non-negative and $S_{\beta_1, \beta_2}$ is positive semi-definite. In the semi-definite order, we have $S_{\beta_1, \beta_2} + \tau \text{Id}_p \succeq \tau \text{Id}_p$. In particular,

\[ \|\beta_1 - \beta_2\| \leq \frac{1}{\tau} \|f(\beta_1) - f(\beta_2)\|. \]

\[ \square \]
Proposition 3.1 yields the following lemma.

**Lemma 3.2** For any $\beta_1$,

$$\|\hat{\beta} - \beta_1\| \leq \frac{1}{\tau} \|f(\beta_1)\|.$$  

The lemma is a simple consequence of Eq. (15) since by definition $f(\hat{\beta}) = 0$.

Our strategy in what follows is to come up with “good candidates” for $\beta_1$, for which we can control $f(\beta_1)$ and transfer the information we will glean about the statistical properties of $\beta_1$ to $\hat{\beta}$ through Lemma 3.2.

**On $\|\hat{\beta}\|$ and $\|\hat{\beta} - \beta_0\|$**

We show in the following lemma that $\|\hat{\beta}\|$ and $\|\hat{\beta} - \beta_0\|$ cannot be too large.

**Lemma 3.3** Let us call $W_n(b) = \frac{1}{n} \sum_{i=1}^{n} X_i \psi_i(\epsilon_i + X_i'b)$, $W_n \in \mathbb{R}^p$.

We have, if $D_{\{\psi_i(Y_i)\}_{i=1}^{n}}$ is the $n \times n$ diagonal matrix with $(i,i)$-entry $\psi_i(Y_i)$ = $\psi_i(\epsilon_i + X_i'\beta_0)$,

$$\|\hat{\beta}\| \leq \frac{1}{\tau} \|W_n(\beta_0)\| = \frac{1}{\tau} \left( \frac{1}{n^2} \sum_{i=1}^{n} \psi_i(X_i'X_i)D_{\{\psi_i(Y_i)\}_{i=1}^{n}} \right)^{1/2}.$$

and if $D_{\{\psi_i(\epsilon_i)\}_{i=1}^{n}}$ is the $n \times n$ diagonal matrix with $(i,i)$-entry $\psi_i(\epsilon_i)$,

$$\|\hat{\beta} - \beta_0\| \leq \|\beta_0\| + \frac{1}{\tau} \|W_n(0)\| = \|\beta_0\| + \frac{1}{\tau} \left( \frac{1}{n^2} \sum_{i=1}^{n} \psi_i(X_i'X_i)D_{\{\psi_i(\epsilon_i)\}_{i=1}^{n}} \right)^{1/2}.$$

Also,

$$\|W_n(\beta_0)\|^2 \leq \frac{1}{n} \sum_{i=1}^{n} \psi_i(X_i'X_i)^{1/2}.$$

Therefore, under our assumptions O1–O6,

$$\mathbb{E}\left(\|\hat{\beta}\|^2\right) \leq \frac{1}{\tau^2} \frac{P}{n} C^2 \text{polyLog}(n), \quad \text{and} \quad (18)$$

$$\mathbb{E}\left(\|\hat{\beta}\|^4\right) \leq \frac{1}{\tau^4} C \text{polyLog}(n). \quad (19)$$

Similarly, for any finite $k$,

$$\mathbb{E}\left(\|\hat{\beta} - \beta_0\|^2\right) \leq C_k \left[ \|\beta_0\|^k + \text{polyLog}(n)/\tau^k \right].$$
In the case \( k = 2 \), we have the more precise bound
\[
\mathbb{E} \left( \| \hat{\beta} - \beta_0 \|_2^2 \right) \leq 2 \left[ \| \beta_0 \|^2 + \frac{p/n}{\tau^2} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left( \psi_i^2(\epsilon_i) \right) \right].
\]

**Proof** The first and key inequality simply comes from applying Lemma 3.2 with \( \beta_1 = 0 \), after noticing that \( f(0) = -W_n(\beta_0) \).

The second one comes from using \( \beta_1 = \beta_0 \) and noticing that \( f(\beta_0) = -W_n(0) + \tau \beta_0 \).

We note that under our assumptions, according to Lemma 3.38,
\[
\| \|X'X/n\|_2 \| = O_{L_k}(\text{polyLog}(n)), \quad \text{and} \quad \frac{1}{n} \sum_{i=1}^{n} \psi_i^2(Y_i) \leq \frac{1}{n} \sum_{i=1}^{n} \| \psi_i \|_\infty = O(1),
\]
which gives all the results about \( L_k \) bounds.

The last result about \( k = 2 \) follows from computing
\[
\mathbb{E} \left( \| W_n(0) \|_2^2 \right) = \frac{p}{n} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left( \psi_i^2(\epsilon_i) \right)
\]
and using the bound
\[
\| \hat{\beta} - \beta_0 \|_2^2 \leq 2 \| \beta_0 \|^2 + \frac{2}{\tau^2} \| W_n(0) \|_2^2.
\]

**About \( \mathbb{E} \left( \| W_n(0) \|_2^2 \right) **

For the sake of clarity, we now explain in detail why \( \mathbb{E} \left( \| W_n(0) \|_2^2 \right) = \frac{p}{n} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left( \psi_i^2(\epsilon_i) \right) \).

Recall that
\[
W_n(0) = \frac{1}{n} \sum_{i=1}^{n} X_i \psi_i(\epsilon_i).
\]

Hence,
\[
\| W_n(0) \|_2^2 = W_n(0)' W_n(0) = \frac{1}{n^2} \sum_{i,j} X_i' X_j \psi_i(\epsilon_i) \psi_j(\epsilon_j).
\]

Since \( \epsilon_i \)'s and \( X_i \)'s are independent, using the fact that \( \mathbb{E} (X_i) = 0 \) and that \( \psi_i \)'s are bounded, we have
\[
\mathbb{E} \left( \| W_n(0) \|_2^2 \right) = \frac{1}{n^2} \sum_{i=1}^{n} \mathbb{E} \left( \| X_i \|_2^2 \right) \mathbb{E} \left( \psi_i^2(\epsilon_i) \right).
\]

Now, \( \mathbb{E} \left( \| X_i \|_2^2 \right) = \mathbb{E} \left( \text{trace} (X_i'X_i) \right) = \mathbb{E} \left( \text{trace} (X_i X_i') \right) \). Since \( \mathbb{E} (X_i) = 0 \), \( \mathbb{E} (X_i X_i') = \text{cov} (X_i) = \text{Id}_p \). Hence, \( \mathbb{E} (\| X_i \|_2^2) = p \). And we conclude that
\[
\mathbb{E} \left( \| W_n(0) \|_2^2 \right) = \frac{1}{n^2} \sum_{i=1}^{n} p \mathbb{E} \left( \psi_i^2(\epsilon_i) \right).
\]

This gives the announced result. \( \square \)
Appendix 3: Approximating \( \hat{\beta} \) by \( \hat{\beta}(i) \): leave-one-observation-out

We consider the situation where we leave the \( i \)th observation, \((X_i, Y_i)\), out. By definition,

\[
\hat{\beta}(i) = \arg\min_{\beta \in \mathbb{R}^p} F_i(\beta), \quad \text{where} \quad F_i(\beta) = \frac{1}{n} \sum_{j \neq i} \rho_j \left( \epsilon_j + X'_j \beta_0 - X'_j \beta \right) + \frac{\tau}{2} \| \beta \|^2.
\]

We call

\[
\tilde{r}_{j,(i)} = \epsilon_j - X'_j(\hat{\beta}(i) - \beta_0) \quad \text{and} \quad S_i = \frac{1}{n} \sum_{j \neq i} \psi_j(\tilde{r}_{j,(i)}) X_j X'_j.
\]

We also call

\[
f_i(\beta) = -\frac{1}{n} \sum_{j \neq i} X'_j \psi_j \left( \epsilon_j + X'_j \beta_0 - X'_j \beta \right) + \tau \beta = f(\beta) + \frac{1}{n} X_i \psi_i \left( \epsilon_i - X'_i (\beta - \beta_0) \right).
\]

We have of course

\[
f_i(\hat{\beta}(i)) = 0.
\]

Let us consider

\[
\tilde{\beta}_i = \hat{\beta}(i) + \frac{1}{n} (S_i + \tau \text{Id})^{-1} X_i \psi_i (\text{prox}(c_i \rho_i)(\tilde{r}_{i,(i)})) \triangleq \hat{\beta}(i) + \eta_i, \quad (20)
\]

where

\[
c_i = \frac{1}{n} X'_i (S_i + \tau \text{Id})^{-1} X_i, \quad \text{and} \quad (21)
\]

\[
\eta_i = \frac{1}{n} (S_i + \tau \text{Id})^{-1} X_i \psi_i (\text{prox}(c_i \rho_i)(\tilde{r}_{i,(i)})). \quad (22)
\]

These definitions and the approximations they will imply can be understood in light of the probabilistic heuristics we derived for a related problem in [16, 17]. The interested reader is referred to those papers—where we made a large effort to explain our intuitive ideas—for more information and intuition; given page limit requirements, we do not give a complete heuristic derivation of our results and refer the reader to Sect. 2.2.2 for a detailed explanation of our strategy. We note however that the rigorous proof requires refinements over the intuitive ideas. Those aspects are of a more technical nature and become apparent only through the analysis that we present here.

One of our aims is to show Theorem 3.9 below, which shows that we can very accurately approximate \( \hat{\beta} \) by \( \tilde{\beta}_i \). Note that the statistical properties of \( \tilde{\beta}_i \) are easier to understand that those of \( \hat{\beta} \); our high-quality approximations will allow us to transfer our understanding of \( \tilde{\beta}_i \) to \( \hat{\beta} \).
Deterministic bounds

**Proposition 3.4** We have, with $\tilde{\beta}_i$ defined in Eq. (20),

$$\|\hat{\beta} - \tilde{\beta}_i\| \leq \frac{1}{\tau} \|\mathcal{R}_i\|.$$

where

$$\mathcal{R}_i = \frac{1}{n} \sum_{j \neq i} \left[ \psi_j' \left( \gamma^*(X_j, \hat{\beta}_i, \eta_i) \right) - \psi_j' \left( \tilde{r}_{j, (i)} \right) \right] X_j X'_j \eta_i.$$  \hfill (24)

and $\gamma^*(X_j, \hat{\beta}_i, \eta_i)$ is in the (“unordered”) interval $(\tilde{r}_{j, (i)}, \tilde{r}_{j, (i)} - X'_i \eta_i) = (\tilde{r}_{j, (i)} - X'_i \eta_i).$

**Proof** The proof and strategy are similar to the corresponding ones in [15]. However, since there are delicate cancellations in the argument, we give all the details.

We recall that $Y_i = \epsilon_i + X'_i \beta_0.$

Since $f_i(\hat{\beta}_i) = 0$, and $\hat{\beta}_i = \tilde{\beta}_i + \eta_i,$

$$f(\tilde{\beta}_i) = f(\tilde{\beta}_i) - f_i(\tilde{\beta}_i) = -\frac{1}{n} X_i \psi_i(Y_i - X'_i \tilde{\beta}_i)$$

$$+ \frac{1}{n} \sum_{j \neq i} X_j \left[ \psi_j \left( Y_j - X'_i \tilde{\beta}_i \right) - \psi_j \left( Y_j - X'_i \left( \tilde{\beta}_i + \eta_i \right) \right) \right] + \tau \eta_i.$$

By the mean-value theorem, we also have

$$\psi_j \left( Y_j - X'_i \tilde{\beta}_i \right) - \psi_j \left( Y_j - X'_i \left( \tilde{\beta}_i + \eta_i \right) \right) = \psi_j' \left( \tilde{r}_{j, (i)} \right) X'_j \eta_i$$

$$+ \left[ \psi_j' \left( \gamma^*(X_j, \hat{\beta}_i, \eta_i) \right) - \psi_j' \left( \tilde{r}_{j, (i)} \right) \right] X'_j \eta_i,$$

where $\gamma^*(X_j, \hat{\beta}_i, \eta_i)$ is in the (“unordered”) interval $(Y_j - X'_i \tilde{\beta}_i, Y_j - X'_i \left( \tilde{\beta}_i + \eta_i \right))$, i.e. $(\tilde{r}_{j, (i)}, \tilde{r}_{j, (i)} - X'_i \eta_i).$

Hence, if $\mathcal{R}_i$ is the quantity defined in Eq. (24),

$$\frac{1}{n} \sum_{j \neq i} X_j \left[ \psi_j(Y_j - X'_i \tilde{\beta}_i) - \psi_j(Y_j - X'_i \left( \tilde{\beta}_i + \eta_i \right)) \right]$$

$$= \frac{1}{n} \sum_{j \neq i} \psi_j' \left( \tilde{r}_{j, (i)} \right) X_j X'_j \eta_i + \mathcal{R}_i,$$

$$= S_i \eta_i + \mathcal{R}_i.$$

In light of the previous simplifications, we have, using

$$f(\beta) = f_i(\beta) - \frac{1}{n} X_i \psi_i(Y_i - X'_i \beta)$$

and $f_i(\tilde{\beta}_i) = 0,$
the equality
\[ f(\tilde{\beta}_i) = -\frac{1}{n} X_i \psi_i(Y_i - X'_i \tilde{\beta}_i) + (S_i + \tau \text{Id})\eta_i + \mathcal{R}_i. \]

Since by definition, \( \eta_i = \frac{1}{n} (S_i + \tau \text{Id})^{-1} X_i \psi_i(\text{prox}(c_i \rho_i)(\tilde{r}_{i,(i)})) \),

\[ (S_i + \tau \text{Id})\eta_i = \frac{1}{n} X_i \psi_i(\text{prox}(c_i \rho_i)(\tilde{r}_{i,(i)})). \]

In other respects,

\[ Y_i - X'_i \tilde{\beta}_i = \tilde{r}_{i,(i)} - c_i \psi_i(\text{prox}(c_i \rho_i)(\tilde{r}_{i,(i)})). \]

When \( \rho \) is differentiable, \( x - c \psi(\text{prox}(c \rho)(x)) = \text{prox}(c \rho)(x) \) almost by definition of the proximal mapping (see Lemma 3.31). Therefore, \( Y_i - X'_i \tilde{\beta}_i = \text{prox}(c_i \rho_i)(\tilde{r}_{i,(i)}) \) and

\[ -\frac{1}{n} X_i \psi_i(Y_i - X'_i \tilde{\beta}_i) + (S_i + \tau \text{Id})\eta_i \]
\[ = \frac{1}{n} X_i \left[ -\psi_i(\text{prox}(c_i \rho_i)(\tilde{r}_{i,(i)})) + \psi_i(\text{prox}(c_i \rho_i)(\tilde{r}_{i,(i)})) \right] = 0. \]

We conclude that
\[ f(\tilde{\beta}_i) = \mathcal{R}_i. \]

Applying Lemma 3.2, we see that
\[ \|\hat{\beta} - \tilde{\beta}_i\| \leq \frac{1}{\tau} \|\mathcal{R}_i\|. \]

\[ \square \]

**On \( \mathcal{R}_i \)**

Clearly, controlling \( \mathcal{R}_i \) is the key to controlling \( \|\hat{\beta} - \tilde{\beta}_i\| \), so we need to develop insights into \( \mathcal{R}_i \).

**Lemma 3.5** We have
\[ \|\eta_i\| \leq \frac{1}{\sqrt{n} \tau} \frac{\|X_i\|}{\sqrt{n}} |\psi_i(\tilde{r}_{i,(i)})|, \quad (25) \]

and
\[ \|\mathcal{R}_i\| \leq \|\tilde{\Sigma}\|_2 \sup_{j \neq i} \left| \psi'_j \left( y^*(X_j, \tilde{\beta}_{(i)}, \eta_i) \right) - \psi'_j (\tilde{r}_{j,(i)}) \right| \frac{1}{\sqrt{n} \tau} \frac{\|X_i\|}{\sqrt{n}} |\psi_i(\tilde{r}_{i,(i)})|. \]

\[ (26) \]
We note that under our assumptions, we have
\[ \left| \psi_i(\tilde{r}_{i,(i)}) \right| \leq \| \psi_i \|_\infty \leq C \text{polyLog}(n) \]
and, using Lemma 3.35
\[ \sup_i \frac{\|X_i\|}{\sqrt{n}} = O_{L_k}(\sup_i |\lambda_i|). \]

Proof This proof is essentially obvious. We refer the reader to a corresponding one given in [15] in case details are needed.
\[ \square \]

On \( \gamma^*(X_j, \hat{\beta}(i), \eta_i) \) and related quantities

We now show how to control \( \frac{1}{\sqrt{n}} \sup_{j \neq i} \left| \psi_{j}'(\gamma^*(X_j, \hat{\beta}(i), \eta_i)) - \psi_{j}'(\tilde{r}_{j,(i)}) \right| \), which is essential for turning Eq. (23) into a useful bound. We proceed by first getting a better understanding of Eq. (26).

Lemma 3.6 Suppose, as in our assumption O3, that \( \psi_i' \) is \( L_i(n) \)-Lipschitz. Then,
\[ \sup_{j \neq i} \left| \psi_{j}'(\gamma^*(X_j, \hat{\beta}(i), \eta_i)) - \psi_{j}'(\tilde{r}_{j,(i)}) \right| \leq \left[ \sup_{1 \leq i \leq n} L_i(n) \right] \sup_{j \neq i} |X_j'\eta_i|. \]

Proof By definition, we have
\[ |\gamma^*(X_j, \hat{\beta}(i), \eta_i) - \tilde{r}_{j,(i)}| \leq |X_j'\eta_i|. \]
The bound follows immediately, using the fact that \( \psi_i' \) is \( L_i(n) \)-Lipschitz.
\[ \square \]

Stochastic aspects

Recall that we have by definition
\[ X_j'\eta_i = \psi_i(\text{prox}(c_i \rho_i)(\tilde{r}_{i,(i)})) \frac{1}{n} X_j'(S_i + \tau \text{Id}_p)^{-1} X_i. \]

We can therefore bound \( \|R_i\| \) by
\[ \|R_i\| \leq \left[ \sup_{j \neq i} \frac{|X_j'(S_i + \tau \text{Id}_p)^{-1} X_i|}{n} \right] \sup_{1 \leq i \leq n} L_i(n) \frac{\|X_i\|}{\sqrt{n\tau}} \frac{\|\hat{\Sigma}\|_2}{\sqrt{n}} \|\psi_i(\text{prox}(c_i \rho_i)(\tilde{r}_{i,(i)}))\|^2 \|\psi_i\|_\infty. \]
Therefore, we also have
\[ \|R_i\| \leq \left[ \sup_{j \neq i} \frac{|X_j'(S_i + \tau \text{Id}_p)^{-1} X_i|}{n} \right] \frac{\sup_{1 \leq i \leq n} L_i(n)}{\sqrt{n\tau}} \frac{\|X_i\|}{\sqrt{n}} \frac{\|\hat{\Sigma}\|_2}{\sqrt{n}} \|\psi_i\|_\infty^2. \]
This bound on $\|R_i\|$ shows that we can control $\|\hat{\beta} - \tilde{\beta}_i\|$ in $L_k$ provided we can control each term in the above product in $L_{3k}$, by appealing to Holder’s inequality and Proposition 3.4.

We now turn our attention to the various elements of the bound on $\|R_i\|$ and show that we can control them under our assumptions.

On $\sup_{j \neq i} |X_j'(S_i + \tau \mathrm{Id})^{-1} X_i / n|$ We will control $X_j'(S_i + \tau \mathrm{Id})^{-1} X_i / n$ by appealing to Lemma 3.36.

**Lemma 3.7** Suppose $X_i$’s are independent and satisfy Assumption O4; suppose $\lambda_i$’s satisfy O6. Then

$$\sup_{j \neq i} |X_j'(S_i + \tau \mathrm{Id})^{-1} X_i / n| \leq \frac{1}{\sqrt{n}} \sup_{j \neq i} \frac{\|X_j\|}{\tau \sqrt{n}} \text{polyLog}(n)$$

in $L_k$, for any finite $k$. Note also that under our Assumption O4, for any finite $k$,

$$\sup_{j \neq i} \|X_j\|/\sqrt{n} = O_{L_k}(1).$$

**Proof** The proof follows from that of Lemma 2.3 in [15]. Indeed,

$$|X_j'(S_i + \tau \mathrm{Id})^{-1} X_i / n| = |\lambda_i \lambda_j| |X_j'(S_i + \tau \mathrm{Id})^{-1} X_i / n|.$$  

The proof of Lemma 2.3 in [15] shows that

$$\sup_{j \neq i} |X_j'(S_i + \tau \mathrm{Id})^{-1} X_i / n| \leq \frac{1}{\sqrt{n}} \sup_{j \neq i} \frac{\|X_j\|}{\tau \sqrt{n}} \text{polyLog}(n)/c_n^{1/2}$$

in $L_k$, when $\sup_{j \neq i} \frac{\|X_j\|}{\tau \sqrt{n}} = O_{L_k}(1)$; this latter result is shown in [15] (see the discussion after Lemma 2.3 or Lemma 3.35 in the current “Appendix” applied to $F_i(\mathcal{X}_i) = \|\mathcal{X}_i\|$ and noting that $E(\|\mathcal{X}_i\|) \leq \sqrt{E(\|\mathcal{X}_i\|^2)} = \sqrt{p}$).

Now our assumptions O6 concerning $\sup_i |\lambda_i| = O_{L_k}(\text{polyLog}(n))$ guarantee that the bounds we announced are valid. \qed

**Consequences**

We have the following result. Recall that $\psi'_i$ is assumed to be Lipschitz with Lipschitz constant $L_i(n)$.

**Proposition 3.8** Under Assumptions O1–O6, we have

$$\|R_i\| = O_{L_k} \left( \frac{\sup_{1 \leq i \leq n} L_i(n) ||\psi_i\|_\infty^2}{n \tau \text{polyLog}(n)} \right).$$
Furthermore, the same bound holds for \( \sup_{1 \leq i \leq n} \| R_i \| \) with \( \sup_{1 \leq i \leq n} \| \psi_i \|_\infty^2 \) (instead of \( \| \psi_i \|_\infty^\infty \)) in the right-hand side.

**Proof** The proof follows by aggregating all the intermediate results we had, using Holder's inequality and noticing that under our assumptions, \( \| \hat{\Sigma} \|_2 = O_Lk (\sup_i \lambda^2 c_i^{-1}) = O_Lk (\text{polyLog}(n)) \). This latter result is shown in Lemma 3.38.

The statement concerning \( \sup_{1 \leq i \leq n} \| R_i \| \) follows by the same arguments. \( \blacksquare \)

We can now prove and state the following result, which relates residuals to leave-one-out-prediction errors and give a way to do online update from \( \hat{\beta}_{(i)} \) to \( \hat{\beta} \).

We recall that \( \tilde{\beta}_i \) is defined in Eq. (20).

