Can we trust the bootstrap in high-dimension?

Noureddine El Karoui*
and
Elizabeth Purdom
Department of Statistics, University of California, Berkeley

February 4, 2015

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 residuals 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 bootstrap procedures based on our theoretical results that mitigate these problems. However, the corrections depend on assumptions regarding the underlying data-generation model, suggesting that in high-dimensions it may be difficult to have universal, robust bootstrapping techniques.

Keywords: Resampling, high-dimensional inference, bootstrap, random matrices

*The authors gratefully acknowledge grants NSF DMS-1026441 and NSF DMS-0847647 (CAREER). They would also like to thank Peter Bickel and Jorge Banuelosfor discussions.

1 Introduction

The bootstrap (Efron, 1979) is a ubiquitous tool in applied statistics, allowing for inference when very little is known about the statistical properties of the data. The bootstrap is a powerful tool in applied settings because it does not make the strong assumptions common to classical statistical theory regarding the distribution of the data. Instead, the bootstrap resamples the observed data to create an estimate, $\hat{F}$, of the unknown distribution of the data, F, which 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 $\hat{F}$, the estimate of 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, and 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 $\hat{F}$ 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 quantiles, 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 in the setting of standard linear models where data y_i are drawn from the linear model

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

The focus of our work is on the bootstrap or resampling properties of the estimator defined as

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

We will consider the setting where the number of predictors, p is of the same order of magnitude as the number of observations, n. 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), Mammen (1992), Mammen (1993), Section 3.9 of Koenker (2005), Parzen et al. (1994)). However, they are not high-dimensional in the modern sense of the word because in the above cited works, the assumption of high-dimensionality requires $p/n \to 0$ even though p is allowed to grow with n.

One early work considered least-squares regression in the setting where $p/n \to \kappa \in (0, 1)$: the paper Bickel and Freedman (1983). This paper showed convincingly that there was serious problems with the bootstrap in that setting. One striking result indicates that even for "smart" bootstraps, there is always one projection direction of $\widehat{\beta}$ where the bootstrap fails (see Theorem 3.1 in that paper).

In the current paper, we take a more minimalist point of view in terms of what we require of the bootstrap. We do not require the bootstrap distribution of the estimate to converge conditionally almost surely to the sampling distribution of the estimator. This latter requirement is usually what is required to say that the bootstrap "works" (see van der Vaart (1998) for instance). We simply ask whether we can build, using bootstrap or other resampling method, trustworthy confidence intervals for *one* coordinate of our estimator, or equivalently for the projection of our estimator on a pre-specified direction. In particular, some of the important and interesting problems pointed out in Bickel and Freedman (1983) disappear if we ask the types of questions we are interested in here, because our requirements are less stringent. We think that our requirements are the minimal ones a practitioner would require from the bootstrap or other resampling plans.

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*, a semi-parametric method where the bootstrapped data consists of $y_i^* = \widehat{\beta}' X_i + \hat{\epsilon}_i^*$, where $\hat{\epsilon}_i^*$ is drawn from the empirical distribution of the estimated residuals, e_i. Both of these methods are extremely

flexible for linear models regardless of the method of fitting β or the error distribution of the ϵ_i.

Contributions of the paper We show, via a mixture of simulation and theoretical study, that the performance of either of these bootstrap procedures for inference becomes highly suspect when the dimension of X_i's, given by p, grows with the sample size, n, in such a way that $p/n \to \kappa > 0$. In particular, pairs resampling becomes highly conservative, to the point of being non-informative, while residual resampling of the observed residuals becomes highly anti-conservative, wrongly rejecting the null hypothesis at very high rates. This is in sharp contrast to the setting where p is small relative to n, which is the context for many theoretical treatments of the bootstrap (see references above).

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 resolve the problem. We propose a different method of resampling, based on scaled leave-one-out predicted errors, that seems to perform better in our simulations.

For pairs bootstrapping we show that the expected bootstrap variance of the estimator is, even in simple cases, very different from the variance of the sampling distribution of the estimator. This can be explained in part by the fact that the spectral properties of weighted and unweighted high-dimensional sample covariance matrices are very different. We demonstrate that a different resampling scheme can alleviate the problems to a certain extent, but we also highlight the practical limitations in such a strategy, since it relies heavily on having strong knowledge about the data-generating model.

