

Can We Trust the Bootstrap in High-dimensions? The Case of Linear Models

Noureddine El Karoui

NKAROUI@BERKELEY.EDU, N.ELKAROUI@CRITEO.COM

*Criteo AI Lab
32 Rue Blanche
75009 Paris, France
and*

*Department of Statistics
University of California
Berkeley, CA 94270, USA*

Elizabeth Purdom

EPURDOM@STAT.BERKELEY.EDU

*Department of Statistics
University of California
Berkeley, CA 94270, USA*

Editor: Guy Lebanon

Abstract

We consider the performance of the bootstrap in high-dimensions for the setting of linear regression, where $p < n$ but p/n is not close to zero. We consider ordinary least-squares as well as robust regression methods and adopt a minimalist performance requirement: can the bootstrap give us good confidence intervals for a single coordinate of β (where β is the true regression vector)?

We show through a mix of numerical and theoretical work that the bootstrap is fraught with problems. Both of the most commonly used methods of bootstrapping for regression—residual bootstrap and pairs bootstrap—give very poor inference on β as the ratio p/n grows. We find that the residual bootstrap tend to give anti-conservative estimates (inflated Type I error), while the pairs bootstrap gives very conservative estimates (severe loss of power) as the ratio p/n grows. We also show that the jackknife resampling technique for estimating the variance of $\hat{\beta}$ severely overestimates the variance in high dimensions.

We contribute alternative procedures based on our theoretical results that result in dimensionality adaptive and robust bootstrap methods.

Keywords: Bootstrap, high-dimensional inference, random matrices, resampling

1. Introduction

The bootstrap (Efron, 1979) is a ubiquitous tool in applied statistics, allowing for inference when very little is known about the properties of the data-generating distribution. The bootstrap is a powerful tool in applied settings because it does not make the strong assumptions common to classical statistical theory regarding this data-generating distribution. Instead, the bootstrap resamples the observed data to create an estimate, \hat{F} , of the unknown data-generating distribution, F . The distribution \hat{F} then forms the basis of further inference.

Since its introduction, a large amount of research has explored the theoretical properties of the bootstrap, improvements for estimating F under different scenarios, and how to most effectively estimate different quantities from \hat{F} (see the pioneering Bickel and Freedman, 1981 for instance and many many more references in the book-length review of Davison and Hinkley, 1997, as well as van der Vaart, 1998 for a short summary of the modern point of view on these questions). Other resampling techniques exist of course, such as subsampling, m-out-of-n bootstrap, and jackknifing, all of which have been studied and much discussed (see Efron, 1982; Hall, 1992; Politis et al., 1999; Bickel et al., 1997; and Efron and Tibshirani, 1993 for a practical introduction).

An important limitation for the bootstrap is the quality of \hat{F} . The standard bootstrap estimate of F based on the empirical distribution of the data may be a poor estimate when the data has a non-trivial dependency structure, when the quantity being estimated, such as a quantile, is sensitive to the discreteness of \hat{F} , or when the functionals of interest are not smooth (see e.g., Bickel and Freedman, 1981 for a classic reference, as well as Beran and Srivastava, 1985 or Eaton and Tyler, 1991 in the context of multivariate statistics).

An area that has received less attention is the performance of the bootstrap in high dimensions and this is the focus of our work. In particular, we consider the setting of standard linear models where data y_i are drawn from the linear model

$$\forall i, y_i = \beta' X_i + \epsilon_i, 1 \leq i \leq n, \text{ where } X_i \in \mathbb{R}^p.$$

We are interested in the bootstrap or resampling properties of the estimator defined as

$$\hat{\beta}_\rho = \operatorname{argmin}_{b \in \mathbb{R}^p} \sum_{i=1}^n \rho(y_i - X_i' b), \text{ where } \rho \text{ is a convex function.}$$

We consider the two standard methods for resampling to create a bootstrap distribution in this setting. The first is *pairs resampling*, where bootstrap samples are drawn from the empirical distribution of the pairs (y_i, X_i) . The second resampling method is *residual resampling*, where the bootstrapped data consists of $y_i^* = \hat{\beta}_\rho' X_i + \hat{\epsilon}_i^*$, where $\hat{\epsilon}_i^*$ is drawn from the empirical distribution of the estimated residuals, e_i . We also consider the jackknife, a resampling method focused specifically on estimating the variance of functionals of $\hat{\beta}_\rho$. These three methods are extremely flexible for linear models regardless of the method of fitting β or the error distribution of the ϵ_i .

The high dimensional setting: $p/n \rightarrow \kappa \in (0, 1)$ In this work we call a high-dimensional setting one where the number of predictors, p , is of the same order of magnitude as the number of observations, n , formalized mathematically by assuming that $p/n \rightarrow \kappa \in (0, 1)$. Several reasons motivate our theoretical study in this regime. The asymptotic behavior of the estimate $\hat{\beta}_\rho$ is known to depend heavily on whether one makes the classical theoretical assumption that $p/n \rightarrow 0$ or instead assumes $p/n \rightarrow \kappa \in (0, 1)$ (see Section 1.2 and Appendix A and references therein). But from the standpoint of practical usage on moderate-sized data sets (i.e., n and p both moderately sized with $p < n$), it is not always obvious which assumption is justified. Working in the high-dimensional regime of $p/n \rightarrow \kappa \in (0, 1)$ captures better the complexity encountered even in reasonably low-dimensional practice than using the classical assumption $p/n \rightarrow 0$. In fact, asymptotic predictions based on the

high-dimensional assumption can work surprisingly well in very low-dimension (see Johnstone, 2001). Furthermore, in these high-dimensional settings, where much is still unknown theoretically, the bootstrap is a natural and compelling alternative to asymptotic analysis.

Another motivation for our investigation is that of very large scale applications (Chapelle et al., 2014; Criteo, 2017; Langford et al., 2007), where one might resort to subsampling methods or recent variants like the bag-of-little-bootstraps (Kleiner et al., 2014) for uncertainty assessment. Subsampling is also very commonly used in this setting for simple computational speed-up. In such cases, even if one had started with a data set where $p \ll n$, after subsampling one often ends up with p comparable to n on the subsamples where bootstrap-like computations are performed. It is therefore important to know if the bootstrap and other resampling plans perform well when p is comparable to n .

Defining success: accurate inference on β_1 The common theoretical definition of whether the bootstrap “works” is that the bootstrap distribution of the entire bootstrap estimate $\hat{\beta}_\rho^*$ converges conditionally almost surely to the sampling distribution of the estimator $\hat{\beta}_\rho$ (see e.g., van der Vaart, 1998). The work of Bickel and Freedman (1983) on the residual bootstrap for least squares regression, which we discuss in the background section 1.2, shows that this theoretical requirement is not fulfilled even for the simple problem of least squares regression.

In this paper, we choose to focus only on accurate inference for the projection of our parameter on a pre-specified direction v . More specifically, we concentrate only on whether the bootstrap gives accurate confidence intervals for $v'\beta$. We think that this is the absolute minimal requirement we can ask of a bootstrap inferential method, as well as one that is meaningful from the standpoint of applied statistics. This is of course a much less stringent requirement than performing well on complicated functionals of the whole parameter vector, which is the implicit demand of standard definitions of bootstrap success. For this reason, we focus throughout the exposition on inference for β_1 (the first element of β) as an example of a pre-defined direction of interest (where β_1 corresponds to choosing $v = e_1$, the first canonical basis vector).

We note that considering the asymptotic behavior of $v'\beta$ as $p/n \rightarrow \kappa \in (0, 1)$ implies that $v = v(p)$ changes with p . By “pre-defined” we will mean simply a deterministic sequence of directions $v(p)$. We will continue to suppress the dependence on p in writing v in what follows for the sake of clarity.

1.1 Organization and Main Results of the Paper

In Section 2 we demonstrate that in high dimensions residual-bootstrap resampling results in extremely poor inference on the coordinates of β_ρ with error rates much higher than the reported Type I error. We show that the error in inference based on residual bootstrap resampling is due to the fact that the distribution of the residuals e_i are a poor estimate of the distribution of ϵ_i ; we further illustrate that common methods of standardizing the e_i do not solve the problem for general ρ . We propose two new dimension-adaptive methods of residual resampling that appear promising for use in bootstrapping linear models. We also provide some theoretical results for the behavior of this method as $p/n \rightarrow 1$.

In Section 3 we examine *pairs-bootstrap resampling* and show that confidence intervals based on bootstrapping the pairs also perform very poorly. Unlike in the residual-bootstrap

case discussed in Section 2, the confidence intervals obtained from the pairs-bootstrap are instead conservative to the point of being non-informative. This results in a dramatic loss of power. We prove in the case of L_2 loss, i.e., $\rho(x) = x^2$, that the variance of the bootstrapped $v'\widehat{\beta}^*$ is greater than that of $v'\widehat{\beta}$, leading to the overly conservative performance we see in simulations. We demonstrate that a different resampling scheme we propose can provide accurate confidence intervals in moderately high dimensions.

In Section 4, we discuss another resampling scheme, the jackknife. We focus on the jackknife estimate of variance and show that it has similarly poor behavior in high dimensions. In the case of L_2 loss with Gaussian design matrices, we further prove that the jackknife estimator over estimates the variance of our estimator by a factor of $1/(1 - p/n)$; we also provide corrections for other losses that improve the jackknife estimate of variance in moderately high dimensions.

We rely on simulation results to demonstrate the practical impact of the failure of the bootstrap. The settings for our simulations and corresponding theoretical analyses are idealized, without many of the common settings of heteroskedasticity, dependency, outliers and so forth that are known to be a problem for bootstrapping. This is intentional, since even these idealized settings are sufficient to demonstrate that the standard bootstrap methods have poor performance. For brevity, we give only brief descriptions of the simulations in what follows; detailed descriptions can be found in AppendixD.1.

Similarly, we focus on the basic implementations of the bootstrap for linear models. While there are many proposed alternatives (often for specific loss functions or types of data), the standard methods we study are most commonly used and recommended in practice. Furthermore, to our knowledge none of the alternative bootstrap methods we have seen specifically address the underlying theoretical problems that appear in high dimensions without making low-dimensional assumptions about either the design matrix or the sparsity of β , and therefore are likely to suffer from the same fate as standard methods. We note that in truly large scale applications, sparsity assumptions are not always made by practitioners (Chapelle et al., 2014; Langford et al., 2007; Criteo, 2017) and it is hence natural to study the performance of estimators outside of sparse settings. We have also experimented with more complicated ways to build confidence intervals (e.g., bias correction methods), but have found their performance to be erratic in high-dimension and offer no improvement.

We first give some background regarding the bootstrap and estimation of linear models in high dimensions before presenting our new results.

1.2 Background: Inference Using the Bootstrap

We consider the setting $y_i = \beta'X_i + \epsilon_i$, where $E(\epsilon_i) = 0$ and $\text{var}(\epsilon_i) = \sigma_\epsilon^2$. The vector β is estimated as minimizing the average loss,

$$\widehat{\beta}_\rho = \operatorname{argmin}_{b \in \mathbb{R}^p} \sum_{i=1}^n \rho(y_i - X_i'b), \quad (1)$$

where ρ defines the loss function for a single observation. The function ρ is assumed to be convex in all the paper. Common choices are $\rho(x) = x^2$, i.e., least-squares, $\rho(x) = |x|$, which defines L_1 regression, or Huber $_k$ loss where $\rho(x) = (x^2/2)1_{|x| < k} + (k|x| - k^2/2)1_{|x| \geq k}$.

Bootstrap methods are used in order to estimate the distribution of the estimate $\widehat{\beta}_\rho$ under the true data-generating distribution, F . The bootstrap estimates this distribution with the distribution obtained when the data is drawn from an estimate \widehat{F} of F . Following standard convention, we designate this bootstrapped estimator $\widehat{\beta}_\rho^*$ to note that this is an estimate of β using loss function ρ when the data-generating distribution is known to be exactly equal to \widehat{F} . Since \widehat{F} is completely specified, we can in principle exactly calculate the distribution of $\widehat{\beta}_\rho^*$ and use it as an approximation of the distribution of $\widehat{\beta}_\rho$ under F . In practice, we simulate B independent draws of size n from the distribution \widehat{F} and perform inference based on the empirical distribution of $\widehat{\beta}_\rho^{*b}$, $b = 1, \dots, B$.

In bootstrap inference for the linear model, there are two common methods for re-sampling, which results in different estimates \widehat{F} . In the first method, called the residual bootstrap, \widehat{F} is an estimate of the conditional distribution of y_i given β and X_i . In this case, the corresponding resampling method consists of resampling ϵ_i^* from an estimate of the distribution of ϵ and forming data $y_i^* = X_i' \widehat{\beta}_\rho + \epsilon_i^*$, from which $\widehat{\beta}_\rho^*$ is computed. This method of bootstrapping assumes that the linear model is correct for the mean of y (i.e., that $\mathbf{E}(y_i) = X_i' \beta$); it also assumes fixed X_i design vectors because the sampling is conditional on the X_i . In the second method, called pairs bootstrap, \widehat{F} is an estimate of the joint distribution of the vector $(y_i, X_i) \in R^{p+1}$ given by the empirical joint distribution of $\{(y_i, X_i)\}_{i=1}^n$; the corresponding resampling method resamples the pairs (y_i, X_i) . This method makes no assumption about the mean structure of y and, by resampling the X_i , also does not condition on the values of X_i . For this reason, pairs resampling is often considered to be more generally applicable than residuals resampling (see e.g., Davison and Hinkley, 1997).

1.3 Background: High-dimensional Inference of Linear Models

Recent research shows that $\widehat{\beta}_\rho$ has very different asymptotic properties when p/n has a limit κ that is bounded away from zero than it does in the classical setting where $p/n \rightarrow 0$ (see e.g., Huber, 1973; Huber and Ronchetti, 2009; Portnoy, 1984, 1985, 1986, 1987; Mammen, 1989 for $\kappa = 0$; El Karoui et al., 2013 for $\kappa \in (0, 1)$). A simple example is that the vector $\widehat{\beta}_\rho$ is no longer consistent in Euclidean norm when $\kappa > 0$. We should be clear, however, that projections on fixed non-random directions such as we consider, i.e., $v' \widehat{\beta}_\rho$, are \sqrt{n} consistent for $v' \beta$, even when $\kappa > 0$. In particular, the coordinates of $\widehat{\beta}_\rho$ are \sqrt{n} -consistent for the coordinates of β (El Karoui et al., 2013, Lemma 1). Hence, in practice the estimator $\widehat{\beta}_\rho$ is still a reasonable quantity to consider.

Bootstrap in high-dimensional linear models Very interesting work exists already in the literature about bootstrapping regression estimators when p is allowed to grow with n (Shorack, 1982; Wu, 1986; Mammen, 1989, 1992, 1993; Parzen et al., 1994; Koenker, 2005, Section 3.9). With a few exceptions, this work has been in the classical, low-dimensional setting where either p is held fixed or p grows slowly relative to n (i.e., $\kappa = 0$ in our notation). For instance, in Mammen (1993), it is shown that under mild technical conditions and assuming that $p^{1+\delta}/n \rightarrow 0$, $\delta > 0$, the pairs bootstrap distribution of linear contrasts $v'(\widehat{\beta}_\rho^* - \widehat{\beta}_\rho)$ is in fact very close to the sampling distribution of $v'(\widehat{\beta}_\rho - \beta)$ with high-probability, when using least-squares. Other results, such as Shorack (1982) and Mammen (1989),

also allow for increasing dimensions, for example in the case of linear contrasts in robust regression, by making assumptions on the diagonal entries of the hat matrix. In our context, these assumptions would be satisfied only if $p/n \rightarrow 0$. Hence those interesting results do not apply to the present study. We also note that Hall (1992, p. 167) contains cautionary notes about using the bootstrap in high-dimension.

While there has not been much theoretical work on the bootstrap in the setting where $p/n \rightarrow \kappa \in (0, 1)$, one early work of Bickel and Freedman (1983) considered bootstrapping scaled residuals for least-squares regression when $\kappa > 0$. They show that when $p/n \rightarrow \kappa \in (0, 1)$, there exists a data-dependent direction c , such that $c' \widehat{\beta}_{LS}^*$ does not have the correct asymptotic distribution (Bickel and Freedman, 1983, Theorem 3.1, p.39), i.e., its distribution is not conditionally in probability close to the sampling distribution of $c' \widehat{\beta}_{LS}$. Furthermore, they show that when the errors in the model are Gaussian, under the assumption that the diagonal entries of the hat matrix are not all close to a constant, the empirical distribution of the residuals is a scaled-mixture of Gaussian, which is not close to the original error distribution.

As we previously explained, in this work we instead only consider inference for *predefined* contrasts $v'\beta$. The important and interesting problems pointed out in Bickel and Freedman (1983) disappear if we focus on fixed, non-data-dependent projection directions. Hence, our work complements the work of Bickel and Freedman (1983) and is not redundant with it.

There has been some recent interest in residual bootstrap methods for penalized likelihood methods in high-dimensions (often proposed for the case when $p \gg n$), for example lasso estimates (Chatterjee and Lahiri, 2010, 2011), adaptive lasso estimates (Chatterjee and Lahiri, 2013), de-biased lasso estimates (Belloni et al., 2015; Dezeure et al., 2017), and ridge regression (Lopes, 2014). These bootstrap results make the assumption of sparsity of some form, generally in terms of the number of non-zero components of β , but in the case of Lopes (2014) by the assumption that the design matrix X is nearly low-rank. As explained previously, our work is focused on a very different line of inquiry: the case of a comparatively diffuse signal in β , where there is no reduction of the high-dimensional problem to a low-dimensional approximation.

The role of the distribution of X An important consideration in interpreting theoretical work on linear models in high dimensions is the role of the design matrix X . In classical asymptotic theory, the results can be stated conditionally on X so that the assumptions can be stated in terms of conditions that can be evaluated on any observed design matrix X . In the high dimensional setting, the available theoretical tools do not yet allow for an asymptotic analysis conditional on X ; instead the results make assumptions about the distribution of the entries of X . Theoretical work in the nascent literature for the high dimensional setting usually allows for a fairly general class of distributions for the individual elements of X_i and can handle covariance between the predictor variables. However, the X_i 's are generally considered i.i.d., which limits the ability of any X_i to be too influential in the fit of the model (see AppendixA for more detail). For discussion of limitations of the corresponding models for statistical purposes, see Diaconis and Freedman (1984); Hall et al. (2005); El Karoui (2009).

1.4 Notations and Default Conventions

When referring to the Huber loss in a numerical context, we refer (unless otherwise noted) to the default implementation in the `r1m` package in R, where the transition from quadratic to linear behavior is at $k = 1.345$. We call X the design matrix and $\{X_i\}_{i=1}^n$ its rows. We have $X_i \in \mathbb{R}^p$. β denotes the true regression vector, i.e., the population parameter. $\widehat{\beta}_\rho$ refers to the estimate of β using loss ρ ; from this point on, however, we will often drop the ρ and refer to simply $\widehat{\beta}$. The i -th residual is denoted as e_i , i.e., $e_i = y_i - X_i' \widehat{\beta}$. Throughout the paper, we assume that the linear model holds, i.e., $y_i = X_i' \beta + \epsilon_i$ for some fixed $\beta \in \mathbb{R}^p$ and that ϵ_i 's are i.i.d. with mean 0 and $\text{var}(\epsilon_i) = \sigma_\epsilon^2$. We call G the distribution of ϵ . When we need to stress the impact of the error distribution on the distribution of $\widehat{\beta}_\rho$, we will write $\widehat{\beta}_\rho(G)$ or $\widehat{\beta}_\rho(\epsilon)$ to denote our estimate of β obtained assuming that ϵ_i 's are i.i.d. G .

We denote generically by $\kappa = \lim_{n \rightarrow \infty} p/n$. We restrict ourselves to $\kappa \in (0, 1)$. The standard notation $\widehat{\beta}_{(i)}$ refers to the leave-one-out estimate of $\widehat{\beta}$ where the i -th pair (y_i, X_i) is excluded from the regression, and $\tilde{e}_{i(i)} \triangleq y_i - X_i' \widehat{\beta}_{(i)}$ is the i -th predicted error (based on the leave-one-out estimate of $\widehat{\beta}$). We also use the notation $\tilde{e}_{j(i)} \triangleq y_j - X_j' \widehat{\beta}_{(i)}$. The hat matrix is of course $H = X(X'X)^{-1}X'$. o_P denotes a “little-oh” in probability, a standard notation (see van der Vaart, 1998). When we say that we work with a Gaussian design with covariance Σ , we mean that $X_i \stackrel{iid}{\sim} \mathcal{N}(0, \Sigma)$. Throughout the paper, the loss function ρ is assumed to be convex, $\mathbb{R} \mapsto \mathbb{R}^+$. We use the standard notation $\psi = \rho'$. We finally assume that ρ is such that there is a unique solution to the robust regression problem—an assumption that applies to all classical losses in the context of our paper.

2. Residual Bootstrap

We first focus on the method of bootstrap resampling where \widehat{F} is the conditional distribution $y|\widehat{\beta}, X$. In this case the distribution of $\widehat{\beta}^*$ under \widehat{F} is formed by independent resampling of ϵ_i^* from an estimate \widehat{G} of the distribution G that generated ϵ_i . Then new data y_i^* are formed as $y_i^* = X_i' \widehat{\beta} + \epsilon_i^*$ and the model is fitted to this new data to get $\widehat{\beta}^*$. Generally the estimate of the error distribution, \widehat{G} , is taken to be empirical distribution of the observed residuals, so that the ϵ_i^* are found by sampling with replacement from the e_i .

Yet, even a cursory evaluation of e_i in the simple case of least-squares regression ($\rho(x) = x^2$) reveals that the empirical distribution of the e_i may be a poor approximation to the error distribution of ϵ_i . In particular, it is well known that e_i has variance equal to $\sigma_\epsilon^2(1 - h_i)$ where h_i is the i th diagonal element of the hat matrix. This problem becomes particularly pronounced in high dimensions. For instance, if $X_i \stackrel{iid}{\sim} \mathcal{N}(0, \Sigma)$, $h_i = p/n + o_P(1)$ so that e_i has variance approximately $\sigma_\epsilon^2(1 - p/n)$, i.e., generally much smaller than the true variance of ϵ for $\lim p/n > 0$. This fact is also true when assuming more general distributions for the design matrix X (see e.g., Wachter, 1978; Haff, 1979; Silverstein, 1995; Pajor and Pastur, 2009; El Karoui and Koesters, 2011, where the main results of some of these papers require minor adjustments to get the approximation of h_i we just mentioned).

In Figure 1, we plot the error rate of 95% bootstrap confidence intervals based on resampling from the residuals for different loss functions, based on a simulation when the entries of X are i.i.d. $\mathcal{N}(0, 1)$ and $\epsilon \sim N(0, 1)$. Even in this idealized situation, as the ratio of

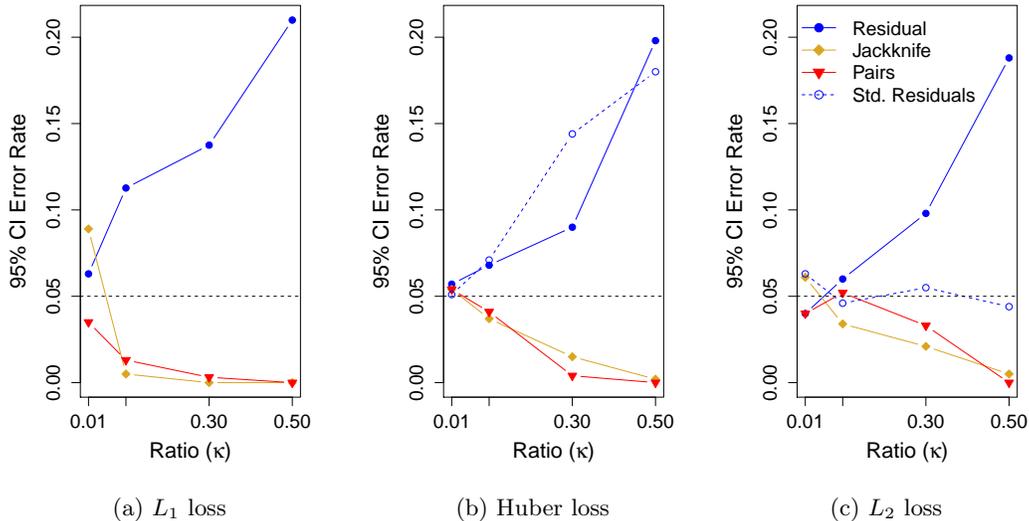


Figure 1: **Performance of 95% confidence intervals of β_1** : Here we show the coverage error rates for 95% confidence intervals for $n = 500$ based on applying common resampling-based methods to simulated data: pairs bootstrap (red), residual bootstrap (blue), and jackknife estimates of variance (yellow). These bootstrap methods are applied with three different loss functions shown in the three plots above: (a) L_1 , (b) Huber, and (c) L_2 . For L_2 and Huber loss, we also show the performance of methods for standardizing the residuals before bootstrapping described in the text (blue, dashed line). If accurate, all of these methods should have an error rate of 0.05 (shown as a horizontal black line). Error rates above 5% correspond to anti-conservative methods. Error rates below 5% correspond to conservative methods. The error rates are based on 1,000 simulations, with $N(0, 1)$ error, and entries of the design matrix i.i.d $N(0, 1)$; see the description in Appendix D.1 for more details. The exact values plotted here are given in Table A-1 in Appendix I.

p/n increases the error rate of the confidence intervals in least squares regression increases well beyond the expected 5%: we observe error rates of 10-15% for $p/n = 0.3$ and approximately 20% for $p/n = 0.5$. We see similar error rates for other robust-regression methods, such as L_1 and Huber loss, and also for different error distributions and distributions of X (Supplementary Figures A-1 and A-2). We explain some of the reasons for these problems in Subsection 2.2 below.

2.1 Bootstrapping from Corrected Residuals

While resampling directly from the uncorrected residuals is widespread and often given as a standard bootstrap procedure (e.g., Koenker, 2005; Chernick, 1999), the discrepancy between the distribution of ϵ_i and e_i has spurred more refined recommendations in the

case of least-squares: form corrected residuals $r_i = e_i/\sqrt{1 - h_i}$ and sample the ϵ_i^* from the empirical distribution of the $r_i - \bar{r}$ (see e.g., Davison and Hinkley, 1997).

This correction is known to exactly align the variance of r_i with that of ϵ_i regardless of the design vectors X_i or the true error distribution, using simply the fact that the hat matrix is a rank $\min(n, p)$ orthogonal projection matrix. We see that for L_2 loss it corrects the error in bootstrap inference in our simulations (Figure 1). This is not so surprising, given that with L_2 loss, the error distribution G impacts the inference on β only through σ_ϵ^2 , in the case of homoskedastic errors (see Section 2.4 for much more detail).