**Theorem 3.9** Under Assumptions O1–O7, we have, for any fixed \( k \), when \( \tau \) is held fixed and \( L_i(n) \leq Cn^\alpha \),

\[
\sup_{1 \leq i \leq n} \| \hat{\beta} - \tilde{\beta}_i \| = O_Lk \left( \frac{\text{polyLog}(n)}{n^{1-\alpha}} \right).
\]

In particular, we have

\[
\forall 1 \leq i \leq n, \quad E \left( \| \hat{\beta} - \tilde{\beta}_i \|^2 \right) = O(\text{polyLog}(n)/n^{2-2\alpha}).
\]

Also,

\[
\sup_{1 \leq i \leq n} \sup_{j \neq i} | \tilde{r}_{j,(i)} - R_j | = O_Lk \left( \frac{\text{polyLog}(n)}{n^{1/2-\alpha}} \right).
\]

Finally,

\[
\sup_{1 \leq i \leq n} | R_i - \text{prox}(c_i \rho_i) (\tilde{r}_{i,(i)}) | = O_Lk \left( \frac{\text{polyLog}(n)}{n^{1/2-\alpha}} \right). \tag{27}
\]

We note that we could state a slightly finer result involving \( L_i(n) \) and various powers of \( \| \psi_i \|_\infty \). However, we will not need such fine results in what follows, so we opt for slightly coarser but easier-to-state statements.

**Proof** The first two results simply follow from our work on \( \| R_i \| \).

The third result follows from the coarse bound

\[
\sup_{j \neq i} | \tilde{r}_{j,(i)} - R_j | = \sup_{j \neq i} \left| X_j ' (\hat{\beta} - \tilde{\beta}_{(i)}) \right| \leq \sup_{j \neq i} \left| X_j ' (\hat{\beta} - \tilde{\beta}_i) \right| + \sup_{j \neq i} | X_j ' (\tilde{\beta}_i - \tilde{\beta}_{(i)}) |,
\]

\[
\leq \left( \sup_{1 \leq j \leq n} \frac{\| X_j \|}{\sqrt{n}} \right) \sqrt{n} \| \hat{\beta} - \tilde{\beta}_i \| + \sup_{j \neq i} | X_j ' \eta_i |,
\]

and the fact that \( \left( \sup_{1 \leq j \leq n} \frac{\| X_j \|}{\sqrt{n}} \right) = O_Lk (\text{polyLog}(n)) \) under our assumptions. Our results on \( \| \hat{\beta} - \tilde{\beta}_i \| \) give control of the first term. Control of the second term follows from Lemma 3.7 and the assumption that \( \psi_i \) is bounded by \( C \text{polyLog}(n) \).
Let us now turn to the final result, i.e. the approximation of the residual \( R_i \) by a non-linear function of the leave-one-out prediction error \( \tilde{r}_{i,(i)} \). Recall that

\[
R_i = \epsilon_i + X_i' \beta_0 - X_i' \tilde{\beta} = \epsilon_i + X_i' \beta_0 - X_i' \tilde{\beta}_i - X_i'(\tilde{\beta} - \tilde{\beta}_i).
\]

Now, given the definition of \( \tilde{\beta}_i \), we have

\[
X_i' \tilde{\beta}_i = X_i' \tilde{\beta}_{i,(i)} + c_i \psi_i [\text{prox}(c_i \rho_i)(\tilde{r}_{i,(i)})].
\]

Hence, since almost by definition, if \( y = \text{prox}(c \rho)(x) \), \( y + c \psi(y) = x \), we get

\[
\epsilon_i + X_i' \beta_0 - X_i' \tilde{\beta}_i = \tilde{r}_{i,(i)} - c_i \psi_i [\text{prox}(c_i \rho_i)(\tilde{r}_{i,(i)})] = \text{prox}(c_i \rho_i)(\tilde{r}_{i,(i)}).
\]

So we have established that

\[
sup_i |R_i - \text{prox}(c_i \rho_i)(\tilde{r}_{i,(i)})| = sup_i |X_i'(\tilde{\beta}_i - \tilde{\beta})|
\]

and the result follows from our previous bounds. \( \square \)

**On the limiting variance of \( \| \hat{\beta} \|^2 \) and \( \| \hat{\beta} - \beta_0 \|^2 \)**

An interesting consequence of our leave-one-observation-out work is that we can use the ideas and approximations developed above to show that \( \| \hat{\beta} - \beta_0 \| \) and \( \| \hat{\beta} \| \) are asymptotically deterministic (in other words, they can be approximated asymptotically by deterministic sequences).

**Proposition 3.10** Under our assumptions \( O1\text{--}O7 \),

\[
\text{var} \left( \| \hat{\beta} \|^2 \right) \to 0 \text{ as } n \to \infty.
\]

Therefore \( \| \hat{\beta} \|^2 \) has a deterministic equivalent in probability and in \( L_2 \).

More precisely, we have

\[
\text{var} \left( \| \hat{\beta} \|^2 \right) = O \left( \frac{\text{polyLog}(n)}{n^{1-2\alpha}} \right).
\]

The same results are true for \( \text{var} \left( \| \hat{\beta} - \beta_0 \|^2 \right) \) provided \( \| \beta_0 \| = O(\text{polyLog}(n)) \), as in Assumption \( O7 \).

**Proof** We use the Burkholder/Efron–Stein inequality to show that \( \text{var} \left( \| \hat{\beta} \|^2 \right) \) goes to 0 as \( n \to \infty \). In what follows, we rely on our approximations and our assumptions to have enough moments for all the expectations of the type \( \mathbb{E} \left( \| \hat{\beta} \|^{2k} \right) \) to be bounded like \( \text{polyLog}(n)/\tau^{2k} \). Note that this the content of our Lemma 3.3.
Recall the Efron–Stein inequality [11]: if \( W \) is a function of \( n \) independent random variables, and \( W_{(i)} \) is any function of all those random variables except the \( i \)th,

\[
\text{var}(W) \leq \sum_{i=1}^{n} \text{var}(W - W_{(i)}) \leq \sum_{i=1}^{n} E\left((W - W_{(i)})^2\right).
\]

In our arguments below, \( \|\hat{\beta}\|^2 \) plays the role of \( W \) and \( \|\hat{\beta}_{(i)}\|^2 \) plays the role of \( W_{(i)} \).

We first observe that

\[
E\left(\|\hat{\beta}\|^2 - \|\hat{\beta}_{(i)}\|^2\right) \leq 2\left[E\left(\|\hat{\beta}\|^2 - \|\hat{\beta}_{i}\|^2\right) + E\left(\|\hat{\beta}_{i}\|^2 - \|\hat{\beta}_{(i)}\|^2\right)\right].
\]

Of course, using the fact that \( \hat{\beta} = \hat{\beta} - \hat{\beta}_{i} + \hat{\beta}_{i} \) and \( \|\hat{\beta}\|^2 - \|\hat{\beta}_{i}\|^2 = [(\hat{\beta} - \hat{\beta}_{i})(\hat{\beta} + \hat{\beta}_{i})]^2 \), and hence \( (\hat{\beta} - \hat{\beta}_{i})(\hat{\beta} + \hat{\beta}_{i}) = 2(\hat{\beta} - \hat{\beta}_{i})\hat{\beta} - \|\hat{\beta} - \hat{\beta}_{i}\|^2 \), we have

\[
\|\hat{\beta}\|^2 - \|\hat{\beta}_{i}\|^2 = O_{\beta}(\|\hat{\beta} - \hat{\beta}_{i}\|^4) + \sqrt{O_{\beta}((\text{polyLog}(n))\|\hat{\beta} - \hat{\beta}_{i}\|^4)},
\]

by the Cauchy–Schwarz inequality, since \( E\left(\|\hat{\beta}\|^k\right) \) exists and is bounded by \( K\text{polyLog}(n)/\tau^k \).

Using the results of Theorem 3.9, we see that

\[
E\left(\|\hat{\beta}\|^2 - \|\hat{\beta}_{i}\|^2\right) = O\left(\frac{\text{polyLog}(n)}{n^{2-2\alpha}}\right) = o(n^{-1}),
\]

provided \( \alpha < 1/2 \).

On the other hand, given the definition in Eq. (20),

\[
\|\hat{\beta}_{i}\|^2 - \|\hat{\beta}_{(i)}\|^2 = \frac{1}{n} \hat{\beta}'_{(i)}(S_{i} + \tau\text{Id})^{-1}X_{i}\psi_{i}(\text{prox}(c_{i}\rho_{i})(\hat{r}_{i,(i)})) + \frac{1}{n^2}X'_{i}(S_{i} + \tau\text{Id})^{-2}X_{i}\psi_{i}^{2}(\text{prox}(c_{i}\rho_{i})(\hat{r}_{i,(i)})).
\]

Since \( \hat{\beta}_{(i)} \) and \( S_{i} \) are independent of \( X_{i} \), and \( \|\hat{\beta}_{(i)}(S_{i} + \tau\text{Id})^{-1}\|_{2} \leq 1/\tau \), \( \hat{\beta}'_{(i)}(S_{i} + \tau\text{Id})^{-1}X_{i} = O_{\beta}(\lambda_{i}\|\hat{\beta}_{(i)}\|/\mathfrak{c}_{n}^{1/2}) \), using our assumptions \( \text{O4} \) on \( X_{i} \) applied to linear forms. Recall also that \( \sup_{i}\|\psi_{i}\|_{\infty} = O(\text{polyLog}(n)) \). Therefore, we see that both terms are \( O_{\beta}(\text{polyLog}(n)/n\mathfrak{c}_{n}^{1/2}) \).

We conclude that then

\[
E\left(\|\hat{\beta}_{i}\|^2 - \|\hat{\beta}_{(i)}\|^2\right) = O\left(\frac{\text{polyLog}(n)}{n^{2}}\right).
\]
Taking $W = \|\hat{\beta}\|^2$ and $W(i) = \|\hat{\beta}(i)\|^2$ in the Efron-Stein inequality, we clearly see that

\[
\text{var} \left( \|\hat{\beta}\|^2 \right) = O \left( \frac{\text{polyLog}(n)}{n^{1-2\alpha}} \right) = o(1).
\]

This shows that $\|\hat{\beta}\|^2$ has a deterministic equivalent in probability and in $L_2$.

- About $\|\hat{\beta} - \beta_0\|$. The results are obtained in a similar fashion using our bound on $\|\hat{\beta} - \beta_0\|$ and replacing everywhere in the arguments $\|\hat{\beta}\|$ by $\|\beta - \beta_0\|$, and $\|\tilde{\beta}_i\|$ by $\|\tilde{\beta}_i - \beta_0\|$. The condition $\|\beta_0\| = O(\text{polyLog}(n))$ plays a role to guarantee in Lemma 3.3 that $\|\hat{\beta} - \beta_0\| = O_{L_k}(\text{polyLog}(n))$.

Let us now provide some more details. As above, we write

\[
E \left( \|\hat{\beta} - \beta_0\|^2 - \|\hat{\beta}(i) - \beta_0\|^2 \right)^2 \leq 2 \left[ E \left( \|\hat{\beta} - \beta_0\|^2 - \|\tilde{\beta}_i - \beta_0\|^2 \right)^2 \right. + E \left. \left( \|\tilde{\beta}_i - \beta_0\|^2 - \|\hat{\beta}(i) - \beta_0\|^2 \right)^2 \right].
\]

Of course, using the fact that $\hat{\beta} = \hat{\beta}_i + \tilde{\beta}_i$ and $\|\hat{\beta} - \beta_0\|^2 - \|\tilde{\beta}_i - \beta_0\|^2 = [(\hat{\beta} - \tilde{\beta}_i)((\hat{\beta} + \tilde{\beta}_i - 2\beta_0))^2$, and hence $(\hat{\beta} - \tilde{\beta}_i)((\hat{\beta} + \tilde{\beta}_i - 2\beta_0) = 2(\hat{\beta} - \tilde{\beta}_i)(\beta - \beta_0) - \|\hat{\beta} - \tilde{\beta}_i\|^2$, we have

\[
\|\hat{\beta} - \beta_0\|^2 - \|\tilde{\beta}_i - \beta_0\|^2 = O_{L_1}(\|\hat{\beta} - \tilde{\beta}_i\|^4) + \sqrt{O_{L_1}(\text{polyLog}(n))}\|\hat{\beta} - \tilde{\beta}_i\|^4,
\]

by the Cauchy–Schwarz inequality, since $E \left( \|\hat{\beta} - \beta_0\|^k \right)$ exists and is bounded by $K\text{polyLog}(n)/\tau^k$. This latter fact follows from Assumption O7 and Lemma 3.3.

Using the results of Theorem 3.9, we see that

\[
E \left( \|\hat{\beta} - \beta_0\|^2 - \|\tilde{\beta}_i - \beta_0\|^2 \right)^2 = O \left( \frac{\text{polyLog}(n)}{n^{2-2\alpha}} \right) = o(n^{-1}),
\]

provided $\alpha < 1/2$.

As above, given the definition in Eq. (20),

\[
\|\tilde{\beta}_i - \beta_0\|^2 - \|\tilde{\beta}(i) - \beta_0\|^2 = \frac{1}{n} \left( \hat{\beta}(i) - \beta_0 \right)'(S_i + \tau \text{Id})^{-1} X_i \psi_i(\text{prox}(c_i \hat{\rho}_i)(\tilde{r}_i(i))) + \frac{1}{n^2} X_i'(S_i + \tau \text{Id})^{-2} X_i \psi_i^2(\text{prox}(c_i \hat{\rho}_i)(\tilde{r}_i(i))).
\]

Since $\hat{\beta}(i) - \beta_0$ and $S_i$ are independent of $X_i$, and $\|(S_i + \tau \text{Id})^{-1}\|_2 \leq 1/\tau$, $(\hat{\beta}(i) - \beta_0)'(S_i + \tau \text{Id})^{-1} X_i = O_{L_2}(\|\hat{\beta}(i) - \beta_0\|/c_i^{n/2})$, using our assumptions O4 on $X_i$ applied to linear forms. Recall also that $\sup_i \|\psi_i\|_\infty = O(\text{polyLog}(n))$. Therefore, we see that both terms are $O_{L_2}(\text{polyLog}(n)/nc_i^{n/2})$.
We conclude that then
\[
E \left( \left| \| \hat{\beta}_i - \beta_0 \| - \| \hat{\beta}_{(i)} - \beta_0 \| \right|^2 \right) = O \left( \frac{\text{polyLog}(n)}{n^2} \right).
\]

Taking now \( W = \| \hat{\beta} - \beta_0 \| \) and \( W_{(i)} = \| \hat{\beta}_{(i)} - \beta_0 \| \) in the Efron-Stein inequality, we clearly see that
\[
\text{var} \left( \| \hat{\beta} - \beta_0 \| \right) = O \left( \frac{\text{polyLog}(n)}{n^{1-2\alpha}} \right) = o(1).
\]

This shows that \( \| \hat{\beta} - \beta_0 \| \) has a deterministic equivalent in probability and in \( L_2 \).

\[\square\]

Appendix 4: Leaving out a predictor

In this second step of the proof, we do need at various points that the entries of the vector \( X_i \) be independent, whereas as we showed before, it is not important when studying what happens when we leave out an observation.

Let \( V \) be the \( n \times (p - 1) \) matrix corresponding to the first \( (p - 1) \) columns of the design matrix \( X \). We call \( V_i \) in \( \mathbb{R}^{p-1} \) the vector corresponding to the first \( p - 1 \) entries of \( X_i \), i.e. \( V_i' = (X_i(1), \ldots, X_i(p-1)) \). We call \( X(p) \) the vector in \( \mathbb{R}^n \) with \( j \)th entry \( X_j(p) \), i.e. the \( p \)-th entry of the vector \( X_j \). When this does not create problems, we also use the standard notation \( X_{j,p} \) for \( X_j(p) \).

We use the notation \( \beta_0 = (\gamma_0', \beta_0(p))' \), i.e. \( \gamma_0 \) is the vector corresponding to the first \( p - 1 \) coordinates of \( \beta_0 \).

Let us call \( \hat{\gamma} \) the solution of
\[
\hat{\gamma} = \arg\min_{\gamma \in \mathbb{R}^{p-1}} \frac{1}{n} \sum_{i=1}^{n} \rho_i (\epsilon_i - V_i'(\gamma - \gamma_0)) + \frac{\tau}{2} \| \gamma \|_2^2.
\] (28)

Note that \( \left( \hat{\gamma}, 0 \right) \) is the solution of the original optimization problem (3) when \( X_i(p) \) is replaced by 0.

In this part of the paper, we will rely heavily on the following definitions:

**Definition** We call the residuals corresponding to this optimization problem \( \{ r_i, [p] \}_{i=1}^{n} \), in other words
\[
r_i, [p] = \epsilon_i + V_i' \gamma_0 - V_i' \hat{\gamma}.
\]

We call
\[
u_p = \frac{1}{n} \sum_{i=1}^{n} \psi_i'(r_i, [p]) V_i X_i(p), \quad \text{and} \quad \mathcal{S}_p = \frac{1}{n} \sum_{i=1}^{n} \psi_i'(r_i, [p]) V_i V_i'.
\]
Note that $u_p \in \mathbb{R}^{p-1}$ and $\mathcal{S}_p$ is $(p-1) \times (p-1)$. We call

$$
\xi_n \triangleq \frac{1}{n} \sum_{i=1}^{n} X_i^2(p) \psi'_i(r_i, [p]) - u'_p (\mathcal{S}_p + \tau \text{Id})^{-1} u_p,
$$

and

$$
N_p \triangleq \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i(p) \psi_i(r_i, [p]).
$$

We will show later, in “On $\xi_n$” of Appendix 4 that $\xi_n \geq 0$. However, we will use this information from the beginning and there are no circular arguments. We note that $\xi_n$ depends on the coordinate we are considering, here $p$. So it should be written $\xi_n(p)$. However, to avoid cumbersome notations, we keep the notation $\xi_n$ unless there are ambiguities or we need to stress what coordinate we are referring to. This happens only in parts of “About $c_i$’s, $\xi_n$, $N_p$, and the limiting distribution of $\hat{\beta}(p)$” of Appendix 5 below.

We consider

$$
b_p \triangleq \beta_0(p) - \frac{\xi_n}{\tau + \xi_n} + \frac{1}{\sqrt{n}} \frac{N_p}{\tau + \xi_n}.
$$

Note that when $\xi_n > 0$, we have

$$
b_p - \beta_0(p) = \frac{n^{-1/2} N_p - \tau b_p}{\xi_n} = \frac{1}{n} \sum_{i=1}^{n} X_i^2(p) \psi'_i(r_i, [p]) - u'_p (\mathcal{S}_p + \tau \text{Id})^{-1} u_p.
$$

Indeed, essentially by definition, $\beta_0(p) = [(\tau + \xi_n) b_p - n^{-1/2} N_p] / \xi_n$; hence $b_p - \beta_0(p) = [n^{-1/2} N_p - \tau b_p] / \xi_n$. We call

$$
\tilde{b} = \begin{bmatrix} \hat{\gamma} \\ \beta_0(p) \end{bmatrix} + [b_p - \beta_0(p)] \begin{bmatrix} - (\mathcal{S}_p + \tau \text{Id})^{-1} u_p \\ 1 \end{bmatrix}.
$$

The aim of our work in the second part of this proof is to establish Theorem 3.20 on p.39, which shows that $\| \tilde{b} - \hat{\beta} \| = O(\text{polyLog}(n)/n)$ in $L_k$. Because the last coordinate of $\tilde{b}$, $b_p$, has a reasonably simple probabilistic structure and our approximations are sufficiently good, we will be able to transfer our insights about this coordinate to $\hat{\beta}_p$, the last coordinate of $\hat{\beta}$. This is also true when considering $\sqrt{n}(b_p - \hat{\beta}_p)$, so our approximations will be interesting at that scale, too.

The approach and approximating quantities we choose—as well as the intuition behind those choices—can be understood by using variants of the ideas discussed in our work in [3,16] and [17].
Deterministic aspects

Proposition 3.11 We have

\[ \| \hat{\beta} - \tilde{b} \| \leq \frac{1}{\tau} |b_p - \beta_0(p)| \sup_{1 \leq i \leq n} |d_{i,p}| \| \Sigma \|_2 \sqrt{\| (\xi \tau_1^p + \tau \text{Id})^{-1} u_p \|^2 + 1}. \]  

(33)

where \( d_{i,p} = [\psi'_i(\gamma_i^* - \psi'_i(\xi, p))] \) and \( \gamma_i^* \) is in the interval \( (\epsilon_i + V_i' \gamma_0 - V_i' \hat{\gamma}, \epsilon_i + X'_i \beta_0 - X'_i \tilde{b}) \).

Furthermore,

\[ \| (\xi \tau_1^p + \tau \text{Id})^{-1} u_p \|^2 \leq \frac{1}{n \tau} \sum_{i=1}^{n} X_i^2(p) \psi'_i(\xi, p) = \frac{1}{n \tau} \sum_{i=1}^{n} \lambda_i^2 \psi'_i(\xi, p) \lambda_i^2(p). \]  

(34)

As we saw in Eq. (15) and Lemma 3.2, we have

\[ \| \hat{\beta} - \tilde{b} \| \leq \frac{1}{\tau} \| f(\tilde{b}) \|, \]

where

\[ f(\tilde{b}) = -\frac{1}{n} \sum_{i=1}^{n} X_i \psi_i (\epsilon_i + X'_i \beta_0 - X'_i \tilde{b}) + \tau \tilde{b}. \]

We note furthermore that, by definition of \( \hat{\gamma} \),

\[ g(\hat{\gamma}) \triangleq -\frac{1}{n} \sum_{i=1}^{n} V_i \psi_i (\epsilon_i + V'_i \gamma_0 - V'_i \hat{\gamma}) + \tau \hat{\gamma} = 0_{p-1}. \]

The strategy of the proof is to control \( f(\tilde{b}) \) by using \( g(\hat{\gamma}) \) to create good approximations and then recalling that \( g(\hat{\gamma}) = 0_{p-1} \).