Finally, we briefly study an alternative resampling strategy, the jackknife, and also show that it misbehaves, similarly over-estimating the variance even in simple situations.

These results have important practical implications. Robust methods for regression based on alternative loss functions, such as L_1 or Huber loss, usually rely on resampling methods for inference. This is especially true in high-dimension where, until recently, there was essentially no theory about these estimators (see El Karoui et al. (2011)). Yet our results show that even in idealized settings, the bootstrap fails in high dimensions.

Why use the framework of $p/n \to \kappa \in (0,1)$? Several reasons motivate our theoretical study in this regime. From the standpoint of moderate-sized statistical analysis (i.e n and p of a similar order of magnitude but not extremely large), it is not always obvious whether the classical theoretical assumption that $p/n \to 0$ is justified, yet the assumption is known theoretically to have huge impact on the behavior of estimators. We think that working in the high-dimensional regime captures better the complexity encountered even in reasonably low-dimensional practice. In fact, asymptotic predictions based on the high-dimensional assumption can work surprisingly well in very low-dimension (see Johnstone (2001)). We also think that in these high-dimensional settings – where much is still unknown theoretically – the bootstrap is a intuitive alternative. Hence, it is natural to study how it performs in high-dimension in the simple problems we are starting to understand theoretically.

We first give some basic notation and background regarding the bootstrap and estimation of linear models in high dimensions.

1.1 Inference using the Bootstrap

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

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

where ρ defines the loss function for a single observation. ρ is assumed to be convex in all the paper. Robust regression often involves choosing the loss function ρ so as to be less sensitive to outliers or non-normal error distributions, as compared to ordinary least squares where $\rho(x) = x^2$. Common choices are $\rho(x) = |x|$, which defines L_1 regression, a.k.a Least-Absolute-Deviation (LAD), or Huber loss where $\rho(x) = x^2$ for $|x| < k$ and $\rho(x) = |x|$ for values greater than k. Bootstrap inference is particularly common in robust regression because asymptotic inference for these losses (when available) often will require more assumptions about the data-generating process than ordinary least squares – which may in turn defeat the desired goal of robustness.

Bootstrap methods are used in order to estimate the distribution of the estimate $\widehat{\beta}_\rho$ under the true distribution of the data, F. The bootstrap estimates this distribution with the distribution that

would occur if the data was in fact drawn from an estimate $\hat{F}$ of the data distribution. Following standard convention, we designate this estimate $\widehat{\beta}^*_\rho$ to note that this is an estimate of β using loss function ρ when the data distribution is known to be exactly equal to $\hat{F}$. Since $\hat{F}$ is completely specified, we can in principle exactly calculate the distribution of $\widehat{\beta}^*_\rho$ and use this distribution as an approximation of the distribution of $\widehat{\beta}_\rho$ under F. In practice calculation of the distribution of $\widehat{\beta}^*_\rho$ under $\hat{F}$ cannot be done explicitly. Instead we simulate B independent draws of size n from the distribution $\hat{F}$ and perform inference based on the distribution of $\widehat{\beta}^{*b}_\rho$, $b = 1, \ldots, B$.

In bootstrap inference for the linear model, there are two common methods for resampling, which results in different estimates $\hat{F}$. In the first method, called the residual bootstrap, $\hat{F}$ is an estimate of the conditional distribution of y_i given β and X_i. In this case, the corresponding resampling method is to resample ϵ^*_i from an estimate of the distribution of ϵ and form data $y^*_i = X_i'\widehat{\beta} + \epsilon^*_i$. This method of bootstrapping assumes that the linear model is correct for the mean of y; it is also assuming fixed X_i design points by sampling conditional on the X_i. In the second method, $\hat{F}$ 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); 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 robust than residuals resampling - see e.g Davison and Hinkley (1997).