However, this adjustment of the residuals is a correction specific to the least-squares problem. Similar corrections for robust estimation procedures using a loss function ρ are given by McKean et al. (1993) with standardized residuals r_i given by,

$$r_i = \frac{e_i}{\sqrt{1 - dh_i}}, \text{ where } d = \frac{2 \sum e'_j \psi(e'_j)}{\sum \psi(e'_j)} - \frac{\sum \psi(e'_j)^2}{(\sum \psi(e'_j))^2}, \quad (2)$$

where h_i is the i -th diagonal entry of the hat matrix, $e'_j = e_j/s$, s is a estimate of σ , and ψ is the derivative of ρ , assuming ψ is a bounded and odd function (see Davison and Hinkley, 1997 for a complete description of its implementation for the bootstrap and McKean et al., 1993 for a full description of the regularity conditions).

Unlike the correction for L_2 loss mentioned earlier, however, the scaling described in Equation (2) for the residuals is an approximate variance correction, and the approximation depends on assumptions that do not hold true in higher dimensions. The error rate of confidence intervals in our simulations based on this rescaling show no improvement in high dimensions over that of simple bootstrapping of the residuals. This could be explained by the fact that standard perturbation analytic methods used for the analysis of M-estimators in low-dimension, which are at the heart of the correction in Equation (2), fail in high-dimensions.

2.2 Understanding the Behavior of the Residual Bootstrap

This misbehavior of the residual bootstrap can be explained by the fact that in high-dimension, the residuals tend to have a very different distribution from that of the true errors. Their distributions differ not only in simple properties, such as their variances, but in more general aspects, such as their marginal distributions. To make these statements precise, we make use of the previous work of El Karoui et al. (2013) and El Karoui (2013). These papers do not discuss bootstrap or resampling issues, but rather are entirely focused on providing asymptotic theory for the behavior of $\hat{\beta}_\rho$ as $p/n \rightarrow \kappa \in (0, 1)$; in the course of doing so, they characterize the asymptotic relationship of e_i to ϵ_i in high-dimensions. We make use of this relationship to characterize the behavior of the residual bootstrap and to suggest an alternative estimates of \hat{G} for bootstrap resampling.

Behavior of residuals in high-dimensional regression We now summarize the asymptotic relationship between e_i and ϵ_i in high-dimensions given in the above cited work (see AppendixA for a more detailed and technical summary). Let $\hat{\beta}_{(i)}$ be the estimate of β based on fitting the linear model of Equation (1) without using observation i , and $\tilde{e}_{j(i)}$ be the error of observation j from this model (the leave-one-out or predicted error), i.e., $\tilde{e}_{j(i)} = y_j - X_j' \hat{\beta}_{(i)}$

For simplicity of exposition, X_i is assumed to have an elliptical distribution, i.e., $X_i = \lambda_i \Gamma_i$, where $\Gamma_i \sim N(0, \Sigma)$, and λ_i is a scalar random variable independent of Γ_i with $\mathbf{E}(\lambda_i^2) = 1$. For simplicity in restating their results, we will assume $\Sigma = \text{Id}_p$, but equivalent statements can be made for arbitrary Σ ; similar results also apply when $\Gamma_i = \Sigma^{1/2} \xi_i$, with ξ_i having i.i.d. non-Gaussian entries, satisfying a few technical requirements (see AppendixA).

With this assumption on X_i , for any sufficiently smooth loss function ρ and any size dimension where $p/n \rightarrow \kappa < 1$, the relationship between the i -th residual e_i and the true error ϵ_i can be summarized as,

$$\tilde{e}_{i(i)} = \epsilon_i + |\lambda_i| \|\widehat{\beta}_{\rho(i)} - \beta\|_2 Z_i + o_P(u_n) \tag{3}$$

$$e_i + c_i \lambda_i^2 \psi(e_i) = \tilde{e}_{i(i)} + o_P(u_n) \tag{4}$$

where Z_i is a random variable distributed $N(0, 1)$ and independent of ϵ_i . The variable u_n refers to a sequence of numbers tending to 0. The quantities c_i , λ_i and $\|\widehat{\beta}_{\rho(i)} - \beta\|_2$ are all of order 1, i.e., they are not close to 0 in general in the high-dimensional setting. The scalar c_i is given as $\frac{1}{n} \text{trace}(S_i^{-1})$, where $S_i = \frac{1}{n} \sum_{j \neq i} \psi'(\tilde{e}_{j(i)}) X_j X_j'$. For p, n large the c_i 's are approximately equal and $\|\widehat{\beta}_{\rho(i)} - \beta\|_2 \simeq \|\widehat{\beta}_\rho - \beta\|_2 \simeq \mathbf{E}(\|\widehat{\beta}_\rho - \beta\|_2)$; furthermore $c_i \lambda_i^2$ can be approximated by $X_i' S_i^{-1} X_i / n$. Note that when ρ is either non-differentiable at all points (L_1) or not twice differentiable (Huber), arguments can be made that make these expressions valid, using for instance the notion of sub-differential for ψ (Hiriart-Urruty and Lemaréchal, 2001).

Interpretation of Equations (3) and (4) Equation (3) means that the marginal distribution of the leave- i -th-out predicted error, $\tilde{e}_{i(i)}$, is asymptotically a convolution of the true error, ϵ_i , and an independent scale mixture of Normals. Furthermore, Equation (4) means that the i -th residual e_i can be understood as a non-linear transformation of $\tilde{e}_{i(i)}$. As we discuss below, these relationships are qualitatively very different from the classical case $p/n \rightarrow 0$.

2.2.1 CONSEQUENCE FOR THE RESIDUAL BOOTSTRAP

We apply these results to the question of the residual bootstrap to give an understanding of why bootstrap resampling of the residuals can perform so badly in high-dimension. The distribution of the e_i is far removed from that of the ϵ_i , and hence bootstrapping from the residuals effectively amounts to sampling errors from a distribution that is very different from the original error distribution, ϵ .

The impact of these discrepancies for bootstrapping is not equivalent for all dimensions, error distributions, or loss functions. It depends on the constant c_i and the risk, $\|\widehat{\beta}_{\rho(i)} - \beta\|_2$, both of which are highly dependent on the dimensions of the problem, the distribution of the errors and the choice of loss function. We now discuss some of these issues.

Least Squares regression In the case of least squares regression, the relationships given in Equation (3) are exact, i.e., $u_n = 0$. Further, $\psi(x) = x$, and $c_i = h_i / (1 - h_i)$, giving the well known linear relationship $e_i = (1 - h_i) \tilde{e}_{i(i)}$ (see, e.g., the standard reference Weisberg, 2014). This linear relationship is exact regardless of dimension, though the dimensionality aspects are captured by h_i . This expression can be used to show that asymptotically $\mathbf{E}(\sum_{i=1}^n e_i^2) = \sigma_\epsilon^2 (n - p)$, when ϵ_i 's have the same variance. Hence, sampling at random

from the residuals results in a distribution that underestimates the variance of the errors by a factor $1 - p/n$. The corresponding bootstrap confidence intervals are then naturally too small, and hence the error rate increases far from the nominal 5% - as we observed in Figure 1c.

More general robust regression The situation is much more complicated for general robust regression estimators. One clear implication of Equations (3) and (4) is that simply rescaling the residuals e_i should not in general result in an estimated error distribution \hat{G} that will have similar properties to those of G . The relationship between the residuals and the errors is very non-linear in high-dimensions. This is why in what follows we will propose to work with leave-one-out predicted errors $\tilde{e}_{i(i)}$ instead of the residuals e_i .

The classical case of $p/n \rightarrow 0$: In this setting, $c_i \rightarrow 0$ and therefore Equation (3) shows that the residuals e_i are approximately equal in distribution to the predicted errors, $\tilde{e}_{i(i)}$. Similarly, $\hat{\beta}_\rho$ is L_2 consistent when $p/n \rightarrow 0$, so $\|\hat{\beta}_{\rho(i)} - \beta\|_2^2 \rightarrow 0$ and Equation (4) gives $\tilde{e}_{i(i)} \simeq \epsilon_i$. Hence, the residuals should be fairly close to the true errors in the model when p/n is small. This dimensionality assumption is key to many theoretical analyses of robust regression, and underlies the derivation of corrected residuals r_i of McKean et al. (1993) given in Equation (2) above for losses other than L_2 .

2.3 Alternative Residual Bootstrap Procedures

We propose two methods for improving the performance of confidence intervals obtained through the residual bootstrap. Both do so by providing alternative estimates of \hat{G} from which bootstrap errors ϵ_i^* can be drawn. They estimate a \hat{G} appropriate for the setting of high-dimensional data by accounting for relationship of the distribution of ϵ and $\tilde{e}_{i(i)}$.

Method 1: Deconvolution The relationship in Equation (3) says that the distribution of $\tilde{e}_{i(i)}$ is a convolution of the correct G distribution and a normal distribution. This suggests applying techniques for deconvolving a signal from Gaussian noise. Specifically, we propose the following bootstrap procedure: **1)** calculate the predicted errors, $\tilde{e}_{i(i)}$; **2)** estimate the variance of the normal (i.e., $|\lambda_i| \|\hat{\beta}_{\rho(i)} - \beta\|_2^2$); **3)** deconvolve in $\tilde{e}_{i(i)}$ the error term ϵ_i from the normal term; **4)** Use the resulting estimate \hat{G} to draw errors ϵ_i^* for residual bootstrapping.

Deconvolution problems are known to be very difficult (see Fan, 1991, Theorem 1 p. 1260, that gives $1/\log(n)^\alpha$ rates of convergence when convolving with a Gaussian distribution). The resulting deconvolved errors are likely to be quite noisy estimates of ϵ_i . However, it is possible that while individual estimates are poor, the distribution of the deconvolved errors is estimated well-enough to form a reasonable \hat{G} for the bootstrap procedure.

We used the deconvolution algorithm in the `decon` package in R (Wang and Wang, 2011) to estimate the distribution of ϵ_i . The deconvolution algorithm requires knowledge of the variance of the Gaussian that is convolved with the ϵ_i , i.e., estimation of $|\lambda_i| \|\hat{\beta}_{\rho(i)} - \beta\|_2^2$ term. In what follows, we assume a Gaussian design, i.e., $\lambda_i = 1$, so that we need to estimate only the term $\|\hat{\beta}_{\rho(i)} - \beta\|_2^2$. An estimation strategy for the more general setting of $|\lambda_i| \neq 1$ is presented in Appendix B.5. We use the fact that $\|\hat{\beta}_{\rho(i)} - \beta\|_2^2 \simeq \|\hat{\beta}_\rho - \beta\|_2^2$ for all i and estimate $\|\hat{\beta}_{\rho(i)} - \beta\|_2^2$ as $\widehat{var}(\tilde{e}_{i(i)}) - \hat{\sigma}_\epsilon^2$, where $\widehat{var}(\tilde{e}_{i(i)})$ is the empirical variance of the $\tilde{e}_{i(i)}$ and $\hat{\sigma}_\epsilon^2$ is an estimate of the variance of G , which we discuss below. We note

that the deconvolution strategy we employ makes assumptions of homoskedastic errors ϵ_i 's, which is true in our simulations but may not be true in practice. See AppendixB for details regarding the implementation of Method 1.

Method 2: Bootstrapping from standardized $\tilde{e}_{i(i)}$ A simpler alternative is bootstrapping from the predicted error terms, $\tilde{e}_{i(i)}$, without deconvolution. Specifically, we propose to bootstrap from a scaled version of $\tilde{e}_{i(i)}$,

$$\tilde{r}_{i(i)} = \frac{\hat{\sigma}_\epsilon}{\sqrt{\widehat{\text{var}}(\tilde{e}_{i(i)})}} \tilde{e}_{i(i)},$$

where $\widehat{\text{var}}(\tilde{e}_{i(i)})$ is the standard estimate of the variance of $\tilde{e}_{i(i)}$ and $\hat{\sigma}_\epsilon$ is an estimate of σ_ϵ . This scaling aligns the first two moments of $\tilde{e}_{i(i)}$ with those of ϵ_i . On the face of it, resampling from $\tilde{r}_{i(i)}$ seems problematic, since Equation (3) demonstrates that $\tilde{e}_{i(i)}$ does not have the same distribution as ϵ_i , even if the first two moments are the same. However, as we demonstrate in simulations, this distributional mismatch appears to have limited practical effect on our bootstrap confidence intervals.

Estimation of σ_ϵ^2 Both methods described above require an estimator of σ_ϵ that is consistent regardless of dimension and error distribution. As we have explained earlier, for general ρ we cannot rely on the observed residuals e_i nor on $\tilde{e}_{i(i)}$ for estimating σ_ϵ (see Equations (3) and (4)). The exception is the standard estimate of σ_ϵ^2 from least-squares regression, i.e., $\rho(x) = x^2$,

$$\hat{\sigma}_{\epsilon,LS}^2 = \frac{1}{n-p} \sum_i e_{i,L_2}^2.$$

$\hat{\sigma}_{\epsilon,LS}^2$ is a consistent estimator of σ^2 for any error distribution G , assuming i.i.d. errors and mild moment requirements. In implementing the two alternative residual-bootstrap methods described above, we use $\hat{\sigma}_{\epsilon,LS}$ as our estimate of σ_ϵ , including for bootstrapping robust regression where $\rho(x) \neq x^2$.

Performance in bootstrap inference In Figure 2 we show the error rate of confidence intervals based on the two residual-bootstrap methods we proposed above. We see that both methods control the Type I error, unlike bootstrapping directly from the residuals, and that both methods are conservative. There is little difference between the two methods with this sample size ($n = 500$), though with $n = 100$, we observe the deconvolution performance to be worse in L_1 (data not shown).

The deconvolution strategy, however, depends on the distribution of the design matrix, which in these simulations we assumed was Gaussian (so we did not have to estimate λ_i 's). For elliptical designs ($\lambda_i \neq 1$), the error rate of the deconvolution method described above, with no adaptation for the design, was similar to that of uncorrected residuals in high dimensions (i.e., > 0.25 for $p/n = 0.5$). Individual estimates of λ_i might improve the deconvolution strategy, but this problem points to the general reliance of the deconvolution method on precise knowledge about the design matrix. The bootstrap using standardized predicted errors, on the other hand, had a Type I error for an elliptical design only slightly higher than the target 0.05 (around 0.07, data not shown), suggesting that it might be less sensitive to the properties of the design matrix.

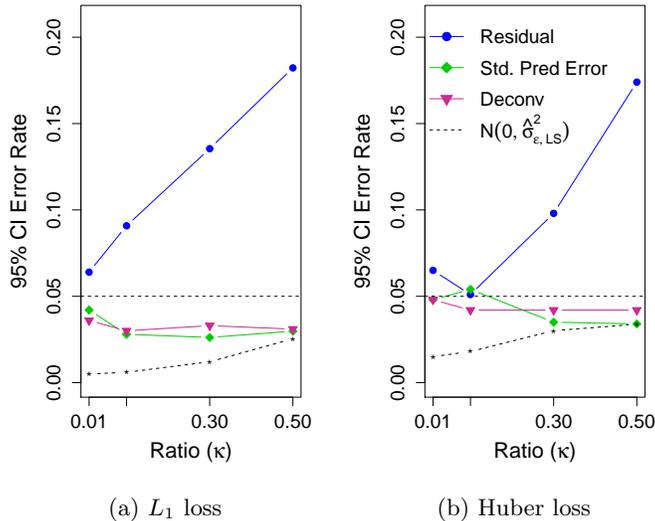


Figure 2: **Bootstrap based on predicted errors:** We plotted the error rate of 95% confidence intervals for the alternative bootstrap methods described in Section 2.3: bootstrapping from standardized predicted errors (green) and from deconvolution of predicted error (magenta). We demonstrate its improvement over the standard residual bootstrap (blue) for (a) L_1 loss and (b) Huber loss. The error distribution is double exponential, but otherwise the simulations parameters are as in Figure 1. The error rates on confidence intervals based on bootstrapping from a $N(0, \hat{\sigma}_{\epsilon, LS}^2)$ (dashed curve) are as a lower bound on the problem. For the precise error rates see Appendix, Table A-3.

Given our previous discussion of the behavior of $\tilde{e}_{i(i)}$, it is somewhat surprising that resampling from the distribution of $\tilde{r}_{i(i)}$ performed well in our simulations. Clearly a few cases exist where $\tilde{r}_{i(i)}$ should work well as an approximation of ϵ_i . We have already noted that as $p/n \rightarrow 0$, the effect of the convolution with the Gaussian disappears since $\|\hat{\beta}_\rho - \beta\| \rightarrow 0$; in this case both e_i and $\tilde{r}_{i(i)}$ should be good estimates of ϵ_i . Similarly, in the case $\epsilon_i \sim N(0, \sigma^2)$, Equation (3) tells us that $\tilde{e}_{i(i)}$ are also asymptotically marginally normally distributed, so that correcting the variance should result in $\tilde{r}_{i(i)}$ having the same distribution as ϵ_i , at least when $X_{i,j}$ are i.i.d.

Surprisingly, for larger p/n we do not see a deterioration of the performance of bootstrapping from $\tilde{r}_{i(i)}$. This is unexpected, since as $p/n \rightarrow 1$ the risk $\|\hat{\beta}_\rho - \beta\|_2^2$ grows to be much larger than σ_ϵ^2 (a claim we will make more precise in the next section); together with Equation (3), this implies that $\tilde{r}_{i(i)}$ is essentially distributed $N(0, \hat{\sigma}_{\epsilon, LS}^2)$ as $p/n \rightarrow 1$ regardless of the original distribution of ϵ_i . This is confirmed in Figure 2 where we superimpose the results of bootstrap confidence intervals from when we simply estimate \hat{G} with $N(0, \hat{\sigma}_{\epsilon, LS}^2)$; we see the Type I error rate of the confidence intervals based on bootstrapping from $\tilde{r}_{i(i)}$ do indeed approach that of $N(0, \hat{\sigma}_{\epsilon, LS}^2)$. Putting these two pieces of information

together leads to the conclusion that as $p/n \rightarrow 1$ we can estimate \hat{G} simply as $N(0, \hat{\sigma}_{\epsilon, LS})$ regardless of the actual distribution of ϵ .

In the next section we give some theoretical results that seek to understand this phenomenon.

2.4 Behavior of the Risk of $\hat{\beta}$ When $\kappa \rightarrow 1$

In the previous section we saw even if the distribution of the bootstrap errors ϵ_i^* , given by \hat{G} , is not close to that of G , we can sometime get accurate bootstrap confidence intervals. For example, in least squares Equation (3) makes clear that even the standardized residuals, r_i , do not have the same marginal distribution as ϵ_i , yet they still provide accurate bootstrap confidence intervals in our simulations. We would like to understand for what choice of distributions \hat{G} will we see the same performance in our bootstrap confidence intervals of $\hat{\beta}_1$?

When working conditional on X as in residual resampling, the statistical properties of $(\hat{\beta}^* - \hat{\beta})$ differ from that of $(\hat{\beta} - \beta)$ only because the errors are drawn from a different distribution: \hat{G} rather than G . Then to understand whether the distribution of $\hat{\beta}_1^*$ matches that of $\hat{\beta}_1$ we can ask, what are the distributions of errors, G , that yield the same distribution for the resulting $\hat{\beta}_1(G)$? In this section, we narrow our focus on understanding not the entire distribution of $\hat{\beta}_1$, but only its variance. We do so because under assumptions on the design matrix X , it is known that $\hat{\beta}_1$ is asymptotically normally distributed. This is true for both the classical setting of $\kappa = 0$ and the high-dimensional setting of $\kappa \in (0, 1)$ (see AppendixA for a review of these results and a more technical discussion). Our previous question is then reduced to understanding which distributions G give the same $\text{var}(\hat{\beta}_1(G))$.

In the setting of least squares, it is clear that the only property of $\epsilon_i \stackrel{iid}{\sim} G$ that matters for the variance of $\hat{\beta}_{1, L_2}$ is σ_ϵ^2 , since $\text{var}(\hat{\beta}_{1, L_2}) = (X'X)^{-1}(1, 1)\sigma_\epsilon^2$. For general ρ , if we assume $p/n \rightarrow 0$, then $\text{var}(\hat{\beta}_{1, \rho})$ will depend on features of G beyond the first two moments (specifically through $\mathbf{E}(\psi^2(\epsilon)) / [\mathbf{E}(\psi'(\epsilon))]^2$, (Huber, 1973)). If we assume instead $p/n \rightarrow \kappa \in (0, 1)$, then it has been shown (El Karoui et al., 2013) that $\text{var}(\hat{\beta}_{1, \rho}(G))$ depends on G only by the effect of G on the squared risk of the vector $\hat{\beta}_\rho(G)$, i.e., through $\mathbf{E}(\|\hat{\beta}_\rho(G) - \beta\|_2^2)$ (for the convenience of the reader we give a review of these results, which are a bit scattered in the literature, in AppendixA).

For this reason, in the setting of $p/n \rightarrow \kappa \in (0, 1)$, we need to characterize the risk of $\hat{\beta}_\rho$ to understand when different distributions of ϵ result in the same variance of $\hat{\beta}_{1, \rho}$. In what follows, we denote by $r_\rho^2(\kappa; G)$ the asymptotic squared risk of $\hat{\beta}_\rho(G)$ as p and n tend to ∞ ,

$$r_\rho^2(\kappa; G) = \lim_{n, p \rightarrow \infty, \frac{p}{n} \rightarrow \kappa} E\|\hat{\beta}_\rho(G) - \beta\|^2.$$

The dependence of $r_\rho^2(\kappa; G)$ on ϵ is characterized, under appropriate technical conditions on X , ρ and ϵ_i 's, by a system of two non-linear equations (El Karoui et al., 2013). Specifically, if we define $\hat{z}_\epsilon = \epsilon + r_\rho(\kappa; G)Z$, where $Z \sim \mathcal{N}(0, 1)$ is independent of ϵ , and ϵ has the same distribution G as the ϵ_i 's, then there exists a constant c such that the pair of positive,

finite, and deterministic scalars $(c, r_\rho(\kappa; G))$ satisfy the following system of equations:

$$\begin{aligned} \mathbf{E} ((\text{prox}(c\rho))'(\hat{z}_\epsilon)) &= 1 - \kappa, \\ \kappa r_\rho^2(\kappa; G) &= \mathbf{E} ([\hat{z}_\epsilon - \text{prox}(c\rho)(\hat{z}_\epsilon)]^2). \end{aligned} \tag{5}$$

In this system, $\text{prox}(c\rho)$ refers to Moreau’s proximal mapping of the convex function $c\rho$ (see Moreau, 1965; Hiriart-Urruty and Lemaréchal, 2001).

It is therefore not entirely trivial to characterize those distributions Γ for which $r_\rho^2(\kappa; G) = r_\rho^2(\kappa; \Gamma)$. In the following theorem, however, we show that as $\kappa \rightarrow 1$, $r_\rho^2(\kappa; G)$ converges to a constant that depends only on σ_ϵ^2 . This implies that when $\kappa \rightarrow 1$, two different error distributions that have the same variances will result in estimators $\hat{\beta}_{1,\rho}$ with the same variance. Before stating our theorem formally, however, we will review the necessary assumptions for the system of equations in (5) to hold. For a precise statement of the assumptions, see El Karoui (2017)

Assumptions for Equation 5: The proof of (5) provided in El Karoui (2013) assumes that the X_i ’s have mean 0, $\text{cov}(X_i) = \text{Id}_p$, and they satisfy sub-Gaussian concentration assumptions (with constants dependent on n). El Karoui (2013) further assumes that the ϵ_i have a unimodal density, are independent from the X_{ij} , $\sup_{1 \leq i \leq n} |\epsilon_i| = O_P(\text{polyLog}(n))$, and that similar bounds also hold for a few moments of ϵ_i (the number of such moments depends on the loss function ρ). Log-concave densities such as those corresponding to double exponential or Gaussian errors used in the current paper fall within the scope of this theorem. The reader interested in generalizations and truly heavy-tailed situation is referred to El Karoui (2017) and references therein.

The loss function ρ is assumed by El Karoui (2013) (in the unpenalized case) to be non-negative, twice differentiable, strongly convex, non-linear, taking value 0 at 0, and with a derivative that grows at most polynomially at infinity and a second derivative that is locally Lipschitz, with local Lipschitz constant that grow at most polynomially at infinity. It should be noted that distributions with sufficiently many moments, the condition of strong convexity of ρ can be obtained by adding $\delta x^2/2$ to the initial ρ , with δ “small”, e.g., $\delta = 10^{-10}$, and that modification will change very little or anything to the estimator. Furthermore, the requirement of strong convexity of ρ , though superficially limiting, is likely an artifact of the proof, where the main motivation was log-concave distributions with an eye towards optimality (Bean et al., 2013). In fact, the theoretical predictions of (5) were verified numerically in El Karoui et al. (2011) outside of the assumptions stated above, and the predictions of Equation (5) were found to be very accurate in simulations even for non-smoothed ℓ_1 and Huber losses with certain error distributions.

We now state the theorem formally; see AppendixE for the proof of this statement.

Theorem 1 *Suppose we are working with robust regression estimators with loss ρ , and $p/n \rightarrow \kappa$. Suppose that $r_\rho^2(\kappa; G)$ is characterized by the system of equations in (5). Then,*

$$\lim_{\kappa \rightarrow 1} \frac{1 - \kappa}{\sigma_\epsilon^2} r_\rho^2(\kappa; G) = 1,$$

provided ρ is additionally differentiable near 0 and $\rho'(x) \sim x$ near 0.

Implications for the Bootstrap For the purposes of the residual-bootstrap, Theorem 1 implies that different methods of estimating the residual distribution \hat{G} will result in similar residual-bootstrap confidence intervals as $p/n \rightarrow 1$, if \hat{G} has the same variance. This agrees with our simulations, where both of our proposed bootstrap strategies set the variance of \hat{G} equal to $\hat{\sigma}_{\epsilon,LS}^2$ and both had similar performance in our simulations for large p/n . Furthermore, as we noted, for p/n closer to 1, they both had similar performance to a bootstrap procedure that simply sets $\hat{G} = \mathcal{N}(0, \hat{\sigma}_{\epsilon,LS}^2)$ (Figure 2) (see also Appendix A.3 for further discussion of residual bootstrap methods which draw from the “wrong” distribution, i.e., forms of wild bootstraps (Wu, 1986)).