Proof The proof strategy and ideas are tied to the technique developed in [15]; however, because there are a number of delicate cancellations in the argument, we give it in full details. (Naturally, coming up with good approximating quantities required much work.)

a. Work on the first \((p-1)\) coordinates of \( f(\tilde{b}) \)

We call \( f_{p-1}(\beta) \) the first \( p-1 \) coordinates of \( f(\beta) \). We call \( \hat{\gamma}_{\text{ext}} \) the \( p \)-dimensional vector whose first \( p - 1 \) coordinates are \( \hat{\gamma} \) and last coordinate is \( \beta_0(p) \), i.e.

\[ \hat{\gamma}_{\text{ext}} = \left[ \begin{array}{c} \hat{\gamma} \\ \beta_0(p) \end{array} \right]. \]

For a vector \( v \), we use the notation \( v_{\text{comp}, k} \) to denote the \( p - 1 \) dimensional vector consisting of all the coordinates of \( v \) except the \( k \)th.
Clearly,
\[ f_{p-1}(\tilde{b}) = f_{p-1}(\tilde{b}) - g(\tilde{\gamma}) \]
\[ = -\frac{1}{n} \sum_{i=1}^{n} V_i \left[ \psi_i(\epsilon_i + X_i'\beta_0 - X_i'\tilde{b}) - \psi_i(\epsilon_i + V_i'\gamma_0 - V_i'\tilde{\gamma}) \right] + \tau(\tilde{b}_{\text{comp},p} - \tilde{\gamma}). \]

We can write by using the mean value theorem, for \( \gamma_{i,p}^* \) in the interval \( (\epsilon_i - V_i'(\tilde{\gamma} - \gamma_0), \epsilon_i - X_i'(\tilde{b} - \beta_0)) \),
\[ \psi_i(\epsilon_i + X_i'\beta_0 - X_i'\tilde{b}) - \psi_i(\epsilon_i + V_i'\gamma_0 - V_i'\tilde{\gamma}) \]
\[ = \psi_i(\gamma_{i,p}^*) X_i'(\tilde{\gamma}_{\text{ext}} - \tilde{b}), \]
\[ = \psi_i'(r_{i,[p]}) X_i'(\tilde{\gamma}_{\text{ext}} - \tilde{b}) + [\psi_i'(\gamma_{i,p}^*) - \psi_i'(r_{i,[p]})] X_i'(\tilde{\gamma}_{\text{ext}} - \tilde{b}). \]

Let us call
\[ d_{i,p} = [\psi_i'(\gamma_{i,p}^*) - \psi_i'(r_{i,[p]})], \]
\[ \delta_{i,p} = [\psi_i'(\gamma_{i,p}^*) - \psi_i'(r_{i,[p]})] X_i'(\tilde{\gamma}_{\text{ext}} - \tilde{b}), \]
\[ R_p = -\frac{1}{n} \sum_{i=1}^{n} d_{i,p} V_i X_i'(\tilde{\gamma}_{\text{ext}} - \tilde{b}). \]

We have with this notation
\[ f_{p-1}(\tilde{b}) = -\frac{1}{n} \sum_{i=1}^{n} \psi_i'(r_{i,[p]}) V_i X_i'(\tilde{\gamma}_{\text{ext}} - \tilde{b}) + \tau(\tilde{b}_{\text{comp},p} - \tilde{\gamma}) + R_p \triangleq A_p + R_p. \]

We note that by definition,
\[ \tilde{\gamma}_{\text{ext}} - \tilde{b} = [b_p - \beta_0(p)] \left[ (\mathcal{G}_p + \tau \text{Id})^{-1} u_p \right]_{-1}, \]
\[ \tilde{b}_{\text{comp},p} - \tilde{\gamma} = [-b_p - \beta_0(p)](\mathcal{G}_p + \tau \text{Id})^{-1} u_p. \]

Therefore, \( X_i'(\tilde{\gamma}_{\text{ext}} - \tilde{b}) = [b_p - \beta_0(p)][V_i'(\mathcal{G}_p + \tau \text{Id})^{-1} u_p - X_i(p)], \) and
\[ A_p = -[b_p - \beta_0(p)] \left( \frac{1}{n} \sum_{i=1}^{n} \psi_i'(r_{i,[p]}) V_i \left[ V_i'(\mathcal{G}_p + \tau \text{Id})^{-1} u_p - X_i(p) \right] \right) \]
\[ + \tau(\mathcal{G}_p + \tau \text{Id})^{-1} u_p \right). \]

Recalling the definition of \( \mathcal{G}_p \) and \( u_p \), we see that
\[ A_p = -[b_p - \beta_0(p)] \left( \mathcal{G}_p(\mathcal{G}_p + \tau \text{Id})^{-1} u_p - u_p + \tau(\mathcal{G}_p + \tau \text{Id})^{-1} u_p \right) = 0_{p-1}, \]
since \( \mathcal{G}_p(\mathcal{G}_p + \tau \text{Id})^{-1} + \tau(\mathcal{G}_p + \tau \text{Id})^{-1} = \text{Id}. \)
We conclude that

\[
    f_{p-1}(\tilde{b}) = R_p.
\]

b. Work on the last coordinate of \( f(\tilde{b}) \)

We call \([ f(\tilde{b}) ]_p\) the last coordinate of \( f(\tilde{b}) \). We have shown above that

\[
    \psi_i(\epsilon_i + X'_i \beta_0 - X'_i \tilde{b}) = \psi_i(\epsilon_i + V'_i \gamma_0 - V'_i \hat{\gamma})
    = \psi'_i(r_{i,[p]}) X'_i(\hat{\gamma}_{ext} - \tilde{b}) + [\psi'_i(\gamma_{i,p}^*) - \psi'_i(r_{i,[p]})] X'_i(\hat{\gamma}_{ext} - \tilde{b}).
\]

Recall the notation

\[
    \delta_{i,p} = [\psi'_i(\gamma_{i,p}^*) - \psi'_i(r_{i,[p]})] X'_i(\hat{\gamma}_{ext} - \tilde{b}).
\]

Clearly,

\[
    \psi_i(\epsilon_i + X'_i \beta_0 - X'_i \tilde{b}) = \psi_i(r_{i,[p]}) + \psi'_i(r_{i,[p]}) X'_i(\hat{\gamma}_{ext} - \tilde{b}) + \delta_{i,p},
    = \psi_i(r_{i,[p]}) + \psi'_i(r_{i,[p]})[b_p - \beta_0(p)]\left[V'_i(\mathcal{G}_p + \tau \text{Id})^{-1} u_p - X_i(p)\right] + \delta_{i,p}.
\]

We therefore see that

\[
    [ f(\tilde{b}) ]_p + \frac{1}{n} \sum_{i=1}^n X_i(p) \delta_{i,p}
    = -\frac{1}{n} \sum_{i=1}^n X_i(p) \left(\psi_i(r_{i,[p]}\right)
    + \psi'_i(r_{i,[p]})[b_p - \beta_0(p)]\left[V'_i(\mathcal{G}_p + \tau \text{Id})^{-1} u_p - X_i(p)\right] + \tau \tilde{b}_p,
    = -\frac{1}{n} \sum_{i=1}^n X_i(p) \psi_i(r_{i,[p]}) - (b_p - \beta_0(p)) u'_p(\mathcal{G}_p + \tau \text{Id})^{-1} u_p
    + (b_p - \beta_0(p)) \left(\frac{1}{n} \sum_{i=1}^n \psi'_i(r_{i,[p]}) X^2_i(p) - u'_p(\mathcal{G}_p + \tau \text{Id})^{-1} u_p\right)
    = -\left[\frac{1}{n} \sum_{i=1}^n X_i(p) \psi_i(r_{i,[p]}) - \tau b_p\right]
    + (b_p - \beta_0(p)) \left(\frac{1}{n} \sum_{i=1}^n \psi'_i(r_{i,[p]}) X^2_i(p) - u'_p(\mathcal{G}_p + \tau \text{Id})^{-1} u_p\right),
    = -\left[\frac{1}{\sqrt{n}} N_p - \tau b_p\right] + (b_p - \beta_0(p)) \xi_n,
    = 0.
\]
We conclude that
\[
[f(\tilde{b})]_p = -\frac{1}{n} \sum_{i=1}^n X_i(p) \delta_{i,p} = -\frac{1}{n} \sum_{i=1}^n d_{i,p} X_i(p) X'_i(\gamma_{\text{ext}} - \tilde{b}).
\]

**Representation of** \( f(\tilde{b}) \)

Aggregating all the results we have obtained so far, we see that
\[
f(\tilde{b}) = \left( -\frac{1}{n} \sum_{i=1}^n d_{i,p} X'_i \right)(\gamma_{\text{ext}} - \tilde{b}),
\]
\[
= -(b_p - \beta_0(p)) \left( -\frac{1}{n} \sum_{i=1}^n d_{i,p} X'_i \right) \left[ (\mathcal{G}_p + \tau \text{Id})^{-1} u_p \right].
\]

We conclude immediately that
\[
\|f(\tilde{b})\| \leq |b_p - \beta_0(p)| \sup_{1 \leq i \leq n} |d_{i,p}| \|I\| |\Sigma| \|
\left( \mathcal{G}_p + \tau \text{Id} \right)^{-1} u_p \|^2 + 1.
\] (35)

This gives Eq. (33). The rest of the proof follows easily with mild modifications from [15] and we do not repeat it here. \( \square \)

**Stochastic aspects**

From now on, we assume that \( \mathcal{X}(p) \), is independent of \( \{\mathcal{V}_i, \epsilon_i\}_{i=1}^n \). This is consistent with Assumption P1. (Recall that \( X_i = \lambda_i \mathcal{X}_i \) and therefore \( V_i = \lambda_i \mathcal{V}_i \).) Note that Assumption O4 is satisfied for \( \mathcal{V}_i \) if it is satisfied for \( \mathcal{X}_i \): convex 1-Lipschitz function of \( \mathcal{V}_i \) can be trivially made to be convex 1-Lipschitz function of \( \mathcal{X}_i \) by simply not acting on the last coordinate of \( \mathcal{X}_i \).

Naturally, a large amount of the rest of the proof consists in showing that we can bound \( \|f(\tilde{b})\| \) sufficiently finely for our results to hold true. So we will work on bounding each term in the product appearing in Eq. (33) in the rest of this section.

The last term is very easy to bound. In fact, using Eq. (34), we have
\[
\|\left( \mathcal{G}_p + \tau \text{Id} \right)^{-1} u_p \|^2 \leq \frac{1}{\tau} \frac{1}{n} \sum_{i=1}^n \|\psi'_i\|_{\infty} \lambda_i^2 \mathcal{X}_i^2(p),
\]
and
\[
\|\left( \mathcal{G}_p + \tau \text{Id} \right)^{-1} u_p \|^2 \leq \frac{\sup_i \|\psi'_i\|_{\infty}}{\tau} \frac{1}{n} \sum_{i=1}^n \lambda_i^2 \mathcal{X}_i^2(p).
\]
Hence, under assumptions O3–O4 and O6, we see that, for any fixed \( k \) and at \( \tau \) fixed,

\[
\left\| (\mathcal{G}_p + \tau \text{Id})^{-1} u_p \right\|_2^2 = O_{L_k} (\text{polyLog}(n)).
\]

As an aside, we note that \( p \) does not play any particular role here. If we considered the same quantity when we remove the \( k \)th predictor, and took the sup over \( 1 \leq k \leq p \) of the corresponding random variables, the same inequality would hold, in light of our work in e.g. Lemma 3.37.

The previous equation guarantees that

\[
\left\| (\mathcal{G}_p + \tau \text{Id})^{-1} u_p \right\|_2 \leq (1 + \left\| (\mathcal{G}_p + \tau \text{Id})^{-1} u_p \right\|_2) = O_{L_k} (\text{polyLog}(n)).
\]

We conclude, using Eq. (33), that

\[
\| \hat{\beta} - \tilde{b} \| = O_{L_k} \left( \frac{1}{\tau} \text{polyLog}(n) | b_p - \beta_0(p) | \sup_{1 \leq i \leq n} | d_{i,p} | \| \hat{\Sigma} \|_2 \right),
\]

provided the terms appearing inside the \( O_{L_k} \) have enough moments to enable us to use Holder’s inequality.

Recall that Lemma 3.38 gives \( \| \hat{\Sigma} \|_2 = O_{L_k} (\text{polyLog}(n)) \) under our assumptions O1–O7. At a high level, we expect \( \sup_{1 \leq i \leq n} | d_{i,p} | \) and \( | b_p - \beta_0(p) | \) to be small, which “should give us” that

\[
\| \hat{\beta} - \tilde{b} \| = O_{L_k} (\text{polyLog}(n) \sup_{1 \leq i \leq n} | d_{i,p} | \| b_p - \beta_0(p) |).
\]

In fact, we will show in Proposition 3.12 that \( b_p - \beta_0(p) = O_{L_k} (\text{polyLog}(n) | n^{-1/2} \vee n^{\alpha - \epsilon} |) \) and in Proposition 3.19 that \( \sup_{1 \leq i \leq n} | d_{i,p} | = O_{L_k} (\text{polyLog}(n) | n^{\alpha - 1/2} \vee n^{\alpha - \epsilon} |) \).

These are the key bounds we will need in showing that \( \| \hat{\beta} - \tilde{b} \| \) is small.

We now turn our attention to showing these two results.

### On \( b_p - \beta_0(p) \)

We recall the notations

\[
N_p = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_i(r_{i,[p]}) X_i(p) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \lambda_i \psi_i(r_{i,[p]}) X_i(p),
\]

\[
\xi_n = \frac{1}{n} \sum_{i=1}^{n} \psi_i'(r_{i,[p]}) X_i^2(p) - u'(\mathcal{G}_p + \tau \text{Id})^{-1} u_p.
\]

Under our assumptions, we have \( \mathbb{E} (X_i) = 0 \) and \( \text{cov} (X_i) = \text{Id}_p \) and hence \( \mathbb{E} (X_i^2(p)) = 1 \). Recall that since we assume that \( X(p) \) is independent of \( \{V_i, \epsilon_i\}_{i=1}^{n} \), \( X(p) \) is independent of \( \{r_{i,[p]}\}_{i=1}^{n} \).

\( \mathbb{E} \) Springer
Proposition 3.12 We have

$$|b_p - \beta_0(p)| \leq \frac{1}{\sqrt{n}\tau} |N_p| + |\beta_0(p)| \leq \frac{1}{\sqrt{n}\tau} |N_p| + \|\beta_0\|_{\infty}.$$ 

Furthermore, under assumptions $O1–O7$ and $P1$, $N_p = O_{L_k}(\text{polyLog}(n))$ and therefore, when $\tau$ is held fixed,

$$|b_p - \beta_0(p)| = O_{L_k}(\text{polyLog}(n)n^{-1/2} + \|\beta_0\|_{\infty}).$$

Proof From the definition of $b_p$, we see that, when $\xi_n \neq 0$

$$b_p - \beta_0(p) = \frac{1}{\sqrt{n}\tau + \xi_n} - \frac{\tau \beta_0(p)}{\tau + \xi_n}.$$ 

We will see later, in “On $\xi_n$” of Appendix 4 that $\xi_n \geq 0$ (there is no circular arguments, it is simply more convenient to postpone the investigation of the properties of $\xi_n$). It immediately then follows that

$$|b_p| \leq \frac{1}{\sqrt{n}\tau} |N_p| + |\beta_0(p)|.$$ 

Using independence of $\mathcal{X}(p)$ and $\{\mathcal{V}_i, \epsilon_i\}_{i=1}^n$, and $E(X_i(p)) = 0$ for all $i$, we have for instance,

$$E(N_p^2) = \frac{1}{n} \sum_{i=1}^n E(X_i^2(p))^2 E(\lambda_i^2 \psi_i^2(r_i,[p])),$$

whether the right-hand side is finite or not. Using our bounds on $\max \lambda_i^2$ and $\sup_i \|\psi_i\|_{\infty}$, we therefore have

$$E(N_p^2) \leq \frac{1}{n} \sum_{i=1}^n E(X_i^2(p))^2 \|\psi_i\|_{\infty}^2 E(\lambda_i^2) = O(1) = O(\text{polyLog}(n)).$$

Simple computations also show that $N_p$ has as many moments as we need and that for any finite $k$, under our assumptions,

$$N_p = O_{L_k}(\text{polyLog}(n)).$$

We therefore have

$$|b_p - \beta_0(p)| \leq \frac{1}{\sqrt{n}\tau} O_{L_k}(\text{polyLog}(n)) + \sup_{1 \leq k \leq p} |\beta_0(k)|.$$
Let us write ξₙ using matrix notations. Let $D_{\psi_i(r_{i,p})}$ be the $n \times n$ diagonal matrix such that

$$D_{\psi_i(r_{i,p})}(i, i) = \psi_i^2(r_{i,p}).$$

The notation $D_{\psi_i(r_{i,p})}$ might make it clearer that we are referring to a unique matrix and not a sequence of matrices indexed by $i$ but we use $D_{\psi_i(r_{i,p})}$ because it is a less cumbersome notation. We also denote by $X(p)$ the last column of the design matrix $X$. Then we have

$$\xi_n = \frac{1}{n} X(p)' D^{1/2}_{\psi_i(r_{i,p})} M D^{1/2}_{\psi_i(r_{i,p})} X(p), \quad (36)$$

where

$$M = \text{Id}_n - \frac{D^{1/2}_{\psi_i(r_{i,p})}}{\sqrt{n}} V \left( \frac{1}{n} V' D_{\psi_i(r_{i,p})} V + \tau \text{Id} \right)^{-1} V' D^{1/2}_{\psi_i(r_{i,p})} . \quad (37)$$

This simply comes from elementary linear algebra and representing $u_p$ and $\Sigma_p$ in matrix form. For example, $n u_p' = X(p)' D_{\psi_i(r_{i,p})} V$.

We are now ready to investigate in more detail the properties of $\xi_n$.

**Lemma 3.13** We have

$$\xi_n \geq 0.$$

Furthermore, under Assumptions O1–O7 and P1, if $D_{\lambda_i}$ is the diagonal matrix with $i$th entry $\lambda_i$,

$$\left| \xi_n - \frac{1}{n} \text{trace} \left( D_{\lambda_i} D^{1/2}_{\psi_i(r_{i,p})} M D^{1/2}_{\psi_i(r_{i,p})} D_{\lambda_i} \right) \right| = O_L \left( \sup_{1 \leq i \leq n} \lambda_i^2 \psi_i^2(r_{i,p}) / (\sqrt{n} \sigma_n) \right). \quad (38)$$

**Proof** Let us first focus on

$$M = \text{Id}_n - \frac{1}{n} D^{1/2}_{\psi_i(r_{i,p})} V \left( \frac{V' D_{\psi_i(r_{i,p})} V}{n} + \tau \text{Id} \right)^{-1} V' D^{1/2}_{\psi_i(r_{i,p})}.$$

The first part of the proof is very similar to the corresponding arguments in [15]. When $\tau > 0$, it is clear that all the eigenvalues of $M$ are strictly positive, i.e. $M$ is positive definite. Indeed, if the singular values of $n^{-1/2} D^{1/2}_{\psi_i(r_{i,p})} V$ are denoted by $\sigma_i$, the eigenvalues of $M$ are $\tau / (\sigma_i^2 + \tau)$.

Therefore, since $\xi_n = \frac{1}{n} v' M v$ with $v = D^{1/2}_{\psi_i(r_{i,p})} X(p)$, $\xi_n \geq 0$.

Since $M$ is symmetric and has eigenvalues between 0 and 1, we also have, using e.g. Lemma V.1.5 in [6],
\[
0 \leq D_{\psi_n}^{1/2} M D_{\psi_n}^{1/2} \leq D_{\psi_n}(r, \|p\|) = D_{\psi_n}(r, \|p\|)M D_{\psi_n}^{1/2} \psi_n(r, \|p\|).
\]

The matrix \(M\) is independent of \(\mathcal{X}(p)\) under Assumption \(\textbf{P1}\). \(D_{\psi_n}(r, \|p\|)\) is also independent of \(\mathcal{X}(p)\). Of course, we have \(X(p) = D_{\lambda_p} \mathcal{X}(p)\), where \(D_{\lambda_p}\) is the diagonal matrix with \((i, i)\) entry \(\lambda_{ip}\).

Since \(\mathcal{X}_p\) satisfy the necessary concentration assumptions under Assumption \(\textbf{P1}\), we can now appeal to Lemma 3.37 to obtain

\[
\left\| \frac{1}{n} X(p)' D_{\psi_n}^{1/2} M D_{\psi_n}^{1/2} \psi_n(r, \|p\|) X(p) - \frac{1}{n} \text{trace} \left( D_{\lambda_p} D_{\psi_n}^{1/2} M D_{\psi_n}^{1/2} \psi_n(r, \|p\|) D_{\lambda_p} \right) \right\| = O_L \left( \frac{1}{\sqrt{n} \sigma_n} \sup_{1 \leq i \leq n} \lambda_{ip}^2 \psi_n(r, \|p\|) \right).
\]

\(\Box\)

We now take a slight detour from the aim of showing that we have a very good approximation of \(\hat{\beta}\) through \(\hat{b}\) by working on finer properties of \(\dot{\xi}_n\) and \(b_p\). These properties will be essential in establishing the validity of the system (4).

To get a finer understanding of \(\dot{\xi}_n\), we now focus on the properties of

\[
\frac{1}{n} \text{trace} \left( D_{\lambda_p} D_{\psi_n}^{1/2} M D_{\psi_n}^{1/2} \psi_n(r, \|p\|) D_{\lambda_p} \right).
\]

The previous lemma shows clearly why this is natural.

\textbf{About} \(\frac{1}{n} \text{trace} \left( D_{\lambda_p} D_{\psi_n}^{1/2} M D_{\psi_n}^{1/2} \psi_n(r, \|p\|) D_{\lambda_p} \right)\)

\textbf{Lemma 3.14} Let us call \(\mathcal{S}_p = \frac{1}{n} \sum_{i=1}^n \psi_n(r, \|p\|) V_i V_i'\) and \(\mathcal{S}_p(i) = \mathcal{S}_p - \frac{1}{n} \psi_n(r, \|p\|) V_i V_i'\). Let us also call

\[
c_{\tau, p} = \frac{1}{n} \text{trace} \left( (\mathcal{S}_p + \tau \text{Id})^{-1} \right),
\]

\[
\dot{\xi}_i = \frac{1}{n} V_i' (\mathcal{S}_p(i) + \tau \text{Id})^{-1} V_i - \lambda_{ip}^2 c_{\tau, p}.
\]

Then we have under Assumptions \(\textbf{O1–O7}\) and \(\textbf{P1}\), if \(M\) is the matrix defined in Eq. (37),

\[
\left\| \frac{1}{n} \text{trace} \left( \text{Id}_n - M \right) - \left( \frac{1}{n} \text{trace} \left( D_{\lambda_p} D_{\psi_n}^{1/2} M D_{\psi_n}^{1/2} \psi_n(r, \|p\|) D_{\lambda_p} \right) \right) c_{\tau, p} \right\| \leq \left[ \sup_i |\dot{\xi}_i| \right] \frac{1}{n} \sum_{i=1}^n \psi_n(r, \|p\|).
\]

(39)

We also have

\[
\frac{1}{n} \text{trace} \left( \text{Id}_n - M \right) = \frac{p - 1}{n} - \tau c_{\tau, p}.
\]
Proof We call \( d_{i,i} = \psi_i'(r_{i,[p]})/n \). Of course, by using the Sherman-Morrison-Woodbury formula (see e.g. [21], p.140)

\[
M_{i,i} = 1 - d_{i,i} V_i' (V'D_{\psi_i'(r_{i,[p]})} V + \tau \text{Id})^{-1} V_i,
\]

\[
= 1 - d_{i,i} \frac{V_i' (\mathbb{S}_p(i) + \tau \text{Id})^{-1} V_i}{1 + d_{i,i} V_i' (\mathbb{S}_p(i) + \tau \text{Id})^{-1} V_i},
\]

\[
= 1 + d_{i,i} V_i' (\mathbb{S}_p(i) + \tau \text{Id})^{-1} V_i.
\]

It immediately follows that \(\frac{1}{n} \sum d_{i,i} \leq 1\) and that \(\frac{1}{n} \sum_{i} |\psi_i'(r_{i,[p]}\text{Id})| \leq 1\).