1.2 High-dimensional inference of linear models

Recent research shows that $\widehat{\beta}_\rho$ has very different properties when p/n has a limit κ that is bounded away from zero than it does in the classical setting where $p/n \to 0$. A simple example is that the vector $\widehat{\beta}_\rho$ is no longer consistent in Euclidean norm. This has important impact on theoretical analysis of linear models, since traditional theoretical results come from perturbation analyses that assume that $\widehat{\beta}_\rho$ is close to β in Euclidean norm for large enough n – an assumption that does not hold true for high dimensional problems. We should be clear, however, that projections on fixed non-random directions, i.e $v'\widehat{\beta}_\rho$ are still $\sqrt{n}$ consistent. This includes estimates of each coordinate entry of $\widehat{\beta}_\rho$, meaning that in practice the estimate of $\widehat{\beta}_\rho$ is still a reasonable quantity of interest (moreover, the estimator is generally consistent in $\|\cdot\|_{2+\epsilon}$ for $\epsilon > 0$).

Another important impact of $\kappa > 0$ which is of particular interest for robust regression is that the optimal loss function ρ for a given error distribution is no longer given by the log-likelihood of the error distribution (Bean et al., 2013). For example, when the errors are double-exponential the optimal loss function in high dimensions is not the L_1 penalty, and in fact ordinary least squares has better performance than L_1, provided κ is large enough (Bean et al. (2013) gives an expression for the optimal loss-function under assumptions about the behavior of the design matrix).

The theoretical consistency of bootstrap estimates has been extensively studied, as mentioned above. 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 ($\kappa = 0$). For instance, in Mammen (1993), it is shown that under mild technical conditions and assuming that $p^{1+\delta}/n \to 0$, $\delta > 0$, the bootstrap distribution of linear contrasts $v'(\widehat{\beta}^* - \widehat{\beta})$ is in fact very close to the sampling distribution of $v'(\widehat{\beta} - \beta)$ with high-probability, when using least-squares. Other results such as Shorack (1982) and Mammen (1989), also allow for increasing dimensions for e.g linear contrasts in robust regression, by making assumptions on the diagonal entries of the hat matrix – which in our context would be true only if $p/n \to 0$ – hence those interesting results do not apply to the present study. We also note that Hall (1992) contains on p. 167 cautionary notes about using the bootstrap in high-dimension.

Directly relevant to the problem we study is Bickel and Freedman (1983). In that paper, the authors consider the least-squares problem, bootstrapping scaled residuals. They show (see Theorem 3.1 p.39 in Bickel and Freedman (1983)) that when $p/n \to \kappa \in (0, 1)$, there exists a direction v, such that $v'\widehat{\beta}^*$ does not have the correct distribution, i.e its distribution is not conditionally in probability close to the sampling distribution of $v'\widehat{\beta}$. 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 discuss below, we have less stringent requirements for the bootstrap to work and in the case of the design matrix X having i.i.d Gaussian entries, the diagonal entries of the hat matrix are actually close to a constant. Hence, our work complements the work of Bickel and Freedman (1983) and is not redundant with it.

Comment An important consideration in interpreting theoretical work on linear models in high dimensions is the role of the design matrix X. Unlike much classical theory, the assumptions in most theoretical results in the high dimensional setting are not stated as conditions of a specific design matrix X, but instead are assumptions that X is generated according to certain classes of distributions. Theoretical work in the literature usually allows for a fairly general class of distributions for the individual elements of X_i and handle covariance between the predictor variables. However, the X_i are generally considered i.i.d., which limits the ability of any X_i to be too influential in the fit of the model. El Karoui (2009) shows that many theoretical results in random matrix theory can be quite fragile to violations of these geometric properties; for example, simply scaling each X_i by a different value λ_i (hence getting an elliptical distribution) can alter the asymptotic properties of estimators (see also in different statistical contexts Diaconis and Freedman (1984); Hall et al. (2005)).

1.3 Results of Paper

The rest of the paper is laid out as follows. In Section 2 we demonstrate that in high dimensions bootstrapping the residuals – or even appropriately standardized residuals – results in extremely poor inference on β with error rates much higher than the reported Type I error. In Section 2.2 we give theoretical results that help explain this behavior and in 2.3 we introduce alternative bootstrap methods based on leave-one-out predicted errors that appear to be more robust in high dimensions. In Section 3 we examine bootstrapping pairs and show that bootstrapping the pairs has very poor performance as well but in the opposite direction. We prove in the case of L_2 loss, that the variance of the bootstrapped $\widehat{\beta}^*$ is greater than that of $\widehat{\beta}$, leading to the overly conservative performance we see in simulations. We propose some remedies for these problems based on our theoretical understanding and discuss the limits of our solutions. In Section 4, we discuss an alternative resampling technique, the jackknife estimate of variance, and we 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 by a factor of $n/(n-p)$; we briefly mention other corrections for other losses.