We return specifically to the bootstrap based on $\tilde{r}_{i(i)}$, the standardized predicted errors. Equation (3) tells us that the marginal distribution of $\tilde{e}_{i(i)}$ is a convolution of the distribution of ϵ_i and a normal, with the variance of the normal governed by the term $\|\hat{\beta}_\rho - \beta\|_2$. Theorem 1 makes rigorous our previous assertion that as $p/n \rightarrow 1$, the normal term will dominate and the marginal distribution of $\tilde{e}_{i(i)}$ will approach normality, regardless of the distribution of ϵ . However, Theorem 1 also implies that as $p/n \rightarrow 1$, inference for the coordinates of β will be increasingly less reliant on features of the error distribution beyond the variance, implying that our standardized predicted errors, $\tilde{r}_{i(i)}$, will still result in an estimate \hat{G} that will give accurate confidence intervals. Conversely, as $p/n \rightarrow 0$ classical theory tells us that the inference of β relies heavily on the distribution G beyond the first two moments, but in that case the distribution of $\tilde{r}_{i(i)}$ approaches the correct distribution as we explained earlier. So bootstrapping from the marginal distribution of $\tilde{r}_{i(i)}$ also makes sense when p/n is small.

For κ between these two extremes it is difficult to theoretically predict the risk of $\hat{\beta}_\rho(\hat{G})$ when the distribution \hat{G} is given by resampling from the $\tilde{r}_{i(i)}$. We turn to numerical simulations to evaluate this risk. Specifically, for $\epsilon_i \sim G$, we simulated data that is a convolution of G and a normal with variance equal to $r_\rho^2(\kappa; G)$; we then scale this simulated data to have variance σ_ϵ^2 . The scaled data are the ϵ_i^* and we refer to the distribution of ϵ_i^* as the convolution distribution, denoted G_{conv} . Then, G_{conv} is the asymptotic version of the marginal distribution of the standardized predicted errors, $\tilde{r}_{i(i)}$, used in our bootstrap method proposed above.

In Figure 3 we plot for both Huber loss and L_1 loss the average risk $r_\rho(\kappa; G_{conv})$ (i.e., errors given by G_{conv}) relative to the average risk $r_\rho(\kappa; G)$ (i.e., errors distributed according to G), where G has a double exponential distribution. We also plot the relative average risk $r_\rho(\kappa; G_{norm})$, where $G_{norm} = \mathcal{N}(0, \sigma_\epsilon^2)$. As predicted by Theorem 1, for κ close to 1, $r_\rho(\kappa; G_{conv})/r_\rho(\kappa; G)$ and $r_\rho(\kappa; G_{norm})/r_\rho(\kappa; G)$ converge to 1. Conversely, as $\kappa \rightarrow 0$, $r_\rho(\kappa; G_{norm})/r_\rho(\kappa; G)$ diverges dramatically from 1, while $r_\rho(\kappa; G_{conv})/r_\rho(\kappa; G)$ approaches 1, as expected. For Huber, the divergence of $r_\rho(\kappa; G_{conv})/r_\rho(\kappa; G)$ from 1 is at most 8%, but the difference is larger for L_1 (12%), probably due to the fact that the convolution with a normal error has a larger effect on the risk for L_1 .

3. Pairs Bootstrap

As described above, estimating the distribution \hat{F} from the empirical distribution of (y_i, X_i) (*pairs bootstrapping*) is generally considered the most general and widely applicable method of bootstrapping, allowing for the linear model to be incorrectly specified (i.e., $\mathbf{E}(y_i)$ is not a linear function of X_i). It is also considered to be slightly more conservative compared

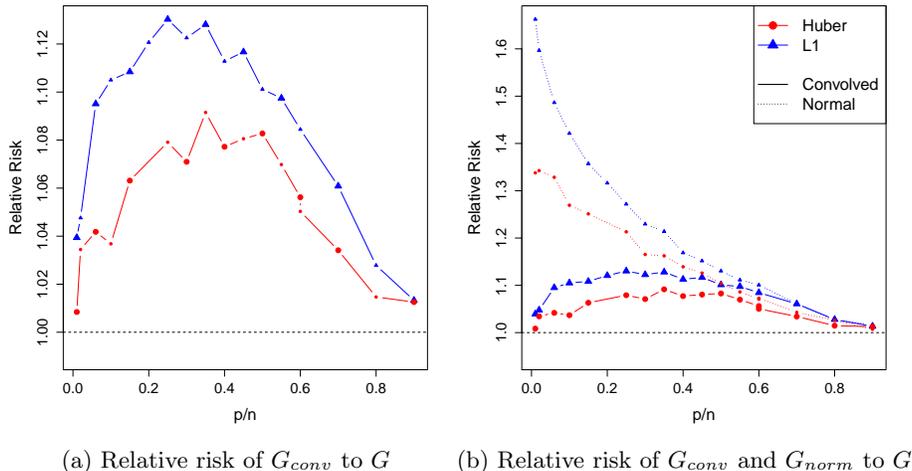


Figure 3: **Relative Risk of $\hat{\beta}$ for scaled predicted errors vs original errors - population version:** (a) Plotted with a solid lines are the ratios of the average risk of $\hat{\beta}(G_{conv})$ to the average risk of $\hat{\beta}(G)$ for Huber and L_1 loss. (b) shows the same plot, but added to the plot (dotted lines) is the relative risk of $\hat{\beta}(G)$ when the errors are distributed $G_{norm} = \mathcal{N}(0, \sigma_\epsilon^2)$. For both figures, the y-axis gives the relative risk, and the x-axis is the ratio p/n , with n fixed at 500. Blue/triangle plotting symbols indicate L_1 loss; red/circle plotting symbols indicate Huber loss. The average risk is calculated over 500 simulations, where the design matrix X has Gaussian entries. The “true” error distribution G is the standard Laplacian distribution with $\sigma_\epsilon^2 = 2$. Each simulation uses the standard estimate of σ_ϵ^2 from the generated ϵ_i ’s. $r_\rho(\kappa; G)$ was computed using a first run of simulations with $\epsilon_i \stackrel{iid}{\sim} G$. The Huber loss in this plot is Huber₁ and not the default Huber_{1.345} of the `rlm` function.

to bootstrapping from the residuals. In the case of random design, it makes also a lot of intuitive sense to use the pairs bootstrap, since resampling the predictors might be interpreted as mimicking the data generating process.

However, as in residual bootstrap, it is clear that the pairs bootstrap will have problems, at least in quite high dimensions. In fact, when resampling the X_i ’s from \hat{F} , the number of times a certain vector X_{i_0} is picked has asymptotically Poisson(1) distribution. So the expected number of different vectors appearing in the bootstrapped design matrix X^* is $n(1 - 1/e)$. When p/n is large, with increasingly high probability the bootstrapped design matrix X^* will no longer be of full rank. For example, if $p/n > (1 - 1/e) \approx 0.63$ then with probability tending to one as $n \rightarrow \infty$, the bootstrapped design matrix X^* is singular, even when the original design matrix X is of rank $p < n$. Bootstrapping the pairs in that situation makes little statistical sense (El Karoui, 2010, Subsection 4.4; Zheng et al., 2014).

For smaller ratios of p/n , we evaluate the performance of pairs bootstrapping on simulated data. We see that the performance of the bootstrap for inference also declines

dramatically as the dimension increases, becoming increasingly conservative (Figure 1). In pairs bootstrapping, the error rates of 95%-confidence-intervals drop far below the nominal 5%, and are essentially zero for the ratio of $p/n = 0.5$. Like residual bootstrap, this overall trend is seen for all the settings we simulated under (Supplemental Figures A-1, A-2). For L_1 loss, even ratios as small as 0.1 yield incredibly conservative bootstrap confidence intervals for $\hat{\beta}_1$, with the error rate dropping to less than 0.01. For Huber and L_2 losses, the severe loss of power in our simulations starts for ratios of 0.3.

A minimal requirement for the distribution of the bootstrapped data to give reasonable inferences is that the variance of the bootstrap estimator $\hat{\beta}_1^*$ needs to be a good estimate of the variance of $\hat{\beta}_1$. This is not the case in high-dimensions. In Figure 5 we plot the ratio of the variance of $\hat{\beta}_1^*$ to the variance of $\hat{\beta}_1$ evaluated over simulations. We see that for $p/n = 0.3$ and design matrices X with i.i.d. $\mathcal{N}(0, 1)$ entries, the average variance of $\hat{\beta}_1^*$ roughly overestimates the true variance of $\hat{\beta}_1$ by a factor 1.3 in the case of least-squares; for Huber and L_1 the bootstrap estimate of variance is roughly twice as large as it should be.

In the case of least-squares, we can further quantify this loss in power by comparing the size of the bootstrap confidence intervals to the size of the correct confidence interval based on theoretical results (Figure 4). We see that even for ratios κ as small as 0.1, the confidence intervals for some design matrices X were 15% larger for pairs bootstrap than the correct size (e.g., the case of elliptical distributions where λ_i is exponential). For much higher dimensions of $\kappa = 0.5$, the simple case of i.i.d. normal entries for the design matrix gives intervals that are 80% larger than needed; for the elliptical distributions we simulated, the width of the bootstrap confidence interval was as much as 3.5 times larger than that of the correct confidence interval. Furthermore, as we can see in Figure 1, least-squares regression represents the best case scenario; L_1 and Huber will have even worse loss of power and at smaller values of κ .

3.1 Theoretical Analysis for Least-Squares

In the setting of least-squares, we can for some distributions of the design matrix X theoretically determine the asymptotic expectation of the variance of $v'\hat{\beta}^*$ and show that it is a severe over-estimate of the true variance of $v'\hat{\beta}$.

We first setup some notation for the theorem that follows. Define $\hat{\beta}_w$ as the result of regressing y on X with random weight w_i for each observation (y_i, X_i) . In other words,

$$\hat{\beta}_w = \operatorname{argmin}_{u \in \mathbb{R}^p} \sum_{i=1}^n w_i (y_i - X_i' u)^2 .$$

We assume that the weights are independent of $\{y_i, X_i\}_{i=1}^n$ and define $\hat{\beta}_w^*$ to be the random variable with distribution equal to that of $\hat{\beta}_w$ conditional on the data $\{y_i, X_i\}_{i=1}^n$, i.e., $\hat{\beta}_w^* \stackrel{\mathcal{L}}{=} \hat{\beta}_w | \{y_i, X_i\}_{i=1}^n$. For the standard pairs bootstrap, the distribution of $\hat{\beta}^*$ from resampling from the pairs (y_i, X_i) is equivalent to the distribution of $\hat{\beta}_w^*$, where w is drawn from a multinomial distribution with expectation $1/n$ for each entry. In which case, the variance of $v'\hat{\beta}_w^*$ refers to the standard bootstrap estimate of variance given by the distribution of $v'\hat{\beta}^*$ over repeated resampling from the pairs (y_i, X_i) .

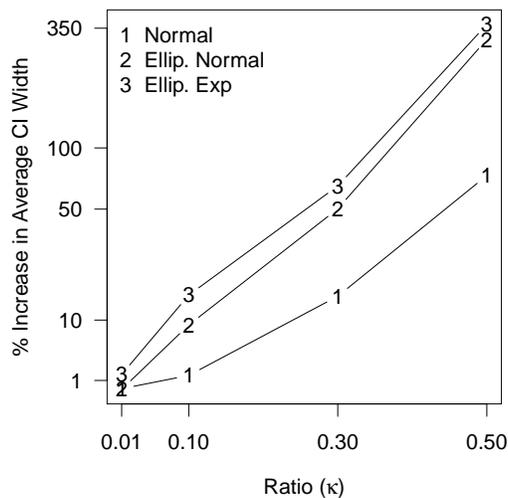


Figure 4: **Comparison of width of 95% confidence intervals of β_1 for L_2 loss:** Here we demonstrate the increase in the width of the confidence interval due to pairs bootstrapping. Shown on the y-axis is the percent increase of the average confidence interval width based on simulation ($n = 500$), as compared to the average for the standard confidence interval based on normal theory in L_2 ; the percent increase is plotted against the ratio $\kappa = p/n$ (x-axis). Shown are three different choices in simulating the entries of the design matrix X : (1) $X_{ij} \sim N(0, 1)$ (2) elliptical X_{ij} with $\lambda_i \sim N(0, 1)$ and (3) elliptical X_{ij} with $\lambda_i \sim Exp(\sqrt{2})$. The methods of simulation are the same as described in Figure 1; exact values are given in Table A-2 in Appendix I.

We have the following result for the expected value of the bootstrap variance of any contrast $v'\widehat{\beta}_w^*$ where v is deterministic, assuming independent weights with a Gaussian design matrix X and some mild conditions on the distribution of the w 's.

Theorem 2 *Let the weights $(w_i)_{i=1}^n$ be i.i.d. and without loss of generality that $\mathbf{E}(w_i) = 1$; we suppose that the w_i 's have 8 moments and for all i , $w_i > \eta > 0$. Suppose X_i 's are i.i.d. $\mathcal{N}(0, \Sigma)$, Σ is positive definite and the vector v is deterministic with $\|v\|_2 = 1$.*

Suppose $\widehat{\beta}$ is obtained by solving a least-squares problem and $y_i = X_i'\beta + \epsilon_i$, ϵ_i 's being i.i.d. mean 0, with $\text{var}(\epsilon_i) = \sigma_\epsilon^2$.

If $\lim p/n = \kappa < 1$ then the expected variance of the bootstrap estimator, asymptotically as $n \rightarrow \infty$, is given by

$$p \frac{\mathbf{E}\left(\text{var}\left(v'\widehat{\beta}_w^*\right)\right)}{v'\Sigma^{-1}v} = p \frac{\mathbf{E}\left(\text{var}\left(v'\widehat{\beta}_w|\{y_i, X_i\}_{i=1}^n\right)\right)}{v'\Sigma^{-1}v} \rightarrow \sigma_\epsilon^2 \left[\frac{\kappa}{1 - \kappa - f(\kappa)} - \frac{1}{1 - \kappa} \right],$$

where $f(\kappa) = \mathbf{E}\left(\frac{1}{(1+cw_i)^2}\right)$ and c is the unique solution of $\mathbf{E}\left(\frac{1}{1+cw_i}\right) = 1 - \kappa$.

We note that $\mathbf{E}\left(\frac{1}{(1+cw_i)^2}\right) \geq \left[\mathbf{E}\left(\frac{1}{1+cw_i}\right)\right]^2 = (1 - \kappa)^2$ - where the first inequality comes from Jensen's inequality, and therefore the expression we give for the asymptotic limit of the expected bootstrap variance is non-negative. For a proof of this theorem and a consistent estimator of this limit, see AppendixF.

In light of previous work on model robustness issues in high-dimensional statistics (see e.g., (Diaconis and Freedman, 1984; Hall et al., 2005; El Karoui, 2009, 2010)), it is natural to ask whether the central results of Theorem 2 still apply when X_i is not Gaussian but has an elliptical distribution. The formula in Theorem 2 does not apply directly to this latter case. However, the proof given in AppendixF extends to that setting, and we refer the interested reader to the AppendixF.1 where we give the necessary details of how to change the formulas and proof to encompass the elliptical case (we do not provide them in rigorous mathematical detail in this work as they are substantially more cumbersome than those in Theorem 2 and do not give enough additional insights to justify inclusion). On the other hand, a number of the quantities appearing in the proof of Theorem 2 will converge to the same limit as that given in Theorem 2 when i.i.d. Gaussian predictors are replaced by i.i.d. predictors with mean 0 and variance 1 and sufficiently many moments (an example being bounded random variables). Thus the results we present here should be fairly robust to changing i.i.d. normality assumptions for the entries of the design matrix X , but again the technical work necessary for making this rigorous is beyond the scope of this paper.

Implications for Pairs Bootstrap In the standard pairs bootstrap, the weights are chosen according to a Multinomial($n, 1/n$) distribution. This violates two conditions in the previous theorem: independence of w_i 's and the condition $w_i > 0$. In AppendixF.2, we give the technical details for how to extend the proof of Theorem 2 to multinomial weights. In what follows, however, we use i.i.d. Poisson(1) weights, which asymptotically and marginally correspond to the Multinomial($n, 1/n$) weights, to develop intuition about the bootstrap. In this case, we can apply the formula in Theorem 2 to explain why pairs bootstrap confidence intervals perform poorly in high-dimensions, at least for least squares regression with Gaussian design matrix.

When $X_i \stackrel{iid}{\sim} \mathcal{N}(0, \Sigma)$, it is well known in the least-squares case that the quantity $p \text{ var} \left(v' \widehat{\beta} \right) / v' \Sigma^{-1} v$ converges asymptotically to $\kappa / (1 - \kappa) \sigma_\epsilon^2$ (this can be shown through simple Wishart computations Haff, 1979; Mardia et al., 1979). If the variance of $v' \widehat{\beta}_w^*$ converged to the variance of $v' \widehat{\beta}$, we should be able to equate this latter quantity to the limit given in Theorem 2, i.e.,

$$\left[\frac{\kappa}{1 - \kappa - f(\kappa)} - \frac{1}{1 - \kappa} \right] = \frac{\kappa}{1 - \kappa},$$

and hence should have

$$f(\kappa) = \mathbf{E} \left(\frac{1}{(1 + cw_i)^2} \right) = \frac{1 - \kappa}{1 + \kappa}.$$

However, this relationship does not hold for most weight distributions, and in particular does not hold for weights following a Poisson(1) distribution (which asymptotically corresponds to the standard pairs bootstrap, as explained above). Thus the pairs bootstrap does not correctly estimate the variance of $v' \widehat{\beta}$. In Figure 5a we calculate the theoretical predictions of $\mathbf{E} \left(\text{var} \left(\widehat{\beta}_w^* \right) \right)$ given by Theorem 2 (using Poisson(1) weights and $\Sigma = \text{Id}_p$), and we compare them to the asymptotic variance of $\widehat{\beta}_1$ given by $\kappa / (1 - \kappa) \sigma_\epsilon^2 / p$. We see that Theorem 2 predicts that the pairs bootstrap overestimates the variance of the estimator by a factor that ranges from 1.2 to 3 as κ varies between 0.3 and 0.5. These theoretical predictions correspond to the level of overestimation of the variance seen in our bootstrap simulations (Figure 5b).

3.2 Alternative Weight Distributions for Resampling

The formula given in Theorem 2 suggests that resampling from a distribution \widehat{F} defined using weights other than i.i.d. Poisson(1) (or, equivalently for our asymptotics, Multinomial($n, 1/n$)) should give us better bootstrap estimators than using the standard pairs bootstrap. In fact, we should require, at least, that the bootstrap expected variance of these estimators match the correct variance $\text{var} \left(v' \widehat{\beta} \right) = \kappa / (1 - \kappa) \sigma_\epsilon^2 / p$ (for the Gaussian design, when $\Sigma = \text{Id}_p$). We focus our discussion on the case $\Sigma = \text{Id}_p$; see AppendixC for the case $\Sigma \neq \text{Id}_p$.

We note that if we use $w_i = 1, \forall i$, the bootstrap variance will be 0, since with such a resampling scheme the resampled data set is always the original data set. On the other hand, we have seen that with $w_i \sim \text{Poisson}(1)$, the expected bootstrap variance was too large compared to $\kappa / (1 - \kappa) \sigma_\epsilon^2 / p$. Hence, we tried to find alternative weights via calculating a parameter α such that if

$$w_i \stackrel{iid}{\sim} 1 - \alpha + \alpha \text{Poisson}(1), \tag{6}$$

the expected bootstrap variance would match the theoretical value of $\kappa / (1 - \kappa) \sigma_\epsilon^2 / p$.

We numerically solved this problem to find $\alpha(\kappa)$ (for details of computation see AppendixC). We then used these values and performed bootstrap resampling using the weights defined in Equation (6). We evaluated bootstrap estimate of $\text{var} \left(\widehat{\beta}_1 \right)$ as well as the confidence interval coverage of the true β_1 . We find that this adjustment of the weights in

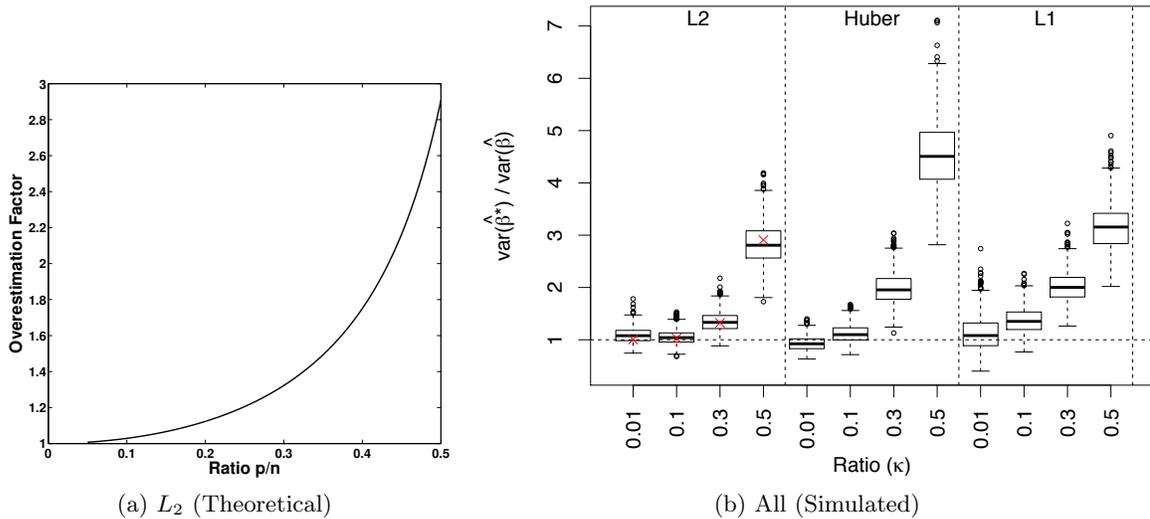


Figure 5: **Factor by which standard pairs bootstrap over-estimates the variance:** (a) plotted is the ratio of the value of the expected bootstrap variance computed from Theorem 2 using Poisson(1) weights to the asymptotic variance $\kappa/(1-\kappa)\sigma_{\epsilon}^2$. (b) boxplots of the ratio of the bootstrap variance of $\hat{\beta}_1^*$ to the variance $\hat{\beta}_1$, as calculated over 1000 simulations (i.e., $\text{var}(\hat{\beta})$ is estimated across simulated design matrices X , and not conditional on X). The theoretical prediction for the mean of the distribution from Theorem 2 is marked with a ‘X’ for L_2 regression. Simulations were performed with normal design matrix X and normal error ϵ_i with values of $n = 500$. For the median values of each boxplot, see Table A-6 in Appendix I.

estimating \hat{F} results in accurate bootstrap estimates of variance and appropriate levels of confidence interval coverage (Table 1).

However, small changes in the choice of α can result in fairly large changes in $\mathbf{E}\left(\text{var}\left(v'\hat{\beta}_w|X,\epsilon\right)\right)$. For instance, for $\kappa = 0.5$, using the value of $\alpha = 0.95$ which is close to the correct value of $\alpha(0.5) = 0.92$ results in an expected bootstrap variance roughly 30% larger than it should be.

Moreover, this strategy for finding a good weight distribution requires knowing a great deal about the distribution of the design matrix. Hence the work we just presented on finding new weight distributions for bootstrapping is a proof of principle that alternative weighting schemes could be used for pairs bootstrapping in high-dimension, but important practical details would depend strongly on the statistical model that is assumed. This is in sharp contrast with the low-dimensional situation, where a unique and model-free bootstrap resampling technique works in a broad variety of situations.

	κ			
	.1	.2	.3	.5
α	.9875	.9688	.9426	.9203
Error Rate of 95% CIs	0.051	0.06	0.061	0.057
Ratio of Variances	1.0119	1.0236	0.9931	0.9992

Table 1: **Summary of weight-adjusted bootstrap simulations for L_2** : Given are the results of performing bootstrap resampling for $n = 500$ according to the estimate of \hat{F} given by the weights in Equation (6). “Error Rate of 95% CIs” denotes the percent of bootstrap confidence intervals that did not contain the correct value of the parameter β_1 . “Ratio of Variances” gives the ratio of the empirical expected bootstrap variance over our simulations divided by the theoretical value $\sigma_\epsilon^2 \kappa / (1 - \kappa)$. Results are based on 1000 simulations, with a Gaussian random design and errors distributed as double exponential.

4. The Jackknife

In the context we are investigating, where we know that the distribution of $\hat{\beta}_1$ is asymptotically normal, it is natural to ask whether we could simply use the jackknife to estimate the variance of $\hat{\beta}_1$. The jackknife relies on leave-one-out procedures to estimate $\text{var}(\hat{\beta}_1)$. More specifically, for a fixed vector v , the jackknife estimate of $\text{var}(v'\hat{\beta})$ is given by:

$$\widehat{\text{var}}_{JACK}(v'\hat{\beta}) = \frac{n-1}{n} \sum_{i=1}^n (v'[\hat{\beta}_{(i)} - \tilde{\beta}])^2 \quad (7)$$

where $\tilde{\beta} = \frac{1}{n} \sum_{i=1}^n \hat{\beta}_{(i)}$. The case of $\hat{\beta}_1$ corresponds to picking $v = e_1$, i.e., the first canonical basis vector. The Efron-Stein inequality guarantees in general that the expectation of the jackknife estimate of variance gives an upper-bound on the variance of the statistic under consideration (Efron and Stein, 1981).

Given the problems we just documented with the pairs bootstrap, it is natural to ask whether confidence intervals based on the jackknife estimate of variance perform better than pairs bootstrap intervals in high-dimensions. The jackknife is known to have problems (Efron, 1982 or Koenker, 2005, p.105), but the reliance of the jackknife on leave-one-out estimates $\hat{\beta}_{(i)}$ might suggest it could be more robust to dimensionality issues than other methods.