Recall that we are interested in

\[
\frac{1}{n} \sum_i \lambda_i^2 \psi_i'(r_{i,[p]}M_{i,i} = \frac{1}{n} D_{\lambda_{i}} D_{\psi_i'(r_{i,[p]})} MD_{\psi_i'(r_{i,[p]})} D_{\lambda_{i}}.\]

Note that, since trace \((AB) = \text{trace} \((BA)\),

\[
\text{trace} \((\text{Id}_n - M) = \text{trace} \((\mathbb{S}_p + \tau \text{Id})^{-1} \mathbb{S}_p\)
\]

\[
= p - 1 - \tau \text{trace} \((\mathbb{S}_p + \tau \text{Id})^{-1}\) = p - 1 - n \tau \mathcal{C}_{\tau,p}.
\]

This shows the second result of the lemma.

On the other hand,

\[
\text{trace} \((\text{Id}_n - M) = \sum_i (1 - M_{i,i})
\]

\[
= \sum_i d_{i,i} \frac{V_i' (\mathbb{S}_p(i) + \tau \text{Id})^{-1} V_i}{1 + d_{i,i} V_i' (\mathbb{S}_p(i) + \tau \text{Id})^{-1} V_i}.
\]

With our definitions, we have, since \(\lambda_i^2 \mathcal{C}_{\tau,p} + \zeta_i = \frac{1}{n} V_i' (\mathbb{S}_p(i) + \tau \text{Id})^{-1} V_i\),

\[
\frac{1}{n} \text{trace} \((\text{Id}_n - M) = \left( \frac{1}{n} \sum_i \lambda_i^2 \psi_i'(r_{i,[p]}M_{i,i} \right) \mathcal{C}_{\tau,p}
\]

\[
+ \frac{1}{n} \sum_i \psi_i'(r_{i,[p]} \frac{\zeta_i}{1 + d_{i,i} V_i' (\mathbb{S}_p(i) + \tau \text{Id})^{-1} V_i}.
\]

It immediately follows that

\[
\left| \frac{1}{n} \text{trace} \((\text{Id}_n - M) - \left( \frac{1}{n} \sum_i \lambda_i^2 \psi_i'(r_{i,[p]}M_{i,i} \right) \mathcal{C}_{\tau,p} \right| \leq \left( \sup_i |\zeta_i| \right) \frac{1}{n} \sum_i \psi_i'(r_{i,[p]}),
\]

as announced. \(\square\)

The previous result will be especially useful as an approximation result if we can show that \(\zeta_i\)'s are small, since assumption P2—which we will use later—implies that \(\frac{1}{n} \sum_i V_i' \|\psi_i'\|_{\infty} \leq 1\) cannot be too large. This is what we do in the next few pages.
Controlling $\zeta_i$

The main problem that arises when trying to control $\zeta_i$ is the fact that $r_{j,[p]}$ appearing in $\mathcal{S}_p(i)$ depend on $V_j$. This prevents us from using concentration of quadratic forms results such as those shown in Lemma 3.37. So further approximations arguments are needed. Of course, the idea of using a leave-two-out residuals to approximate $\{r_{j,[p]}\}_{j \neq i}$ immediately comes to mind. Hence our work in “Appendix 3” will later play a key role in showing that $\zeta_i$’s are small.

Lemma 3.15 Suppose we can find $\{r_{j,[p]}(i)\}_{j \neq i}$ independent of $(\lambda_i, V_i)$ and $K_n$ such that

$$\sup_i \sup_{j \neq i} |\psi_j'(r_{j,[p]}(i)) - \psi_j'(r_{j,[p]})| \leq K_n.$$

Then

$$\sup_i |\zeta_i| = O_{L_k}\left(\left[\frac{1}{\tau^2} K_n |||\hat{\Sigma}|||_2 + \frac{\text{polyLog}(n)}{\tau \sqrt{n c_n}} + \frac{1}{n \tau} \right] \text{polyLog}(n)\right), \quad (41)$$

provided $K_n$ has $3k$ uniformly bounded moments.

Proof We call

$$AM_{i,p} = \frac{1}{n} \sum_{j \neq i} \psi_j'(r_{j,[p]}(i)) V_j V_j'.$$

Then, using for instance the first resolvent identity, i.e. $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$, we see that

$$|||((\mathcal{S}_p(i) + \tau \text{Id})^{-1} - (AM_{i,p} + \tau \text{Id})^{-1})|||_2 \leq \frac{1}{\tau^2} K_n |||\hat{\Sigma}|||_2.$$

since $|||\frac{1}{n} \sum_i V_i V_i'|||_2 \leq |||\hat{\Sigma}|||_2$. In particular,

$$\left|\frac{1}{n} V_i'((\mathcal{S}_p(i) + \tau \text{Id})^{-1} V_i - \frac{1}{n} V_i'(AM_{i,p} + \tau \text{Id})^{-1} V_i\right| \leq \frac{||V_i||^2}{n} \frac{1}{\tau^2} K_n |||\hat{\Sigma}|||_2.$$

However, since $AM_{i,p}$ is independent of $(\lambda_i, V_i)$, we can use Lemma 3.37 and see that, since $V_i = \lambda_i V_i$,

$$\sup_{1 \leq i \leq n} \left|\frac{1}{n} V_i'(AM_{i,p} + \tau \text{Id})^{-1} V_i - \frac{\lambda_i^2}{n} \text{trace}((AM_{i,p} + \tau \text{Id})^{-1})\right|$$

$$= O_{L_k}\left(\frac{\text{polyLog}(n)}{\tau \sqrt{n c_n}} \sup_{1 \leq i \leq n} \lambda_i^2\right),$$

by using the fact that $\lambda_{\max}((AM_{i,p} + \tau \text{Id})^{-1}) \leq \frac{1}{\tau}$. 

$\square$ Springer
Using the operator norm bound we gave above, we also have

$$\left| \frac{1}{n} \text{trace} \left( (AM_{i, p} + \tau \text{Id})^{-1} \right) - \frac{1}{n} \text{trace} \left( (\mathcal{G}_{p}(i) + \tau \text{Id})^{-1} \right) \right| \leq \frac{1}{\tau^2} K_n \|\hat{\Sigma}\|_2 \frac{p}{n}. $$

We conclude that

$$\sup_{1 \leq i \leq n} \left| \frac{1}{n} V_i'(\mathcal{G}_{p}(i) + \tau \text{Id})^{-1} V_i - \frac{\lambda_i^2}{n} \text{trace} \left( (\mathcal{G}_{p}(i) + \tau \text{Id})^{-1} \right) \right| \leq O_{L_k} \left( \frac{1}{\tau^2} K_n \|\hat{\Sigma}\|_2 \frac{p}{n} \right).$$

Now, it is clear that under $O_1$ and $O_4$, $\sup_{1 \leq i \leq n} \|V_i\|^2 / n = O_{L_k}(1)$ and finally

$$\sup_{1 \leq i \leq n} \left| \frac{1}{n} V_i'(\mathcal{G}_{p}(i) + \tau \text{Id})^{-1} V_i - \frac{\lambda_i^2}{n} \text{trace} \left( (\mathcal{G}_{p}(i) + \tau \text{Id})^{-1} \right) \right| \leq O_{L_k} \left( \frac{1}{\tau^2} \|\hat{\Sigma}\|_2 + \frac{\text{polyLog}(n)}{\tau \sqrt{n\mathcal{C}_n}} \left[ \sup_{1 \leq i \leq n} \lambda_i^2 \lor 1 \right] \right).$$

**Control of** $\frac{1}{n} \text{trace} \left( (\mathcal{G}_{p}(i) + \tau \text{Id})^{-1} \right) - \frac{1}{n} \text{trace} \left( (\mathcal{G}_{p} + \tau \text{Id})^{-1} \right)$

Using the Sherman-Woodbery-Morrison formula, we have

$$(\mathcal{G}_{p}(i) + \tau \text{Id})^{-1} - (\mathcal{G}_{p} + \tau \text{Id})^{-1} = \frac{1}{n} V_i'(r_{i, |p|}) (\mathcal{G}_{p}(i) + \tau \text{Id})^{-1} V_i V_i'(\mathcal{G}_{p}(i) + \tau \text{Id})^{-1}.$$ 

After taking traces, we see that

$$0 \leq \text{trace} \left( (\mathcal{G}_{p}(i) + \tau \text{Id})^{-1} \right) - \text{trace} \left( (\mathcal{G}_{p} + \tau \text{Id})^{-1} \right) \leq \frac{1}{\tau},$$

since $V_i'(\mathcal{G}_{p}(i) + \tau \text{Id})^{-2} V_i \leq \frac{1}{\tau} V_i'(\mathcal{G}_{p}(i) + \tau \text{Id})^{-1} V_i$.

Therefore,

$$0 \leq \frac{1}{n} \text{trace} \left( (\mathcal{G}_{p}(i) + \tau \text{Id})^{-1} \right) - \frac{1}{n} \text{trace} \left( (\mathcal{G}_{p} + \tau \text{Id})^{-1} \right) \leq \frac{1}{n \tau},$$

We conclude that

$$\sup_{1 \leq i \leq n} |\xi_i| = O_{L_k} \left( \frac{1}{\tau^2} K_n \|\hat{\Sigma}\|_2 + \frac{\text{polyLog}(n)}{\tau \sqrt{n\mathcal{C}_n}} + \frac{1}{n \tau} \right) \left[ \sup_{1 \leq i \leq n} \lambda_i^2 \lor 1 \right].$$
provided we can use Holder’s inequality. In effect, this requires $K_n$ to have $3k$ moments.

\[ \square \]

Control of $K_n$

A natural choice for $r_{j,[p]}^{(i)}$ defined in Lemma 3.15 is to use a leave one out estimator of $\hat{\gamma}$, where the $i$th observation (and hence $V_i$) is omitted. Hence, all the work done in Theorem 3.9 becomes immediately relevant.

**Lemma 3.16** Suppose we use for $\{r_{j,[p]}^{(i)}\} \neq i$ the residuals we would get by using a leave-one-out estimator of $\hat{\gamma}$, i.e. excluding $(V_i, \epsilon_i)$ from problem (28).

With the notations of Lemma 3.15, we have under assumptions $O1–O7$ and $P1$

\[ K_n = O_{L_k} \left( n^{2\alpha-1/2} \text{polyLog}(n) \right). \]

In particular, for any fixed $\tau$,

\[ \sup_i |\zeta_i| = O_{L_k} \left( n^{2\alpha-1/2} \text{polyLog}(n) \right). \]

**Proof** Let us call $\delta_n(i)$ random variables such that

\[ \sup_{j \neq i} |r_{j,[p]}^{(i)} - r_{j,[p]}| \leq \delta_n(i). \]

Applying Theorem 3.9 with $R_j = r_{j,[p]}$ and $\tilde{r}_{j,(i)} = r_{j,[p]}^{(i)}$, we get

\[ \sup_i (\delta_n(i)) = O_{L_k} \left( \frac{\text{polyLog}(n)}{n^{1/2-\alpha}} \right). \]

The control of $K_n$ follows immediately by using our assumptions on $\psi'_i$, specifically the fact that it is $Cn^\alpha$-Lipschitz.

\[ \square \]

**Important remark:** the previous remark has important consequences for $c_i$ defined in Eq. (21). Indeed, we have the following corollary.

**Corollary 3.17** Let $c_i$ be defined as in Eq. (21) and $c_\tau$ be defined as in Eq. (14). Then, under assumptions $O1–O7$ and $P1$, we have

\[ \sup_i |c_i - \lambda_i^2 c_\tau| = O_{L_k} \left( n^{2\alpha-1/2} \text{polyLog}(n) \right). \] (44)

The corollary follows from drawing analogy between these quantities and the situation investigated in Lemmas 3.14, 3.15, and 3.16; we now give a detailed proof.
Proof We have now established that
\[
\sup_i \left| \frac{1}{n} V'_i(\hat{G}_p(i) + \tau \text{Id})^{-1} V_i - \lambda_i^2 c_{\tau,p} \right| = O_Lk \left( \frac{\text{polyLog}(n)}{n^{1/2-2\alpha}} \right).
\]

Recalling the notation
\[
c_{\tau} = \frac{1}{n} \text{trace} \left( \left[ \frac{1}{n} \sum_{i=1}^{n} \psi'_i(R_i) X_i X'_i + \tau \text{Id} \right]^{-1} \right),
\]
we see that this quantity is the analog of $c_{\tau,p}$ when we use all the predictors and not only $(p-1)$ of them.

Indeed, $c_i$ in Eq. (21) is defined, in the notation of the proof of Lemma 3.15 as an analog of $\frac{1}{n} V'_i(AM_i,p + \tau \text{Id})^{-1} V_i$, with the role of $\{r^{(i)}_{j,\{p\}}\}_{j \neq i}$ being played by the residuals obtained from the leave-one-out estimate of $\hat{\beta}$, excluding $(X_i, Y_i)$ from the problem. Lemma 3.15 in connection with Theorem 3.20 shows that $\sup_i \left| \frac{1}{n} V'_i(AM_i,p + \tau \text{Id})^{-1} V_i - \lambda_i^2 c_{\tau,p} \right| = O_Lk (\text{polyLog}(n)/n^{1/2-2\alpha})$ under our assumptions. Passing from the $(p-1)$ dimensional version of this result, i.e. Lemma 3.15, to the $p$-dimensional version gives the approximation stated in the corollary.

We therefore see that
\[
\sup_i |c_i - \lambda_i^2 c_{\tau}| = O_Lk (n^{2\alpha-1/2}\text{polyLog}(n)).
\]

Further results on $\xi_n$ and $b_p$

We can combine all the results we have obtained so far in the following proposition.

**Proposition 3.18** We have, under Assumptions O1–O7 and P1,
\[
\left| c_{\tau,p}(\xi_n + \tau) - \frac{p-1}{n} \right| = O_Lk \left( \frac{\text{polyLog}(n)}{n^{1/2-2\alpha}} \right). \tag{45}
\]

Furthermore, under Assumptions O1–O7 and P1–P3, since $\|\beta_0\|_\infty = O(n^{-e})$,
\[
\left( \frac{p}{n} \right)^2 n \mathbb{E} \left( \left| b_p - \beta_0(p) \right|^2 \right) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left( \left| c_{\tau,p} \lambda_i \psi_i(r_{i,\{p\}}) \right|^2 \right) + n\tau^2 \beta_0^2(p) \mathbb{E} \left( c_{\tau,p}^2 \right) + o(1). \tag{46}
\]

Both equations in this proposition are very important for this paper. The first one gives us a very precise idea of the behavior of $\xi_n$ in terms of $c_{\tau,p}$, which we will see in
“Appendix 5” is relatively easy to understand. This first equation is also a stepping stone towards the first equation of the System (4).

The second equation, on the other hand, is a stepping stone towards the second equation of System (4) in our main theorem, Theorem 2.1.

**Proof**  • **First equation**

The proof of Eq. (45) consists just in aggregating all the previous results and noticing that $c_{\tau, p} \leq (p - 1)/(n \tau)$ and therefore remains bounded. Indeed, we have

$$\frac{p - 1}{n} - \tau c_{\tau, p} = \frac{1}{n} \text{trace} (\text{Id} - M) \geq 0.$$

This latter quantity was approximated in Lemma 3.14 by

$$\left(\frac{1}{n} \text{trace} \left(D_{1/2}^{1/2} \psi_i(r_i, p) M D_{1/2}^{1/2} \psi_i(r_i, p)\right)\right) c_{\tau, p}.$$

And in Lemma 3.13, we approximated $\xi_n$ by $\left(\frac{1}{n} \text{trace} \left(D_{1/2}^{1/2} \psi_i(r_i, p) M D_{1/2}^{1/2} \psi_i(r_i, p)\right)\right)$. This gives the result of Eq. (45), by simply keeping track of the approximation errors we make at each step.

• **Second equation**

Recall that by definition (see Eqs. (31) and (30)),

$$\sqrt{n} \left[ (\tau + \xi_n) b_p - \xi_n \beta_0(p) \right] = N_p = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \lambda_i \psi_i(r_i, p) \mathcal{X}_i(p).$$

Therefore,

$$c_{\tau, p} \sqrt{n} \left[ (\tau + \xi_n) [b_p - \beta_0(p)] + \tau \beta_0(p) \right] = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} c_{\tau, p} \lambda_i \psi_i(r_i, p) \mathcal{X}_i(p),$$

or

$$c_{\tau, p} \sqrt{n} (\tau + \xi_n) [b_p - \beta_0(p)] = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} c_{\tau, p} \lambda_i \psi_i(r_i, p) \mathcal{X}_i(p) - c_{\tau, p} \sqrt{n} \tau \beta_0(p).$$

We note that $c_{\tau, p} \lambda_i \psi_i(r_i, p)$, which depends only on $\{\lambda_i, \psi_i, \epsilon_i\}_{i=1}^{n}$, is independent of $\{\mathcal{X}_i(p)\}_{i=1}^{n}$. (If needed, see the definition of $c_{\tau, p}$ in Lemma 3.14.)

Since $\mathcal{X}_i(p)$’s are independent with mean 0 and variance 1, we conclude that

$$\mathbb{E} \left( c_{\tau, p}^2 n (\tau + \xi_n)^2 [b_p - \beta_0(p)]^2 \right) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left( [c_{\tau, p} \lambda_i \psi_i(r_i, p)]^2 \right) + n \tau^2 \beta_0^2(p) \mathbb{E} \left( c_{\tau, p}^2 \right).$$
Given the result in Eq. (45) and our bound on $\sqrt{n [b_p - \beta_0(p)]}$ in Proposition 3.12, this means that
\[
\left( \frac{P}{n} \right)^2 n E \left( [b_p - \beta_0(p)]^2 \right) = \frac{1}{n} \sum_{i=1}^{n} E \left( [c_{\tau, p} \psi_i(r_i, [p])]^2 \right) + n \tau^2 \beta_0^2(p) E \left( c_{\tau, p}^2 \right) + o(1).
\]
In this last equation, we make use of Proposition 3.12 and Assumption P3 since under this assumption $n\|\beta_0\|_{\infty} \text{polyLog}(n) n^{2\alpha - 1/2} \to 0$. This is what allows us to replace $c_{\tau, p}(\tau + \xi_n)$ by $p/n$ without loss of accuracy in going from the second-to-last to the last equation.

We now need to control $d_{i, p}$ to show that our approximation of $\hat{\beta}$ by $\tilde{b}$ in Proposition 3.11 will yield sufficiently good results that they can be used to prove Theorem 2.1.

On $d_{i, p}$

Recall the definition
\[
d_{i, p} = [\psi'_i(\gamma^*_i, p) - \psi'_i(r_i, [p])],
\]
where $\gamma^*_i, p \in (r_i, [p], r_i, [p] + \nu_i)$, with
\[
\nu_i = [b_p - \beta_0(p)] X'_i \left[ (\mathcal{G}_p + \tau \text{Id})^{-1} u_p \right] \triangleq [b_p - \beta_0(p)] \pi_i.
\]
(The fact that $\gamma^*_i, p \in (r_i, [p], r_i, [p] + \nu_i)$ follows from writing the definition of $Y_i - X'_i \tilde{b}$.)

We have the following result.

**Proposition 3.19** We have, under Assumptions O1–O7 and P1–P3, at fixed $\tau$,
\[
\sup_i |d_{i, p}| = O_L \left( \text{polyLog}(n) n^\alpha \right).
\]

**Proof** Note that we can rewrite
\[
\pi_i = X'_i \left[ (\mathcal{G}_p + \tau \text{Id})^{-1} u_p \right] = V'_i (\mathcal{G}_p + \tau \text{Id})^{-1} u_p - X_i(p).
\]
Recall that $u_p = \frac{1}{n} V' D_{\psi_i(r_i, [p])} X(p)$. We can also rewrite it as
\[
u_p = \frac{1}{n} \nu'_i D_{\psi_i(r_i, [p])} X(p).
\]
On the impact of predictor geometry on the performance on...

147

Recalling that $\mathcal{X}(p)$ with $\{(\nu_i, \epsilon_i)\}_{i=1}^{n}$, and our concentration assumptions on $\mathcal{X}(p)$ formulated in P1, we see that according to Lemma 3.36, we have

$$\sup_i |V_i' (\mathcal{S}_p + \tau \text{Id})^{-1} u_p|$$

$$= O_L \left( \frac{\text{polyLog}(n)}{c_n^{1/2}} \sup_i \frac{1}{n} D_{\lambda_i^2 \psi'_i(r_{i,p})} \mathcal{V}(\mathcal{S}_p + \tau \text{Id})^{-1} V_i \right),$$

where we look at $V_i' (\mathcal{S}_p + \tau \text{Id})^{-1} u_p$ as a linear form in $\mathcal{X}(p)$. Note that we have absorbed the $\sup_i |\lambda_i|$ in the polyLog($n$) term.