We focus throughout the exposition on inference of β_1 (the first element of β) as exemplary of a contrast of interest, rather than the entire vector of β (which in general in high dimensions does

not have a consistent estimator in L_2). Another reason to focus on this quantity is because we feel that the minimal requirement we can ask of an inferential method is to perform correctly on one coordinate of the parameter of interest or on any pre-specified contrast vector. This is a much less stringent requirement than doing well on complicated functionals of the whole parameter vector.

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 problems of heteroskedasticity, dependency, outliers and so forth that are known to be a problem for robust 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 Supplementary Text, Section S1.

Similarly, we focus on the basic implementations of the bootstrap for linear models. While there are many alternatives proposed – often for specific loss functions or for specific settings – the standard methods are most commonly used in practice. Furthermore, to our knowledge none of the alternatives specifically address the underlying theoretical problems that appear in high dimensions and therefore are likely to suffer from the same fate.

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 `rlm` package in R, where the transition from quadratic to linear behavior is at $x = 1.345$. 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 estimate of β using loss ρ; from this point on, however, we will often drop the ρ, unless for clarity we need to emphasize that the estimate could be from any loss ρ. We denote generically by $\kappa = \lim_{n\to\infty} p/n$. 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. 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 $\mathrm{var}(\epsilon_i) = \sigma_\epsilon^2$. e_i denotes the i-th residual, i.e $e_i = y_i - X_i'\widehat{\beta}$. $\tilde{e}_{i(i)} \triangleq y_i - X_i'\widehat{\beta}_{(i)}$ 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, we mean that

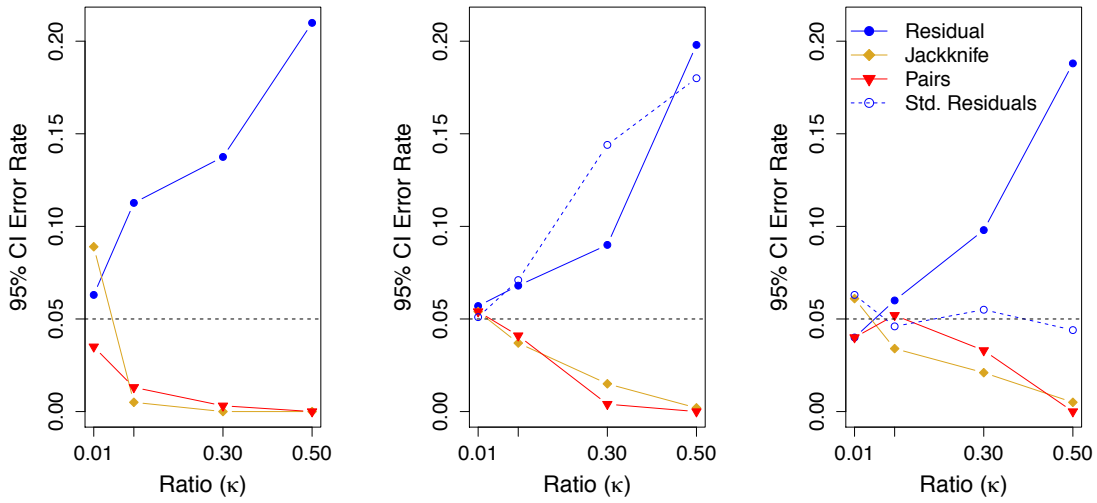
$X_i \overset{iid}{\sim} \mathcal{N}(0, \mathrm{Id}_p)$. Throughout the paper, the loss function ρ is assumed to be convex, $\mathbb{R} \mapsto \mathbb{R}^+$. We use the standard notation $\psi = \rho'$. We finally assume that ρ is such that there is a unique solution to the robust regression problem - an assumption that applies to all classical losses in the context of our paper.