Empirical findings As in the pairs bootstrap case, simulations show that confidence intervals based on the jackknife estimate of variance lead to extremely poor inference for β_1 (Figure 1) and that the jackknife dramatically overestimates the variance of $\hat{\beta}_1$ (Figure 6). For L_2 and Huber loss, the jackknife estimate of variance is 10-15% too large for $p/n = 0.1$, and for $p/n = 0.5$ the jackknife estimate of variance is 2-2.5 times larger than it should be. In the case of L_1 loss, the jackknife variance is completely erratic, even in low dimensions; this is not completely surprising given the known problems with the jackknife for the median (Koenker,

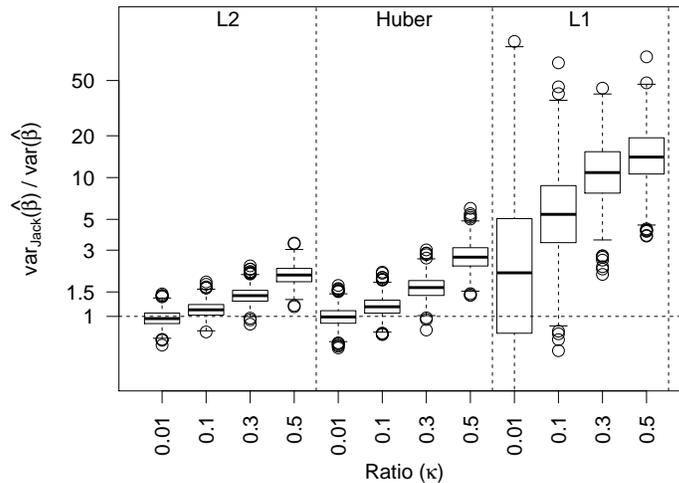


Figure 6: **Factor by which jackknife over-estimates the variance:** boxplots of the ratio of the jackknife estimate of the variance $\widehat{\beta}_1$ to the variance of $\widehat{\beta}_1$ as calculated over 1000 simulations. Simulations were with normal design matrix X and normal error ϵ_i with values of $n = 500$. Note that because the L_1 jackknife estimates so wildly overestimate the variance, in order to put all the methods on the same plot the boxplot of ratio is on log-scale; y-axis labels give the corresponding ratio to which the log values correspond. For the median values of each boxplot, see Table A-6 in AppendixI.

2005). Even for $p/n = 0.01$, the estimate is not unbiased for L_1 , with median estimates twice as large as they should be and enormous variance in the estimates of variance. Higher dimensions only worsen the behavior with jackknife estimates being 15 times larger than they should.

4.1 Theoretical Results

Again, in the case of least-squares regression with a Gaussian design matrix, we can theoretically evaluate the behavior of the jackknife. The proof of the following theorem is given in AppendixG (when the observations have covariance Id) and in AppendixH (to show how to extend the results to general covariance).

Theorem 3 *Let us call var_{JACK} the jackknife estimate of variance of $v'\widehat{\beta}$ given in (7), where v is any deterministic vector with $\|v\|_2 = 1$. Suppose the design matrix X is such that $X_i \stackrel{iid}{\sim} \mathcal{N}(0, \Sigma)$, $\widehat{\beta}$ is computed using least-squares, and the errors ϵ have a variance. Then we have, as $n, p \rightarrow \infty$ and $p/n \rightarrow \kappa < 1$,*

$$\frac{\mathbf{E}(\text{var}_{JACK})}{\text{var}(v'\widehat{\beta})} \rightarrow \frac{1}{1 - \kappa}.$$

As in Theorem 2, the proof of Theorem 3 is based on random matrix techniques where further technical work should allow an extension for the entries of $X_{i,j}$ to be i.i.d. from a distribution other than Gaussian, provided $X_{i,j}$'s have sufficiently many moments. This is also beyond the scope of our work, but interested readers can see AppendixG.3 for more details.

Correcting the Jackknife in Least Squares Theorem 3 implies that scaling the jackknife estimate of variance by multiplying it by $1 - p/n$ will result in an estimate of $\text{var}(\widehat{\beta}_1)$ with the correct expectation; simulations shown in Figure 7 confirm that confidence intervals based on this corrected estimate of variance yield correct confidence intervals for least-squares estimates of $\widehat{\beta}$ when the design matrix X is Gaussian. However this scaling factor is not robust to violations of these assumptions. In particular when the X matrix follows an elliptical distribution the correction of $1 - p/n$ from Theorem 3 gives little improvement even when the loss is still L_2 (Figure 7).

Corrections for more general settings For the more general setting of an elliptical design matrix X and loss function ρ , preliminary computations suggest an alternative result. Let \mathcal{S} be the random matrix defined by

$$\mathcal{S} = \frac{1}{n} \sum_{i=1}^n \psi'(e_i) X_i X_i'.$$

Then in our asymptotic regime, and when $\Sigma = \text{Id}_p$, preliminary heuristic calculations suggest that we can estimate the amount by which $\mathbf{E}(\text{var}_{JACK})$ overestimates the variance of $\widehat{\beta}_1$ by $\mathbf{E}(\widehat{\gamma})$, where

$$\widehat{\gamma} \triangleq \frac{\text{trace}(\mathcal{S}^{-2})/p}{[\text{trace}(\mathcal{S}^{-1})/p]^2}. \quad (8)$$

Note that when applied to least-squares regression with $X \sim \mathcal{N}(0, \text{Id}_p)$ this conforms to our result in Theorem 3. Theoretical considerations suggest that in our asymptotics, for smooth ρ , $\widehat{\gamma} \simeq \mathbf{E}(\widehat{\gamma})$, which suggests a data-driven correction to the jackknife estimate of variance; however that correction depends having information about the distribution of the design matrix.

Equation (8) assumes that the loss function can be twice differentiated, which is not the case for either Huber or L_1 loss. In the case of non-differentiable ρ and ψ , we can use appropriate regularizations to make sense of those functions. For $\rho = \text{Huber}_k$, i.e., a Huber function that transitions from quadratic to linear at $|x| = k$, ψ' should be understood as $\psi'(x) = 1_{|x| \leq k}$. For L_1 loss, ψ' should be understood as $\psi'(x) = 1_{x=0}$.

In Figure 7 we show simulation results for confidence intervals created based on rescaling the jackknife estimate of variance by $\mathbf{E}(\widehat{\gamma})$ defined in Equation (8). In the case of least-squares with an elliptical design matrix, this correction—which directly uses the distribution of the observed X matrix—leads to a definite improvement in our jackknife confidence intervals. Similarly, for the Huber loss we see a definite improvement as compared to the standard jackknife estimate, as well as an improvement over the simpler correction of $1 - p/n$ that would be appropriate for squared error loss.

It should be noted that the quality of this proposed correction seems to depend on how smooth is the function ψ . In particular, even using the previous interpretations, the

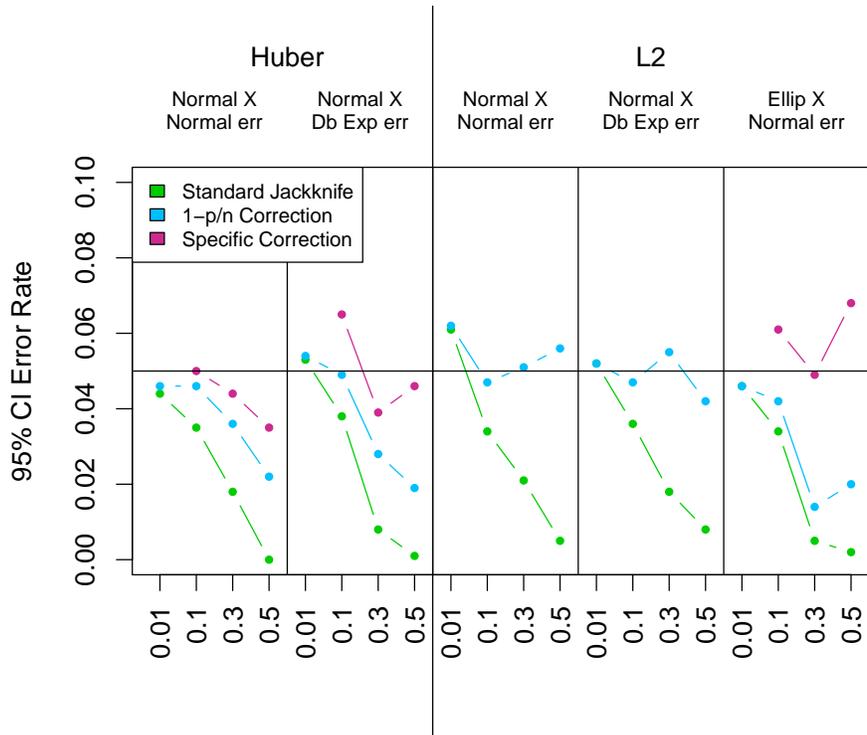


Figure 7: **Rescaling jackknife estimate of variance:** Shown are the error rates for confidence intervals for different re-scalings of the jackknife estimate of variance: the standard jackknife estimate (green); re-scaling using $1 - p/n$ as given in Theorem 3 for the L_2 case with normal design matrix X (blue); and re-scaling based on the heuristic in Equation (8) for those settings not covered by the assumptions of Theorem 3 (magenta). The Huber loss in this plot is Huber_1 rather than the default $\text{Huber}_{1.345}$; Huber_1 is further from L_2 than $\text{Huber}_{1.345}$ and therefore better shows the improvement gained by using the heuristic in Equation (8).

correction does not perform well for L_1 (at least for $n = 1000$ and $\kappa = 0.1, 0.3, 0.5$, data not shown) - though as we mentioned Figure 6 shows that jackknifing in L_1 -regression is probably not a good idea; see also Koenker (2005, Section 3.9).

We also note that the assumption of $\text{cov}(X_i) = \text{Id}_p$ is essential to the jackknife correction proposed in Equation (8). Let $\hat{\beta}_\rho(\Sigma)$ denotes our estimator of β when $\text{cov}(X_i) = \Sigma$. $\hat{\gamma}$ in equation (8) is an estimate of

$$R(\hat{\beta}_\rho) = \frac{\mathbf{E} \left(\text{var}_{JACK}(v' \hat{\beta}_\rho(\text{Id}_p)) \right)}{\text{var} \left(v' \hat{\beta}_\rho(\text{Id}_p) \right)}.$$

Invariance arguments applied to $\text{cov}(X_i)$ can be used to show that when X_i has an elliptical distribution, then $R(\hat{\beta}_\rho(\text{Id}_p)) = R(\hat{\beta}_\rho(\Sigma))$ for all convex loss functions ρ (see AppendixH).

However, even though this population quantity is unchanged when $\text{cov}(X_i) = \Sigma$ changes, the estimate $\hat{\gamma}$ we give above depends crucially on knowing that $\text{cov}(X_i) = \text{Id}_p$ and cannot be used as-is when $\Sigma \neq \text{Id}_p$.

5. Conclusion

In this paper, we have studied various resampling plans in the high-dimensional setting where p/n is not close to zero. One of our main findings is that the two most widely-used and advocated bootstraps will yield either highly conservative or highly anti-conservative confidence intervals. This is in sharp contrast to the low-dimensional setting where p is fixed and $n \rightarrow \infty$ or $p/n \rightarrow 0$. Under various assumptions underlying our simulations, we explained theoretically the phenomena we were observing in our numerical work.

Beside our theoretical contributions, we propose improved and dimension-adaptive bootstrap methods for both pairs and residual bootstraps, as well as jackknife corrections, that give confidence intervals with approximately correct coverage probability. These methods are novel dimensionality-robust resampling techniques for linear models. The resulting modified pairs bootstrap gives a principled method for pairs bootstrapping regardless of the value of $p/n < 1$, and avoids the problem of non-invertible bootstrapped design matrices X^* that commonly result from the standard pairs bootstrap. The most promising of our proposed resampling schemes is our proposed residual bootstrapping method that resamples from appropriately scaled predicted errors. This bootstrap routine performed well without distribution-specific corrections that some of our other methods require. It has the greatest potential to be a general-purpose bootstrap method for linear models in high dimensions.

This work has focused on estimation of the linear model, where there are theoretical benchmarks. The real practical power of the bootstrap lays in giving the ability to perform inference in complicated settings involving sophisticated statistical procedures for which we do not even begin to have theoretical results for the behavior of our estimators. Yet our work shows that even for the simple case of inference in the linear model and for the simplest inferential question, the two most common and natural resampling techniques perform very poorly in only moderately high-dimensions. More importantly, these two equally intuitive methods have completely divergent statistical behavior with one being extremely conservative and the other anti-conservative. This casts serious doubts about the reliability, interpretability and accuracy of inferential statements made through generic resampling methods in moderate and high dimensions. This is troubling for more complicated problems where resampling techniques are the only inference tools currently available. Our findings also suggest that appropriate resampling methods for high-dimensional problems may not be able to rely on generic resampling procedures but rather need to be tailored to the statistical problem of interest. This raises many interesting new theoretical and methodological questions for the future.

Acknowledgments

The authors gratefully acknowledge grants NSF DMS-1026441, NSF DMS-0847647 (CAREER), and NSF DMS-1510172 and support of the ENS-CFM Data Science Chair. They

would also like to thank Peter Bickel and Jorge Banuelos for discussions. N. El Karoui would like to thank Criteo AI Lab for providing a great and very stimulating research environment.

Appendices

Notations : in these appendices, we use e_i to denote the i -th residual, i.e., $e_i = y_i - X_i' \widehat{\beta}$. We use $\tilde{e}_{i(i)}$ to denote the i -th prediction error, i.e., $\tilde{e}_{i(i)} = y_i - X_i' \widehat{\beta}_{(i)}$, where $\widehat{\beta}_{(i)}$ is the estimate of $\widehat{\beta}$ with the i -th pair (y_i, X_i) left out. We assume that the linear model holds so that $y_i = X_i' \beta + \epsilon_i$. We assume that the errors ϵ_i are i.i.d with mean 0.

A	Technical Background on Existing Literature on Robust Regression	30
A.1	Classical Results and Asymptotic Normality	30
A.2	Summary of recent results on high-dimensional robust regression	30
A.3	Consequences for the residual bootstrap	33
B	Deconvolution Bootstrap	34
B.1	Estimating $\ \widehat{\beta}_\rho - \beta\ $ and the Variance of the Z_i	35
B.2	Estimating \widehat{G}	35
B.3	Random Draws from \widehat{G}	36
B.4	Bootstrap Estimates $\widehat{\beta}^*$ from \widehat{G}	37
B.5	Estimation of λ_i^2	37
C	Alternative Weight Distributions for Pairs Bootstrap	38
C.1	Case $\Sigma \neq \text{Id}_p$	38
D	Description of Numerics	39
D.1	Simulation Description	39
D.2	Values of Simulation Parameters	40
D.3	Details of Additional Numerics	40
E	Proof of Theorem 1 (Residual bootstrap, p/n close to 1)	41
F	Proof of Theorem 2 (Expected Variance of the Pairs Bootstrap Estimator)	42
F.1	Extension: Elliptical Design	48
F.2	Extension: Multinomial($n, 1/n$) weights	49
G	Proof of Theorem 3 (Jackknife Variance)	50
G.1	Dealing with Centering	52
G.2	Putting Everything Together	53
G.3	Extension: More Involved Designs and Different Loss Functions	53
H	Going from $\Sigma = \text{Id}_p$ to $\Sigma \neq \text{Id}_p$	54
H.1	Consequences for the Jackknife	54
H.2	Consequences for the Pairs Bootstrap	55
H.3	Rotational Invariance Arguments and Consequences	56
I	Supplementary Tables & Figures	57

Appendix A. Technical Background on Existing Literature on Robust Regression

Recall that we consider

$$\widehat{\beta}_\rho = \operatorname{argmin}_{u \in \mathbb{R}^p} \sum_{i=1}^n \rho(y_i - X_i' u), \text{ where } y_i = \epsilon_i + X_i' \beta.$$

The ϵ_i 's are assumed to be i.i.d with mean 0 here.

A.1 Classical Results and Asymptotic Normality

Least-squares In this case $\rho(x) = x^2/2$ and we have of course

$$\widehat{\beta}_{LS} - \beta = (X'X)^{-1} X' \epsilon.$$

Hence,

$$\operatorname{cov}(\widehat{\beta}_{LS}) = (X'X)^{-1} \operatorname{var}(\epsilon).$$

Robust regression We recall the classic result of Huber (Huber, 1973) and (Huber and Ronchetti, 2009), Chapter 7: when p is fixed and $n \rightarrow \infty$, the limiting covariance of $\widehat{\beta}_\rho$ is, with a slight abuse of notation,

$$\operatorname{cov}(\widehat{\beta}_\rho) = \frac{1}{n} \left(\frac{X'X}{n} \right)^{-1} \frac{\mathbf{E}(\psi^2(\epsilon))}{[\mathbf{E}(\psi'(\epsilon))]^2}.$$

See also the papers Portnoy (1984, 1985, 1987); Mammen (1989) for the situation where $p \rightarrow \infty$ and $p/n \rightarrow 0$ at various rates.

Asymptotic normality questions and impact on confidence intervals: $p/n \rightarrow 0$ In the case of least-squares, the Lindeberg-Feller theorem (Stroock, 1993) guarantees that under mild conditions on the $p \times n$ matrix X , the coordinates of $\widehat{\beta}_{LS}$ are asymptotically Normal. Similarly if the $1 \times n$ vector $v'(X'X)^{-1} X'$ satisfy the conditions of the the Lindeberg-Feller theorem, then $v'(\widehat{\beta}_{LS} - \beta)$ is asymptotically normal. Similarly, under mild conditions on X , the classic papers mentioned above guarantee asymptotic normality of the coordinates of $\widehat{\beta}_\rho$ when $p/n \rightarrow 0$. In these cases, the width of confidence intervals for the coordinates of β are hence only dependent asymptotically on the variance of the coordinates of $\widehat{\beta}_\rho$.

A.2 Summary of recent results on high-dimensional robust regression

We summarize in this section the key results we use from the recent papers El Karoui et al. (2011); El Karoui et al. (2013); El Karoui (2013, 2017). The third paper is a mathematically rigorous version of the heuristic arguments of the first two; the first paper is the long-form version of the second one. Those papers are concerned with the asymptotic properties of $\widehat{\beta}_\rho$ when $p/n \rightarrow \kappa \in (0, 1)$. The predictor vectors X_i 's are assumed to be random and independent, with $X_i = \lambda_i \Sigma^{1/2} \tilde{X}_i$, where \tilde{X}_i has i.i.d (not necessarily Gaussian) entries with mean 0 and variance 1. λ_i 's are independent random variables with $\mathbf{E}(\lambda_i^2) = 1$. (The

$n \times p$ design matrix X is full rank with probability 1. Σ has only positive eigenvalues.) X_i 's are independent of ϵ_i 's.

Role of $\text{cov}(X_i) = \Sigma$ It is shown in these papers that, if $\widehat{\beta}(\beta; \Sigma)$ is the regression vector corresponding to the situation where $y_i = X_i' \beta + \epsilon_i$ and $\text{cov}(X_i) = \Sigma$ for all i ,

$$\widehat{\beta}_\rho(\beta; \Sigma) = \beta + \Sigma^{-1/2} \widehat{\beta}(0; \text{Id}_p) .$$

This follows from a simple change of variable. It also means that to understand the properties of $\widehat{\beta}_\rho(\beta; \Sigma)$, it is enough to understand the “null case” $\beta = 0$ and $\Sigma = \text{Id}_p$.

Consequence for leave-one-out-predicted errors The result we just mentioned has an important consequence for our leave-one-out predicted error, i.e $\tilde{\epsilon}_{i(i)} = y_i - X_i' \widehat{\beta}_{(i)}$: $\tilde{\epsilon}_{i(i)}(\beta; \Sigma) = \tilde{\epsilon}_{i(i)}(0; \text{Id}_p)$. In other words, we can assume without loss of generality that $\beta = 0$ and $\Sigma = \text{Id}_p$ when working with leave-one-out-predicted errors.

A non-asymptotic and exact stochastic representation in the elliptical case When $X_i \stackrel{iid}{\sim} \lambda_i v_i$, where $v_i \sim \mathcal{N}(0, \Sigma)$ and λ_i is a random variable independent of v_i , it is shown that

$$\widehat{\beta}_\rho(\beta; \Sigma) \stackrel{\mathcal{L}}{=} \beta + \|\widehat{\beta}_\rho(0; \text{Id}_p)\|_2 \Sigma^{-1/2} u ,$$

where u is uniformly distributed on the unit sphere in \mathbb{R}^p and $\|\widehat{\beta}_\rho(0; \text{Id}_p)\|_2$ is independent of u . $\|\widehat{\beta}_\rho(0; \text{Id}_p)\|_2$ is simply the norm of $\widehat{\beta}_\rho$ when $\beta = 0$ and $\text{cov}(X_i) = \text{Id}_p$. Note that u has the stochastic representation $u \stackrel{\mathcal{L}}{=} Z_p / \|Z_p\|_2$, where $Z_p \sim \mathcal{N}(0, \text{Id}_p)$.

Consequence of the previous representation for large p Since $\|Z_p\|_2$ has χ_p distribution, it is clear that as $p \rightarrow \infty$, if v is a deterministic vector,

$$\sqrt{p} \frac{v'(\widehat{\beta}_\rho(\beta; \Sigma) - \beta)}{\|\widehat{\beta}_\rho(0; \text{Id}_p)\|_2} \implies \mathcal{N}(0, v' \Sigma^{-1} v) ,$$

where \implies denotes weak convergence of distributions. Hence, provided $\|\widehat{\beta}_\rho(0; \text{Id}_p)\|_2$ and $v' \Sigma^{-1} v$ remain bounded, $v' \widehat{\beta}_\rho(\beta; \Sigma)$ is \sqrt{p} -consistent for $v' \beta$.

Properties of $\|\widehat{\beta}_\rho(0; \text{Id}_p)\|_2$ It is shown, under various technical assumptions, that as p and n tend to infinity with $p/n \rightarrow \kappa$, the variance of the random variable $\|\widehat{\beta}_\rho(0; \text{Id}_p)\|_2$ goes to zero. Hence, for practical matters, $\|\widehat{\beta}_\rho(0; \text{Id}_p)\|_2$ can be considered non-random. In particular, that implies that

$$\sqrt{p} v' (\widehat{\beta}_\rho(\beta; \Sigma) - \beta) \text{ is approximately Normal as } p/n \rightarrow \kappa .$$

Of great importance is the characterization of $\|\widehat{\beta}_\rho(0; \text{Id}_p)\|_2$, since it will affect the width of confidence intervals. It can be characterized, in the case where $\lambda_i = 1$ (see the papers for the case $\lambda_i \neq 1$) in the following way: $\|\widehat{\beta}_\rho(0; \text{Id}_p)\|_2 \rightarrow r_\rho(\kappa)$. The non-random scalar $r_\rho(\kappa)$ can be characterized through a system of two non-linear equations, involving another constant, c . The pair of positive and deterministic scalars $(c, r_\rho(\kappa))$ satisfy: if $\hat{z}_\epsilon = \epsilon + r_\rho(\kappa) Z$, where $Z \sim \mathcal{N}(0, 1)$ is independent of ϵ , and ϵ has the same distribution as ϵ_i 's:

$$\begin{cases} \mathbf{E}((\text{prox}(c\rho))'(\hat{z}_\epsilon)) &= 1 - \kappa , \\ \kappa r_\rho^2(\kappa) &= \mathbf{E}([\hat{z}_\epsilon - \text{prox}(c\rho)(\hat{z}_\epsilon)]^2) . \end{cases} \quad (9)$$

In this system, $\text{prox}(c\rho)$ refers to Moreau's proximal mapping of the convex function $c\rho$ - see Moreau (1965) or Hiriart-Urruty and Lemaréchal (2001). (The system is rigorously shown in El Karoui (2013) under the assumption that the X_i 's have i.i.d entries with mean 0 and variance 1, as well as a few other minor requirements; these assumptions are satisfied when $X_{i,j}$ have a Gaussian distribution, or are bounded, or do not have heavy tails, the latter requiring appeal to various truncation arguments. Another proof of the validity of this system, which first appeared in El Karoui et al. (2011), can be found in Donoho and Montanari (2013). That proof is limited to the case of X_i 's having i.i.d Gaussian entries.) The assumptions on ϵ_i 's and ρ are relatively mild. See El Karoui (2017) for the latest, handling the situation where ϵ_i 's have for instance a Cauchy distribution. We note that some of the results in El Karoui (2013) are stated with ρ strongly convex (and ϵ_i 's having many moments). While the proof in that paper suggests several ways of removing this assumption, it is also possible to change ρ in to $\rho + \eta x^2/2$ with η very small (e.g $\eta = 10^{-100}$) to satisfy this technical assumption and change essentially nothing to the statistical problem at hand.

Consequences for the distribution of $\widehat{\beta}_1$ or other contrasts of interest In our simulation setup, the previous results imply that the distribution of $\widehat{\beta}_1$ (or any other coordinates or contrasts $v'\widehat{\beta}$ for v deterministic) is asymptotically normal. In the case where $\Sigma = \text{Id}_p$, the variance of $\sqrt{p}(\widehat{\beta}_1 - \beta_1)$ is roughly $\mathcal{N}(0, r_\rho^2(\kappa))$. See Bean et al. (2013) and its supplementary material for a longer discussion and questions related to building confidence intervals.

Asymptotic normality questions and impact on confidence intervals: $p/n \rightarrow \kappa \in (0, 1)$ Because we know that, in the Gaussian design case, the coordinates of $\widehat{\beta}_\rho$ are asymptotically normal, the width of these intervals is completely determined by the variance of the coordinates of $\widehat{\beta}_\rho$. We explain above how these variances depend on the distribution of ϵ and the loss function ρ : basically through $\|\widehat{\beta}(\rho; \text{Id})\|_2$ and hence $r_\rho(\kappa)$. Therefore, as was the case in the low-dimensional situation, the variance of the coordinates of $\widehat{\beta}_\rho$ can be used as a proxy for the width of the confidence interval in the high-dimensional case where $p/n \rightarrow \kappa$, $0 < \kappa < 1$.

In (Bean et al., 2013), these asymptotic normality results are used to create confidence intervals for $v'\beta$ in the Gaussian design case: if $z_{1-\alpha/2}$ is the $(1 - \alpha/2)$ -quantile of the Gaussian distribution a $100(1 - \alpha)\%$ confidence interval for $v'\beta$ is

$$v'\widehat{\beta} \pm \frac{z_{1-\alpha/2}}{\sqrt{p}} \widehat{r} \sqrt{(1 - p/n)v'\widehat{\Sigma}^{-1}v},$$

where \widehat{r} is a consistent estimator of $\|\widehat{\beta}_\rho(0; \text{Id}_p)\|_2$. In (Bean et al., 2013), it is said without more precision that leave-one-techniques can be used to come up with \widehat{r} ; we propose in the current paper estimates \widehat{r} based on leave-one-out predicted errors that can therefore be used for the purpose of building those confidence intervals. (See Section 2.3 in the main paper)

Leave-one-out approximations for $\widehat{\beta}$ It is shown in the aforementioned papers that

$$\widehat{\beta} \simeq \widehat{\beta}_{(i)} + \frac{1}{n} S_i^{-1} X_i \psi(e_i),$$

where \simeq means that we are neglecting a quantity that is negligible for all our mathematical and statistical purposes (see the papers for very precise bounds on the quantity we are

neglecting). This approximation is the key to the approximations in Equations (3) and (4) which we use in the main paper. Recall that $S_i = \frac{1}{n} \sum_{j \neq i} \psi'(\tilde{e}_{j(i)}) X_j X_j'$.