Now,

$$\frac{1}{n} D_{\lambda_i^2 \psi'_i(r_{i,p})} \mathcal{V}(\mathcal{S}_p + \tau \text{Id})^{-1} V_i$$

$$= \frac{1}{n} V_i'(\mathcal{S}_p + \tau \text{Id})^{-1} \mathcal{V}' D_{\lambda_i^2 \psi'_i(r_{i,p})} \frac{\mathcal{V}}{n} (\mathcal{S}_p + \tau \text{Id})^{-1} V_i.$$

Notice that $\mathcal{S}_p = \frac{\mathcal{V}' D_{\lambda_i^2 \psi'_i(r_{i,p})} \mathcal{V}}{n}$. Hence, $\frac{\mathcal{V}' D_{\lambda_i^2 \psi'_i(r_{i,p})} \mathcal{V}}{n} \leq \frac{\mathcal{V}_i}{n \tau} \parallel D_{\lambda_i^2 \psi'_i(r_{i,p})} \parallel_2$ and we conclude that

$$\frac{1}{n} V_i'(\mathcal{S}_p + \tau \text{Id})^{-1} \mathcal{V}' D_{\lambda_i^2 \psi'_i(r_{i,p})} \frac{\mathcal{V}}{n} (\mathcal{S}_p + \tau \text{Id})^{-1} V_i \leq \frac{\mathcal{V}_i}{n \tau} \parallel D_{\lambda_i^2 \psi'_i(r_{i,p})} \parallel_2$$

$$= \parallel V_i \parallel_2 \sup_i \lambda_i^2 \psi'_i(r_{i,p}).$$

We also note that $\sup_i |X_i(p)| = O_L (\text{polyLog}(n)/\sqrt{c_n})$ under O4, O6 and P1, using the results of “Appendix 7”. So we conclude that

$$\sup_i |\pi_i| = O_L \left( \frac{\text{polyLog}(n)}{c_n^{1/2}} \left[ 1 + \sqrt{\sup_i \lambda_i^2 \psi'_i(r_{i,p})} \frac{\mathcal{V}_i}{n \tau} \right] \right),$$

$$= O_L \left( \frac{\text{polyLog}(n)}{c_n^{1/2}} \left[ 1 + \sqrt{\sup_i \lambda_i^2 \psi'_i(r_{i,p})} \right] \right),$$

$$= O_L (\text{polyLog}(n)).$$

Recalling that $|b_p - \beta_0(p)| = O_L (n^{-1/2} \text{polyLog}(n) + \parallel \beta_0 \parallel_\infty)$, we finally see that

$$\sup_i |v_i| = O_L \left( \frac{\text{polyLog}(n)}{\sqrt{n} \wedge n^e} \right).$$
Under our assumption that $\psi_i'$ is $Cn^\alpha$-Lipschitz, we see that

$$\sup_i |d_{i,p}| = O_{L_k} \left( \frac{\text{polyLog}(n)n^\alpha}{n^{1/2} \land n^e} \right).$$

3.1 Final conclusions

We can now gather together our approximation results in the following Theorem.

**Theorem 3.20** Under Assumptions O1–O7 and P1–P3, we have, for any fixed $\tau > 0$,

$$\|\hat{\beta} - \tilde{b}\| \leq O_{L_k} \left( \frac{\text{polyLog}(n)n^\alpha}{[n^{1/2} \land n^e]^2} \right).$$

In particular,

$$\sqrt{n}(\hat{\beta}_p - b_p) = O_{L_k} \left( \frac{\text{polyLog}(n)n^{\alpha+1/2}}{[n^{1/2} \land n^e]^2} \right),$$

$$\sup_i |X_i'(\hat{\beta} - \tilde{b})| = O_{L_k} \left( \frac{\text{polyLog}(n)n^{\alpha+1/2}}{[n^{1/2} \land n^e]^2} \right),$$

$$\sup_i |R_i - r_{i,[p]}| = O_{L_k} \left( \left[ \frac{\text{polyLog}(n)}{\sqrt{n} \land n^e} \right] \lor \left[ \frac{\text{polyLog}(n)n^{\alpha+1/2}}{[n^{1/2} \land n^e]^2} \right] \right).$$

We note that the index $p$ in the previous theorem plays no particular role and similar results holds when $p$ is replaced by any $k$, $1 \leq k \leq p$.

**Proof** The Theorem is just the aggregation of all of our results, using the key bound on $\|\hat{\beta} - \tilde{b}\|$ in Proposition 3.11.

The last statement is the only one that might need an explanation. With the notations of the proof of Proposition 3.19, we have $R_i - r_{i,[p]} = X_i'(\tilde{b} - \hat{\beta}) - v_i$. The results on $\sup_i |v_i|$ in the proof of Proposition 3.19 as well as the bound on $\sup_i |X_i'(\tilde{b} - \hat{\beta})|$ give us the announced result.

Combining the results of Eq. (46) and the previous theorem, we see that under Assumptions O1–O7 and P1–P4,

$$\left( \frac{p^2}{n} \right) n E \left( (\hat{\beta}_p - \beta_0(p))^2 \right)$$

$$= \frac{1}{n} \sum_{i=1}^n E \left( [c_{\tau,p} \lambda_i \psi_i(r_{i,[p]})]^2 \right) + n \tau^2 \beta_0^2(p) E \left( c_{\tau,p}^2 \right) + o(1).$$

Since $p$ did not play any particular role as compared to any other index in our analysis, the same result holds when $p$ is replaced by $k$, $1 \leq k \leq p$. 

\(\odot\) Springer
Dividing the previous expression by $n$ on both sides and summing over all the indices $1 \leq k \leq p$, we finally get

$$
\left( \frac{p}{n} \right)^2 \mathbf{E} \left( \| \hat{\beta} - \beta_0 \|_2^2 \right)
= \frac{1}{n} \sum_{k=1}^{p} \left[ \frac{1}{n} \sum_{i=1}^{n} \mathbf{E} \left( \| \mathbf{c}_{\tau,k} \lambda_i \psi_i (r_i, [k]) \|_2^2 \right) \right] + \tau^2 \sum_{k=1}^{p} \beta_0^2 (k) \mathbf{E} \left( \mathbf{c}_{\tau,k}^2 \right) + o(1). \quad (47)
$$

Our aim now is to further simplify the above expression to get the second equation of our system.

### 3.1.1 On $c_{\tau,p}$ and $c_{\tau}$

We now show that $c_{\tau,k}$’s are all close to the same quantity, which turns out to be $c_{\tau}$.

**Proposition 3.21** We have, under Assumptions $O1$–$O7$ and $P1$–$P3$,

$$
\sup_{1 \leq k \leq p} | c_{\tau} - c_{\tau,k} | = O_{L_k} \left( \sqrt{\frac{\text{polyLog}(n)n^\alpha}{\sqrt{n} \wedge n^\theta}} \right) \vee \left[ \frac{\text{polyLog}(n)n^{2\alpha+1/2}}{[n^{1/2} \wedge n^\theta]^2} \right] \vee \text{polyLog}(n). 
$$

Of course, we also have $0 \leq c_{\tau} \leq p/(n\tau)$ and $0 \leq c_{\tau,k} \leq p/(n\tau)$.

**Proof** Let us recall the notation

$$
S = \frac{1}{n} \sum_{i=1}^{n} \psi'_i (R_i) X_i X'_i.
$$

If we call $\Gamma = \frac{1}{n} \sum_{i=1}^{n} \psi'_i (R_i) V_i V'_i$ and $a = \frac{1}{n} \sum_{i=1}^{n} \psi'_i (R_i) X_i^2 (p)$, we see that

$$
S = \left( \Gamma \bigvee a \right).
$$

According to Lemma 3.40, we see that, since $c_{\tau} = \frac{1}{n} \text{trace} \left( (S + \tau \text{Id}_p)^{-1} \right)$,

$$
| c_{\tau} - \frac{1}{n} \text{trace} \left( (\Gamma + \tau \text{Id}_{p-1})^{-1} \right) | \leq \frac{1}{n} \frac{1 + a/\tau}{\tau}.
$$

It is clear that under our assumptions, $a = O_{L_k} (\text{polyLog}(n))$, since

$$
a = \frac{1}{n} \sum_{i=1}^{n} \lambda_i^2 X_i^2 (p) \psi'_i (R_i) \leq \text{polyLog}(n) \frac{1}{n} \sum_{i=1}^{n} \lambda_i^2 X_i^2 (p) = O_{L_k} (\text{polyLog}(n)),
$$

using e.g. our work in “Appendix 7”. Since $\psi'_i$ is $Cn^\alpha$-Lipschitz and

$$
\sup_{i} | R_i - r_i, [p] | = O_{L_k} \left( \left[ \frac{\text{polyLog}(n)}{\sqrt{n} \wedge n^\theta} \right] \vee \left[ \frac{\text{polyLog}(n)n^{\alpha+1/2}}{[n^{1/2} \wedge n^\theta]^2} \right] \right),
$$
we have
\[
\sup_i |\psi_i'(R_i) - \psi_i'(r_i[p])| = O_{L_k}\left(\left[\frac{\text{polyLog}(n)n^\alpha}{\sqrt{n} \wedge n^\theta}\right] \lor \left[\frac{\text{polyLog}(n)n^{2\alpha+1/2}}{[n^{1/2} \wedge n^\theta]^2}\right]\right).
\]

Hence, using arguments similar to the ones we have used in the proof of Lemma 3.15 (i.e. first resolvent identity, etc...), we see that
\[
1/n \text{trace}\left((\Gamma + \tau \text{Id})^{-1}\right) - 1/n \text{trace}\left((\mathcal{S}_p + \tau \text{Id})^{-1}\right) = O_{L_k}\left(\left[\frac{\text{polyLog}(n)n^\alpha}{\sqrt{n} \wedge n^\theta}\right] \lor \left[\frac{\text{polyLog}(n)n^{2\alpha+1/2}}{[n^{1/2} \wedge n^\theta]^2}\right]\right).
\]

Since \(c_{\tau,p} = 1/n \text{trace}\left((\mathcal{S}_p + \tau \text{Id})^{-1}\right)\), the result we announced follows immediately.

We note that \(p\) did not play a particular role here and hence taking the sup over those indices only adds a polyLog\((n)\) term to the approximation. Hence our approximation is valid also for \(\sup_{1 \leq k \leq n} |c_{\tau} - c_{\tau,k}|\). \(\square\)

We are now ready to prove the last proposition of this section, which will help us get the second equation of our System (4).

**Proposition 3.22** Under Assumptions O1–O7 and P1–P4,
\[
\left(\frac{p}{n}\right)^2 \mathbb{E}\left(\|\hat{\beta} - \beta_0\|_2^2\right) = \frac{p}{n} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left((c_{\tau} \lambda_i \psi_i(\text{prox}(c_{\tau} \lambda_i^2 \rho_i)(\tilde{r}_{i,(i)})))^2\right) + \tau^2 \|\beta_0\|_2^2 \mathbb{E}\left(c_{\tau}^2\right) + o(1).
\]

Furthermore, when all \(\lambda_i\)'s are non-zero,
\[
\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left((c_{\tau} \lambda_i \psi_i(\text{prox}(c_{\tau} \lambda_i^2 \rho_i)(\tilde{r}_{i,(i)})))^2\right) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left(\frac{[\tilde{r}_{i,(i)} - \text{prox}(c_{\tau} \lambda_i^2 \rho_i)(\tilde{r}_{i,(i)})]^2}{\lambda_i^2}\right).
\]

**Proof** In light of the result in Proposition 3.21 and Assumption P3 which guarantees that \(\|\beta_0\|_2^2\) is uniformly bounded in \(p\) and \(n\), we see that
\[
\sum_{k=1}^{p} \beta_0^2(k) \mathbb{E}\left(c_{\tau,k}^2\right) = \sum_{k=1}^{p} \beta_0^2(k) \mathbb{E}\left(c_{\tau}^2\right) + o(1) = \|\beta_0\|_2^2 \mathbb{E}\left(c_{\tau}^2\right) + o(1).
\]

Therefore, Eq. (47) implies that
\[
\left(\frac{p}{n}\right)^2 \mathbb{E}\left(\|\hat{\beta} - \beta_0\|_2^2\right) = \frac{p}{n} \frac{1}{p} \sum_{k=1}^{p} \left[\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left((c_{\tau,k} \lambda_i \psi_i(r_{i,(k)}))^2\right)\right] + \tau^2 \|\beta_0\|_2^2 \mathbb{E}\left(c_{\tau}^2\right) + o(1).
\]
Using Theorem 3.20 and our bound on \( \| \psi_i' \|_\infty \) from Assumption O3, we see that
\[
\frac{1}{p} \sum_{k=1}^{p} E \left( \left[ c_{\tau,k} \lambda_i \psi_i (r_{i,[k]}) \right]^2 \right) = \frac{1}{p} \sum_{k=1}^{p} E \left( \left[ c_{\tau,k} \lambda_i \psi_i (R_i) \right]^2 \right) + o(1).
\]

Thanks to Proposition 3.21 we also have
\[
\frac{1}{p} \sum_{k=1}^{p} E \left( \left[ c_{\tau,k} \lambda_i \psi_i (R_i) \right]^2 \right) = \frac{1}{p} \sum_{k=1}^{p} E \left( \left[ c_{\tau,k} \lambda_i \psi_i (R_i) \right]^2 \right) + o(1)
\]
\[
= E \left( \left[ c_{\tau,k} \lambda_i \psi_i (R_i) \right]^2 \right) + o(1).
\]

In light of Eq. (27) and Assumption O3, we have
\[
\frac{1}{n} \sum_{i=1}^{n} E \left( \left[ c_{\tau} \lambda_i \psi_i (R_i) \right]^2 \right) = \frac{1}{n} \sum_{i=1}^{n} E \left( \left[ c_{\tau} \lambda_i \psi_i (\text{prox}(c_i \rho_i)(\tilde{r}_{i,(i)})) \right]^2 \right) + o(1).
\]

Lemma 3.32 and specifically the computation of the derivative of \( \text{prox}(c\rho)(x) \) with respect to \( c \), allows us to bound the error \( |\psi_i(\text{prox}(c_i \rho_i)(x)) - \psi_i(\text{prox}(c \lambda_i^2 \rho_i)(x))| \) for all \( x \). In light of this, we see that, by using Corollary 3.17, we can re-express the previous equation as
\[
\frac{1}{n} \sum_{i=1}^{n} E \left( \left[ c_{\tau} \lambda_i \psi_i (R_i) \right]^2 \right) = \frac{1}{n} \sum_{i=1}^{n} E \left( \left[ c_{\tau} \lambda_i \psi_i (\text{prox}(c \lambda_i^2 \rho_i)(\tilde{r}_{i,(i)})) \right]^2 \right) + o(1),
\]

since Eq. (44) in Corollary 3.17, gives \( \sup_i |c_i - \lambda_i^2 c_{\tau}| = O_{\lambda_k}(n^{2a-1/2}\text{polyLog}(n)) \).

When \( \lambda_i \)'s are all different from 0, we can rewrite this equation as
\[
\frac{1}{n} \sum_{i=1}^{n} E \left( \left[ c_{\tau} \lambda_i \psi_i (R_i) \right]^2 \right) = \frac{1}{n} \sum_{i=1}^{n} E \left( \left[ \frac{c_{\tau} \lambda_i^2 \psi_i (\text{prox}(c \lambda_i^2 \rho_i)(\tilde{r}_{i,(i)})))}{\lambda_i^2} \right]^2 \right) + o(1).
\]

Finally, since almost by definition,
\[
\forall x \in \mathbb{R}, x = \text{prox}(c\rho)(x) + c\psi(\text{prox}(c\rho)(x)),
\]

we have
\[
\frac{1}{n} \sum_{i=1}^{n} E \left( \left[ \frac{c_{\tau} \lambda_i^2 \psi_i (\text{prox}(c \lambda_i^2 \rho_i)(\tilde{r}_{i,(i)})))}{\lambda_i^2} \right]^2 \right)
\]
\[
= \frac{1}{n} \sum_{i=1}^{n} E \left( \frac{\tilde{r}_{i,(i)} - \text{prox}(c \lambda_i^2 \rho_i)(\tilde{r}_{i,(i)}))}{\lambda_i^2} \right)^2 .
\]
\[\Box\]
Appendix 5: Last steps of the proof

We now reach the last steps of the proof and two important tasks remain to be completed. The first one is understanding the limiting behavior of \( \tilde{r}_{i,(i)} \) and showing that it behaves like \( \epsilon_i + \lambda_i r_p(\epsilon) Z_i \) in the limit, where \( Z_i \sim N(0, 1) \). With a little bit of further work, the corresponding results will give us in connection with Proposition 3.22 the second equation of our main system (4).

The second main task is then to show that \( c_T \) is asymptotically deterministic, i.e. it converges towards a non-random number.

On the asymptotic distribution of \( \tilde{r}_{i,(i)} \)

We have the following lemma.

**Lemma 3.23** Under Assumptions O1–O7 and P1–P4, as \( n \) and \( p \) tend to infinity, \( \tilde{r}_{i,(i)} = Y_i - X_i^\prime \hat{\beta}(i) \) behaves like \( \epsilon_i + \lambda_i \sqrt{\mathbf{E} (\| \hat{\beta} - \beta_0 \|^2)} Z_i \), where \( Z_i \sim N(0, 1) \) is independent of \( \epsilon_i \) and \( \lambda_i \), in the sense of weak convergence.

Furthermore, if \( i \neq j \), \( \tilde{r}_{i,(i)} \) and \( \tilde{r}_{j,(j)} \) are asymptotically (pairwise) independent. The same is true for the pairs \( (\tilde{r}_{i,(i)}, \lambda_i) \) and \( (\tilde{r}_{j,(j)}, \lambda_j) \).

**Proof** We recall that \( X_i = \lambda_i X_i^0 \) and hence \( \tilde{r}_{i,(i)} = Y_i - X_i^\prime \hat{\beta}(i) = \epsilon_i - \lambda_i X_i^0 (\hat{\beta}(i) - \beta_0) \).

- **First part** The only problem is of course showing that \( (\hat{\beta}(i) - \beta_0)^\prime X_i \) is approximately \( N(0, \mathbf{E} (\| \hat{\beta} - \beta_0 \|^2)) \). Recall that \( \hat{\beta}(i) \) is independent of \( X_i^0 \) and that \( X_i^0 \) has mean 0, \( \text{cov} (X_i^0) = \text{Id}_p \) and that, for any finite \( k \), the first \( k \) absolute moments of its entries are assumed to be bounded uniformly in \( n \).

Recall that we showed in Proposition 3.10 that \( \text{var} (\| \hat{\beta} - \beta_0 \|^2) \to 0 \). Thanks to Lemma 3.3, we also know that \( \mathbf{E} (\| \hat{\beta} - \beta_0 \|^2) \) is uniformly bounded. Furthermore, in the proof of Proposition 3.10, we showed that \( \mathbf{E} (\| \hat{\beta} - \beta_0 \|^2) \to 0 \) and that \( \mathbf{E} (\| \hat{\beta} - \beta_0 \|^2 - \| \hat{\beta}(i) - \beta_0 \|^2) \to 0 \).

Let us now show that \( (\hat{\beta}(i) - \beta_0)^\prime X_i \) behaves like \( \mathcal{N}(0, \mathbf{E} (\| \hat{\beta}(i) - \beta_0 \|^2)) \). We employ a similar strategy as was done in [15] but give the argument in details since it requires some new work.

We need a simple generalization of the standard Lindeberg-Feller theorem (see e.g. [39]). Indeed, if \( a_{n,p}(k) \) are random variables with \( \sqrt{\sum_{k=1}^p a_{n,p}(k)^2} \sim A_n, \mathbf{E} (A_n^2) \) remains bounded in \( n \), and \( a_{n,p}(k) \)'s are independent of \( X_i \), we see that: a) if \( Z \sim \mathcal{N}(0, \text{Id}_p) \), independent of \( a_{n,p}(k) \), then \( a_{n,p}^\prime Z \sim A_n N \) where \( N \sim \mathcal{N}(0, 1) \) and independent of \( A_n \) (conditionally and unconditionally on \( a_{n,p} \)); b) Theorem 2.1.5 and its proof in [39] hold provided \( \sum_{i=1}^n \mathbf{E} (|a_{n,p}(k)|^3) = o(1) \). The proof simply needs to be started conditionally on \( a_{n,p} \), and the final moment bounds are then taken unconditionally. This very mild generalization gives, if \( \phi \) is a \( C^3 \) function, with bounded 2nd and third derivatives,
\[ \forall \varepsilon > 0, \quad \left| E \left( \phi(a_{n,p}X_i) \right) - E (\phi(A_nN)) \right| \leq K \left( \varepsilon \|\phi^{(3)}\|_\infty E \left( \sum_{k=1}^p a_{n,p}(k)^2 \right) + \frac{\|\phi^{(2)}\|_\infty}{\varepsilon} \sum_{k=1}^p E \left( |a_{n,p}(k)|^3 \right) \right), \]

where \( K \) is a constant that depends on the second and third absolute moments of the entries of \( X_i \). It is therefore independent of \( n \) and \( p \) under our assumptions on \( X_i \).

To make matters clearer, we allow ourselves to use the notation \( \nu_k \) or \( \nu(k) \) to refer to the \( k \)th coordinate of the vector \( \nu \).

In our setting, \( a_{n,p}(k) = \hat{\beta}(i)(k) - \beta_0(k) \). Recall that we have shown that

\[ \hat{\beta}(p) - \beta_0(p) = O_{L_k} \left( \frac{\text{polyLog}(n)n^{3\alpha}}{[n^{1/2} \wedge n^\theta]^2} \right). \]

The same arguments we used apply also to \( \hat{\beta}(i)(p) \), the \( p \)th coordinate of the leave-one-out estimate \( \hat{\beta}(i) \). So it is clear that

\[ E \left( |\hat{\beta}(i)(p) - \beta_0(p)|^3 \right) = O \left( \frac{\text{polyLog}(n)n^{3\alpha}}{[n^{1/2} \wedge n^\theta]^6} \right). \]

We conclude that

\[ E \left( \sum_{k=1}^p |\hat{\beta}(i)(k) - \beta_0(k)|^3 \right) = O \left( \frac{\text{polyLog}(n)n^{3\alpha+1}}{[n^{1/2} \wedge n^\theta]^6} \right) = o(1). \]

This, in connection with Corollary 2.1.9 in [39], shows that \( (\hat{\beta}(i) - \beta_0)^\prime X_i \) behaves asymptotically like \( \|\hat{\beta}(i) - \beta_0\|N \) in the sense of weak convergence.

Since \( \|\hat{\beta}(i) - \beta_0\| - E (\|\hat{\beta}(i) - \beta_0\|) \rightarrow 0 \) in probability and \( E (\|\hat{\beta}(i) - \beta_0\|) \) remains bounded under our assumptions, Slutsky’s lemma guarantees that

\[ (\hat{\beta}(i) - \beta_0)^\prime X_i \text{ behaves like } E (\|\hat{\beta}(i) - \beta_0\|)N \]

asymptotically, in the sense of weak convergence—by which we mean that the difference of their characteristic functions goes to 0. Using the fact, which can be shown using results in the proof of Proposition 3.10, that \( E (\|\hat{\beta}(i) - \beta_0\|) - E (\|\beta - \beta_0\|) \rightarrow 0 \) and Slutsky’s lemma, we see that

\[ (\hat{\beta}(i) - \beta_0)^\prime X_i \text{ behaves like } E (\|\beta - \beta_0\|)N. \]

in the sense of weak convergence.

We note that the same reasoning applies when replacing \( a_{n,p}(k) = \hat{\beta}(i)(k) - \beta_0(k) \) by \( \tilde{a}_{n,p}(k) = \lambda_i [\hat{\beta}(i)(k) - \beta_0(k)] \), provided \( \lambda_i \) has 3 moments. This shows that

\[ \lambda_i (\hat{\beta}(i) - \beta_0)^\prime X_i = (\hat{\beta}(i) - \beta_0)^\prime X_i \text{ behaves like } E (\|\beta - \beta_0\|) \lambda_i N. \]
This shows the first part of the lemma, since
\[ E \left( \| \hat{\beta} - \beta_0 \| \right) = \sqrt{E \left( \| \hat{\beta} - \beta_0 \|^2 \right) + o(1) } \]
by Proposition 3.10.