2 Bootstrapping the error distribution

We first focus on the method of bootstrap resampling where $\hat{F}$ is the conditional distribution $y|\widehat{\beta}, X$. In this case the distribution of $\widehat{\beta}^*$ under $\hat{F}$ is formed by repeated resampling of ϵ_i^* from an estimate of the distribution of ϵ. 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 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 (L_2 loss) 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 ith diagonal element of the hat matrix. This problem becomes particularly pronounced in high dimensions. For instance, if $X_i \sim \mathcal{N}(0, \Sigma)$, $h_i = p/n + \mathrm{o}_P(1)$ 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 in much greater generality for the distribution of the design matrix X (see e.g Wachter (1978), Haff (1979), Silverstein (1995), Pajor and Pastur (2009), El Karoui (2010), El Karoui and Koesters (2011), where the main results of some of these papers require minor adjustments to get this result).

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 p/n increases the error rate of the confidence intervals in least squares regression increases well beyond the expected 5%: error rates of 10-15% for $p/n = 0.3$ and approximately 20% for $p/n = 0.5$ (Table S1). We see similar error rates for other robust methods, such as L_1 and Huber loss, and also for different error distributions and distributions of X (Supplementary Figures S1 and S2). We



(a) L_1 loss (b) Huber loss (c) L_2 loss

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). The error rates are based on 1,000 simulations, see the description in Supplementary Text, Section S1 for more details; exact values are given in Table S1. Error rates above 5% correspond to anti-conservative methods. Error rates below 5% correspond to conservative methods.

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 led more refined recommendation in the case of L_2 loss: 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 values 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 impacts the estimator only through σ_ϵ^2 (in the case of homoskedastic errors; we note that the standard representation of $\widehat{\beta} - \beta$ in the case of least-squares in connection with the Lindeberg-Feller Theorem (Stroock, 1993) show that $v'(\widehat{\beta} - \beta)$ is asymptotically normal in the situations we consider in this paper and many other).

However, this adjustment of the residuals is a correction specific to using a L_2 loss function. 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,

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

where $e'_j = e_j/s$, 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 regularity conditions)

Unlike the correction in L_2 loss, however, the above correction for the residuals is an approximate correction and the approximation depends on assumptions that that do not hold true in higher dimensions. The error rate of confidence intervals in our simulations based on this correction show no improvement in high dimensions over that of simple bootstrapping of the residuals, unlike that of L_2 (Figure 1). This could be explained by the fact that standard perturbation analytic methods used for the analysis of M-estimators in low-dimension fail in high-dimension (see El Karoui et al. (2013); El Karoui (2013) and compare to e.g van der Vaart (1998)).

2.2 Understanding the behavior of residual bootstrap

We can understand the behavior of residual bootstrapping in high dimensions better by making use of previous work on the behavior of robust estimation procedures in high dimensions (El Karoui et al., 2013; El Karoui, 2013).

At a high-level, 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. This is in general true both in terms of simple properties such as variance and in terms of

more general aspects, such as the whole marginal distribution. Let us now be more precise.

Let $\widehat{\beta}_{(i)}$ 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 $\tilde{e}_{j(i)} = y_j - X_j'\widehat{\beta}_{(i)}$

In general in high-dimension, the distribution of $\tilde{e}_{i(i)}$ is going to be more closely related to the distribution of the ϵ_i than that of e_i. In the above cited work, the authors gave asymptotic approximations to the distribution of $\tilde{e}_{i(i)}$ and a general asymptotic relationship between e_i and $\tilde{e}_{i(i)}$ for any sufficiently smooth loss function ρ and any size dimension where $p/n \to \kappa < 1$. There, X_i is assumed for simplicity of exposition to have an elliptical distribution, $X_i = \lambda_i \chi_i$, where $\chi_i \sim N(0, \Sigma)$, though similar results apply when $\chi_i = \Sigma^{1/2}\xi_i$, with ξ_i having i.i.d non-Gaussian entries, satisfying a few technical requirements.

For simplicity in restating their results, we will assume $\Sigma = \mathrm{Id}_p$, but equivalent statements can be made for arbitrary Σ. With this assumption, the relationship between e_i and the true error ϵ_i can be summarized as,

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

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

where Z_i is a random variable distributed $N(0,1)$ and independent of ϵ_i. u_n is a sequence of numbers tending to 0 - see El Karoui (2013) for details. The scalar c_i is given as $\frac{1}{n}\text{trace}\left(S_i^{-1}\right)$, 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; furthermore c_i 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 (see El Karoui et al. (2013)), using for instance the notion of sub-differential for ψ (Hiriart-Urruty and Lemaréchal, 2001).