A.3 Consequences for the residual bootstrap

We call $\{\epsilon_i^*\}_{i=1}^n$ the estimated errors used in the residual bootstrap. When doing a residual bootstrap, we are effectively sampling from a model with fixed design X , “true β ” taken to be equal to $\hat{\beta}_\rho$ and i.i.d errors sampled according to the empirical distribution of the $\{\epsilon_i^*\}_{i=1}^n$. As a shortcut, we call this distribution ϵ^* in what follows. We call $\hat{\beta}_\rho^*$ the bootstrapped version of $\hat{\beta}$.

Case $p/n \rightarrow 0$ Naturally, the classic results mentioned above imply that the distribution of $v'(\hat{\beta}_\rho^* - \hat{\beta}_\rho)$ is going to be asymptotically normal (under mild conditions on X that are satisfied in our simulations); the variance of the coordinates of $\hat{\beta}_\rho^*$, on the other hand depends on $\frac{\mathbf{E}(\psi^2(\epsilon^*))}{[\mathbf{E}(\psi'(\epsilon^*))]^2}$. Hence, even if the distribution of the estimated errors ϵ^* is very different from that of the “true” errors, ϵ , the residual bootstrap may work very well: indeed, if ϵ and ϵ^* have two very different distribution but

$$\frac{\mathbf{E}(\psi^2(\epsilon^*))}{[\mathbf{E}(\psi'(\epsilon^*))]^2} = \frac{\mathbf{E}(\psi^2(\epsilon))}{[\mathbf{E}(\psi'(\epsilon))]^2},$$

using a residual bootstrap with “the wrong error distribution”, ϵ^* , will give us bootstrap confidence intervals of the right width. An important question then becomes, when p/n is small: what class of distributions ϵ^* is such that $\frac{\mathbf{E}(\psi^2(\epsilon^*))}{[\mathbf{E}(\psi'(\epsilon^*))]^2} = \frac{\mathbf{E}(\psi^2(\epsilon))}{[\mathbf{E}(\psi'(\epsilon))]^2}$, as this class defines all acceptable error distributions from the point of view of our residual bootstrap.

Case $p/n \rightarrow \kappa \in (0, 1)$ We note that at this point in the case $p/n \rightarrow \kappa \in (0, 1)$ we are not aware of central limit theorems for the coordinates of $\hat{\beta}$ that are valid conditional on the design matrix X . However, it is expected that such theorems will hold if the design matrix results from a draw of a random design matrix similar to the ones we consider (with very high-probability with respect to the sampling of the design matrix). The discussions above make then clear that the key quantity to describe the width of the residual bootstrap confidence intervals becomes the risk $\|\hat{\beta}_\rho(0; \text{Id}_p; \epsilon^*)\|_2$, i.e the risk $\|\hat{\beta}_\rho(0; \text{Id}_p)\|_2$ when the error distribution is ϵ^* . A “good” error distribution is therefore one for which $r_\rho(\kappa; \epsilon^*) \simeq r_\rho(\kappa; \epsilon)$. (We used the notation $r_\rho(\kappa; \epsilon) = \lim_{n \rightarrow \infty} \|\hat{\beta}_\rho(0; \text{Id}_p; \epsilon)\|$, when $p/n \rightarrow \kappa$.)

The case of least squares Let us call $\hat{G}_{n,p}$ the distribution of the errors we use in our residual bootstrap. We assume that $\hat{G}_{n,p}$ has mean 0. Let us call $w' = v'(X'X)^{-1}X'$ - where we choose to not index v and w by p for the sake of clarity. v is a deterministic sequence of p -dimensional vectors. Assume that w and $\hat{G}_{n,p}$ satisfy the conditions of the Linderberg-Feller theorem for triangular arrays, and that $\lim_{n \rightarrow \infty} \text{var}(\hat{G}_{n,p}) = \sigma_\epsilon^2$. Then the Lindeberg-Feller theorem guarantees that

$$\frac{v'(\hat{\beta}^* - \hat{\beta})}{\|w\|} \implies \mathcal{N}(0, \sigma_\epsilon^2).$$

Note that it also guarantees, under the same assumptions on w that

$$\frac{v'(\widehat{\beta} - \beta)}{\|w\|} \implies \mathcal{N}(0, \sigma_\epsilon^2).$$

These results do not depend on the size of κ , the limit of the ratio p/n .

Informally, what this means is that provided that the entries of w are all relatively small, that $\widehat{G}_{n,p}$ has mean 0 and $\text{var}(\widehat{G}_{n,p})$ is close to σ_ϵ^2 , then bootstrapping from the residuals in least-squares works for approximating the distribution $v'(\widehat{\beta} - \beta)$.

Conclusion for the purposes of the main paper In our discussions we use $\|\widehat{\beta}_\rho(0; \text{Id}_p; \epsilon^*)\|$ and its closeness to its value under the correct error distribution, $\|\widehat{\beta}_\rho(0; \text{Id}_p; \epsilon)\|$, as a proxy to understand a priori the quality of residual bootstrap confidence intervals when using ϵ^* to sample the errors instead of ϵ . The previous discussion explains why we do so. Our numerical work in Section 2.3 of the main text shows numerically that this yields valuable insights. This is why our discussion in Section 2.4 is focused on understanding $\|\widehat{\beta}(0; \text{Id}_p; \epsilon)\|_2$ for various error distributions. In particular, Theorem 1 shows that when p/n is close to 1, if ϵ^* has approximately the same two first moments as ϵ , $\|\widehat{\beta}(0; \text{Id}_p; \epsilon^*)\|/\|\widehat{\beta}(0; \text{Id}_p; \epsilon)\| \simeq 1$. This explains why the scaled $\tilde{r}_{i(i)}$ is probably a good error distribution ϵ^* to use in the residual bootstrap when κ is close to 0 or 1. We note that when κ is close to 1, $\tilde{r}_{i(i)}$ gives an error distribution that is in general very different from the distribution of ϵ . Our numerical work of Section 2.3 shows that it is nonetheless a good error distribution from the point of view of the residual bootstraps we consider.

Appendix B. Deconvolution Bootstrap

In the main text, we considered situations where our predictors X_i are i.i.d with an elliptical distribution and assume for instance that $X_i = \lambda_i \xi_i$, where $\xi_i \sim \mathcal{N}(0, \Sigma)$ and λ_i are i.i.d scalar random variables with $\mathbf{E}(\lambda_i^2) = 1$. As described in the main text, if X is elliptical, $\tilde{e}_{i(i)}$ is a convolution of the correct G distribution and a Normal distribution,

$$\tilde{e}_{i(i)} \simeq \epsilon_i + \tilde{Z}_i,$$

where

$$\tilde{Z}_i \overset{iid}{\sim} \mathcal{N}(0, \lambda_i^2 \|\widehat{\beta}_{\rho(i)} - \beta\|_2^2)$$

and are independent of ϵ_i .

We proposed in Section 2.3 of the main text an alternative bootstrap method based on using deconvolution techniques to estimate G (Method 1). Specifically, we proposed the following bootstrap procedure:

1. Calculate the predicted errors, $\tilde{e}_{i(i)}$
2. Estimate $|\lambda_i| \|\widehat{\beta}_{\rho(i)} - \beta\|_2$ (the standard deviation of the \tilde{Z}_i)
3. Deconvolve in $\tilde{e}_{i(i)}$ the error term ϵ_i from the \tilde{Z}_i term ;
4. Use the resulting estimates of G as the estimate of \widehat{G} in residual bootstrapping.

B.1 Estimating $\|\widehat{\beta}_\rho - \beta\|$ and the Variance of the Z_i

Deconvolution methods that deconvolve ϵ from the \tilde{Z}_i require an estimate of the variance of the \tilde{Z}_i . Equation (3) gives the variance as $\lambda_i^2 \|\widehat{\beta}_{\rho(i)} - \beta\|_2^2$, and we need to estimate this quantity from the data. We use the approximation

$$\|\widehat{\beta}_{\rho(i)} - \beta\|_2 \simeq \|\widehat{\beta}_\rho - \beta\|_2.$$

See Appendix A and references therein for justification of this approximation.

Furthermore, as we note in the main text, in our implementation of this deconvolution in simulations we assume $X \sim \mathcal{N}(0, Id_p)$ so that $\lambda_i = 1$ (see Section B.5 below for estimating λ_i in the elliptical case). This means we are estimating the variance of \tilde{Z}_i as $\|\widehat{\beta} - \beta\|_2^2$ for all i . We estimate this as

$$\widehat{\text{var}}(\tilde{Z}_i) = \widehat{\text{var}}(\tilde{e}_{i(i)}) - \hat{\sigma}_\epsilon^2,$$

where $\widehat{\text{var}}(\tilde{e}_{i(i)})$ is the standard estimate of variance and $\hat{\sigma}_\epsilon^2$ is the estimate of variance from the least squares fit, $\hat{\sigma}_{\epsilon, LS}^2$, defined in the main text.

In the case where $\widehat{\text{var}}(\tilde{e}_{i(i)}) \leq \hat{\sigma}_\epsilon^2$, we do not do a deconvolution, but simply bootstrap from the $\tilde{e}_{i(i)}$. This is generally only the case when p/n is quite small.

B.2 Estimating \hat{G}

We used the deconvolution algorithm in the `decon` package in R (Wang and Wang, 2011) to estimate the distribution of ϵ_i . Deconvolution algorithms require selection of a bandwidth in the kernels that make up the functional basis of the estimate. The appropriate bandwidth parameter in deconvolution problems is tied intrinsically to the use of the estimate, with optimal bandwidths depending on what functional of the distribution is wanted (e.g. the pdf versus the cdf). Moreover, the optimal bandwidth depends on the distribution of \tilde{Z}_i with which the signal is being convolved. Ultimately, our procedure resamples from the distribution \hat{G} , requiring estimates of $G^{-1}(y)$, and the distribution of \tilde{Z}_i is Gaussian. There is no specific theory for the optimal bandwidth in this setting (though see the work of Hall and Lahiri (2008) for optimal bandwidth selection for estimations of the quantiles of \hat{G} if the \tilde{Z}_i are distributed according to a distribution whose characteristic function decays polynomially at infinity - see Assumption (A.11) on p.2133; this is clearly violated in our case where \tilde{Z}_i are normally distributed.)

We used the bandwidth estimation procedure `bw.dboot2` provided in the package `decon`. Delaigle (2014) outlines problems in the estimation of bandwidth parameter in `decon`; specifically that the implementation in `decon` of existing bandwidth estimation procedures does not match their published descriptions. `bw.dboot2` was not one of the bandwidth procedures with these discrepancies. However, we also compared our results with a bandwidth selected via the bandwidth selection method of Delaigle and Gijbels (2002, 2004) and used the R code implementation provided by the authors on <http://www.ms.unimelb.edu.au/~aurored/links.html#Code>. The two different choices in bandwidth, however, had little effect on the coverage of the bootstrap confidence intervals (Supplemental Figure 8). The results in Figure 2 in the main text make use of the bandwidth parameter of Delaigle and Gijbels (2002, 2004).

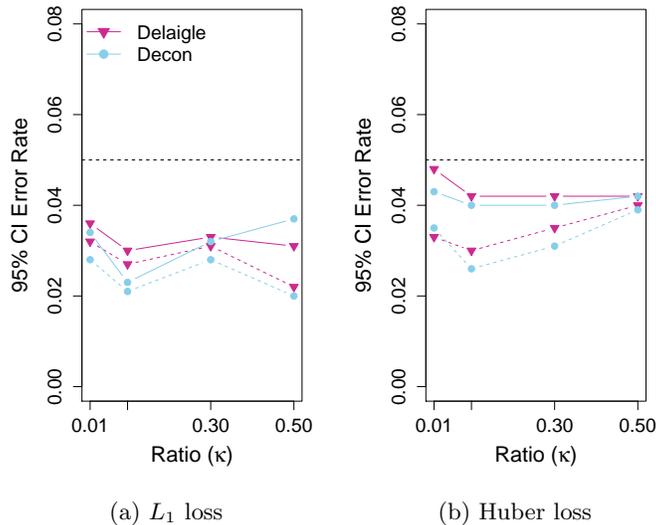


Figure 8: **Different bandwidths for Method 1:** We plotted the error rate of 95% confidence intervals for the deconvolution bootstrap (Method 1) using two different choices of bandwidth: the `bw.dboot2` in `decon` (light blue) or that of Delaigle and Gijbels (2002, 2004) (maroon). The solid lines refer to bootstrapping by drawing $\{\epsilon_i^*\}_{i=1}^n$ as i.i.d draws from \hat{G} ; the dashed lines refer to $\{\epsilon_i^*\}_{i=1}^n$ drawn from repeated resampling of a single draw ($\{\hat{\epsilon}_i\}_{i=1}^n$) from \hat{G} . See section B.4 below. Note that the y-axis for these plots is different than that shown in the main text.

For both bandwidth selections, we estimated the cdf using the function `DeconCdf` provided in the `decon` package and provided the bandwidth parameters described above. We specified the error distribution as ‘Normal’ and set the variance of \tilde{Z}_i as described above in Section B.1. The number of grid points for evaluating the cdf (the ‘ngrid’ argument) was set to be the number needed to get a space of 0.01 across the range of observed $\tilde{e}_{i(i)}$, with a lower bound of 512 grid points (the default of ‘ngrid’ given by the `DeconCdf` function). Other options were set to the default of `DeconCdf`.

B.3 Random Draws from \hat{G}

The end result of the `DeconCdf` function was values of the \hat{G} evaluated at specific grid points x . The resulting $\hat{G}(x)$ was not always guaranteed to be ≤ 1 nor monotonically decreasing; this is likely due to the fact that use of higher-order kernels estimates (which is standard practice in deconvolution literature) does not constrain the estimate be a proper density. Furthermore, the tail ends of the cdf are based on little data and unlikely to be reliable, as well as having problems either non-monotonicity or extending beyond the boundaries of $(0, 1)$. We truncated the left tail of $\hat{G}(x)$ to be within 0.001 by finding the largest such x_0 such that $\hat{G}(x_0) \leq 0.001$ and setting $\hat{G}(x) = 0.001$ for $x \leq x_0$; and we similarly trimmed the right tail based on $1 - 0.001$. We then calculated the differences $d_i = \hat{G}(x_i) - \hat{G}(x_{i-1})$

and for $d_i < 0$ set $d_i = 0$. We then defined a monotone cdf based on the cumulative sum of the d_i ,

$$C(x_j) = \sum_{i=1}^j d_i.$$

We then renormalized the values $C(x_j)$ so that they extend from 0 to 1, giving the final monotone estimate of $\hat{G}(x_j)$ as

$$\hat{G}(x_j) = \frac{C(x_j) - \min_i C(x_i)}{\max_i C(x_i) - \min_i C(x_i)}$$

To randomly sample from \hat{G} , we needed to be able to evaluate \hat{G} for all x . We did this by linearly interpolating between the $\hat{G}(x_j)$ values. In what follows, we consider the values $\hat{G}(x)$ based on this smoothed and monotone version of the original output of the `DeconCdf` function.

We create random draws from \hat{G} by drawing random variables U_i from a $Unif(0, 1)$ and calculating $E_i = \hat{G}(U_i)$. We further centered and standardized the draws E_j from \hat{G} to get

$$\epsilon_j^* = (E_j - \text{mean}_i(E_j)) \frac{\hat{\sigma}_{\epsilon, LS}}{\sqrt{\text{var}(E_i)}}$$

so that the resulting ϵ_j^* have mean zero and variance $\hat{\sigma}_{\epsilon, LS}^2$. This was done because the variance of \hat{G} was not guaranteed to have the correct variance, despite the fact we had to pre-specify the variance in the deconvolution call. Ensuring the correct moments of ϵ_j^* was a critical component for reasonable coverage of the bootstrap confidence intervals. When we did not standardize the results and just took the draws from E_j , the resulting bootstrap confidence intervals became more and more conservative as p/n grew. This again highlights the results of Theorem 1 – the variance of \hat{G} is the most important feature of the distribution in order to have accurate confidence intervals.

B.4 Bootstrap Estimates $\hat{\beta}^*$ from \hat{G}

We used \hat{G} to create bootstrap errors, $\{\epsilon_i^*\}_{i=1}^n$ in two ways. For the first method we estimated $\{\epsilon_i^*\}_{i=1}^n$ as i.i.d draws from \hat{G} , and repeatedly drew such samples from \hat{G} , B times. In the second method, we drew one single estimate $\{\hat{\epsilon}_i\}_{i=1}^n$ as i.i.d draws from \hat{G} and then created $\{\epsilon_i^*\}_{i=1}^n$ from resampling from the empirical distribution of the $\{\hat{\epsilon}_i\}_{i=1}^n$, and repeated this resampling from the empirical distribution of $\{\hat{\epsilon}_i\}_{i=1}^n$ B times. For both methods, we then calculated $\hat{\beta}^*$ from the data (X_i, y_i^*) where $y_i^* = X_i' \hat{\beta} + \epsilon_i^*$, as in the standard residual bootstrap. The first method seems to do slightly better in simulations, see Figure 8.

B.5 Estimation of λ_i^2

To extend the deconvolution bootstrapping method to the elliptical case when $p/n \rightarrow \kappa \in (0, 1)$, one needs to be able to estimate λ_i , at least up to sign. In which case, one could estimate individually the variance of \tilde{Z}_i and feed these individual estimates into the deconvolution method described above.

We recall a simple proposal from the paper (El Karoui, 2010) to solve this problem. Specifically, the author proposes to use

$$\widehat{\lambda}_i^2 = \frac{\|X_i^2\|_2/p}{\frac{1}{p}\text{trace}(\widehat{\Sigma})} = \frac{\|X_i\|_2^2}{\frac{1}{n}\sum_{i=1}^n\|X_i\|_2^2},$$

where $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n X_i X_i'$. Under mild conditions on Σ and λ_i , it can be shown that when $n \rightarrow \infty$ and $p/n \rightarrow \kappa \in (0, \infty)$

$$\sup_{1 \leq i \leq n} |\lambda_i^2 - \widehat{\lambda}_i^2| \rightarrow 0 \text{ in probability.}$$

The intuition and proof are as follows. Concentration of measure arguments (Ledoux, 2001) show that if $\xi_i \sim \mathcal{N}(0, \Sigma)$, $\|\xi_i\|^2/p \simeq \text{trace}(\Sigma)/p$ and hence $\|X_i\|^2/p \simeq \lambda_i^2 \text{trace}(\Sigma)/p$. The law of large numbers and a little bit of further technical work then imply that $\frac{1}{n}\sum_{i=1}^n\|X_i\|^2/p \simeq \mathbf{E}(\lambda_i^2) \text{trace}(\Sigma)/p = \text{trace}(\Sigma)/p$.

Appendix C. Alternative Weight Distributions for Pairs Bootstrap

The formula given in Theorem 2 suggests resampling from a distribution \widehat{F} defined using weights other than i.i.d. Poisson(1). An acceptable weight distribution is such that the variance of the resampled estimator is equal to the variance of the sampling distribution of the original estimator. In Section 3.2 we consider the least-squares estimator where the variance, in the case where $\Sigma = \text{Id}_p$ is asymptotically $\kappa/(1-\kappa)\sigma_\epsilon^2/p$.

In the main text, we proposed to use

$$w_i \stackrel{iid}{\sim} 1 - \alpha + \alpha \text{Poisson}(1)$$

We determined α numerically so that

$$\left[\kappa \frac{1}{1 - \kappa - \mathbf{E}w_i \left[\frac{1}{(1+cw_i)^2} \right]} - \frac{1}{1 - \kappa} \right] = \frac{\kappa}{1 - \kappa}. \quad (10)$$

This was done via a simple dichotomous search for α over the interval $[0, 1]$. Our initial α was .95. We specified a tolerance of 10^{-2} for the results reported in the paper in Table 1. This means that we stopped the algorithm when the ratio of the two terms in Equation (10) was within 1% of 1. We used a sample size of 10^6 to estimate all the expectations. Table 2 gives the values of $\alpha(\kappa)$ found.

C.1 Case $\Sigma \neq \text{Id}_p$

In the case where $\Sigma \neq \text{Id}$, both $\mathbf{E}(\text{var}(v'\widehat{\beta}^*))$ and $\text{var}(v'\widehat{\beta})$ depend on $v'\Sigma^{-1}v$. It is therefore natural to ask how we could estimate this quantity. If we are able to do so, it is clear that we could follow the same strategy as above to find α from the data. Standard Wishart results (Mardia et al. (1979), Theorem 3.4.7) give that

$$\frac{v'\Sigma^{-1}v}{v'\widehat{\Sigma}^{-1}v} \sim \frac{\chi_{n-p}^2}{n-1} \rightarrow (1-\kappa) \text{ in probability.}$$

κ	0.05	0.10	0.15	0.20	0.25
$\alpha(\kappa)$	0.9938	0.9875	0.9812	0.9688	0.9562
κ	0.30	0.35	0.40	0.45	0.50
$\alpha(\kappa)$	0.9426	0.9352	0.9277	0.9222	0.9203

Table 2: Values of $\alpha(\kappa)$ to use to fix the variance estimation issue in high-dimensional pairs-bootstrap

This of course suggests using $(1 - p/n)v'\widehat{\Sigma}^{-1}v$ as an estimator of $v'\Sigma^{-1}v$ and solves the question we were discussing above.

However, we note that since

$$\frac{\mathbf{E}\left(\text{var}\left(v'\widehat{\beta}^*\right)\right)}{\text{var}\left(v'\widehat{\beta}\right)}$$

does not depend on Σ when the design is Gaussian or Elliptical, the same α should work regardless of Σ , provided it is positive definite. In particular, an acceptable weight distribution for resampling as defined above could be computed by assuming $\Sigma = \text{Id}_p$ and would work for any positive definite Σ .

Appendix D. Description of Numerics

Here we describe the implementation of various computational numerics used in the paper.

D.1 Simulation Description

In the simulations described in the paper, we explored variations in the distribution of the design matrix X , the error distribution, the loss function, the sample size (n), and the ratio of $\kappa = p/n$, detailed below.

All results in the paper were based upon 1,000 replications of our simulation routine for each combination of these values. Each simulation consisted of

1. Simulation of data matrix X , $\{\epsilon_i\}_{i=1}^n$ and construction of data $y_i = X'\beta + \epsilon_i$. However, for our simulations, $\beta = 0$ (without loss of generality for the results, which are shift equivariant), so $y_i = \epsilon_i$.
2. Estimate $\widehat{\beta}$ using the corresponding loss function. For L_2 this was via the `lm` command in R, for Huber via the `rlm` command in the `MASS` package with default settings ($k = 1.345$) (Venables and Ripley, 2002), and for L_1 via an internal program making use of MOSEK optimization package and accessed in R using the `Rmosek` package (MOSEK, 2014). The internal L_1 program was checked to give the same results as the `rq` function that is part of the R package `quantreg` (Koenker, 2013), but was much faster for simulations.
3. Bootstrapping according to the relevant bootstrap procedure (using the `boot` package) and estimating $\widehat{\beta}^*$ for each bootstrap sample. Each bootstrap resampling consisted of

$R = 1,000$ bootstrap samples, the minimum generally suggested for 95% confidence intervals (Davison and Hinkley, 1997). For jackknife resampling and for calculating leave-one-out prediction errors $\tilde{e}_{i(i)}$, we wrote an internal function that left out each observation in turn and recalculated $\hat{\beta}_{(i)}$.

4. Construction of confidence intervals for $\hat{\beta}_1$. For bootstrap resampling, we used the function `boot.ci` in the `boot` package to calculate confidence intervals. We calculated “basic”, “percentile”, “normal”, and “BCA” confidence intervals (see help of `boot.ci` and Davison and Hinkley (1997) for details about each of these), but all results shown in the manuscript rely on only the percentile method. The percentile method calculates the boundaries of the confidence intervals as the estimates of 2.5% and 97.5% percentiles of $\hat{\beta}_1^*$ (note that the estimate is not exactly the *observed* 2.5% and 97.5% of $\hat{\beta}_1^*$, since there is a correction term for estimating the percentile, again see Davison and Hinkley (1997)). For the jackknife confidence intervals, the confidence interval calculated was a standard normal confidence interval $(\pm 1.96\sqrt{\widehat{var}_{Jack}(\hat{\beta}_1)})$

D.2 Values of Simulation Parameters

Design Matrix For the design matrix X , we considered the following designs for the distribution of an element X_{ij} of the matrix X , ensuring that the vectors X_i had covariance Id_p in all cases :

- Normal: X_{ij} are i.i.d $N(0, 1)$
- Double Exp: X_{ij} are i.i.d. double exponential with variance 1.
- Elliptical: $X_{ij} \sim \lambda_i Z_{ij}$ where the Z_{ij} are i.i.d $N(0, 1)$ and the λ_i are i.i.d according to
 - $\lambda_i \sim \text{Exp}(\sqrt{2})$ (i.e. mean $1/\sqrt{2}$), so $\mathbf{E}(\lambda_i^2) = 1$
 - $\lambda_i \sim N(0, 1)$
 - $\lambda_i \sim \text{Unif}(\sqrt{\frac{12}{13}}0.5, \sqrt{\frac{12}{13}}1.5)$ so that $\mathbf{E}(\lambda_i^2) = 1$

Error Distribution We used two different distributions for the i.i.d errors ϵ_i : $N(0, 1)$ and standard double exponential (with variance 2).

Dimensions We simulated from $n = 100, 500, \text{ and } 1,000$ though we showed only $n = 500$ in our results for simplicity. Except where noted, no significant difference in the results was seen for varying sample size. The ratio κ was simulated at 0.01, 0.1, 0.3, 0.5.

D.3 Details of Additional Numerics

In Tables A-1 to A-5 in AppendixI we give the precise numerical results from our simulations that are plotted in both the main text and supplementary figures.

Calculating Correction Factors for Jackknife We computed these quantities using the formula we mentioned in the text and Matlab. We solve the associated regression problems with `cvx` (Grant and Boyd, 2014, 2008), running `Mosek` (ApS, 2015) as our optimization

engine. We used $n = 500$ and 1,000 simulations to compute the mean of the quantities we were interested in.