**Second part** For the second part, we use a leave-two-out approach, namely we use the approximation
\[ \tilde{r}_{i,(j)} = \epsilon_i + X_i' \beta_0 - \tilde{\beta}_i X_i = \epsilon_i + X_i' \beta_0 - \tilde{\beta}_{i(i)} X_i + o_L(1) \]
and similarly for \( \tilde{r}_{j,(j)} \) (this is clear from Theorem 2.2; \( \tilde{\beta}_{i(i)} \) is computed by solving Problem (3) without \( (X_i, Y_i) \) nor \( (X_j, Y_j) \)). It is clear that \( \tilde{r}_{i,(i)} \) and \( \tilde{r}_{j,(j)} \) are asymptotically independent conditional on \( X_{i(i)} \), i.e. all the predictors except \( X_i \) and \( X_j \). But because their dependence on \( X_{i(i)} \) is only through \( \| \tilde{\beta}_{i(i)} - \beta_0 \| \), which is asymptotically deterministic by arguments similar to those used in the proof of Proposition 3.10, we see that \( \tilde{r}_{i,(i)} \) and \( \tilde{r}_{j,(j)} \) are asymptotically independent.

After this high-level explanation, let us now give a detailed proof. The arguments we gave above apply to \( \tilde{\beta}_{i(i)} \) as they did to \( \hat{\beta}_{i(i)} \). In particular, since
\[ E \left( \sum_{k=1}^{p} | \tilde{\beta}_{i(i)}(k) - \beta_0(k) |^3 \right) = O \left( \frac{\text{polyLog}(n)^{3\alpha + 1}}{n^{1/2} \wedge \text{n}^{\eta} \text{e}^6} \right) = o(1), \]
we also have
\[ \sum_{k=1}^{p} | \tilde{\beta}_{i(i)}(k) - \beta_0(k) |^3 = o_P(1). \]

Of course, \( \tilde{\beta}_{i(i)} \) depends only on \( \{X_{i(i)}, \epsilon_{i(i)}\} \). We call \( P_{i(i)} \) the joint probability measure \( P_{i(i)} = \prod_{k \neq (i,i)} P_{X_k, \epsilon_k} \), i.e. probability computed with respect to all our random variables except \( (X_i, \epsilon_i) \) and \( (X_j, \epsilon_j) \) (we slightly abuse notation and do not index this probability measure by \( n \) for the sake of clarity).

So we have found \( E_{i(i)}^{n} \), depending only on \( (X_{i(i)}, \epsilon_{i(i)}) \), such that \( P_{i(i)}(E_{i(i)}^{n}) \rightarrow 1 \) and \( \sum_{k=1}^{p} | \tilde{\beta}_{i(i)}(k) - \beta_0(k) |^3 = o(1) \) when \( (X_{i(i)}, \epsilon_{k})_{k \neq (i,i)} \in E_{i(i)}^{n} \). The arguments we gave above (treating \( a_nP \)'s as deterministic quantities) then imply that, when \( (X_{i(i)}, \epsilon_{i(i)}) \in E_{i(i)}^{n} \),
\[ (\tilde{\beta}_{(i)} - \beta_0)'X_{i}(i)|X_{i(i), \epsilon_{i(i)}} \] behaves like \( \| \tilde{\beta}_{i(i)} - \beta_0 \| \mathcal{N} \).

Let us now use characteristic function arguments. Let \( \alpha_i = (\tilde{\beta}_{i(i)} - \beta_0)'X_{i} \) and \( \alpha_j = (\tilde{\beta}_{i(i)} - \beta_0)'X_{j} \).

Let \( (w_i, w_j) \in \mathbb{R}^2 \) be fixed and
\[ \chi(w_i, w_j) = E \left( e^{(w_1 \alpha_{i} + w_2 \alpha_{j})} \right) = E \left( e^{(w_1 \alpha_{i} + w_2 \alpha_{j})} 1_{E_{i(i)}^{n}} + 1_{E_{i(i)}^{n}}'c \right). \]

Since \( P(\{E_{i(i)}^{n}\}) = P_{i(i)}(\{E_{i(i)}^{n}\}) \rightarrow 0 \), we can just focus on \( E \left( e^{(w_1 \alpha_{i} + w_2 \alpha_{j})} 1_{E_{i(i)}^{n}} \right) \), since the modulus of the functions we are integrating is bounded by 1.

Now
\[ E \left( e^{(w_1 \alpha_{i} + w_2 \alpha_{j})} 1_{E_{i(i)}^{n}} \right) = E \left( 1_{E_{i(i)}^{n}} E \left( e^{(w_1 \alpha_{i} + w_2 \alpha_{j})} |X_{i(i), \epsilon_{i(i)}}\right) \right), \]
since $1_E^n_{(ij)}$ is a deterministic function of $(X_{(ij)}, \epsilon_{(ij)})$. Independence of $\mathcal{X}_i$ and $\mathcal{X}_j$ implies that

$$E \left( e^{t(w_1\alpha_i + w_2\alpha_j)} | X_{(ij)}, \epsilon_{(ij)} \right) = E \left( e^{t w_1\alpha_i} | X_{(ij)}, \epsilon_{(ij)} \right) E \left( e^{t w_2\alpha_j} | X_{(ij)}, \epsilon_{(ij)} \right).$$

Also, our conditional asymptotic normality arguments above imply that

$$1_{E^n_{(ij)}} \left[ E \left( e^{t w_1\alpha_i} | X_{(ij)}, \epsilon_{(ij)} \right) - e^{-w_1^2/2}\|\widehat{\beta}_{(ij)} - \beta_0\|^2 \right] \to 0$$

in $P_{(ij)}$-probability. We therefore have

$$1_{E^n_{(ij)}} \left[ E \left( e^{t(w_1\alpha_i + w_2\alpha_j)} | X_{(ij)}, \epsilon_{(ij)} \right) - e^{-(w_1^2/2+w_2^2/2)}\|\widehat{\beta}_{(ij)} - \beta_0\|^2 \right] \to 0$$

in $P_{(ij)}$-probability.

So we conclude that

$$E \left( 1_{E^n_{(ij)}} e^{t(w_1\alpha_i + w_2\alpha_j)} \right) - E \left( 1_{E^n_{(ij)}} e^{-(w_1^2/2+w_2^2/2)}\|\widehat{\beta}_{(ij)} - \beta_0\|^2 \right) \to 0.$$

Since $P(E^n_{(ij)}) \to 1$ and $\|\widehat{\beta}_{(ij)} - \beta_0\|^2$ is asymptotically deterministic by arguments similar to those used in the proof of Proposition 3.10, we see that

$$E \left( 1_{E^n_{(ij)}} e^{-(w_1^2/2+w_2^2/2)}\|\widehat{\beta}_{(ij)} - \beta_0\|^2 \right) - e^{-(w_1^2/2+w_2^2/2)}E([\|\widehat{\beta}_{(ij)} - \beta_0\|^2]) \to 0.$$

Therefore,

$$E \left( e^{t(w_1\alpha_i + w_2\alpha_j)} \right) - E \left( e^{t w_1\alpha_i} \right) E \left( e^{t w_2\alpha_j} \right) \to 0.$$

This proves that $\alpha_i$ and $\alpha_j$ are asymptotically independent. It easily follows that the same is true for $\tilde{\alpha}_{i,(i)}$ and $\tilde{\alpha}_{j,(i)}$.

The same leave-two-out approach also shows asymptotic pairwise independence of the pairs $(\lambda_i, \tilde{\beta}_{i,(i)})$ and $(\lambda_j, \tilde{\beta}_{j,(j)})$, since $\widehat{\beta}_{(ij)}$ is independent of $\lambda_i$ and $\lambda_j$ under Assumption O6, which guarantees independence of the $\lambda_i$’s.

The lemma is shown. 

\[ \square \]

**On the asymptotic behavior of $c_\tau$**

We are now in position to show that $c_\tau = \frac{1}{n} \text{trace} \left( (S + \tau \text{Id}_p)^{-1} \right)$ is asymptotically deterministic. This result will require several steps.

**Lemma 3.24** We work under Assumptions O1–O7, P1–P4 and F2–F4.

Consider the random function

$$g_n(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{x + \lambda_i^2 \bar{\psi}'_i(\text{prox}(x\lambda_i^2 \rho_i))}, \text{ defined for } x \geq 0.$$
Let $B > 0$ be in $\mathbb{R}_+$. We have, for any $(x, y) \in \mathbb{R}_+^2$, and $0 \leq x \leq B$, $0 \leq y \leq B$,

\[
\sup_{(x,y):|x-y| \leq \eta, 0 \leq x \leq B, 0 \leq y \leq B} |g_n(x) - g_n(y)| \leq \eta \frac{1}{n} \sum_{i=1}^{n} \left( \lambda_i^2 \| \psi_i \|_{\infty} + B \mathbb{L}_i(n) \lambda_i^4 \| \psi_i \|_{\infty} \right).
\]

In particular, under $P_2$ and $F_3$–$F_4$ we have, for $C$ a constant independent of $n$ and $p$,

\[
P^* \left( \sup_{(x,y):|x-y| \leq \eta, 0 \leq x \leq B, 0 \leq y \leq B} |g_n(x) - g_n(y)| > \delta \right) \leq \frac{\eta}{\delta} C.
\tag{49}
\]

Hence, $g_n$ is stochastically equicontinuous on $[0, B]$ for any $B > 0$ given.

We used the notation $P^*$ above to denote outer probability and avoid a discussion of potential measure theoretic issues associated with taking a supremum over a non-countable collection of random variables (see e.g. [41, Sect. 18.2]). We refer the reader to e.g. [30] for more details on stochastic equicontinuity. While we could probably avoid appealing to abstract concepts like outer measures here, we use this approach because it is a standard tool in empirical process theory and helps us avoid side measurability discussions that would distract us from the main focus of our proof.

**Proof** Let us consider the function, defined for $x \geq 0$,

\[
h_{u}^{(i)}(x) = \frac{1}{1 + x \lambda_i^2 \psi_i'(\text{prox}(x \lambda_i^2 \rho_i)(u))} = \frac{\partial}{\partial u} \text{prox}(x \lambda_i^2 \rho_i)(u).
\]

The last equality comes from Lemma 3.33.

We have, since $\psi_i'$ is non-negative because $\rho_i$ is convex,

\[
\forall u, \quad \left| h_{u}^{(i)}(x) - h_{u}^{(i)}(y) \right| \leq |x \lambda_i^2 \psi_i'(\text{prox}(x \lambda_i^2 \rho_i)(u)) - y \lambda_i^2 \psi_i'(\text{prox}(y \lambda_i^2 \rho_i)(u))| \wedge 1.
\]

Therefore, since $x, y \geq 0$, for all $u$,

\[
\left| h_{u}^{(i)}(x) - h_{u}^{(i)}(y) \right| \leq \lambda_i^2 |x - y| \psi_i'(\text{prox}(x \lambda_i^2 \rho_i)(u)) + \lambda_i^2 y |\psi_i'(\text{prox}(x \lambda_i^2 \rho_i)(u))
\]

\[
- \psi_i'(\text{prox}(y \lambda_i^2 \rho_i)(u))|.
\]

In particular, if $|x - y| \leq \eta$, and $x \vee y \leq B$, with $x, y \geq 0$, for all $u$,

\[
\sup_{y:|x-y| \leq \eta; x \vee y \leq B} \left| h_{u}^{(i)}(x) - h_{u}^{(i)}(y) \right| \leq \lambda_i^2 \eta \psi_i'(\text{prox}(x \lambda_i^2 \rho_i)(u))
\]

\[
+ B \lambda_i^2 \sup_{y:|x-y| \leq \eta; x \vee y \leq B} |\psi_i'(\text{prox}(x \lambda_i^2 \rho_i)(u))
\]

\[
- \psi_i'(\text{prox}(y \lambda_i^2 \rho_i)(u))|.\]
Under assumption O3, $\psi'_i$ is $L_i(n)$-Lipschitz. Therefore, for $x_i = x\lambda_i^2, y_i = y\lambda_i^2 \geq 0$,

$$\forall u, \ |\psi'_i(\text{prox}(x_i\rho_i)(u)) - \psi'_i(\text{prox}(y_i\rho_i)(u))| \leq L_i(n)|\text{prox}(x_i\rho_i)(u) - \text{prox}(y_i\rho_i)(u)|.$$ 

We recall that, according to Lemma 3.32,

$$\frac{\partial}{\partial x} \text{prox}(x\rho_i)(u) = -\frac{\psi_i(\text{prox}(x\rho_i)(u))}{1 + x\psi'_i(\text{prox}(x\rho_i)(u))}.$$ 

Hence,

$$\sup_x |\frac{\partial}{\partial x} \text{prox}(x\rho_i)(u)| \leq \|\psi_i\|_{\infty}.$$ 

We finally conclude that

$$\forall u, \ |\psi'_i(\text{prox}(x_i\rho)(u)) - \psi'_i(\text{prox}(y_i\rho)(u))| \leq [L_i(n)\|\psi_i\|_{\infty}|x_i - y_i|] \wedge 2\|\psi'_i\|_{\infty}.$$ 

We therefore have, when $x \lor y \leq B$, and $x, y \geq 0$,

$$\forall u, \sup_{y:|x-y| \leq \eta} \left| h^{(i)}_u (x) - h^{(i)}_u (y) \right| \leq \lambda_i^2 \eta \|\psi'_i(\text{prox}(x\lambda_i^2\rho_i)(u))\| + B\lambda_i^4 L_i(n)\|\psi_i\|_{\infty}\eta.$$ 

Therefore, for $x, y \geq 0$,

$$\forall u, \sup_{(x, y):|x-y| \leq \eta, x \lor y \leq B} \left| h^{(i)}_u (x) - h^{(i)}_u (y) \right| \leq \lambda_i^2 \eta \|\psi'_i\|_{\infty} + \eta BL_i(n)\lambda_i^4 \|\psi_i\|_{\infty}.$$ 

Since the right-hand side does not depend on $u$, we also have

$$\sup_u \sup_{(x, y):|x-y| \leq \eta, x \lor y \leq B} \left| h^{(i)}_u (x) - h^{(i)}_u (y) \right| \leq \lambda_i^2 \eta \|\psi'_i\|_{\infty} + \eta BL_i(n)\lambda_i^4 \|\psi_i\|_{\infty}.$$ 

Naturally, $g_n(x)$ can be written as

$$g_n(x) = \frac{1}{n} \sum_{i=1}^{n} h^{(i)}_{r_{i,i}} (x).$$ 

Therefore, for any $x, y \geq 0$, we have

$$|g_n(x) - g_n(y)| \leq \frac{1}{n} \sum_{i=1}^{n} |h^{(i)}_{r_{i,i}} (x) - h^{(i)}_{r_{i,i}} (y)|.$$
The bound we have obtained above on \( \sup_{u} |h_u^{(i)}(x) - h_u^{(i)}(y)| \) when \( x \) and \( y \) are sufficiently close to one another can now be used. This shows that for \( x \) given, if \( x, y \geq 0, |x - y| \leq \eta \) and \( x \lor y \leq B \), we have

\[
\sup_{(x, y): |x-y| \leq \eta, 0 \leq x \leq B, 0 \leq y \leq B} |g_n(x) - g_n(y)| \leq \eta \frac{1}{n} \sum_{i=1}^{n} \left( \lambda_i^2 \|\psi_i\|_{\infty} + B L_{\lambda i}(n) \lambda_i^4 \|\psi_i\|_{\infty} \right).
\]

Under assumptions \( P_2 \) and \( F_3-F_4 \), we can now take expectations and get the result in \( L_1 \), since all the terms on the right hand side are bounded in \( L_1 \) under those assumptions.

We have established stochastic equicontinuity of \( g_n(x) \) on \([0, B]\).

**Lemma 3.25** Let us call \( G_n(x) = E(g_n(x)) \). Let \( B > 0 \) be given. For any given \( x_0 \leq B \),

\[
g_n(x_0) - G_n(x_0) = o_{L_2}(1).
\]

Under Assumptions \( O_1-O_7, P_1-P_4 \) and \( F_1-F_5 \), we also have

\[
E^{*} \left( \sup_{0 \leq x \leq B} |g_n(x) - G_n(x)| \right) \to 0.
\]

**Proof** Under assumptions \( F_1 \) and \( F_5 \), we can divide the index set \([1, \ldots, n]\) into \( K \) subsets \( A_1, \ldots, A_K \), where \( K \) is finite (with \( n \)), in which \((X_i, \epsilon_i)_{i \in A_j}\) play a symmetric role. Hence, \( \text{var}(g_n(x_0)) \) can be expressed as a sum of variances and covariances of finitely many functions of finitely many random variables \((\lambda_i, \tilde{r}_{i,(i)})\); for those random variables, we just need to pick a representative in each subset \( \{A_j\}_{j=1}^{K} \).

We note that since \( \psi_i' \) is Lipschitz and hence continuous, \( g_n \) is an average of bounded continuous functions of the random variables of interest to us.

Asymptotic pairwise independence of \((\lambda_i, \tilde{r}_{i,(i)})\)'s, and the fact that \( \psi_i' \) can only be one of finitely many functions imply that

\[
\text{var}(g_n(x_0)) \to 0
\]

and therefore gives the first result.

Let us now pick \( \epsilon > 0 \). By the stochastic equicontinuity of \( g_n \) and our bound in Eq. (49), we can find \( x_1, \ldots, x_K \), independent of \( n \), such that for all \( x \in [0, B] \), there exists \( l \) such that, when \( n \) is large enough,

\[
E(|g_n(x) - g_n(x_l)|) \leq \epsilon.
\]

Note that

\[
|g_n(x) - G_n(x)| \leq |g_n(x) - g_n(x_l)| + |g_n(x_l) - G_n(x_l)| + |G_n(x_l) - G_n(x)|.
\]
We immediately get
\[ E^* \left( \sup_{0 \leq x \leq B} |g_n(x) - G_n(x)| \right) \leq 2\epsilon + E \left( \sup_{1 \leq l \leq K} |g_n(x_l) - G_n(x_l)| \right). \]

Because \( K \) is finite, the fact that for all \( l \), \( |g_n(x_l) - G_n(x_l)| \to 0 \) in \( L_2 \) implies that \( \sup_{1 \leq l \leq K} |g_n(x_l) - G_n(x_l)| \to 0 \) in \( L_2 \). In particular, if \( n \) is sufficiently large,
\[ E \left( \sup_{1 \leq l \leq K} |g_n(x_l) - G_n(x_l)| \right) \leq \epsilon. \]
The lemma is shown. \( \square \)

**Lemma 3.26** Assume \( O1–O7, P1–P4 \) and \( F1–F5 \). Call \( c_\tau = \frac{1}{n} \text{trace} \left( (S + \tau \text{Id}_p)^{-1} \right) \).

Call as before
\[ g_n(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{1 + x\lambda_i^2 \psi_i'(\text{prox}(x\lambda_i^2 \rho_i)(\tilde{f}_{i,l}))}. \]

Then \( c_\tau \) is a near solution of
\[ \frac{p}{n} - \tau x - 1 + g_n(x) = 0, \]
i.e. \( \frac{p}{n} - \tau c_\tau - 1 + g_n(c_\tau) = o_{L_k} (1) \), when \( 3\alpha - 1/2 < 0 \).

Asymptotically, near solutions of
\[ \delta_n(x) \triangleq \frac{p}{n} - \tau x - 1 + g_n(x) = 0, \]
are close to solutions of
\[ \Delta_n(x) = \frac{p}{n} - \tau x - 1 + E \left( g_n(x) \right) = 0. \]

More precisely, call \( T_{n,\epsilon} = \{ x : |\Delta_n(x)| \leq \epsilon \} \). Note that \( T_{n,\epsilon} \subseteq (0, \frac{p}{n\tau} + \epsilon/\tau) \).

For any given \( \epsilon \), as \( n \to \infty \), near solutions of \( \delta_n(x_n) = 0 \) belong to \( T_{n,\epsilon} \) with high-probability.

Our assumptions concerning the possible distributions of \( \epsilon_i's \), specifically \( F1 \), guarantee that as \( n \to \infty \), there is a unique solution to \( \Delta_n(x) = 0 \).

Hence \( c_\tau \) is asymptotically deterministic.

**Proof** Note that \( g_n(x) \leq 1 \).

Let \( \delta_n \) be the function
\[ \delta_n(x) = \frac{p}{n} - \tau x - 1 + g_n(x), \]
and $\Delta_n(x) = \mathbb{E}(\delta_n(x))$. Call $x_n$ a solution $\delta_n(x_n) = 0$ and $x_{n,0}$ a solution of $\Delta_n(x_{n,0}) = 0$. Since $0 \leq g_n \leq 1$, we see that $x_n \leq p/(n\tau)$, for otherwise, $\delta_n(x) < 0$. The same argument shows that if $x > (p/n + \epsilon)/\tau$, $\Delta_n(x) < -\epsilon$ and $x \notin T_{n,\epsilon}$.

Similarly, near solutions of $\delta_n(x) = 0$ must be less or equal to $(p/n + \epsilon)/\tau$.