These equations give theoretical underpinnings as to why bootstrap resampling of the residuals can perform so badly: the distribution of the e_i is far removed from that of the ϵ_i. The residuals e_i have a non-linear relationship with the predicted errors, which themselves are not distributed the same as ϵ but are a convolution of the true error distribution and an independent scale mixture of Normals.

The importance 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, $||\widehat{\beta}_{(i)} - \beta||_2$, both of which are highly dependent on the dimensions of the problem, as well as the distribution of the errors and 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 that $e_i = (1 - h_i)\tilde{e}_{i(i)}$. 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 $\mathbf{E}\left(\sum_{i=1}^{n} e_i^2\right) = \sigma_\epsilon^2(n - p)$, if 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. On the other hand, standardizing the residuals yield a distribution with variance σ_ϵ^2, i.e the correct variance, regardless of dimensionality (though of course the standardization factor is itself dimension-dependent). Because the performance of the least-squares estimator depends on the error distribution only through its variance, it is clear that this approach should fix the dimensionality issues for the problems we are considering. (For finer problems with the standardized-residual bootstrap, associated with more demanding inferential requirements for high-dimensional least-squares, we refer to (Bickel and Freedman, 1983))

The case of $p/n \to 0$: In this setting, $c_i \to 0$ and therefore the residuals e_i are approximately equal in distribution to the predicted errors ($\tilde{e}_{i(i)}$). Similarly, $\widehat{\beta}_\rho$ is L_2 consistent when $p/n \to 0$, and so $||\widehat{\beta}_{\rho(i)} - \beta||_2^2 \to 0$. This 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.

More general robust regression The situation is much more complicated for general robust regression estimators, for two reasons. First, as we have discussed above, the relationship between the residuals and the errors is very non-linear. Second, the systems described in El Karoui et al. (2013) show that the characteristics of the error distribution that impact $\|\widehat{\beta}_\rho - \beta\|_2$ go well beyond σ_ϵ^2. In particular, two error distributions with the same variance might yield $\widehat{\beta}_\rho$ with

quite different risks. Hence, simply rescaling the residuals should not in general result in a error distribution that will give $\widehat{\beta}^*_\rho$ with similar characteristics to that of $\widehat{\beta}_\rho$.

2.3 Approximating ϵ_i from scaled predicted errors

The relationship in Equation (3) suggests that an alternative for calculating the error distribution for bootstrapping would be to calculate the predicted errors, $\tilde{e}_{i(i)}$, estimate $|\lambda_i||\widehat{\beta}_{(i)} - \beta||_2$, and deconvolve the error term ϵ_i from the normal Z_i. Deconvolution problems are known to be very difficult (see Fan (1991), Theorem 1 p. 1260, with $1/\log(n)^\beta$ rates of convergence), and 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 $\hat{F}$ for the bootstrap procedure. This would be the case if the distribution of the deconvolved errors captured the key characteristics driving the performance of $\widehat{\beta}_\rho$ in the systems of equations described in El Karoui et al. (2013).

Proposal: bootstrapping from scaled $\tilde{e}_{i(i)}$ A simpler alternative is bootstrapping from the predicted error terms, $\tilde{e}_{i(i)}$, without deconvolution. On the face of it, this seems problematic, since Equation (3) demonstrates that $\tilde{e}_{i(i)}$ has the wrong distribution. However, we might ask how much practical effect does this have – is $\tilde{e}_{i(i)}$ close enough? Specifically, we standardize $\tilde{e}_{i(i)}$ so that its first two moments align with that of ϵ_i,

$$\tilde{r}_{i(i)} = \frac{\hat{\sigma}}{\widehat{var}(\tilde{e}_{i(i)})}\tilde{e}_{i(i)}, \tag{5}$$

where $\widehat{var}\tilde{e}_{i(i)}$ is an estimate of the variance of $\tilde{e}_{i(i)}$ and $\hat{\sigma}$ is an estimate of σ. Will such a transformation result in an estimate $\hat{F}$ of the distribution of ϵ_i that is sufficiently close for the purposes of estimating $\widehat{\beta}^*_\rho$ for bootstrapping inference?