Relative Risk (Figure 3) In Figure 3 we plot for both Huber loss and L_1 loss the average risk $r_\rho(\kappa; G_{conv})$ (i.e. errors given by G_{conv}) relative to the average risk $r_\rho(\kappa; G)$ (i.e. errors distributed according to G), where G has a double exponential distribution. We also plot the relative average risk $r_\rho(\kappa; G_{norm})$, where $G_{norm} = N(0, \sigma_\epsilon^2)$. The values for Figure 3 were generated with Matlab, using cvx and Mosek, as described above. We picked $n = 500$ and did 500 simulations. p was taken in (5, 10, 30, 50, 75, 100, 125, 150, 175, 200, 225, 250, 275, 300, 350, 400, 450). We used our simulations for the case of the original errors to estimate $\mathbf{E}(\|\hat{\beta} - \beta\|_2)$. We used this estimate in our simulation under the convolved error distribution. The Gaussian error simulations were made with $\mathcal{N}(0, 2)$ to match the variance of the double exponential distribution.

Calculation of amount of variance overestimated in pairs bootstrap (Figure 5a)

In Figure 5a, we plot the theoretical factor by which the pairs bootstrap overestimates the actual variance of $\beta_{1\rho}$. This figure was generated by assuming Poisson(1) weights and computing deterministically the expectations of interest. This was easy since if $W \sim \text{Poisson}(1)$, $P(W = k) = \frac{\exp(-1)}{k!}$.

We truncated the expansion of the expectation at $K = 100$, so we neglected terms of order $1/100!$ or lower only. The constant c was found by dichotomous search, with tolerance 10^{-6} for matching the equation $\mathbf{E}(1/(1 + Wc)) = 1 - p/n$. Once c was found, we approximated the expectation in Theorem 2 in the same fashion as we just described.

Once we had computed the quantity appearing in Theorem 2, we divided it by $\kappa/(1 - \kappa)$. We repeated these computations for $\kappa = .05$ to $\kappa = .5$ by increments of 10^{-3} to produce our figure.

Appendix E. Proof of Theorem 1 (Residual bootstrap, p/n close to 1)

Proof

Recall the system describing the asymptotic limit of $\|\hat{\beta}_\rho - \beta\|$ when $p/n \rightarrow \kappa$ and the design matrix has i.i.d mean 0, variance 1 entries, is, under some conditions on ϵ_i 's and some mild further conditions on the design (see Section A above): $\|\hat{\beta}_\rho - \beta\| \rightarrow r_\rho(\kappa)$ and the pair of positive and deterministic scalars $(c, r_\rho(\kappa))$ satisfy: if $\hat{z}_\epsilon = \epsilon + r_\rho(\kappa)Z$, where $Z \sim \mathcal{N}(0, 1)$ is independent of ϵ , and ϵ has the same distribution as ϵ_i 's:

$$\begin{cases} \mathbf{E}((\text{prox}(c\rho))'(\hat{z}_\epsilon)) &= 1 - \kappa, \\ \kappa r_\rho^2(\kappa) &= \mathbf{E}([\hat{z}_\epsilon - \text{prox}(c\rho)(\hat{z}_\epsilon)]^2). \end{cases}$$

In this system, $\text{prox}(c\rho)$ refers to Moreau's proximal mapping of the convex function $c\rho$ - see Moreau (1965) or Hiriart-Urruty and Lemaréchal (2001).

We first give an informal argument to “guess” the correct values of various quantities of interest, namely c and of course, $r_\rho(\kappa)$.

Note that when $|x| \ll c$, and when $\psi(x) \sim x$ at 0, $\text{prox}(c\rho)(x) \simeq \frac{x}{1+c}$. Hence, $x - \text{prox}(c\rho)(x) \simeq xc/(1+c)$. (Note that as long as $\psi(x)$ is linear near 0, we can assume that $\psi(x) \sim x$, since the scaling of ρ by a constant does not affect the performance of the estimators.)

We see that $1 - \kappa \simeq 1/(1 + c)$, so that $c \simeq \kappa/(1 - \kappa)$ - assuming for a moment that we can apply the previous approximations in the system. Hence, we have

$$\kappa r_\rho(\kappa)^2 \simeq (c/(1 + c))^2 [r_\rho(\kappa)^2 + \sigma_\epsilon^2] \simeq \kappa^2 [r_\rho(\kappa)^2 + \sigma_\epsilon^2] .$$

We can therefore conclude (informally at this point) that

$$r_\rho(\kappa)^2 \sim \frac{\sigma_\epsilon^2 \kappa}{1 - \kappa} \sim \frac{\sigma_\epsilon^2}{1 - \kappa} .$$

Once these values are guessed, it is easy to verify that $r_\rho(\kappa) \ll c$ and hence all the manipulations above are valid if we plug these two expressions in the system driving the performance of robust regression estimators described above. We note that our argument is not circular: we just described a way to guess the correct result. Once this has been done, we have to make a verification argument to show that our guess was correct.

In this particular case, the verification is done as follows: we can rewrite the expectations as integrals and split the domain of integration into $(-\infty, -s_\kappa)$, $(-s_\kappa, s_\kappa)$, (s_κ, ∞) , with $s_\kappa = (1 - \kappa)^{-3/4}$. Using our candidate values for c and $r_\rho(\kappa)$, we see that the corresponding \widehat{z}_ϵ has extremely low probability of falling outside the interval $(-s_\kappa, s_\kappa)$ - recall that $1 - \kappa \rightarrow 0$. Coarse bounding of the integrands outside this interval shows the corresponding contributions to the expectations are negligible at the scales we consider. On the interval $(-s_\kappa, s_\kappa)$, we can on the other hand make the approximations for $\text{prox}(c\rho)(x)$ we discussed above and integrate them. That gives us the verification argument we need, after somewhat tedious but simple technical arguments. (Note that the method of propagation of errors in analysis described in (Miller, 2006) works essentially in a similar a-posteriori-verification fashion. Also, s_κ could be picked as $(1 - \kappa)^{-(1/2+\delta)}$ for any $\delta \in (0, 1/2)$ and the arguments would still go through.) ■

Appendix F. Proof of Theorem 2 (Expected Variance of the Pairs Bootstrap Estimator)

In this section, we compute the expected variance of the bootstrap estimator.

We recall that for random variables T, Γ , we have

$$\text{var}(T) = \text{var}(\mathbf{E}(T|\Gamma)) + \mathbf{E}(\text{var}(T|\Gamma)) .$$

In our case, $T = v' \widehat{\beta}_w$, the projection of the regression estimator $\widehat{\beta}_w$ obtained using the random weights w on the contrast vector v . Γ represents both the design matrix and the errors. We assume without loss of generality that $\|v\|_2 = 1$.

Hence,

$$\text{var}(v' \widehat{\beta}_w) = \text{var}(v' \mathbf{E}(\widehat{\beta}_w|\Gamma)) + \mathbf{E}(\text{var}(v' \widehat{\beta}_w|\Gamma)) .$$

In plain English, the variance of $v' \widehat{\beta}_w$ is equal to the variance of the bagged estimator plus the expectation of the variance of the bootstrap estimator (where we randomly weight observation (y_i, X_i) with weight w_i).

As explained in Section H, we can study without loss of generality the case where $\Sigma = \text{Id}_p$ and $\beta = 0$. This is what we do in this proof. Further the rotational invariance arguments we give in Section H mean that we can focus on the case $v = e_p$, the p -th canonical basis vector, without loss of generality.

We consider the case where $X_i \stackrel{iid}{\sim} \mathcal{N}(0, \text{Id}_p)$. This allows us to work with results in El Karoui et al. (2011); El Karoui et al. (2013), El Karoui (2013).

Notational simplification To make the notation lighter, in what follows in this proof we use the notation $\widehat{\beta}$ for $\widehat{\beta}_w$. There are no ambiguities as we are always using a weighted version of the estimator and hence this simplification should not create any confusion.

In particular, we have, using the derivation of Equation (9) in El Karoui et al. (2013) and noting that in the least-squares case all approximations in that paper are actually exact equalities,

$$\widehat{\beta}_p = \widehat{c} \frac{\sum_{i=1}^n w_i X_i(p) e_{i,[p]}}{p} .$$

$e_{i,[p]}$ here are the residuals based on the first $p - 1$ predictors, when $\beta = 0$. We note that, under our assumptions on X_i 's and w_i 's, $\widehat{c} = \frac{1}{n} \text{trace}(S_w^{-1}) + o_{L_2}(1)$, where $S_w = \frac{1}{n} \sum_{i=1}^n w_i X_i X_i'$. It is known from work in random matrix theory (see e.g El Karoui (2009)) that $\frac{1}{n} \text{trace}(S_w^{-1})$ is asymptotically deterministic in the situation under investigation with our assumptions on w and X , i.e $\frac{1}{n} \text{trace}(S_w^{-1}) = c + o_{L_2}(1)$, where $c = \mathbf{E}(\frac{1}{n} \text{trace}(S_w^{-1}))$.

We also recall the residuals representation from El Karoui et al. (2013), which are exact in the case of least-squares : namely here,

$$\widehat{\beta} - \widehat{\beta}_{(i)} = \frac{w_i}{n} S_i^{-1} X_i \psi(e_i) ,$$

which implies that, with $S_i = \frac{1}{n} \sum_{j \neq i} w_j X_j X_j'$,

$$\tilde{e}_{i(i)} = e_i + w_i \frac{X_i' S_i^{-1} X_i}{n} \psi(e_i) .$$

In the case of least-squares, $\psi(x) = x$, so that

$$e_i = \frac{\tilde{e}_{i(i)}}{1 + w_i c_i} ,$$

where

$$c_i = \frac{X_i' S_i^{-1} X_i}{n} .$$

These equalities also follow from simple linear algebra since we are in the least-squares case. We note that $c_i = c + o_P(1)$, where c is deterministic, as explained in e.g El Karoui (2010), El Karoui (2013). Furthermore, here the approximation holds in L_2 because of our assumptions on w 's and existence of moments for the inverse Wishart distribution - see e.g Haff (1979). As explained in El Karoui (2013), the same is true for $c_{i,[p]}$ which is the same quantity computed using the first $(p - 1)$ coordinates of X_i , vectors we denote generically by V_i . We can rewrite

$$\widehat{\beta}_p = \widehat{c} \frac{\sum_{i=1}^n w_i X_i(p) \frac{\tilde{e}_{i(i),[p]}}{1 + w_i c_{i,[p]}}}{p} .$$

Let us call \widehat{b} the bagged estimate. We note that $\tilde{e}_{i(i),[p]}$ is independent of w_i and so is $c_{i,[p]}$. We have already seen that \widehat{c} is close to a constant, c . So taking expectation with respect to the weights, we have, if $w_{(i)}$ denotes $\{w_j\}_{j \neq i}$, and using independence of the weights,

$$\widehat{b}_p = \frac{1}{p} \sum_{i=1}^n \mathbf{E}_{w_i} \left(\frac{cw_i}{1+cw_i} \right) X_i(p) \mathbf{E}_{w_{(i)}} (\tilde{e}_{i(i),[p]}) [1 + o_{L_2}(1)].$$

Now the last term is of course the prediction error for the bagged problem, i.e

$$\mathbf{E}_{w_{(i)}} (\tilde{e}_{i(i),[p]}) = \epsilon_i - V_i'(\widehat{g}_{(i)} - \gamma)$$

where $\widehat{g}_{(i)}$ is the bagged estimate of $\widehat{\gamma}$ and $\widehat{\gamma}$ is the regression vector obtained by regressing y_i on the first $p-1$ coordinates of X_i . (Recall that in these theoretical considerations we are assuming that $\beta = 0$, without loss of generality.)

So we have, since we can work in the null case where $\gamma = 0$ (without loss of generality),

$$\widehat{b}_p = \frac{1}{p} \sum_{i=1}^n \mathbf{E}_{w_i} \left(\frac{cw_i}{1+cw_i} \right) X_i(p) [\epsilon_i - V_i' \widehat{g}_{(i)}] (1 + o_{L_2}(1)).$$

Hence,

$$\mathbf{E} (p \widehat{b}_p^2) = \frac{1}{p} \sum_{i=1}^n \left[\mathbf{E}_{w_i} \left(\frac{cw_i}{1+cw_i} \right) \right]^2 (\sigma_\epsilon^2 + \mathbf{E} (\|\widehat{g}_{(i)}\|_2^2)) (1 + o(1)).$$

Now, in expectation, using e.g El Karoui (2013), $\mathbf{E} (\|\widehat{g}_{(i)}\|_2^2) (1 + o(1)) = \mathbf{E} (\|\widehat{b}\|_2^2) = p \mathbf{E} (\widehat{b}_p^2)$. The last equality comes from the fact that all coordinates play a symmetric role in this problem, so they are all equal in law.

Now, recall that according to e.g El Karoui et al. (2013), top-right equation on p. 14562, or El Karoui (2010)

$$\frac{1}{n} \sum_{i=1}^n \frac{1}{1+cw_i} = 1 - \frac{p}{n} + o_{L_2}(1),$$

since the previous expression effectively relates trace $(D_w X (X' D_w X)^{-1} X')$ to $n-p$, the rank of the corresponding ‘‘hat matrix’’.

Since $\frac{cw_i}{1+cw_i} = 1 - \frac{1}{1+cw_i}$, we see that

$$\mathbf{E}_{w_i} \left(\frac{cw_i}{1+cw_i} \right) = \frac{p}{n} + o(1).$$

Hence, for the bagged estimate, we have the equation

$$\mathbf{E} (\|\widehat{b}\|_2^2) = \frac{p}{n} (\sigma^2 + \mathbf{E} (\|\widehat{b}\|_2^2)) (1 + o(1)).$$

We conclude that

$$\mathbf{E} (\|\widehat{b}\|_2^2) = (1 + o(1)) \frac{\kappa}{1-\kappa} \sigma^2.$$

Note that $\frac{\kappa}{1-\kappa} \sigma^2 = \mathbf{E} (\|\widehat{\beta}_{sLS}\|_2^2)$, where the latter is the standard (i.e non-weighted) least squares estimator.

We note that the rotational invariance argument given in El Karoui et al. (2011); El Karoui et al. (2013) still apply here, so that we have the

$$\widehat{b} - \beta \stackrel{\mathcal{L}}{=} \|\widehat{b} - \beta\|u ,$$

where u is uniform on the sphere and independent of $\|\widehat{b} - \beta\|$ (recall that this simply comes from the fact that if X_i is changed into OX_i , where O is orthogonal, \widehat{b} is changed into $O\widehat{b}$ - and we then apply invariance arguments coming from rotational invariance of the distribution of X_i). Therefore,

$$\text{var} \left(v'(\widehat{b} - \beta) \right) = \frac{\|v\|_2^2}{p} \mathbf{E} \left(\|\widehat{b} - \beta\|_2^2 \right) .$$

So we conclude that

$$p \mathbf{E} \left(\text{var} \left(v' \widehat{\beta}_w | \Gamma \right) \right) = p \text{var} \left(v' \widehat{\beta}_w \right) - \frac{\kappa}{1 - \kappa} \sigma^2 \|v\|_2^2 + o(1) .$$

Now, the quantity $\text{var} \left(v' \widehat{\beta}_w \right)$ is well understood. The rotational invariance arguments we mentioned before give that

$$\text{var} \left(v' \widehat{\beta}_w \right) = \frac{\|v\|_2^2}{p} \mathbf{E} \left(\|\widehat{\beta}_w - \beta\|_2^2 \right) .$$

In fact, using the notation D_w for the diagonal matrix with $D_w(i, i) = w_i$, since

$$\widehat{\beta}_w - \beta = (X' D_w X)^{-1} X' D_w \epsilon ,$$

we see that

$$\mathbf{E} \left(\|\widehat{\beta}_w - \beta\|_2^2 \right) = \sigma_\epsilon^2 \mathbf{E} \left(\text{trace} \left((X' D_w X)^{-2} X' D_w^2 X \right) \right) .$$

(Note that under mild conditions on ϵ , X and w , we also have $\|\widehat{\beta}_w - \beta\|_2^2 = \mathbf{E} \left(\|\widehat{\beta}_w - \beta\|_2^2 \right) + o_{L_2}(1)$ - owing to concentration results for quadratic forms of vectors with independent entries; see Ledoux (2001).)

We now need to simplify this quantity.

Analytical simplification of trace $((X' D_w X)^{-2} X' D_w^2 X)$ Of course,

$$\text{trace} \left((X' D_w X)^{-2} X' D_w^2 X \right) = \text{trace} \left(D_w X (X' D_w X)^{-2} X' D_w \right) = \sum_{i=1}^n w_i^2 X_i' (X' D_w X)^{-2} X_i .$$

Hence, if $\widehat{\Sigma}_w = \frac{1}{n} \sum_{i=1}^n w_i X_i X_i' \triangleq \frac{w_i}{n} X_i X_i' + \widehat{\Sigma}_{(i)}$, we have

$$\text{trace} \left((X' D_w X)^{-2} X' D_w^2 X \right) = \frac{1}{n} \sum_{i=1}^n w_i^2 \frac{X_i' \widehat{\Sigma}_w^{-2} X_i}{n} .$$

Call $\widehat{\Sigma}(z) = \widehat{\Sigma} - z \text{Id}_p$. Using the identity

$$(\widehat{\Sigma} - z \text{Id}_p)(\widehat{\Sigma} - z \text{Id}_p)^{-1} = \text{Id}_p ,$$

we see, after taking traces, that (Silverstein (1995))

$$\frac{1}{n} \sum_{i=1}^n w_i X_i' (\widehat{\Sigma} - z \text{Id}_p)^{-1} X_i - z \text{trace} \left((\widehat{\Sigma} - z \text{Id}_p)^{-1} \right) = p .$$

We call, for $z \in \mathbb{C}$, $c(z) = \frac{1}{n} \text{trace} \left((\widehat{\Sigma} - z \text{Id}_p)^{-1} \right)$ and $c_i(z) = X_i' (\widehat{\Sigma}_{(i)} - z \text{Id}_p)^{-1} X_i$, provided z is not an eigenvalue of $\widehat{\Sigma}$.

Differentiating with respect to z and taking $z = 0$ (we know here that $\widehat{\Sigma}$ is non-singular with probability 1, so this does not create a problem), we have

$$\frac{1}{n} \sum_{i=1}^n w_i X_i' \widehat{\Sigma}^{-2} X_i - \text{trace} \left(\widehat{\Sigma}^{-1} \right) = 0 .$$

Also, since, by the Sherman-Morrison-Woodbury formula (Horn and Johnson (1990)),

$$X_i' \widehat{\Sigma}(z)^{-1} X_i = \frac{X_i' \widehat{\Sigma}_{(i)}(z)^{-1} X_i}{1 + w_i \frac{1}{n} X_i' \widehat{\Sigma}_{(i)}(z)^{-1} X_i} ,$$

we have, after differentiating,

$$\frac{1}{n} X_i' \widehat{\Sigma}^{-2} X_i = \frac{c_i'(0)}{[1 + w_i c_i(0)]^2} ,$$

where of course $c_i'(0) = X_i' \widehat{\Sigma}_{(i)}^{-2} X_i$. Hence,

$$\frac{1}{n} \sum_{i=1}^n w_i^2 \frac{1}{n} X_i' \widehat{\Sigma}^{-2} X_i = \frac{1}{n} \sum_{i=1}^n w_i^2 \frac{c_i'(0)}{[1 + w_i c_i(0)]^2} = c'(0) \frac{1}{n} \sum_{i=1}^n \frac{w_i^2}{[1 + w_i c(0)]^2} .$$

(Note that the arguments given in e.g El Karoui (2010) or El Karoui and Koesters (2011) for why $c_i(z) = c(z)(1 + o_P(1))$ extend easily to c_i' and c' given our assumptions on w 's and the fact that these functions have simple interpretations in terms of traces of powers of inverses of certain well-behaved - under our assumptions - matrices.)

Going back to

$$\frac{1}{n} \sum_{i=1}^n w_i X_i' (\widehat{\Sigma} - z \text{Id}_p)^{-1} X_i - z \text{trace} \left((\widehat{\Sigma} - z \text{Id}_p)^{-1} \right) = p ,$$

and using the previously discussed identity

$$\frac{w_i}{n} X_i' (\widehat{\Sigma} - z \text{Id}_p)^{-1} X_i = 1 - \frac{1}{1 + w_i c_i(z)} ,$$

we have

$$n - \sum_{i=1}^n \frac{1}{1 + w_i c_i(z)} - z n c(z) = p .$$

In other words,

$$1 - \kappa = \frac{1}{n} \sum_{i=1}^n \frac{1}{1 + w_i c_i(z)} + z c(z) .$$

Now,

$$\begin{aligned} c(z) \frac{1}{n} \sum_{i=1}^n \frac{w_i}{1 + w_i c(z)} &= \frac{1}{n} \sum_{i=1}^n \left(1 - \frac{1}{1 + w_i c(z)}\right) \\ &= \frac{1}{n} \sum_{i=1}^n \left(1 - \frac{1}{1 + w_i c_i(z)}\right) + \eta(z) \\ &= \kappa + z c(z) + \eta(z) , \end{aligned}$$

where $\eta(z)$ is such that $\eta(z) = o_P(1)$ and $\eta'(z) = o_P(1)$ (η has an explicit expression which allows us to verify these claims). Therefore, by differentiation, and after simplifications,

$$\frac{1}{n} \sum \left[\frac{w_i}{1 + w_i c(0)} \right]^2 c'(0) = \kappa \frac{c'(0)}{[c(0)]^2} - 1 + o_P(1) .$$

Hence,

$$\text{trace} \left((X' D_w X)^{-2} X' D_{w^2} X \right) = \left[\kappa \frac{\text{trace} \left(\widehat{\Sigma}_w^{-2} \right) / n}{[\text{trace} \left(\widehat{\Sigma}_w^{-1} \right) / n]^2} - 1 \right] + o_P(1) .$$

The fact that we can take expectations on both sides of this equation and that $o_P(1)$ is in fact $o_{L_2}(1)$ come from our assumptions about w_i 's - especially the fact that they are independent and bounded away from 0 - and properties of the inverse Wishart distribution.

Conclusion We can now conclude that a consistent estimator of the expected variance of the bootstrap estimator is

$$\frac{\|v\|_2^2 \sigma_\epsilon^2}{p} \left[\kappa \frac{\text{trace} \left(\widehat{\Sigma}_w^{-2} \right) / n}{[\text{trace} \left(\widehat{\Sigma}_w^{-1} \right) / n]^2} - \frac{1}{1 - \kappa} \right] .$$

Using the fact that

$$1 - \kappa = \frac{1}{n} \sum_{i=1}^n \frac{1}{1 + w_i c(z)} + z c(z) ,$$

we see that, since $\frac{1}{n} \text{trace} \left(\widehat{\Sigma}_w^{-2} \right) = c'(0)$,

$$\frac{1}{n} \text{trace} \left(\widehat{\Sigma}_w^{-2} \right) = \frac{c(0)}{\frac{1}{n} \sum_{i=1}^n w_i / (1 + w_i c(0))^2} .$$

We further note that asymptotically, when w_i are i.i.d and satisfy our assumptions, $c(0) \rightarrow c$, which solves:

$$\mathbf{E}_{w_i} \left[\frac{1}{1 + w_i c} \right] = 1 - \kappa .$$

Hence, asymptotically, when w_i 's are i.i.d and satisfy our assumptions, we have

$$\frac{\text{trace}\left(\widehat{\Sigma}_w^{-2}\right)/n}{\left[\text{trace}\left(\widehat{\Sigma}_w^{-1}\right)/n\right]^2} \rightarrow \frac{1}{c\mathbf{E}_{w_i}\left[w_i/(1+w_ic)^2\right]}.$$

Since $cw_i/(1+cw_i^2) = 1/(1+cw_i) - 1/(1+cw_i)^2$, we finally see that

$$\begin{aligned} c\mathbf{E}_{w_i}\left[\frac{w_i}{(1+w_ic)^2}\right] &= \mathbf{E}_{w_i}\left[\frac{1}{1+cw_i}\right] - \mathbf{E}_{w_i}\left[\frac{1}{(1+cw_i)^2}\right], \\ &= 1 - \kappa - \mathbf{E}_{w_i}\left[\frac{1}{(1+cw_i)^2}\right]. \end{aligned}$$

So asymptotically, the expected bootstrap variance is equivalent to, when $\|v\|_2 = 1$,

$$\frac{\sigma_\epsilon^2}{p} \left[\kappa \frac{1}{1 - \kappa - \mathbf{E}\left(\frac{1}{(1+cw_i)^2}\right)} - \frac{1}{1 - \kappa} \right],$$

where $\mathbf{E}\left(\frac{1}{1+cw_i}\right) = 1 - \kappa$.

In particular, when $w_i = 1$, we see, unsurprisingly, that the above quantity is 0, as it should, given that the bootstrapped estimate does not change when resampling.

We finally make note of a technical point, that is addressed in papers such as El Karoui (2010, 2013) and on which we rely here by using those papers. Essentially, theoretical considerations regarding quantities such as $\frac{1}{p}\text{trace}\left(\widehat{\Sigma}_w^{-k}\right)$ are easier to handle by working rather with $\frac{1}{p}\text{trace}\left(\left(\widehat{\Sigma}_w + \tau\text{Id}_p\right)^{-k}\right)$, for some $\tau > 0$. In the present context, it is easy to show (and done in those papers) that this approximation allows us to take the limit - even in expectation - for $\tau \rightarrow 0$ in all the expressions we get for $\tau > 0$ and that that limit is indeed $\mathbf{E}\left(\frac{1}{p}\text{trace}\left(\widehat{\Sigma}_w^{-k}\right)\right)$. Technical details rely on using the first resolvent identity (Kato, 1995), using moment properties of inverse Wishart distributions and using the fact that w_i 's are bounded below.

F.1 Extension: Elliptical Design

In this case, we have $\tilde{X}_i = \lambda_i X_i$, where $X_i \sim \mathcal{N}(0, \text{Id}_p)$ and $y_i = \epsilon_i + \tilde{X}_i' \beta$. We assume $\lambda_i \neq 0$ for all i , $\mathbf{E}\left(\lambda_i^2\right) = 1$, λ_i 's are i.i.d and bounded away from 0.

We can go through the proof of Theorem 2 and make necessary adjustments.

Of course, we have

$$\mathbf{E}\left(\|\widehat{\beta}_w - \beta\|_2^2\right) = \sigma_\epsilon^2 \mathbf{E}\left(\text{trace}\left(\left(\tilde{X}' D_w \tilde{X}\right)^{-2} \tilde{X}' D_{w^2} \tilde{X}\right)\right).$$

If we reformulate this expression in terms of X we get

$$\mathbf{E}\left(\|\widehat{\beta}_w - \beta\|_2^2\right) = \sigma_\epsilon^2 \mathbf{E}\left(\text{trace}\left(\left(X' D_{\lambda^2 w} X\right)^{-2} X' D_{\lambda^2 w^2} X\right)\right).$$

So this quantity is affected by the distribution of λ_i 's; hence the risk of $\widehat{\beta}_w$ is different in the Gaussian and elliptical design case.