- **Proof of the fact that $c_\tau$ is such that $\delta_n(c_\tau) = o(1)$**

An important remark is that $c_\tau$ is a near solution of $\delta_n(x) = 0$. This follows most clearly from arguments we have developed for $c_{\tau, p}$ so we start by giving details through arguments for this random variable. Recall that in the notation of Lemma 3.14, we had

$$\frac{p - 1}{n} - \tau c_{\tau, p} = \frac{1}{n} \text{trace } (\text{Id}_n - M).$$

Now, according to Eq. (40),

$$\frac{1}{n} \text{trace } (\text{Id}_n - M) = 1 - \frac{1}{n} \sum_{i=1}^{n} \frac{1}{1 + \psi_i'(r_{i,\{p\}})} \frac{1}{\psi_i'(\bar{\mathbb{G}}_p(i) + \tau \text{Id})^{-1} V_i}.$$

According to Lemmas 3.14, 3.15 and 3.16, we have

$$\sup_i \left| \frac{1}{n} \psi_i'(\bar{\mathbb{G}}_p(i) + \tau \text{Id})^{-1} V_i - \lambda_i^2 c_{\tau, p} \right| = O_L_k \left( \frac{\text{polyLog}(n)}{n^{1/2-2\alpha}} \right).$$

Of course, when $x \geq 0$ and $y \geq 0$, $|1/(1+x) - 1/(1+y)| \leq |x - y| \wedge 1$. Hence, we see that

$$\left| \frac{1}{n} \sum_{i=1}^{n} \frac{1}{1 + \psi_i'(r_{i,\{p\}})} \frac{1}{\psi_i'(\bar{\mathbb{G}}_p(i) + \tau \text{Id})^{-1} V_i} - \frac{1}{n} \sum_{i=1}^{n} \frac{1}{1 + \psi_i'(r_{i,\{p\}})\lambda_i^2 c_{\tau, p}} \right| \leq \sup_{1 \leq i \leq n} \left| \frac{1}{n} \psi_i'(\bar{\mathbb{G}}_p(i) + \tau \text{Id})^{-1} V_i - \lambda_i^2 c_{\tau, p} \right| \frac{1}{n} \sum_{i=1}^{n} \|\psi_i\|_\infty.$$

We conclude that

$$\frac{p}{n} - \tau c_{\tau, p} - 1 + \frac{1}{n} \sum_{i=1}^{n} \frac{1}{1 + \lambda_i^2 c_{\tau, p} \psi_i'(r_{i,\{p\}})} = O_L_k (n^{-1/2+2\alpha} \text{polyLog}(n)).$$

Exactly the same computations can be made with $c_\tau$, so we have established that

$$\frac{p}{n} - \tau c - 1 + \frac{1}{n} \sum_{i=1}^{n} \frac{1}{1 + c_\tau \lambda_i^2 \psi_i'(R_i)} = O_L_k (n^{-1/2+2\alpha} \text{polyLog}(n)).$$

(50)

Now we have seen in Theorem 2.2 that

$$\sup_i |R_i - \text{prox}(c_i r_i)(\hat{r}_{i,(i)})| = O_L_k (n^{-1/2+2\alpha} \text{polyLog}(n)).$$

$\diamond$ Springer
We have furthermore noted that sup
\[ \sup_i |\psi'_i(R_i) - \psi'_i[\text{prox}(c_i,\rho_i)](\hat{r}_{i,(i)})| = O_{L_k}(n^{-1/2+2\alpha}\text{polyLog}(n)). \]

Through our assumptions on \( \psi'_i \), this of course implies that
\[
\sup_i |\psi'_i(R_i) - \psi'_i[\text{prox}(c_i,\rho_i)](\hat{r}_{i,(i)})| = O_{L_k}(n^{-1/2+2\alpha}\text{polyLog}(n)).
\]

Gathering everything together, we get
\[
\left| \psi'_i[\text{prox}(c_i,\rho_i)](\hat{r}_{i,(i)}) - \psi'_i[\text{prox}(\lambda_i c_{T})(\hat{r}_{i,(i)})] \right| \leq \|\psi_i\|_{\infty}|c_i - \lambda_i c_{T}|
\]
and hence
\[
\left| \psi'_i[\text{prox}(c_i,\rho_i)](\hat{r}_{i,(i)}) - \psi'_i[\text{prox}(\lambda_i c_{T}, \rho_i)](\hat{r}_{i,(i)}) \right| = O_{L_k}(\|\psi_i\|_{\infty}n^{-1/2+3\alpha}\text{polyLog}(n)).
\]

So we have established that \( \delta_n(c_{T}) = O_{L_k}(n^{-1/2+3\alpha}\text{polyLog}(n)) \).

**Final details**

Note that for any given \( x \), \( \delta_n(x) - \Delta_n(x) = o_P(1) \) by using Lemma 3.25. In our case, with the notation of this lemma, \( B = p/(n\tau) + \eta/\tau \), for \( \eta > 0 \) given.

This implies that, for any given \( \epsilon > 0 \)
\[
\sup_{x \in (0, p/(n\tau) + \eta/\tau]} |\delta_n(x) - \Delta_n(x)| < \epsilon,
\]
with high-probability when \( n \) is large. Therefore, for any \( \epsilon > 0 \), if \( x_n \) is a solution of \( \delta_n(x_n) = 0 \),
\[
|\Delta_n(x_n)| \leq \epsilon \text{ with high-probability.}
\]

This exactly means that \( x_n \in T_{n,\epsilon} \) with high-probability. The same argument applies for near solutions of \( \delta_n(x) = 0 \), which, for any \( \epsilon > 0 \) must belong to \( T_{n,\epsilon} \) as \( n \to \infty \) with high-probability. Of course, there is nothing random about \( T_{n,\epsilon} \) which is a deterministic set. Note that \( T_{n,\epsilon} \) is compact because it is bounded and closed, using the fact that \( G_n = E(g_n) \) is continuous.

If \( T_{n,0} \) were reduced to a single point, we would have established the asymptotically deterministic character of \( c_{T} \).

Given our work concerning the limiting behavior of \( \hat{r}_{i,(i)} \) and our assumptions about \( \epsilon_i \)'s, we see that Lemma 3.39 applies to \( \lim_{n \to \infty} \Delta_n(x) \) under assumption \( \text{F1} \). Therefore, as \( n \to \infty \), \( T_{n,0} \) is reduced to a point and \( c_{T} \) is asymptotically non-random. (Note that assumption \( \text{F1} \) is stated in terms of the properties of densities of random variables of the form \( \epsilon_i + rZ_i \), where \( Z_i \) is \( N(0, 1) \), independent of \( \epsilon_i \) and \( r \) is arbitrary; Assumption \( \text{F1} \) also gives us guarantees for \( \epsilon_i + r\lambda_i Z_i \) at \( \lambda_i \) given by a simple change)
of variable. The $W_i$’s appearing in Lemma 3.39 are of the form $\epsilon_i + |\lambda_i| r Z_i$, so assumption $F_1$ is all we need for Lemma 3.39 to apply.) □

Proof of Theorem 2.1

We are now ready to prove Theorem 2.1.

Proof of Theorem 2.1  As we had noted in [16],

$$\frac{\partial}{\partial t} \text{prox}(c\rho)(t) = \text{prox}(c\rho)'(t) = \frac{1}{1 + c\psi'(\text{prox}(c\rho)(t))}.$$ 

So $\Delta_n$ can be interpreted as

$$\Delta_n(x) = \frac{p}{n} - \tau x - 1 + \frac{1}{n} \sum_{i=1}^{n} E \left( \text{prox}(x\lambda_i^2 \rho_i)'(\tilde{r}_{i,(i)}) \right).$$

The fact that $c_\tau$ is asymptotically arbitrarily close to the root of $\Delta_n(x) = 0$ gives us the first equation in the system appearing in Theorem 2.1. The second equation of the system comes from Eq. (48). Theorem 2.1 is shown, with $c_\rho(\kappa)$ being the limit of $c_\tau$. □

About $c_i$’s, $\xi_n$, $N_p$, and the limiting distribution of $\hat{\beta}(p)$

Theorem 2.1 as well as many of our intermediate results have interesting consequences for various quantities we encountered. Let us now state them.

When we use the expression “under our assumptions”, we mean assumptions $O_1$–$O_7$, $P_1$–$P_4$ and $F_1$–$F_5$.

3.1.2 On $c_i$’s

Recall that in Corollary 3.17 we had shown that under our assumptions $O_1$–$O_7$ and $P_1$–$P_4$

$$\sup_i |c_i - \lambda_i^2 c_\tau| = O_{L_k}(\text{polyLog}(n)n^{2\alpha-1/2}).$$

Since we have now shown that $c_\tau$ has a deterministic limit $c_\rho(\kappa)$, we have the following lemma.

Lemma 3.27  We have under our assumptions $O_1$–$O_7$ and $P_1$–$P_4$

$$\sup_i |R_i - \text{prox}(\lambda_i^2 c_\tau \rho_i)(\tilde{r}_{i,(i)})| = O_{L_k}(n^{2\alpha-1/2}\text{polyLog}(n)).$$

Springer
Hence, under (all of) our assumptions we have asymptotically, for any given $i$

$$| R_i - \text{prox}(\lambda_i c_\rho(\kappa)\rho_i)(\tilde{r}_i(i)) | = o_L(1).$$

If we furthermore assume that $\lambda_i$'s are uniformly bounded, we have

$$\sup_{1 \leq i \leq n} | R_i - \text{prox}(\lambda_i^2 c_\rho(\kappa)\rho_i)(\tilde{r}_i(i)) | = o_L(1).$$

**Proof** Lemma 3.32 implies that

$$\sup_{x \in \mathbb{R}} | \text{prox}(c_1 \rho)[x] - \text{prox}(c_2 \rho)[x] | \leq \| \psi \|_\infty | c_1 - c_2 |.$$

Therefore, Corollary 3.17 implies that

$$\sup_{1 \leq i \leq n} | \text{prox}(\lambda_i^2 c_\tau \rho_i)(\tilde{r}_i(i)) - \text{prox}(c_i \rho_i)(\tilde{r}_i(i)) | \leq \sup_{i} \| \psi_i \|_\infty \sup_{i} | c_i - \lambda_i^2 c_\tau | = O_L(\text{polyLog}(n)n^{2\alpha - 1/2}).$$

So in light of Theorem 3.9 we conclude that

$$\sup_{i} | R_i - \text{prox}(\lambda_i^2 c_\tau \rho_i)(\tilde{r}_i(i)) | = O_L(n^{2\alpha - 1/2}\text{polyLog}(n)).$$

Since $c_\tau$ is bounded by $p/(n\tau)$ and therefore so is $c_\rho(\kappa)$, we see that convergence in probability of $c_\tau$ to $c_\rho(\kappa)$ implies convergence in $L_k$ for any $k$. Using Holder's inequality, we therefore see that

$$\mathbb{E} \left( \sup_{x \in \mathbb{R}} | \text{prox}(\lambda_i^2 c_\tau \rho_i)[x] - \text{prox}(\lambda_i^2 c_\rho(\kappa)\rho_i)[x] | \right) \leq \sqrt{\mathbb{E} \left( (c_\tau - c_\rho(\kappa))^2 \right)} \mathbb{E} \left( \lambda_i^4 \right).$$

This gives us the second result of the lemma.

The last result is shown by simply remarking that

$$\sup_{i} \sup_{x \in \mathbb{R}} | \text{prox}(\lambda_i^2 c_\tau \rho_i)[x] - \text{prox}(\lambda_i^2 c_\rho(\kappa)\rho_i)[x] | \leq (\sup_{i} \lambda_i^2) | c_\tau - c_\rho(\kappa) |.$$

\[ \square \]

**On $\xi_n$**

We have the following lemma.

**Lemma 3.28** Under our assumptions, $\xi_n \to \xi$ in probability, where $\xi$ is deterministic. Furthermore, $\xi_n$ is bounded in $L_1$ and hence in probability.
We also have
\[ \xi_n = \frac{p - 1}{nc_{\tau,p}} - \tau + o_P(1) = \frac{p - 1}{nc_\tau} - \tau + o_P(1), \]
and \( c_{\tau,p} \) as well as \( c_\tau \) are bounded away from 0 in probability.

We note that using the last result and arguments in the proof below, we have, with the notations of Theorem 2.1
\[ \xi = \frac{p - 1}{nc_\rho(\kappa)} - \tau. \]

**Proof** The proof follows easily from the result of Proposition 3.18 which gives us that
\[ c_{\tau,p}(\xi_n + \tau) - \frac{p - 1}{n} = O_L(n^{-1/2+2\alpha}\text{polyLog}(n)) = o_P(1). \]
Since we have shown that \( c_{\tau,p} \) converges to a deterministic constant (recall that \( c_\tau - c_{\tau,p} \to 0 \)), we see that it is also the case for \( \xi_n \). Note also that \( \xi_n \leq \frac{1}{n} \sum_{i=1}^{n} X_i^2(p)\|\psi_i'\|_{\infty} \), so \( E(\xi_n) \) remains bounded under our assumptions.

To get the last result of the lemma, we just need to show that we can divide in the above display by \( c_{\tau,p} \) and still have something that converges to 0. We now show that \( c_{\tau,p} \) is bounded below. Note that
\[ c_{\tau,p} \frac{p - 1}{n(\xi_n + \tau)} = o_P(1). \]
Since \( \xi_n \) is bounded in probability, we see that \( \frac{p - 1}{n(\xi_n + \tau)} \) is bounded away from 0 in probability, which guarantees that \( c_{\tau,p} \) is bounded away from 0 in probability.

The results involving \( c_\tau \) immediately follow by appealing to Proposition 3.21. \( \Box \)

**On \( N_p \)**

Recall that by definition, we had
\[ N_p = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i(p)\psi_i(r_i,\{p\}). \]
We have the following result.

**Lemma 3.29** Under our assumptions, \( N_p \) is asymptotically \( N(0, v_n^2) \), with
\[ v_n^2 = \frac{1}{n} \sum_{i=1}^{n} E\left( \lambda_i^2 \psi_i^2(\text{prox}(c_\tau \lambda_i^2 \rho_i)(\tilde{r}_i,i)) \right). \]
Furthermore, there exists $v$ such that $v_n^2 \to v^2$, so that

$$N_p \implies \mathcal{N}(0, v^2).$$

The same result applies to $N_k$, for $1 \leq k \leq p$.

As the proof makes clear, we can replace in the asymptotic statements above $v_n^2$ by

$$\hat{v}_n^2 = \frac{1}{n} \sum_{i=1}^{n} \lambda_i^2 \psi_i^2 \left( \text{prox}(c, \lambda_i \rho_i)(\tilde{r}_{i, (i)}) \right).$$

**Proof** Note that we can write

$$N_p = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i(p) \lambda_i \psi_i(r_{i, [p]}).$$

Under our assumptions $X_i(p)$’s are independent and independent of $\{\lambda_i \psi_i(r_{i, [p]})\}_{i=1}^{n}$. The mild generalization of the Lindeberg-Feller argument given in the proof of Lemma 3.23 now applies, using in the notation of that lemma $a_n, p(k) = n^{-1/2} \lambda_i \psi_i(r_{i, [p]})$ and recalling that $|\psi_i(r_{i, [p]})| \leq \|\psi_i\|_\infty$. Since $\lambda_i$’s have 4 uniformly bounded moments under our assumptions, the fact that

$$N_p$$

behaves like $A_n \mathcal{N}(0, 1)$

follows immediately, where $A_n^2 = \sum_{k=1}^{n} d_{n, p}(k) = \frac{1}{n} \sum_{i=1}^{n} \lambda_i^2 \psi_i^2(r_{i, [p]})$. We note that $\mathbb{E} \left( A_n^2 \right) \leq \frac{1}{n} \sum_{i=1}^{n} \|\psi_i\|^2_\infty = O(1)$ under our assumptions.

Work similar to the one done in the proof of Proposition 3.22 shows that under our assumptions

$$A_n^2 - \frac{1}{n} \sum_{i=1}^{n} \lambda_i^2 \psi_i^2 \left[ \text{prox}(c, \lambda_i \rho_i)(\tilde{r}_{i, (i)}) \right] = o_{L_2}(1).$$

Asymptotic pairwise independence of $(\lambda_i, \tilde{r}_{i, (i)})$ and $(\lambda_j, \tilde{r}_{j, (j)})$—see Lemma 3.23—in connection with Assumption F5 guarantees that

$$\text{var} \left( \frac{1}{n} \sum_{i=1}^{n} \lambda_i^2 \psi_i^2 \left[ \text{prox}(c, \lambda_i \rho_i)(\tilde{r}_{i, (i)}) \right] \right) \to 0.$$

We conclude that $A_n^2$ is asymptotically deterministic and so is $A_n$. By Slutsky’s lemma we have

$$N_p$$

behaves like $\mathcal{N}(0, v_n^2)$. 

\(\square\) Springer
where

\[ v_n^2 = \frac{1}{n} \sum_{i=1}^{n} \mathbf{E} \left( \lambda_i^2 \psi_i^2 \left[ \text{prox}(c \rho(\lambda_i \rho_i)) (\tilde{r}_{i,i}) \right] \right), \]

since \( \mathbf{E} \left( A_n^2 \right) - v_n^2 \to 0. \)

We note that under our assumptions \( c \) has limit \( c \rho(\kappa) \) and \( \psi_i \) is a bounded continuous function (one of only finitely many possible functions). Also, \( \lambda_i \)'s have 4 moments. Therefore, since \( \tilde{r}_{i,i} \) behaves asymptotically like \( \epsilon_i + \lambda_i r(\kappa) Z_i \), where \( Z_i \sim \mathcal{N}(0, 1) \) independent of \( \epsilon_i \) and \( r(\kappa) \) is deterministic, we see that \( v_n^2 \) has a limit \( v^2 \). Of course, \( v_n^2 = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \mathbf{E} \left( \lambda_i^2 \psi_i^2 \left[ \text{prox}(c \rho(\lambda_i \rho_i)) (\tilde{r}_{i,i}) \right] \right) \).

In the notation of Theorem 2.1, we can rewrite \( v^2 \) as

\[ v^2 = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \mathbf{E} \left( \lambda_i^2 \psi_i^2 \left[ \text{prox}(c \rho(\lambda_i \rho_i)) (W_i) \right] \right) \]

• Minor technical point: It is true that \( \lambda_i^2 \psi_i^2 \left[ \text{prox}(c \rho(\lambda_i \rho_i)) (\tilde{r}_{i,i}) \right] \) is not a bounded continuous function of \( (\lambda_i, \tilde{r}_{i,i}) \). However, \( \lambda_i^2 \wedge M \psi_i^2 \left[ \text{prox}(c \rho(\lambda_i \rho_i)) (\tilde{r}_{i,i}) \right] \) is, for any \( M \). Since \( \lambda_i \) has 4 moments, it is easy to see that

\[ \mathbb{E} \left( (\lambda_i^2 \wedge M \psi_i^2 \left[ \text{prox}(c \rho(\lambda_i \rho_i)) (\tilde{r}_{i,i}) \right] \right) - \lambda_i^2 \psi_i^2 \left[ \text{prox}(c \rho(\lambda_i \rho_i)) (\tilde{r}_{i,i}) \right] \right) \leq \frac{\mathbb{E} (\lambda_i^4)}{M^2} \| \psi_i \|_\infty^2. \]

This standard approximation/uniform integrability argument shows that

\[ \mathbb{E} \left( \lambda_i^2 \psi_i^2 \left[ \text{prox}(c \rho(\lambda_i \rho_i)) (\tilde{r}_{i,i}) \right] \right) - \mathbb{E} \left( \lambda_i^2 \psi_i^2 \left[ \text{prox}(c \rho(\lambda_i \rho_i)) (W_i) \right] \right) \to 0, \]

since \( M \) can be chosen arbitrarily large.

• Going from \( N_p \) to \( N_k \)
Our arguments apply also if \( p \) is replaced by \( k \), with \( 1 \leq k \leq p \), since \( p \) does not play a particular role here. We note that as defined above \( A_n^2 \) depends on \( p \) but its approximand \( v_n^2 \) does not. So once again \( p \) does not play any role in the definition of the limiting variance and hence \( N_k \) has the same limit as \( N_p \), for all \( 1 \leq k \leq p \).

Asymptotic normality of \( \hat{\beta}_p \)

One of the aims of the previous results was to lead to a fluctuation result for \( \hat{\beta}_p \).
Proposition 3.30 We have, with the notation of the previous lemmas,
\[ \sqrt{n}[(\tau + \xi_n)\hat{\beta}_p - \beta_0(p)\xi_n] \implies N(0, v^2). \]

Furthermore, provided \( \beta_0(p) = O(n^{-1/2}) \), we also have
\[ \sqrt{n}[(\tau + \xi)\hat{\beta}_p - \beta_0(p)\xi] \implies N(0, v^2). \]

Similarly, we have for all \( 1 \leq k \leq p \): provided \( \beta_0(k) = O(n^{-1/2}) \),
\[ \sqrt{n}[(\tau + \xi)\hat{\beta}_k - \beta_0(k)\xi] \implies N(0, v^2). \]

The proof of the proposition we give below shows that \( \xi \) in the previous display can be replaced by any quantity \( \omega_n \) such that \( \xi_n - \omega_n = o_P(1) \). This in particular the case if we choose \( \omega_n = p/(nc_\tau) - \tau \), according to Lemma 3.28.

The main advantage of this \( \omega_n \) is that it is computable from the data. And we can therefore test the null hypothesis that \( \beta_0(p) = 0 \), since we can approximate \( v^2 \) by \( \frac{1}{n} \sum_{i=1}^{n} \lambda_i^2 \psi_i^2[\mathrm{prox}(c_i\lambda_i^2\rho_i)(\tilde{r}_{i,(i)})] \) according to the proof of Lemma 3.29.

Proof Recall that we have shown in Theorem 3.20 that
\[ \sqrt{n} (\hat{\beta}_p - b_p) = o_P(1). \]

Recall that we showed that \( \xi_n = O_{L_4}(1) \) under our assumptions. It is easy to verify that the same is true for \( \xi \), its limit. We also see that
\[ \sqrt{n}(\tau + \xi_n)(\hat{\beta}_p - b_p) = o_P(1). \]

Recall that by definition,
\[ \sqrt{n}[(\tau + \xi_n)b_p - \xi_n\beta_0(p)] = N_p. \]

So we conclude, using Slutsky’s lemma that
\[ \sqrt{n}[(\tau + \xi_n)\hat{\beta}_p - \xi_n\beta_0(p)] \implies N(0, v^2). \]

When \( \beta_0(p) = O(n^{-1/2}) \), we see that
\[ \sqrt{n}(\xi - \xi_n)(\beta_0(p)) = o_P(1). \]

Furthermore, in this setting \( \sqrt{n}b_p = O_P(1) \) and hence \( \sqrt{n}\hat{\beta}_p = O_P(1) \). We conclude that then
\[ \sqrt{n}\hat{\beta}_p(\xi_n - \xi) = o_P(1). \]
Therefore,
\[
\sqrt{n}[(\tau + \xi_n)\hat{\beta}_p - \xi_n\beta_0(p)] = \sqrt{n}[(\tau + \xi)\hat{\beta}_p - \xi\beta_0(p)] + o_P(1)
\]
and we get the second result of the proposition through Slutsky’s lemma.

**Extending the result from** $\hat{\beta}_p$ **to** $\hat{\beta}_k$

We note that $\xi_n$ in the above argument actually depends on $p$ also—we avoided writing out the dependence earlier to avoid cumbersome notations. We now make it explicit for clarity. Our arguments above guarantee that

\[
\sqrt{n}[(\tau + \xi_n(k))\hat{\beta}_k - \xi_n(k)\beta_0(p)] \Longrightarrow N(0, v^2),
\]

where as explained in Lemma 3.29 $v$ does not depend on $k$. As explained above, if $\beta_0(k) = O(n^{-1/2})$, $\sqrt{n}(\xi - \xi_n(k))(\beta_0(k)) = o_P(1)$ for all $k$’s. Very importantly, $\xi$ does not depend on $p$, since it is by definition $\xi = (p - 1)/(nc_\rho(\kappa)) - \tau$. So we conclude that for all $k$, when $\beta_0(k) = O(n^{-1/2})$,

\[
\sqrt{n}[(\tau + \xi_n(k))\hat{\beta}_k - \xi_n(k)\beta_0(p)] = \sqrt{n}[(\tau + \xi)\hat{\beta}_k - \xi\beta_0(k)] + o_P(1).
\]

Since the left hand side weakly converges to $N(0, v^2)$, we conclude by Slutsky’s lemma that for all $k$,

\[
\sqrt{n}[(\tau + \xi)\hat{\beta}_k - \xi\beta_0(k)] \Longrightarrow N(0, v^2),
\]

provided $\beta_0(k) = O(n^{-1/2})$. 

\[
\square
\]

**Appendix 6: Notes on the proximal mapping**

In this section of the Appendix we remind the reader of elementary properties of the proximal mapping. The proofs, when needed, can be found in e.g. [15].