Clearly a few cases exist where $\tilde{r}_{i(i)}$ should work well. We have already noted that as $p/n \to 0$, the effect of the convolution with the Gaussian disappears since $\widehat{\beta}_\rho$ is consistent, so that $\tilde{r}_{i(i)}$ are good estimates of ϵ_i (as are the original residuals e_i). Similarly, in the case of ϵ_i's having normal errors, $\tilde{e}_{i(i)}$ are also 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.

More surprisingly, as $p/n \to 1$ we find that using $\tilde{r}_{i(i)}$ gives equivalent estimates of $\widehat{\beta}_\rho$ as

when using ϵ_i. This is unexpected, since equation (3) shows that as $p/n \to 1$, $\tilde{r}_{i(i)}$ is essentially distributed $N(0, \|\widehat{\beta} - \beta\|_2^2)$, regardless of the original distribution of ϵ_i - though the distribution of ϵ_i could in principle influence $\|\widehat{\beta} - \beta\|_2^2$. (We note that in the asymptotics we consider $\|\widehat{\beta} - \beta\|_2$ can be treated as non-random.) We can understand this phenomena by looking at the risk of our estimate $\widehat{\beta}_\rho$ under different error distributions, with the idea that a minimal requirement for reasonable inference under our resampling distribution is that the risk be close to that of the risk under ϵ_i. El Karoui et al. (2013) resulted in a system of equations, the solution of which quantifies the risk of $\widehat{\beta}_\rho$ for any loss, error and and dimension (under some technical conditions detailed in El Karoui (2013)). Studying this system when $p/n \to 1$ shows that regardless of the true error distribution ϵ_i, the risk of $\widehat{\beta}_\rho$ will be equivalent so long as the variance of the distribution of ϵ_i's is the same.

Theorem 2.1. *Suppose we are working with robust regression estimators, and $p/n \to \kappa$. Assume for instance that the design is Gaussian. Then, under the assumptions stated in El Karoui (2013),*

$$\|\widehat{\beta}_\rho - \beta\|_2^2 \sim_{\kappa \to 1} \frac{\sigma_\epsilon^2}{1 - \kappa} ,$$

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

See Supplementary Text, Section S2 for the proof of this statement. Note that log-concave densities such as those corresponding to double exponential or Gaussian errors fall within the scope of this theorem.

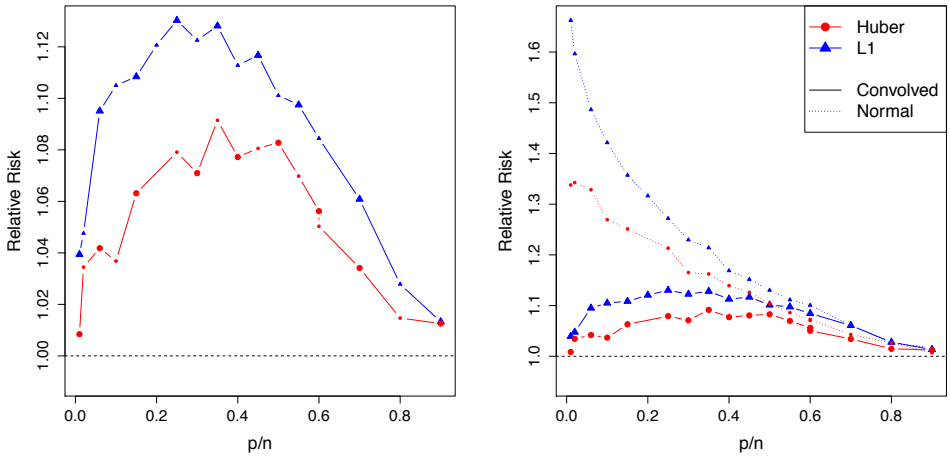
Note that the previous theorem implies that when p/n is close to 1, all robust regression estimators with ρ smooth enough and symmetric at 0 perform like the least-squares estimator, at least at the level of equivalence between two diverging sequences. Note also that the only characteristic of the error distribution that matters in this result is its variance. Hence, two error distributions that are different but have the same variance will result in estimators that perform roughly the same. We should then expect that bootstrapping from the predicted errors should perform reasonably well for p/n quite close to 1.