The other important part of the proof is the computation of the risk of the bagged estimator. In this case, earlier work in random matrix theory (e.g El Karoui (2009); El Karoui and Koesters (2011)) shows that we can use the approximations

$$c_i \simeq \lambda_i^2 c,$$

where $c = \lim_{n,p \rightarrow \infty} \mathbf{E} \left(\frac{1}{n} \text{trace} (S^{-1}) \right)$, where $S = \frac{1}{n} \sum_{i=1}^n \lambda_i^2 w_i X_i X_i'$, i.e $S = \frac{1}{n} X' D_{\lambda^2 w} X$.

If we call

$$g(\lambda_i^2) = \mathbf{E}_{w_i} \left(\frac{c \lambda_i^2}{1 + c \lambda_i^2 w_i} \right),$$

we see by keeping track of changes in the earlier proof that we have asymptotically

$$\mathbf{E} \left(\|\widehat{b}\|_2^2 \right) = \frac{\mathbf{E}_{\lambda_i} (\lambda_i^2 g^2(\lambda_i^2))}{\kappa - \mathbf{E}_{\lambda_i} (\lambda_i^4 g^2(\lambda_i^2))} \sigma_\epsilon^2.$$

The same arguments we used before give that

$$\mathbf{E} (g(\lambda_i^2)) = \kappa.$$

Based on this information, we can compute $\mathbf{E} \left(\text{var} \left(v' \widehat{\beta}_w^* \right) \right)$ as we had in the proof of Theorem 2 and compare it to $\text{var} \left(v' \widehat{\beta} \right)$. The expressions do not seem to simplify much further however in this case, by contrast to the Gaussian design case where $\lambda_i = 1$ for all i . (For instance, when $\lambda_i=1$ for all i 's, $g(\lambda_i) = g(1) = \kappa$ and we recover the results of Theorem 2.)

Importantly, the characteristics of the distribution of λ_i that affect $\mathbf{E} \left(\text{var} \left(v' \widehat{\beta}_w^* \right) \right)$ go beyond $\mathbf{E} (\lambda_i^2)$. And hence the expression we gave in Theorem 2 won't apply directly to the elliptical case.

F.2 Extension: Multinomial($n, 1/n$) weights

A natural question is whether the computations we have made can be extended to w_i 's that are i.i.d *Poisson*(1) and/or Multinomial($n, 1/n$), as in the standard bootstrap.

In both cases, technical issues arise because with asymptotically negligible but non-zero probability, the matrix $X' D_w X$ may be of rank less than p . This can be handled in several ways. A simple one is to replace the weights w_i by $w_i(\tau) = \tau + (1 - \tau)w_i$ and study the problem when $\tau \rightarrow 0$.

Beyond that technicality, an important question is whether one can handle the fact that the weights are dependent in the multinomial case. For quantities of the type $\frac{1}{n} \text{trace} \left((X' D_w X)^{-1} \right)$, it was argued in El Karoui (2010) that one could ignore the dependency issue and treat the problem as if the weights were i.i.d *Po*(1). This type of arguments would be easy to extend where we need them here, for instance in quantities that arise in the computation of $\mathbf{E} \left(\|\widehat{\beta}_w - \beta\|_2^2 \right)$ or to show that we can write $c_i = c + o_P(1)$, where c is deterministic.

The remaining question is therefore the characterization of the risk of the bagged estimator. We have, with a slight modification with respect to the case of independent weights,

$$\hat{b}_p = \frac{1}{p} \sum_{i=1}^n X_i(p) \left[\mathbf{E}_{w_i} \left(\frac{cw_i}{1+cw_i} \right) \epsilon_i - V_i' \mathbf{E}_w \left(\hat{\gamma}_{(i)} \frac{cw_i}{1+cw_i} \right) \right] (1 + o_P(1)).$$

As before, $\mathbf{E}_{w_i} \left(\frac{cw_i}{1+cw_i} \right) = p/n + o(1)$. The problem is the dependence between $\hat{\gamma}_{(i)}$ and w_i . The rotational invariance arguments we invoked before still hold, so that $\hat{\gamma}_i = \|\hat{\gamma}_{(i)}\|_2 u$, where u is uniform on the unit sphere and independent of $\|\hat{\gamma}_{(i)}\|_2$. It is also independent of V_i , since $\hat{\gamma}_{(i)}$ is the leave-one-out estimate of γ . The same rotational invariance arguments hold for the bagged estimate $\mathbf{E}_w \hat{\gamma}_{(i)} \frac{cw_i}{1+cw_i}$. Hence, after a little bit of work we see that

$$\mathbf{E} \left([V_i' \mathbf{E}_w \left(\hat{\gamma}_{(i)} \frac{cw_i}{1+cw_i} \right)]^2 \right) = \mathbf{E} \left(\left\| \mathbf{E}_w \left(\hat{\gamma}_{(i)} \frac{cw_i}{1+cw_i} \right) \right\|_2^2 \right).$$

Using the fact that $w_{(i)}|w_i \sim \text{Multinomial}(n-w_i, 1/(n-1))$, the only real technical hurdle is to show that $\mathbf{E}_{w_i} \hat{\gamma}_{(i)}$ is asymptotically deterministic and independent of w_i . A strategy for this is to create a coupling: one can compare $\hat{\gamma}_{(i)}$ to $\hat{\mathbf{g}}_{(i)}$, where $\hat{\mathbf{g}}_{(i)}$ is computed using a $n-1$ dimensional vector of weights with distribution $\text{Multinomial}(n-1, 1/(n-1))$ - i.e running $w_i - 1$ multinomial trials after having obtained $w_{(i)}$ (the case $w_i = 0$ is easy to handle separately). Clearly, the distribution of $\hat{\mathbf{g}}_{(i)}$ is independent of w_i , by construction. On the other hand, a bit of work on top of the leave-one-observation-out expansions shows that $\|\hat{\mathbf{g}}_{(i)} - \hat{\gamma}_{(i)}\|_2^2$ is roughly of size at most $w_i^2/n \rightarrow 0$. Furthermore, $\|\mathbf{E}_w(\hat{\mathbf{g}}_{(i)}) - \mathbf{E}_w(\hat{\gamma}_{(i)})\|_2 \rightarrow 0$ for the same reason. This suggests that further technical work along those lines will give that

$$\mathbf{E}_w \left(\hat{\gamma}_{(i)} \frac{cw_i}{1+cw_i} \right) \simeq \mathbf{E}_w(\hat{\gamma}_{(i)}) \mathbf{E}_w \left(\frac{cw_i}{1+cw_i} \right),$$

where \simeq means that the approximation is valid in Euclidean norm. The same coupling arguments will give that

$$\|\mathbf{E}_w(\hat{\gamma}_{(i)})\| \simeq \|\hat{b}\|,$$

where \hat{b} is the bagged estimator. This will yield the same results as in the i.i.d Po(1) case.

Numerical results We verified that our theoretical results (i.e Theorem 2) hold for Poisson(1) weights in limited simulations (note that in this case $w_i = 0$ is possible). For Gaussian design matrix, double exponential errors, and ratios $\kappa = .1, .3, .5$ we found that the ratio of the observed bootstrap expected variance of $\hat{\beta}_1^*$ to our theoretical prediction using Poisson(1) weights was 1.0027, 1.0148, and 1.0252, respectively (here $n = 500$, and there were $R = 1000$ bootstrap resamples for each of 1000 simulations).

Appendix G. Proof of Theorem 3 (Jackknife Variance)

As explained in Section H, we can study without loss of generality the case where $\Sigma = \text{Id}_p$ and $\beta = 0$. This is what we do in this proof.

We study it in details in the least-squares case, and postpone a detailed analysis of the robust regression case to future studies.

According to the approximations in El Karoui et al. (2013), which are exact for least squares, or classic results Weisberg (2014) we have:

$$\widehat{\beta} - \widehat{\beta}_{(i)} = \frac{1}{n} \widehat{\Sigma}_{(i)}^{-1} X_i e_i .$$

Recall also that

$$e_i = \frac{\tilde{e}_{i(i)}}{1 + \frac{1}{n} X_i' \widehat{\Sigma}_{(i)}^{-1} X_i} .$$

Hence,

$$v'(\widehat{\beta} - \widehat{\beta}_{(i)}) = \frac{1}{n} v' \widehat{\Sigma}_{(i)}^{-1} X_i \frac{\tilde{e}_{i(i)}}{1 + \frac{1}{n} X_i' \widehat{\Sigma}_{(i)}^{-1} X_i} .$$

Hence,

$$n \sum_{i=1}^n [v'(\widehat{\beta} - \widehat{\beta}_{(i)})]^2 = \frac{1}{n} \sum_{i=1}^n \frac{[v' \widehat{\Sigma}_{(i)}^{-1} X_i \tilde{e}_{i(i)}]^2}{[1 + \frac{1}{n} X_i' \widehat{\Sigma}_{(i)}^{-1} X_i]^2} .$$

Note that at the denominator, we have

$$\begin{aligned} 1 + \frac{1}{n} X_i' \widehat{\Sigma}_{(i)}^{-1} X_i &= 1 + \frac{1}{n} \text{trace} \left(\widehat{\Sigma}^{-1} \right) + o_P(1) , \\ &= 1 + \frac{p}{n} \frac{1}{1 - p/n} + o_P(1) = \frac{1}{1 - p/n} + o_P(1) . \end{aligned}$$

by appealing to standard results about concentration of high-dimensional Gaussian random variables, and standard results in random matrix theory and classical multivariate statistics (see Mardia et al. (1979); Haff (1979)). By the same arguments, this approximation works not only for each i but for all $1 \leq i \leq n$ at once. The approximation is also valid in expectation, using results concerning Wishart matrices found for instance in Mardia et al. (1979).

For the numerator, we see that

$$T_i = v' \widehat{\Sigma}_{(i)}^{-1} X_i \tilde{e}_{i(i)} = v' \widehat{\Sigma}_{(i)}^{-1} X_i (\epsilon_i - X_i' (\widehat{\beta}_{(i)} - \beta)) .$$

Since ϵ_i is independent of X_i and $\widehat{\Sigma}_{(i)}$, we see that

$$\mathbf{E} (T_i^2) = \mathbf{E} (\epsilon_i^2) \mathbf{E} \left((v' \widehat{\Sigma}_{(i)}^{-1} X_i)^2 \right) + \mathbf{E} \left([X_i' (\widehat{\beta}_{(i)} - \beta)]^2 [v' \widehat{\Sigma}_{(i)}^{-1} X_i]^2 \right) .$$

If α and β are fixed vectors, $\alpha' X_i$ and $\beta' X_i$ are Gaussian random variables with covariance $\alpha' \beta$, since we are working under the assumption that $X_i \sim \mathcal{N}(0, \text{Id}_p)$. It is easy to check that if Z_1 and Z_2 are two Gaussian random variables with covariance γ and respective variances σ_1^2 and σ_2^2 , we have

$$\mathbf{E} ((Z_1 Z_2)^2) = \sigma_1^2 \sigma_2^2 + 2\gamma^2 .$$

We conclude that

$$\mathbf{E} ((a' X_i)^2 (b' X_i)^2) = \|a\|_2^2 \|b\|_2^2 + 2(a'b)^2 .$$

We note that

$$\mathbf{E} \left([v' \widehat{\Sigma}_{(i)}^{-1} X_i]^2 \right) = \mathbf{E} \left(v' \widehat{\Sigma}_{(i)}^{-2} v \right) .$$

Classic Wishart computations give (Haff (1979), p.536 (iii)) that as $n, p \rightarrow \infty$,

$$\mathbf{E} \left(\widehat{\Sigma}_{(i)}^{-2} \right) = \left[\frac{1}{(1 - p/n)^3} + o(1) \right] \text{Id}_p .$$

Hence, in our asymptotics,

$$\mathbf{E} \left((v' \widehat{\Sigma}_{(i)}^{-1} X_i)^2 \right) \rightarrow \frac{1}{(1 - p/n)^3} \|v\|_2^2 .$$

We also note that

$$\mathbf{E}_\epsilon \left[(v' \widehat{\Sigma}_{(i)}^{-1} \widehat{\beta}_{(i)})^2 \right] = \frac{1}{n} v' \widehat{\Sigma}_{(i)}^{-3} v .$$

Hence,

$$\mathbf{E} \left((v' \widehat{\Sigma}_{(i)}^{-1} \widehat{\beta}_{(i)})^2 \right) = o(1) \text{ in our asymptotics .}$$

Therefore,

$$\mathbf{E} (T_1^2) = \frac{1}{(1 - p/n)^3} \|v\|_2^2 \sigma_\epsilon^2 \left(1 + \frac{p/n}{1 - p/n} \right) + o(1)$$

since $\mathbf{E} \left(\|\widehat{\beta}_{(i)} - \beta\|_2^2 \right) = \sigma_\epsilon^2 \frac{p/n}{1 - p/n} + o(1)$.

When $v = e_1$, we therefore have

$$\mathbf{E} (T_1^2) = \sigma_\epsilon^2 \frac{1}{(1 - p/n)^4} + o(1) .$$

Therefore, in that situation,

$$\mathbf{E} \left(n \sum_{i=1}^n (v' (\widehat{\beta}_{(i)} - \widehat{\beta}))^2 \right) = \sigma_\epsilon^2 \frac{1}{(1 - p/n)^2} + o(1) .$$

In other words,

$$\mathbf{E} \left(\sum_{i=1}^n (v' (\widehat{\beta}_{(i)} - \widehat{\beta}))^2 \right) = \left[\frac{1}{1 - p/n} + o(1) \right] \text{var} \left(\widehat{\beta}_1 \right)$$

G.1 Dealing with Centering

Let us call $\widehat{\beta}_{(\cdot)} = \frac{1}{n} \sum_{i=1}^n \widehat{\beta}_{(i)}$. We have previously studied the properties of $\sum_{i=1}^n ([v' (\widehat{\beta} - \widehat{\beta}_{(i)})]^2)$ and now need to show that the same results apply to $\sum_{i=1}^n ([v' (\widehat{\beta}_{(\cdot)} - \widehat{\beta}_{(i)})]^2)$.

To show that replacing $\widehat{\beta}$ by $\widehat{\beta}_{(\cdot)}$ does not affect the result, we consider the quantity

$$n^2 [v' (\widehat{\beta} - \widehat{\beta}_{(\cdot)})]^2 .$$

Since $\widehat{\beta} - \widehat{\beta}_{(i)} = \frac{1}{n} \widehat{\Sigma}_{(i)}^{-1} X_i e_i$, we have

$$\widehat{\beta} - \widehat{\beta}_{(\cdot)} = \frac{1}{n^2} \sum_{i=1}^n \widehat{\Sigma}_{(i)}^{-1} X_i e_i .$$

Hence,

$$n^2[v'(\widehat{\beta} - \widehat{\beta}_{(\cdot)})]^2 = \left[\frac{1}{n} \sum_{i=1}^n v' \widehat{\Sigma}_{(i)}^{-1} X_i (\epsilon_i - X_i'(\widehat{\beta} - \beta)) \right]^2 .$$

A simple variance computation gives that $\frac{1}{n} \sum_{i=1}^n v' \widehat{\Sigma}_{(i)}^{-1} X_i \epsilon_i \rightarrow 0$ in L^2 , since each term has mean 0 and the variance of the sum goes to 0.

Recall now that

$$\widehat{\Sigma}^{-1} X_i = \frac{\widehat{\Sigma}_{(i)}^{-1} X_i}{1 + c_i} ,$$

where all c_i 's are equal to $p/n/(1 - p/n) + o_P(1)$. Let us call $\mathbf{c} = p/n/(1 - p/n)$.

We conclude that

$$\frac{1}{n} \sum_{i=1}^n v' \widehat{\Sigma}_{(i)}^{-1} X_i X_i' (\widehat{\beta} - \beta) = v' (\widehat{\beta} - \beta) (1 + \mathbf{c} + o(1)) .$$

When v is given, we clearly have $v'(\widehat{\beta} - \beta) = o_P(p^{-1/2})$, given the distribution of $\widehat{\beta} - \beta$ under our assumptions on X_i 's and ϵ_i 's. So we conclude that

$$n^2[v'(\widehat{\beta} - \widehat{\beta}_{(\cdot)})]^2 \rightarrow 0 \text{ in probability .}$$

Because we have enough moments, the previous result is also true in expectation.

G.2 Putting Everything Together

The jackknife estimate of variance of $v' \widehat{\beta}$ is up to a factor going to 1

$$\begin{aligned} \frac{n}{n-1} \text{JACK}(\text{var}(v' \widehat{\beta})) &= \sum_{i=1}^n [(v' \widehat{\beta}_{(i)} - \widehat{\beta}_{(\cdot)})]^2 \\ &= \sum_{i=1}^n [(v' \widehat{\beta}_{(i)} - \widehat{\beta})]^2 + n[v'(\widehat{\beta} - \widehat{\beta}_{(\cdot)})]^2 . \end{aligned}$$

Our previous analyses therefore imply (using $v = e_1$) that

$$\frac{n}{n-1} \mathbf{E} \left(\text{JACK}(\text{var}(\widehat{\beta}_1)) \right) = \left[\frac{1}{1 - p/n} + o(1) \right] \text{var}(\widehat{\beta}_1) .$$

This completes the proof of Theorem 3

G.3 Extension: More Involved Designs and Different Loss Functions

Our approach could be used to analyze similar problems in the case of elliptical designs. However, in that case, it seems that the factor that will appear in quantifying the amount by which the variance is mis-estimated will depend in general on the ellipticity parameters. We refer to El Karoui (2013) for computations of quantities such as $v' \widehat{\Sigma}^{-2} v$ in that case, which are of course essential to measuring mis-estimation.

We obtained the possible correction we mentioned in the paper for these more general settings following the ideas used in the rigorous proof we just gave, as well as approximation

arguments given in El Karoui et al. (2013) and justified rigorously in El Karoui (2013). Checking fully rigorously all the approximations we made in this Jackknife computation would require a very large amount of technical work, and since this is tangential to our main interests in this paper, we postpone that to a future work of a more technical nature.

It is also clear, since all these results and the proof we just gave rely on random matrix techniques, that a similar analysis could be carried out in the case where $X_{i,j}$ are i.i.d with a non-Gaussian distribution, provided that distribution has enough moments (see e.g Pajor and Pastur (2009) or El Karoui and Koesters (2011) for examples of such techniques, actually going beyond the case of i.i.d entries for the design matrix). The main issues in carrying out this program seem to be technical and not conceptual at this point, so we leave this problem to possible future work.

Appendix H. Going from $\Sigma = \text{Id}_p$ to $\Sigma \neq \text{Id}_p$

As discussed in Section A, we have

$$\widehat{\beta}_\rho(y_i; X_i; \epsilon_i) - \beta = \Sigma^{-1/2} \widehat{\beta}_\rho(\epsilon_i; \Sigma^{-1/2} X_i; \epsilon_i),$$

In other words, $\widehat{\beta}_\rho(\tilde{y}_i; \Sigma^{-1/2} X_i; \epsilon_i)$ is the robust regression estimator in the null case where $\beta = 0$ and X_i is replaced by $\tilde{X}_i = \Sigma^{-1/2} X_i$. Of course, if $\text{cov}(X_i) = \Sigma$, $\text{cov}(\tilde{X}_i) = \text{Id}_p$.

H.1 Consequences for the Jackknife

Naturally the same equality applies to leave-one-out estimators. So, with the notations of Equation (7) in the main text, we have, when $\text{span}(\{X_i\}_{i=1}^n) = \mathbb{R}^p$ and Σ is positive definite,

$$(v'[\widehat{\beta}_{(i)} - \tilde{\beta}])^2 = (v' \Sigma^{-1/2} [\widehat{\beta}_{(i)}(\epsilon_i; \Sigma^{-1/2} X_i; \epsilon_i) - \tilde{\beta}(\epsilon_i; \Sigma^{-1/2} X_i; \epsilon_i)])^2.$$

Let us call $\widehat{\beta}_\rho(\Sigma; \beta)$ our robust regression estimator when $\text{cov}(X_i) = \Sigma$ and $\mathbf{E}(y_i | X_i) = X_i' \beta$. It is clear from the previous display that the properties of $\text{var}_{JACK}(v' \widehat{\beta}_\rho(\Sigma; \beta))$ are the same as those of $\text{var}_{JACK}(v' \Sigma^{-1/2} \widehat{\beta}_\rho(\text{Id}_p; 0))$. So understanding the null case is enough to understand the general case, which is why we focus on the null case in our computations.

Furthermore, by the same arguments, we have

$$\text{var}\left(v' \widehat{\beta}_\rho(y_i; X_i; \epsilon_i)\right) = \text{var}\left(v' \Sigma^{-1/2} \widehat{\beta}_\rho(\text{Id}_p; 0)\right).$$

So we have

$$\frac{\text{var}_{JACK}(v' \widehat{\beta}_\rho(\Sigma; \beta))}{\text{var}\left(v' \widehat{\beta}_\rho(\Sigma; \beta)\right)} = \frac{\text{var}_{JACK}(v' \Sigma^{-1/2} \widehat{\beta}_\rho(\text{Id}_p; 0))}{\text{var}\left(v' \Sigma^{-1/2} \widehat{\beta}_\rho(\text{Id}_p; 0)\right)}.$$

Calling $u_1 = \Sigma^{-1/2} v / \|\Sigma^{-1/2} v\|$, we see that u_1 is a unit vector. And we finally see that

$$\frac{\text{var}_{JACK}(v' \widehat{\beta}_\rho(\Sigma; \beta))}{\text{var}\left(v' \widehat{\beta}_\rho(\Sigma; \beta)\right)} = \frac{\text{var}_{JACK}(u_1' \widehat{\beta}_\rho(\text{Id}_p; 0))}{\text{var}\left(u_1' \widehat{\beta}_\rho(\text{Id}_p; 0)\right)}.$$

Hence, characterizing $\frac{\text{var}_{JACK}(v'\widehat{\beta}_\rho(\text{Id}_p;0))}{\text{var}(v'\widehat{\beta}_\rho(\text{Id}_p;0))}$ for all fixed unit vectors v characterizes

$$\frac{\text{var}_{JACK}(v'\widehat{\beta}_\rho(\Sigma; \beta))}{\text{var}(v'\widehat{\beta}_\rho(\Sigma; \beta))}$$

for all β and invertible Σ . This is why our proof is focused on the null case $\Sigma = \text{Id}_p$ and $\beta = 0$.

H.2 Consequences for the Pairs Bootstrap

Let us call D_w the diagonal matrix with (i, i) -entry $D(i, i) = w_i$. We consider only the case where $w_i > 0$, so we do not have to consider the case where fewer than p X_i 's are assigned positive weights - which would result in $\widehat{\beta}_\rho$ being ill-defined (since infinitely many solutions would then be feasible).

In particular, for least squares, we have in our setting

$$\widehat{\beta}_w = (X'D_w X)^{-1} X'D_w Y = \beta + (X'D_w X)^{-1} X'D_w \epsilon .$$

More generally, by a simple change of variables, since $w_i > 0$ and $\text{span}(\{X_i\}_{i=1}^n) = \mathbb{R}^p$, when Σ is invertible,

$$\widehat{\beta}_{w,\rho}(y_i; \{X_i\}_{i=1}^n; \epsilon_i) - \beta = \Sigma^{-1/2} \widehat{\beta}_{w,\rho}(\epsilon_i; \Sigma^{-1/2} X_i; \epsilon_i) .$$

If b_ρ is the corresponding bagged estimate, obtained by averaging $\widehat{\beta}_{w,\rho}$ over w 's, we also have

$$b_\rho(y_i; \{X_i\}_{i=1}^n; \epsilon_i) - \beta = \Sigma^{-1/2} b_\rho(\epsilon_i; \Sigma^{-1/2} X_i; \epsilon_i) .$$

Hence, we also have

$$\widehat{\beta}_{w,\rho}(y_i; \{X_i\}_{i=1}^n; \epsilon_i) - b_\rho(y_i; \{X_i\}_{i=1}^n; \epsilon_i) = \Sigma^{-1/2} \left[\widehat{\beta}_{w,\rho}(\epsilon_i; \Sigma^{-1/2} X_i; \epsilon_i) - b_\rho(\epsilon_i; \Sigma^{-1/2} X_i; \epsilon_i) \right]$$

We further note that since $y_i = \epsilon_i + X_i' \beta$, $y_i = \epsilon_i + (\Sigma^{-1/2} X_i)' \Sigma^{1/2} \beta$ and hence

$$\widehat{\beta}_{w,\rho}(y_i; \Sigma^{-1/2} X_i; \epsilon_i) = \Sigma^{1/2} \beta + \widehat{\beta}_{w,\rho}(\epsilon_i; \Sigma^{-1/2} X_i; \epsilon_i) .$$

The previous equation clearly implies that, if v is a fixed vector and $u_1 = \Sigma^{-1/2} v$

$$\begin{aligned} & v'(\widehat{\beta}_\rho^*(y_i; \{X_i\}_{i=1}^n; \beta) - b_\rho(y_i; \{X_i\}_{i=1}^n; \beta)) \\ &= u_1' \left[\widehat{\beta}_\rho^*(y_i; \Sigma^{-1/2} X_i; \epsilon_i) - b_\rho(y_i; \Sigma^{-1/2} X_i; \epsilon_i) \right] , \\ &= u_1' \left[\widehat{\beta}_\rho^*(\epsilon_i; \Sigma^{-1/2} X_i; \epsilon_i) - b_\rho(\epsilon_i; \Sigma^{-1/2} X_i; \epsilon_i) \right] . \end{aligned}$$

We note that if $\text{cov}(X_i) = \Sigma$, the last line in the previous display corresponds to the bootstrap distribution of our estimator in the null case where $\beta = 0$ and $\Sigma = \text{Id}_p$, but v has been replaced by $u_1 = \Sigma^{-1/2} v$. This shows that understanding the bootstrap properties of $v'(\widehat{\beta}_\rho^* - b_\rho)$ in the null case $\text{cov}(X_i) = \Sigma$ and $\beta = 0$ gives the result we seek in the general case of $\Sigma \neq \text{Id}_p$ and $\beta \neq 0$. (Here we centered our estimator around the bagged estimator,

because it is natural when computing bootstrap variances. The arguments above show that many other centering choices are possible, however.)