**Lemma 3.31** Almost by definition, we have

\[
prox(c\rho)(x) + c\psi(prox(c\rho)(x)) = x.
\]

Let $\rho$ be differentiable and such that $\psi$ changes sign at 0, i.e. $\text{sign}(\psi(x)) = \text{sign}(x)$ for $x \neq 0$. Then,

\[
prox(c\rho)(0) = 0.
\]

Furthermore,

\[
|\psi(prox(c\rho)(x))| \leq |\psi(x)|.
\]
Also,

\[ |\psi(\text{prox}(c\rho)(x))| \leq |x|/c. \]

We will also need the following simple result.

**Lemma 3.32** Suppose \( x \) is a real and \( \rho \) is twice differentiable and convex. Then, for \( c > 0 \), we have

\[
\frac{\partial}{\partial c} \text{prox}(c\rho)(x) = -\frac{\psi(\text{prox}(c\rho)(x))}{1 + c\psi'(\text{prox}(c\rho)(x))},
\]

and

\[
\frac{\partial}{\partial c} \rho(\text{prox}(c\rho)(x)) = -\frac{\psi^2(\text{prox}(c\rho)(x))}{1 + c\psi'(\text{prox}(c\rho)(x))}.
\]

In particular, at \( x \) given \( c \rightarrow \rho(\text{prox}(c\rho)(x)) \) is decreasing in \( c \).

We also make the following observation, which is useful to obtain a compact representation for the system of Eqs. (4).

**Lemma 3.33** We have

\[
\frac{\partial}{\partial x} \text{prox}(c\rho)(x) = \frac{1}{1 + c\psi'(\text{prox}(c\rho)(x))}.
\]

Moreover, at \( c \) fixed, when \( \psi' \) is continuous, \( x \rightarrow \frac{1}{1 + c\psi'(\text{prox}(c\rho)(x))} \) is a bounded, continuous function of \( x \).

A proof of the first fact follows immediately from the well-known representation (see [29])

\[ \text{prox}(c\rho)(x) = (\text{Id} + c\psi)^{-1}(x). \]

The second result is also immediate, since \( \psi' \geq 0 \).

We finally make notice of the following simple fact.

**Lemma 3.34** The function \( c \rightarrow [c\psi(\text{prox}(c\rho)(x))]^2 \) (defined on \( \mathbb{R}_+ \)) is increasing, for any \( x \).

**Examples** for the sake of concreteness, we now give a couple examples of proximal mappings.

1. if \( \rho(x) = x^2/2 \), \( \text{prox}(c\rho)(x) = \frac{x}{1+cx} \).
2. if \( \rho(x) = |x| \), \( \text{prox}(c\rho)(x) = \text{sgn}(x)(|x| - c)_+ \), i.e. the “soft-thresholding” function.
Appendix 7: On convex Lipschitz functions of random variables

In this section, we provide a brief reminder concerning convex Lipschitz functions of random variables. The proofs can be found in [15].

Lemma 3.35 Suppose that \( \{X_i\}_{i=1}^n \in \mathbb{R}^p \) satisfy the following concentration property: \( \exists C_n, c_n \) such that for any \( G_i \), a convex, 1-Lipschitz (with respect to Euclidean norm) function of \( X_i \),

\[
P(|G_i(X_i) - m_i| \geq t) \leq C_n \exp(-c_n t^2),
\]

where \( m_i \) is deterministic.

Let us now fix \( \{F_i\}_{i=1}^n \), \( n \) functions which are convex and 1-Lipschitz in \( X_i \). Then if \( F_n = \sup_i |F_i(X_i) - m_i| \), we have, even when the \( X_i \)'s are dependent:

1. if \( u_n = \sqrt{\log(n)/c_n} \), \( \mathbb{E}(F_n) \leq u_n + C_n/(2\sqrt{c_n}\sqrt{\log(n)} = \frac{\sqrt{\log n}}{\sqrt{c_n}}(1 + C_n/(2\log n)) \) Similar bounds hold in \( L_k \) for any finite given \( k \).
2. when \( C_n \leq C \), where \( C \) is independent of \( n \), there exists \( K \), independent of \( n \) such that \( F_n/u_n \leq K \) with overwhelming probability, i.e. probability asymptotically smaller than any power of \( 1/n \).
3. \( m_i \) can be chosen to be the mean or the median of \( F_i(X_i) \).

In particular,

\[
F_n = O(\text{polyLog}(n)/\sqrt{c_n})
\]

in probability and any \( L_k \), \( k \) fixed and given.

We note that similar techniques can be used to extend the result to situations where we have \( P(|G_i(X_i) - m_i| \geq t) \leq C_n \exp(-c_n t^\beta) \), with \( \beta \neq 2 \). Of course, the order of magnitudes of the bounds then change: in particular, wherever \( \sqrt{c_n} \) appears, it would have to be replaced by \( c_n^{1/\beta} \). But since under our assumptions \( 1/c \) is at most polyLog(n), this would effectively have no impact on our results.

We now turn our attention to a slightly more complicated setting.

We recall that we denote by \( X(i) = \{X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n\} \). If \( I \) is a subset of \( \{1, \ldots, n\} \) of size \( n-1 \), we call \( X_I \) the collection of the corresponding \( X_i \) random variables. We call \( X_{I^c} \) the remaining random variable.

Lemma 3.36 Suppose \( X_i \)'s are independent and satisfy the concentration inequalities as above. Consider the situation where \( F_{I_k}(\cdot) \) is a convex Lipschitz function of 1 variable; \( F_{I_k}(\xi) \) depends on \( X \) through \( X_{I_k} \) only and we call \( \mathcal{L}_{I_k} \) the Lipschitz constant of \( F_{I_k}(\cdot) \) (at \( X_{I_k} \) given). \( \mathcal{L}_{I_k} \) is assumed to be random, since \( X_{I_k} \) is. Call \( m_{F_{I_k}} = m_{F_{I_k}(X_{I_k})}|X_{I_k} \), \( m \) being the mean or the median. As before, call \( F_n = \sup_{j=1,\ldots,n} |F_j(X_j) - m_{F_j}| \). Then \( F_n = O(\sqrt{\log n/c_n} \sup_{1 \leq j \leq n} \mathcal{L}_{I_j}) \) in probability and in \( \sqrt{L_{2k}} \), i.e. there exists \( K > 0 \), independent of \( n \), such that
\[ \mathbb{E} \left( F_n^k \right) \leq K (\sqrt{\log n / c_n})^k \sqrt{\mathbb{E} \left( \sup_{1 \leq j \leq n} \mathcal{L}_{I_j}^{2k} \right)}. \]

Hence, \( F_n \) is \( \text{polyLog}(n) / c_n^{1/2} \sup_{1 \leq j \leq n} \mathcal{L}_{I_j} \) in \( \sqrt{\mathcal{L}_{2k}} \).

We repeatedly use the following lemma in the proof.

**Lemma 3.37** Suppose the assumptions of the previous Lemma are satisfied. Consider \( Q_{I_j} = \frac{1}{n} X_i' M_{I_j} X_i \), where \( M_{I_j} \) is a random positive-semidefinite matrix depending only on \( X_{I_j} \) whose largest eigenvalue is \( \lambda_{\max, I_j} \). Assume that \( \mathbb{E} (X_i) = 0, \text{cov} (X_i) = \text{Id}_p \) and \( n c_n \to \infty \). Then, we have in \( L_k \),

\[
\sup_{1 \leq j \leq n} \left| Q_{I_j} - \frac{1}{n} \text{trace} (M_{I_j}) \right| = O_{L_k} \left( \frac{\text{polyLog}(n)}{\sqrt{n c_n}} \sup_{1 \leq j \leq n} \lambda_{\max, I_j} \right).
\]

The same bound holds when considering a single \( Q_{I_j} \) without the \( \text{polyLog}(n) \) term.

**On the spectral norm of covariance matrices**

**Lemma 3.38** Suppose \( X_i \)'s are independent random vectors in \( \mathbb{R}^p \), satisfying \( O4 \), and having mean 0 and covariance \( \text{Id}_p \). Suppose that \( \lambda_i \)'s satisfy \( O6 \). Let \( \hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} X_i X_i' \). Then,

\[
||| \hat{\Sigma} |||_2 = O_P \left( \text{polyLog}(n) c_n^{-1} \right).
\]

The results hold also in \( L_k \).

**Proof** The proof is exactly similar to that given in [15], which gives, following a simple adaption of the well-known \( \epsilon \)-net argument explained e.g. in [40], “Appendix 1” that

\[
||| \frac{1}{n} \sum_{i=1}^{n} X_i X_i' |||_2 = O_L (c_n^{-1}).
\]

It is clear that

\[
||| \hat{\Sigma} |||_2 \leq \left( \sup_{1 \leq i \leq n} \lambda_i^2 \right) ||| \frac{1}{n} \sum_{i=1}^{n} X_i X_i' |||_2,
\]

and the result follows immediately. \( \square \)
Appendix 8: Miscellaneous results

An analytic result

We now study the roots of \( F(x) = 0 \), where

\[
F(x) = \frac{p}{n} - \tau x - 1 + \frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left( (\text{prox}(x\lambda_i^2 \rho_i))'(W_i) \right)
\]

where \( W_i \)'s are random variables and \( (\text{prox}(x\rho))'(t) = \frac{\partial}{\partial t} \text{prox}(x\rho)(t) \).

We now show that under mild conditions on \( W_i \)'s this equation has a unique solution. We allow \( W_i \) to depend on the random variables \( \lambda_i \)'s.

Lemma 3.39 Suppose that \( W_i \)'s have smooth densities \( f_i(t, \lambda_i) \) with sign \( f_i'(x, \lambda_i) = -\text{sign}(x) \). Suppose further that \( \lim_{|t| \to \infty} |t| f_i(t, \lambda_i) = 0 \) and that \( \text{sign}(\psi_i(x)) = \text{sign}(x) \). Then, if

\[
F_i(x) = \frac{p}{n} - \tau x - 1 + \mathbb{E} \left( (\text{prox}(x\lambda_i^2 \rho_i))'(W_i) \right),
\]

the function \( F_i \) is decreasing, with \( F_i'(x) \leq -\tau \). Hence, the same applies to \( F \).

In particular, the equation \( F(x) = 0 \) has a unique solution.

Proof We call

\[
G_i(x) \triangleq \mathbb{E} \left( (\text{prox}(x\lambda_i^2 \rho_i))'(W_i) \right),
\]

and

\[
G_i(x, \lambda_i) \triangleq \mathbb{E} \left( (\text{prox}(x\lambda_i^2 \rho_i))'(W_i)|\lambda_i \right).
\]

Of course,

\[
\mathbb{E} \left( (\text{prox}(x\lambda_i^2 \rho_i))'(W_i)|\lambda_i \right) = \int (\text{prox}(x\lambda_i^2 \rho_i))'(t) f_i(t, \lambda_i) dt.
\]

Using contractivity of the proximal mapping (see [29]) we see that \( \lim_{|t| \to \infty} \text{prox}(x\lambda_i^2 \rho_i)(t) f_i(t, \lambda_i) = 0 \) under our assumptions.

Integrating the previous equation by parts, we see that

\[
G_i(x, \lambda_i) = -\int (\text{prox}(x\lambda_i^2 \rho_i))(t) f_i'(t, \lambda_i) dt.
\]
To compute $G'_i(x, \lambda_i)$, we differentiate under the integral sign (under our assumptions the conditions of Theorem 9.1 in [10] are satisfied) to get

$$G'_i(x, \lambda_i) = \int \frac{\psi_i(x \lambda_i^2 \rho_i(t)) f'_i(t, \lambda_i)}{1 + x \lambda_i^2 \psi'_i(x \lambda_i^2 \rho_i(t))} dt.$$ 

Under our assumptions, $\text{sign}(\psi_i(x \lambda_i^2 \rho_i(t))) = \text{sign}(t)$ and $\text{sign}(f'_i(t, \lambda_i)) = -\text{sign}(t)$, so that

$$\forall t \neq 0, \text{sign}(\psi_i(x \lambda_i^2 \rho_i(t))) f'_i(t, \lambda_i) = -1.$$ 

Since the denominator of the function we integrate is positive, we conclude that

$$G'_i(x, \lambda_i) \leq 0 \quad \text{and} \quad G'_i(x) \leq 0.$$ 

Since $F'_i(x) = -\tau + G'_i(x)$, we see that $F'_i(x) \leq -\tau < 0$. Therefore $F_i$ is a decreasing function on $\mathbb{R}_+$. Of course, $\text{prox}(0 \rho)(t) = t$, so that $F_i(0) = p/n$ and $\lim_{x \to \infty} F_i(x) = -\infty$, since, for instance,

$$0 \leq \text{prox}(x \rho)'(t) = \frac{1}{1 + x \psi'[\text{prox}(x \rho)(t)]} \leq 1.$$ 

So we conclude that the equation $F_i(x) = 0$ has a unique root. (Since $F_i$ is differentiable, $F_i$ is of course continuous.)

We note that

$$F(x) = \frac{1}{n} \sum_{i=1}^{n} F_i(x).$$ 

Therefore, $F(0) = p/n$ and $F'(x) \leq -\tau$. So $F$ is decreasing, differentiable and hence has a unique root. $\square$

Remark the conditions on the density of $W$ are satisfied in many situations. For instance if $W_i = \epsilon + r \lambda_i Z$, where $\epsilon$ is symmetric about 0 and log-concave, $Z$ is $\mathcal{N}(0, 1)$, independent of $\lambda_i$ and $\epsilon$, and $r > 0$, it is clear that the density of $W$ satisfies the conditions of our lemma. Similar results hold under weaker assumptions on $\epsilon$ of course. For more details, we refer the reader to e.g. [7] and [25,26,35].

In particular, we recall Theorem 1.6 in [7] which says that the convolution of two symmetric unimodal distributions on $\mathbb{R}$ is unimodal. Hence, when $\epsilon$ has a symmetric and unimodal distribution, so does $W_i = \epsilon + \lambda_i r Z$, for any $r$. This is for instance the case when $\epsilon$ has a Cauchy distribution.

A linear algebraic remark

We need the following lemma at some point in the proof.
Lemma 3.40 Suppose the $p \times p$ matrix $A$ is positive semi-definite and

$$A = \begin{pmatrix} \Gamma & v \\ v' & a \end{pmatrix}.$$ 

Here $a \in \mathbb{R}$. Let $\tau$ be a strictly positive real. Call $\Gamma_\tau = \Gamma + \tau \text{Id}_{p-1}$. Then we have

$$\text{trace}\left( (A + \tau \text{Id}_p)^{-1} \right) = \text{trace}\left( \Gamma_\tau^{-1} \right) + \frac{1 + v'\Gamma_\tau^{-2}v}{a + \tau - v'\Gamma_\tau^{-1}v}.$$ 

In particular,

$$\left| \text{trace}\left( (A + \tau \text{Id}_p)^{-1} \right) - \text{trace}\left( \Gamma_\tau^{-1} \right) \right| \leq \frac{1 + a/\tau}{\tau}.$$ 

The proof is simple and we refer the reader to [15] for details if needed.

References

1. Anderson, T.W.: An Introduction to Multivariate Statistical Analysis. Wiley, New York (1984)
2. Bayati, M., Montanari, A.: The LASSO risk for Gaussian matrices. IEEE Trans. Inform. Theory 58, 1997–2017 (2012). doi:10.1109/TIT.2011.2174612
3. Bean, D., Bickel, P., El Karoui, N., Lim, C., Yu, B.: Penalized Robust Regression in High Dimension. Technical Report, Department of Statistics, UC Berkeley. Submitted to AISTATs in October 2011. Not under consideration anymore (2012)
4. Bean, D., Bickel, P.J., El Karoui, N., Yu, B.: Optimal m-estimation in high-dimensional regression. Proc. Natl. Acad. Sci. 110, 14563–14568 (2013)
5. Beck, A., Teboulle, M.: Convex Optimization in Signal Processing and Communications, chapter Gradient-Based Algorithms with Applications in Signal Recovery Problems, pp. 33–88. Cambridge University Press, Cambridge (2010)
6. Bhatia, R.: Matrix Analysis, Volume 169 of Graduate Texts in Mathematics. Springer, New York (1997)
7. Dharmadhikari, S., Joag-Dev, K.: Unimodality, Convexity, and Applications. Probability and Mathematical Statistics. Academic Press Inc, Boston, MA (1988)
8. Diaconis, P., Freedman, D.: Asymptotics of graphical projection pursuit. Ann. Stat. 12, 793–815 (1984). doi:10.1214/aos/1176346703
9. Donoho, D., Montanari, A.: High Dimensional Robust m-estimation: Asymptotic Variance via Approximate Message Passing. Probab. Theory Relat. Fields. 166(3–4), 935–969 (2016)
10. Durrett, R.: Probability: Theory and Examples, 2nd edn. Duxbury Press, Belmont, CA (1996)
11. Efron, B., Stein, C.: The jackknife estimate of variance. Ann. Stat. 9, 586–596 (1981)
12. Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D.: Cluster analysis and display of genome-wide expression patterns. PNAS 95, 14863–14868 (1998)
13. El Karoui, N.: Concentration of measure and spectra of random matrices: applications to correlation matrices, elliptical distributions and beyond. Annal. Appl. Probab. 19, 2362–2405 (2009)
14. El Karoui, N.: High-dimensionality effects in the Markowitz problem and other quadratic programs with linear constraints: risk underestimation. Ann. Stat. 38, 3487–3566 (2010). doi:10.1214/10-AOS795
15. El Karoui, N.: Asymptotic Behavior of Unregularized and Ridge-Regularized High-Dimensional Robust Regression Estimators: Rigorous Results. arXiv:1311.2445 (2013)
16. El Karoui, N., Bean, D., Bickel, P., Lim, C., Yu, B.: On Robust Regression with High-Dimensional Predictors. Technical Report 811, UC, Berkeley, Department of Statistics. Originally submitted as manuscript AoS1111-009. Not under consideration anymore (2011)
17. El Karoui, N., Bean, D., Bickel, P.J., Lim, C., Yu, B.: On robust regression with high-dimensional predictors. Proc. Natl. Acad. Sci. (2013)
18. El Karoui, N., Koesters, H.: Geometric Sensitivity of Random Matrix Results: Consequences for Shrinkage Estimators of Covariance and Related Statistical Methods. (Submitted to Bernoulli) Available at arXiv:1105.1404 (2011)

19. El Karoui, N., Purdom, E.: Can We Trust the Bootstrap in High-Dimension? Technical Report 824, UC Berkeley, Department of Statistics. Submitted to AoS (2015)

20. Hall, P., Marron, J.S., Neeman, A.: Geometric representation of high dimension, low sample size data. J. R. Stat. Soc. Ser. B Stat. Methodol. 67, 427–444 (2005)

21. Horn, R.A., Johnson, C.R.: Matrix Analysis. Cambridge University Press, Cambridge. (Corrected reprint of the 1985 original) (1990)

22. Huber, P.J.: The 1972 Wald lecture. Robust statistics: a review. Ann. Math. Stat. 43, 1041–1067 (1972)

23. Huber, P.J.: Robust regression: asymptotics, conjectures and Monte Carlo. Ann. Stat. 1, 799–821 (1973)

24. Huber, P.J., Ronchetti, E.M.: Robust Statistics. Wiley Series in Probability and Statistics, 2nd edn. John Wiley & Sons Inc., Hoboken, NJ. doi: 10.1002/9780470434697 (2009)

25. Ibragimov, I.A.: On the composition of unimodal distributions. Teor. Veroyatnost. I Primenen. 1, 283–288 (1956)

26. Karlin, S.: Total Positivity, vol. I. Stanford University Press, Stanford, CA (1968)

27. Ledoux, M.: The Concentration of Measure Phenomenon, Vol 89 of Mathematical Surveys and Monographs. American Mathematical Society, Providence, RI (2001)

28. Mammen, E.: Asymptotics with increasing dimension for robust regression with applications to the bootstrap. Ann. Stat. 17, 382–400 (1989). doi: 10.1214/aos/1176347023

29. Moreau, J.-J.: Proximité et dualité dans un espace hilbertien. Bull. Soc. Math. France 93, 273–299 (1965)

30. Pollard, D.: Convergence of Stochastic Processes. Springer Series in Statistics. Springer, New York (1984)

31. Portnoy, S.: Asymptotic behavior of $M$-estimators of $p$ regression parameters when $p^2/n$ is large I. Consistency. Ann. Stat. 12, 1298–1309 (1984). doi: 10.1214/aos/1176346793

32. Portnoy, S.: Asymptotic behavior of $M$ estimators of $p$ regression parameters when $p^2/n$ is large II. Normal approximation. Ann. Stat. 13, 1403–1417 (1985). doi: 10.1214/aos/1176349744

33. Portnoy, S.: Asymptotic behavior of the empiric distribution of $M$-estimated residuals from a regression model with many parameters. Ann. Stat. 14, 1152–1170 (1986). doi: 10.1214/aos/1176350056

34. Portnoy, S.: A central limit theorem applicable to robust regression estimators. J. Multivar. Anal. 22, 24–50 (1987). doi: 10.1016/0047-259X(87)90073-X

35. Prékopa, A.: On logarithmic concave measures and functions. Acta Sci. Math. (Szeged) 34, 335–343 (1973)

36. Relles, D.: Robust Regression by Modified Least Squares. Ph.D. Thesis, Yale University (1968)

37. Ruszczyński, A.: Nonlinear Optimization. Princeton University Press, Princeton, NJ (2006)

38. Shcherbina, M., Tirozzi, B.: Rigorous solution of the Gardner problem. Comm. Math. Phys. 234, 383–422 (2003). doi: 10.1007/s00220-002-0783-3

39. Stroock, D.W.: Probability Theory, An Analytic View. Cambridge University Press, Cambridge (1993)

40. Talagrand, M.: Spin Glasses: a Challenge for Mathematicians, volume 46 of Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics. Cavity and mean field models [Results in Mathematics and Related Areas. 3rd Series. A Series of Modern Surveys in Mathematics]. Springer, Berlin (2003)

41. van der Vaart, A.W.: Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge (1998)

42. Yin, Y.Q., Bai, Z.D., Krishnaiah, P.R.: On the limit of the largest eigenvalue of the large-dimensional sample covariance matrix. Probab. Theory Related Fields 78, 509–521 (1988)