The last small issue that one needs to handle is the fact that our computations are done for v with unit norm and u_1 may not have unit norm. This is easily handled by simply scaling by the deterministic $\|u_1\|$. In particular, it is easy to see through simple scaling arguments that

$$\frac{\mathbf{E} \left(\text{var} \left(v' \widehat{\beta}_\rho^*(\Sigma; \beta) \right) \right)}{\text{var} \left(v' \widehat{\beta}_\rho(\Sigma; \beta) \right)} = \frac{\mathbf{E} \left(\text{var} \left(\tilde{u}_1' \widehat{\beta}_\rho^*(\text{Id}_p; 0) \right) \right)}{\text{var} \left(\tilde{u}_1' \widehat{\beta}_\rho(\text{Id}_p; 0) \right)},$$

where $\tilde{u}_1 = u_1/\|u_1\|$ has unit norm.

H.3 Rotational Invariance Arguments and Consequences

Motivated by the arguments in the previous two subsections, we now consider the null case where $\beta = 0$ and $\text{cov}(X_i) = \text{Id}_p$. Note that then $y_i = \epsilon_i$. Also, if X_i is replaced by OX_i , where O is an orthogonal matrix, and $\widehat{\beta}$ is replaced by $O\widehat{\beta}$. In other words,

$$\widehat{\beta}_\rho(\epsilon_i; \{OX_i\}_{i=1}^n; \epsilon_i) = O\widehat{\beta}_\rho(\epsilon_i; \{X_i\}_{i=1}^n; \epsilon_i), .$$

Note that when the design matrix is such that $OX_i \stackrel{\mathcal{L}}{=} X_i$ for all i (i.e the distribution of X_i 's is invariant by rotation),

$$\widehat{\beta}_\rho(\epsilon_i; \{OX_i\}_{i=1}^n; \epsilon_i) \stackrel{\mathcal{L}}{=} \widehat{\beta}_\rho(\epsilon_i; \{X_i\}_{i=1}^n; \epsilon_i) .$$

When $w_i > 0$ for all i , we see that exactly the same arguments apply to $\widehat{\beta}_{w,\rho}(\epsilon_i; \{X_i\}_{i=1}^n; \epsilon_i)$ and hence $\widehat{\beta}_\rho^*(\epsilon_i; \{X_i\}_{i=1}^n; \epsilon_i)$. In particular, for any orthogonal matrix O , since $X_i \stackrel{\mathcal{L}}{=} OX_i$,

$$\begin{aligned} \mathbf{E} \left(\text{var} \left(v' \widehat{\beta}_\rho^*(\epsilon_i; \{X_i\}_{i=1}^n; \epsilon_i) \right) \right) &= \mathbf{E} \left(\text{var} \left(v' \widehat{\beta}_\rho^*(\epsilon_i; \{OX_i\}_{i=1}^n; \epsilon_i) \right) \right) \\ &= \mathbf{E} \left(\text{var} \left(v' O \widehat{\beta}_\rho^*(\epsilon_i; \{X_i\}_{i=1}^n; \epsilon_i) \right) \right) . \end{aligned}$$

This implies that for any unit vector v , we have, if e_1 is the first canonical basis vector,

$$\mathbf{E} \left(\text{var} \left(v' \widehat{\beta}_\rho^*(\epsilon_i; X_i; \epsilon_i) \right) \right) = \mathbf{E} \left(\text{var} \left(e_1' \widehat{\beta}_\rho^*(\epsilon_i; X_i; \epsilon_i) \right) \right) .$$

Indeed, we just need to take O to be such that $O'v = e_1$ to prove the above result.

In the case where X_i 's are i.i.d $\mathcal{N}(0, \text{Id}_p)$, we do have $X_i \stackrel{\mathcal{L}}{=} OX_i$, so the arguments above apply. Therefore, to understand $\mathbf{E} \left(\text{var} \left(v' \widehat{\beta}_\rho^* \right) \right)$ in this case it is sufficient to understand $\mathbf{E} \left(\text{var} \left(e_1' \widehat{\beta}_\rho^* \right) \right)$. This latter case is the case tackled in the proof of Theorem 2. (These rotational invariance arguments are closely related to those in El Karoui et al. (2013).)

Appendix I. Supplementary Tables & Figures

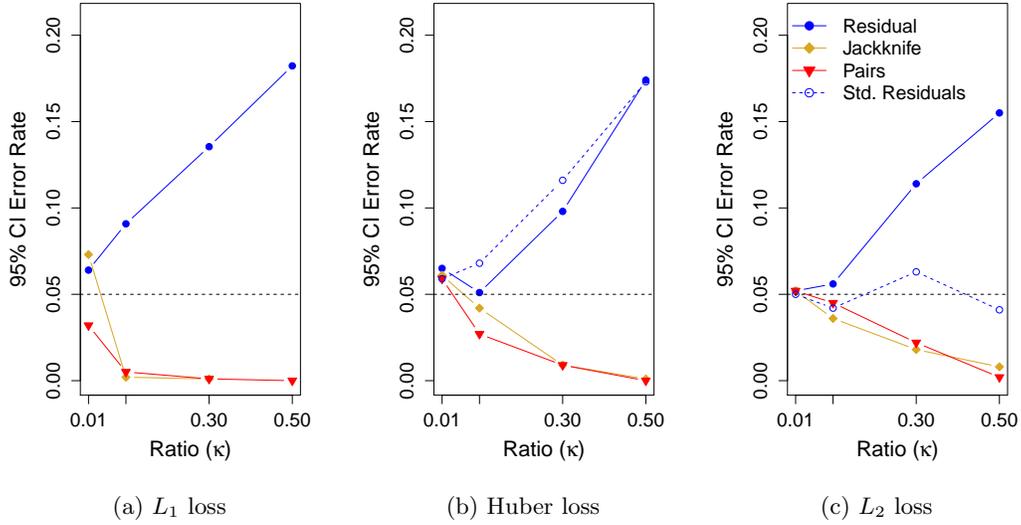


Figure A-1: **Performance of 95% confidence intervals of β_1 (double exponential error):** Here we show the coverage error rates for 95% confidence intervals for $n = 500$ with the error distribution being double exponential (with $\sigma^2 = 2$) and i.i.d. normal entries of X . See the caption of Figure 1 for more details.

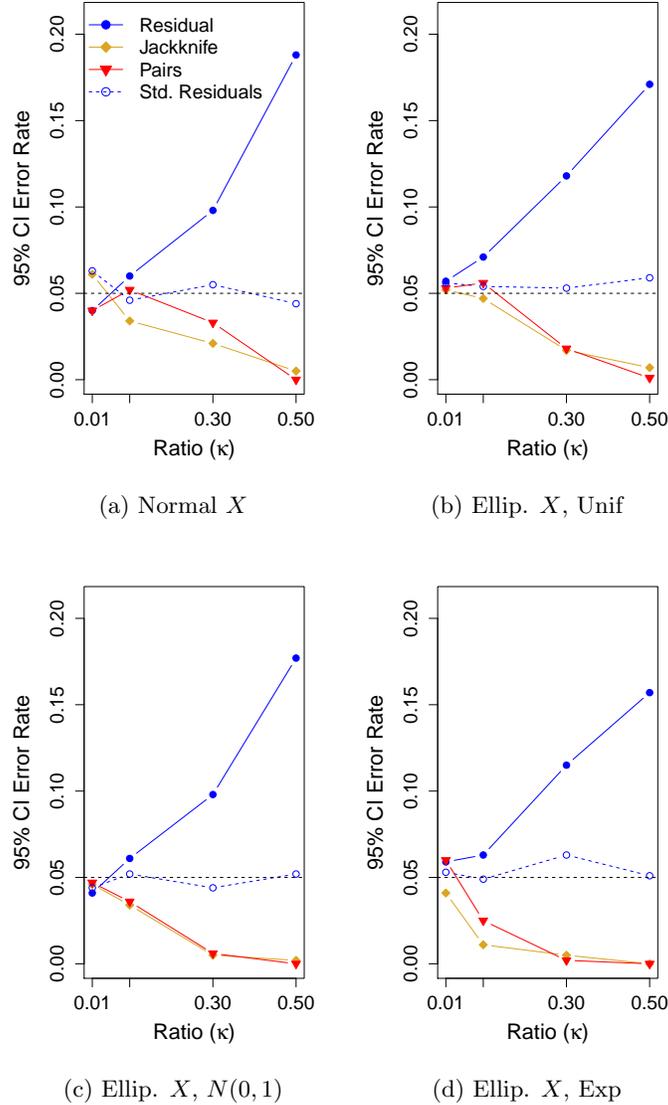


Figure A-2: **Performance of 95% confidence intervals of β_1 for L_2 loss (elliptical design X):** Here we show the coverage error rates for 95% confidence intervals for $n = 500$ with different distributions of the design matrix X using ordinary least squares regression: (a) $N(0, 1)$, (b) elliptical with $\lambda_i \sim U(\sqrt{\frac{12}{13}}0.5, \sqrt{\frac{12}{13}}1.5)$, (c) elliptical with $\lambda_i \sim N(0, 1)$, and (d) elliptical with $\lambda_i \sim Exp(\sqrt{2})$. In all of these plots, the error is distributed $N(0, 1)$ and the loss is L_2 . See the caption of Figure 1 for additional details.

	Residual	Jackknife	Pairs
$\kappa = 0.01$	0.063	0.089	0.035
$\kappa = 0.1$	0.113	0.005	0.013
$\kappa = 0.3$	0.137	0.000	0.003
$\kappa = 0.5$	0.210	0.000	0.000

(a) L_1 loss

	Residual	Jackknife	Pairs
$\kappa = 0.01$	0.057	0.054	0.054
$\kappa = 0.1$	0.068	0.037	0.041
$\kappa = 0.3$	0.090	0.015	0.004
$\kappa = 0.5$	0.198	0.002	0.000

(b) Huber loss

	Residual	Jackknife	Pairs
$\kappa = 0.01$	0.040	0.061	0.040
$\kappa = 0.1$	0.060	0.034	0.052
$\kappa = 0.3$	0.098	0.021	0.033
$\kappa = 0.5$	0.188	0.005	0.000

(c) L_2 loss

Table A-1: **Error rate of 95% confidence intervals of β_1 for $n = 500$** This table gives the exact error rates plotted in Figure 1. $\kappa = p/n$ indicates the ratio of p/n used in the simulation for this and future tables. See Figure 1's caption for more details.

	Normal	Ellip. Normal	Ellip. Exp
$\kappa = 0.01$	1.001	1.001	1.017
$\kappa = 0.1$	1.016	1.090	1.156
$\kappa = 0.3$	1.153	1.502	1.655
$\kappa = 0.5$	1.737	3.123	3.635

Table A-2: **Ratio of CI Width of Pairs compared to Standard.** This table gives the ratio of the average width of the confidence intervals from pairs bootstrapping to the average for the standard interval given by theoretical results, i.e. using $\text{var}(\hat{\beta}) = \sigma^2(X'X)^{-1}$ and creating standard confidence interval. These values were used for Figure 4 in the text.

	Residual	Std. Pred Error	Deconv
$\kappa = 0.01$	0.064	0.042	0.031
$\kappa = 0.1$	0.091	0.028	0.018
$\kappa = 0.3$	0.135	0.026	0.022
$\kappa = 0.5$	0.182	0.030	0.035

(a) L_1 loss

	Residual	Std. Pred Error	Deconv
$\kappa = 0.01$	0.065	0.048	0.036
$\kappa = 0.1$	0.051	0.054	0.039
$\kappa = 0.3$	0.098	0.035	0.037
$\kappa = 0.5$	0.174	0.034	0.036

(b) Huber loss

Table A-3: **Error rate of 95% confidence intervals using predicted errors.** This table gives the exact error rates plotted in Figure 2. See figure caption for more details.

	Residual	Jackknife	Pairs
$\kappa = 0.01$	0.064	0.073	0.032
$\kappa = 0.1$	0.091	0.002	0.005
$\kappa = 0.3$	0.135	0.001	0.001
$\kappa = 0.5$	0.182		0.000

(a) L_1 loss

	Residual	Jackknife	Pairs
$\kappa = 0.01$	0.065	0.061	0.059
$\kappa = 0.1$	0.051	0.042	0.027
$\kappa = 0.3$	0.098	0.009	0.009
$\kappa = 0.5$	0.174	0.001	0.000

(b) Huber loss

	Residual	Jackknife	Pairs
$\kappa = 0.01$	0.052	0.052	0.052
$\kappa = 0.1$	0.056	0.036	0.045
$\kappa = 0.3$	0.114	0.018	0.022
$\kappa = 0.5$	0.155	0.008	0.002

(c) L_2 loss

Table A-4: **Error rate of 95% confidence intervals of β_1 for double exponential error** This table gives the exact error rates plotted in Figure A-1. See figure caption for more details.

	Residual	Jackknife	Pairs
$\kappa = 0.01$	0.057	0.052	0.053
$\kappa = 0.1$	0.071	0.047	0.056
$\kappa = 0.3$	0.118	0.017	0.018
$\kappa = 0.5$	0.171	0.007	0.001

(a) Elliptical, Unif

	Residual	Jackknife	Pairs
$\kappa = 0.01$	0.041	0.046	0.047
$\kappa = 0.1$	0.061	0.034	0.036
$\kappa = 0.3$	0.098	0.005	0.006
$\kappa = 0.5$	0.177	0.002	0.000

(b) Elliptical, Normal

	Residual	Jackknife	Pairs
$\kappa = 0.01$	0.059	0.041	0.060
$\kappa = 0.1$	0.063	0.011	0.025
$\kappa = 0.3$	0.115	0.005	0.002
$\kappa = 0.5$	0.157	0.000	0.000

(c) Elliptical, Exp

Table A-5: **Error rate of 95% confidence intervals of β_1 for elliptical design X** This table gives the exact error rates plotted in Figure A-2. See figure caption for more details.

	L2	Huber	L1
$\kappa = 0.01$	0.964	0.991	2.060
$\kappa = 0.1$	1.115	1.173	5.432
$\kappa = 0.3$	1.411	1.613	10.862
$\kappa = 0.5$	1.986	2.671	14.045

(a) Jackknife

	L2	Huber	L1
$\kappa = 0.01$	1.078	0.923	1.081
$\kappa = 0.1$	1.041	1.098	1.351
$\kappa = 0.3$	1.333	1.954	2.001
$\kappa = 0.5$	2.808	4.507	3.156

(b) Pairs Bootstrap

Table A-6: **Over estimation of variance for Pairs bootstrap and Jackknife** This table gives the median values of the boxplots plotted in Figures 5 and 6. See relevant figure captions for more details.

References

- M. ApS. *The MOSEK optimization toolbox for MATLAB manual. Version 7.1 (Revision 28)*., 2015. URL <http://docs.mosek.com/7.1/toolbox/index.html>.
- D. Bean, P. J. Bickel, N. El Karoui, and B. Yu. Optimal M-estimation in high-dimensional regression. *Proceedings of the National Academy of Sciences*, 110(36):14563–14568, 2013.
- A. Belloni, V. Chernozhukov, and K. Kato. Uniform post-selection inference for least absolute deviation regression and other Z-estimation problems. *Biometrika*, 102(1):77–94, March 2015.
- R. Beran and M. S. Srivastava. Bootstrap tests and confidence regions for functions of a covariance matrix. *Ann. Statist.*, 13(1):95–115, 1985.
- P. J. Bickel and D. A. Freedman. Bootstrapping regression models with many parameters. In *A Festschrift for Erich L. Lehmann*, Wadsworth Statist./Probab. Ser., pages 28–48. Wadsworth, Belmont, Calif., 1983.
- P. J. Bickel, F. Götze, and W. R. van Zwet. Resampling fewer than n observations: gains, losses, and remedies for losses. *Statist. Sinica*, 7(1):1–31, 1997. Empirical Bayes, sequential analysis and related topics in statistics and probability (New Brunswick, NJ, 1995).
- P. J. Bickel and D. A. Freedman. Some asymptotic theory for the bootstrap. *Ann. Statist.*, 9(6):1196–1217, 1981.
- O. Chapelle, E. Manavoglu, and R. Rosales. Simple and scalable response prediction for display advertising. *ACM Trans. Intell. Syst. Technol.*, 5(4):61:1–61:34, December 2014.
- A. Chatterjee and S. N. Lahiri. ASYMPTOTIC PROPERTIES OF THE RESIDUAL BOOTSTRAP FOR LASSO ESTIMATORS. *Proceedings of the American Mathematical Society*, 138(12):4497–4509, December 2010.
- A. Chatterjee and S. N. Lahiri. Rates of convergence of the Adaptive LASSO estimators to the Oracle distribution and higher order refinements by the bootstrap. *The Annals of Statistics*, 41(3):1232–1259, June 2013.
- A. Chatterjee and S. Lahiri. Bootstrapping lasso estimators. *Journal of the American Statistical Association*, 106(494):608–625, 2011.
- M. R. Chernick. *Bootstrap Methods: A Practitioner’s Guide*. Wiley, 1999.
- Criteo. Criteo publicly available datasets, 2017. URL <http://research.criteo.com/outreach/>.
- A. C. Davison and D. V. Hinkley. *Bootstrap methods and their application*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 1997.

- A. Delaigle and I. Gijbels. Estimation of integrated squared density derivatives from a contaminated sample. *Journal of the Royal Statistical Society, B*, 64:869–886, 2002.
- A. Delaigle and I. Gijbels. Practical bandwidth selection in deconvolution kernel density estimation. *Computational Statistics and Data Analysis*, 45:249 – 267, 2004.
- A. Delaigle. Nonparametric kernel methods with errors-in-variables: constructing estimators, computing them, and avoiding common mistakes. *Aust. N. Z. J. Stat.*, 56(2): 105–124, 2014.
- R. Dezeure, P. Bühlmann, and C.-H. Zhang. High-dimensional simultaneous inference with the bootstrap. *TEST*, 26(4):685–719, October 2017.
- P. Diaconis and D. Freedman. Asymptotics of graphical projection pursuit. *Ann. Statist.*, 12(3):793–815, 1984.
- D. Donoho and A. Montanari. High dimensional robust m-estimation: Asymptotic variance via approximate message passing. *arXiv:1310.7320*, 2013.
- M. L. Eaton and D. E. Tyler. On Wielandt’s inequality and its application to the asymptotic distribution of the eigenvalues of a random symmetric matrix. *Ann. Statist.*, 19(1):260–271, 1991.
- B. Efron. Bootstrap methods: another look at the jackknife. *Ann. Statist.*, 7(1):1–26, 1979.
- B. Efron. *The jackknife, the bootstrap and other resampling plans*, volume 38 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, Pa., 1982.
- B. Efron and C. Stein. The jackknife estimate of variance. *Ann. Statist.*, 9(3):586–596, 1981.
- B. Efron and R. J. Tibshirani. *An introduction to the bootstrap*, volume 57 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, New York, 1993.
- N. El Karoui. Concentration of measure and spectra of random matrices: Applications to correlation matrices, elliptical distributions and beyond. *The Annals of Applied Probability*, 19(6):2362–2405, December 2009.
- N. El Karoui. High-dimensionality effects in the Markowitz problem and other quadratic programs with linear constraints: risk underestimation. *Ann. Statist.*, 38(6):3487–3566, 2010.
- N. El Karoui. On the realized risk of high-dimensional markowitz portfolios. *SIAM Journal in Financial Mathematics*, 4(1), 2013.
- N. El Karoui. Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: rigorous results. *arXiv:1311.2445*, 2013. ArXiv:1311.2445.

- N. El Karoui. On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probability Theory and Related Fields*, 2017.
- N. El Karoui and H. Koesters. Geometric sensitivity of random matrix results: consequences for shrinkage estimators of covariance and related statistical methods. *Submitted to Bernoulli*, 2011. Available at arXiv:1105.1404 (68 pages).
- N. El Karoui, D. Bean, P. Bickel, C. Lim, and B. Yu. On robust regression with high-dimensional predictors. Technical Report 811, UC, Berkeley, Department of Statistics, 2011. Originally submitted as manuscript AoS1111-009. Not under consideration anymore.
- N. El Karoui, D. Bean, P. J. Bickel, C. Lim, and B. Yu. On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences*, 2013.
- J. Fan. On the optimal rates of convergence for nonparametric deconvolution problems. *Ann. Statist.*, 19(3):1257–1272, 1991.
- M. Grant and S. Boyd. Graph implementations for nonsmooth convex programs. In V. Blondel, S. Boyd, and H. Kimura, editors, *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, pages 95–110. Springer-Verlag Limited, 2008. http://stanford.edu/~boyd/graph_dcp.html.
- M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 2.1. <http://cvxr.com/cvx>, March 2014.
- L. R. Haff. An identity for the Wishart distribution with applications. *J. Multivariate Anal.*, 9(4):531–544, 1979.
- P. Hall. *The bootstrap and Edgeworth expansion*. Springer Series in Statistics. Springer-Verlag, New York, 1992.
- P. Hall and S. Lahiri. Estimation of distributions, moments and quantiles in deconvolution problems. *The Annals of Statistics*, 36(5):2110–2134, 2008.
- P. Hall, J. S. Marron, and A. Neeman. Geometric representation of high dimension, low sample size data. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 67(3):427–444, 2005.
- J.-B. Hiriart-Urruty and C. Lemaréchal. *Fundamentals of convex analysis*. Grundlehren Text Editions. Springer-Verlag, Berlin, 2001. Abridged version of it Convex analysis and minimization algorithms. I [Springer, Berlin, 1993; MR1261420 (95m:90001)] and it II [ibid.; MR1295240 (95m:90002)].
- R. A. Horn and C. R. Johnson. *Matrix analysis*. Cambridge University Press, Cambridge, 1990. Corrected reprint of the 1985 original.
- P. J. Huber. Robust regression: asymptotics, conjectures and Monte Carlo. *Ann. Statist.*, 1:799–821, 1973.

- P. J. Huber and E. M. Ronchetti. *Robust statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons Inc., Hoboken, NJ, second edition, 2009.
- I. Johnstone. On the distribution of the largest eigenvalue in principal component analysis. *Ann. Statist.*, 29(2):295–327, 2001.
- T. Kato. *Perturbation theory for linear operators*. Classics in Mathematics. Springer-Verlag, Berlin, 1995. Reprint of the 1980 edition.
- A. Kleiner, A. Talwalkar, P. Sarkar, and M. I. Jordan. A scalable bootstrap for massive data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(4): 795–816, 2014.
- R. Koenker. *Quantile regression*, volume 38 of *Econometric Society Monographs*. Cambridge University Press, Cambridge, 2005.
- R. Koenker. *quantreg: Quantile Regression*, 2013. URL <http://CRAN.R-project.org/package=quantreg>. R package version 5.05.
- J. Langford, L. Li, and A. Strehl, 2007. URL https://github.com/JohnLangford/vowpal_wabbit/wiki.
- M. Ledoux. *The concentration of measure phenomenon*, volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2001.
- M. Lopes. A Residual Bootstrap for High-Dimensional Regression with Near Low-Rank Designs. In *Advances in Neural Information Processing Systems NIPS*, pages 3239–3247, 2014.
- E. Mammen. Asymptotics with increasing dimension for robust regression with applications to the bootstrap. *Ann. Statist.*, 17(1):382–400, 1989.
- E. Mammen. Bootstrap, wild bootstrap, and asymptotic normality. *Probab. Theory Related Fields*, 93(4):439–455, 1992.
- E. Mammen. Bootstrap and wild bootstrap for high-dimensional linear models. *Ann. Statist.*, 21(1):255–285, 1993.
- K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate analysis*. Academic Press [Harcourt Brace Jovanovich Publishers], London, 1979. Probability and Mathematical Statistics: A Series of Monographs and Textbooks.
- J. W. McKean, S. J. Sheather, and T. P. Hettmansperger. The Use and Interpretation of Residuals Based on Robust Estimation. *Journal of the American Statistical Association*, 88(424):1254–1263, December 1993.
- P. D. Miller. *Applied asymptotic analysis*, volume 75 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2006.
- J.-J. Moreau. Proximité et dualité dans un espace hilbertien. *Bull. Soc. Math. France*, 93: 273–299, 1965.

- MOSEK. *Rmosek: The R to MOSEK Optimization Interface*, 2014. URL <http://rmosek.r-forge.r-project.org/>, <http://www.mosek.com/>. R package version 7.0.5.
- A. Pajor and L. Pastur. On the limiting empirical measure of eigenvalues of the sum of rank one matrices with log-concave distribution. *Studia Math.*, 195(1):11–29, 2009.
- M. I. Parzen, L. J. Wei, and Z. Ying. A resampling method based on pivotal estimating functions. *Biometrika*, 81(2):341–350, 1994.
- D. N. Politis, J. P. Romano, and M. Wolf. *Subsampling*. Springer Series in Statistics. Springer-Verlag, New York, 1999.
- S. Portnoy. Asymptotic behavior of M -estimators of p regression parameters when p^2/n is large. I. Consistency. *Ann. Statist.*, 12(4):1298–1309, 1984.
- S. Portnoy. Asymptotic behavior of M estimators of p regression parameters when p^2/n is large. II. Normal approximation. *Ann. Statist.*, 13(4):1403–1417, 1985.
- S. Portnoy. Asymptotic behavior of the empiric distribution of M -estimated residuals from a regression model with many parameters. *Ann. Statist.*, 14(3):1152–1170, 1986.
- S. Portnoy. A central limit theorem applicable to robust regression estimators. *J. Multivariate Anal.*, 22(1):24–50, 1987.
- G. R. Shorack. Bootstrapping robust regression. *Comm. Statist. A—Theory Methods*, 11(9):961–972, 1982.
- J. W. Silverstein. Strong convergence of the empirical distribution of eigenvalues of large-dimensional random matrices. *J. Multivariate Anal.*, 55(2):331–339, 1995.
- D. W. Stroock. *Probability theory, an analytic view*. Cambridge University Press, Cambridge, 1993.
- A. W. van der Vaart. *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 1998.
- W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer, New York, fourth edition, 2002. ISBN 0-387-95457-0.
- K. W. Wachter. The strong limits of random matrix spectra for sample matrices of independent elements. *Annals of Probability*, 6(1):1–18, 1978.
- X. Wang and B. Wang. Deconvolution estimation in measurement error models: The `r` package `decon`. *Journal of Statistical Software*, 39(10):1–24, 2011.
- S. Weisberg. *Applied linear regression*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, fourth edition, 2014.
- C.-F. J. Wu. Jackknife, bootstrap and other resampling methods in regression analysis. *Ann. Statist.*, 14(4):1261–1350, 1986. With discussion and a rejoinder by the author.
- S. Zheng, D. Jiang, Z. Bai, and X. He. Inference on multiple correlation coefficients with moderately high dimensional data. *Biometrika*, 101:748–754, 2014.