Thus, as $p/n \to 1$, $\tilde{r}_{i(i)}$ will diverge from the correct distribution, but inference of β will be increasingly less reliant on features of the distribution beyond the first two moments; and as $p/n \to 0$ the inference of β relies heavily on the distribution beyond the first two moments,

but the distribution of $\tilde{r}_{i(i)}$ approaches the correct distribution. Between these two extremes it is difficult to predict the tradeoff. In Figure 2 we plot the average risk of $\widehat{\beta}$ based on simulated data under the convolution distribution of $\tilde{r}_{i(i)}$ in Equation (3) relative to the true risk of $\widehat{\beta}$ under ϵ_i (with accurate risk of $\widehat{\beta}$ again being a proxy for the performance of inference procedures). In these simulations, ϵ_i's have a double exponential, aka Laplacian, distribution. As expected, we see that for large p/n both the convolution and a pure normal distribution with the right variance converge to the true risk. As $p/n \to 0$, the risk of a normal distribution with the correct variance diverges dramatically from the true risk while that of the convolution approaches the truth. We see that for Laplacian errors, L_1 and Huber (which is Huber$_1$ in this simulation) both show that the risk using $\tilde{r}_{i(i)}$ converges to the true risk on the extremes but varies from the truth in the range of $0.2 - 0.5$. For Huber, the divergence is at most 8%, but the difference is larger for L_1 (12%), probably due to the fact that normal error has a larger effect on the risk for L_1.

Performance in bootstrap inference To further quantify the performance of these error distributions, we implement bootstrap procedures for both the deconvolution of the predicted errors and the standardized predicted errors. Both methods require an estimator of σ that is consistent irregardless of dimension and error distribution. As we have seen, we cannot generally rely on the observed residuals e_i nor on $\tilde{e}_{i(i)}$. The exception is L_2, where the standard estimate of σ_ϵ^2 that includes a degrees of freedom correction is always a consistent estimator of σ, assuming i.i.d errors and mild moment requirements. In what follows, therefore, we estimate σ using the standard estimate of σ based on an L_2 fit. We calculate predicted errors by manually rerunning fits leaving out the ith residual; in practice equation (3) gives an approximation that could be used for those e_i where ρ is twice differentiable to speed up calculation.

For the deconvolution strategy, we used the deconvolution algorithm in the `decon` package in R (Wang and Wang, 2011) to estimate the distribution of ϵ_i. For simplicity in reusing existing bootstrapping code, for each simulation we draw a single sample from this distribution and then bootstrap the errors from this draw. The deconvolution algorithm requires the value of the variance of the Gaussian that is convolved with the ϵ_i, which we estimate as $\widehat{var}(\tilde{e}_{i(i)}) - \hat{\sigma}^2$. We note that this makes assumptions of homoskedastic errors, which is true in our simulations but may not be true in practice. We further note that we used a Gaussian design, and hence did not have to estimate λ_i's in the notation of Equation (3). In El Karoui (2010), the author proposed estimators

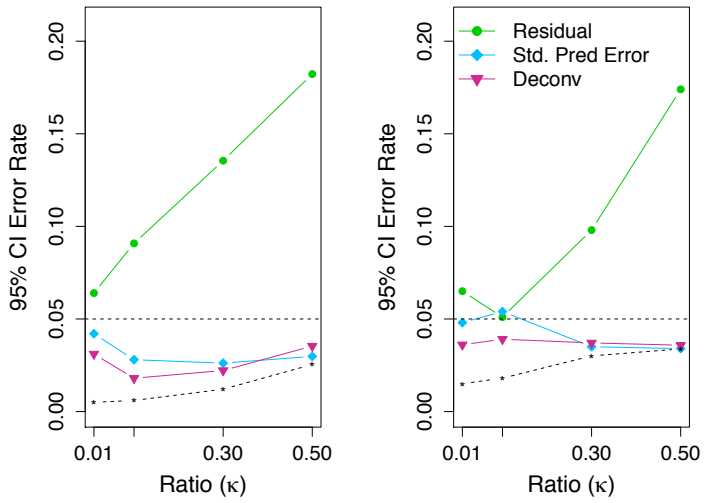


(a) (b)

Figure 2: **Relative Risk of $\widehat{\beta}$ for scaled predicted errors vs original errors - population version: (a)** Plotted with a solid line is the ratio of the average of $\|\widehat{\beta}_\rho - \beta\|_2$ under the "convolution error distribution", i.e errors of the form $\eta_i = \sigma_\epsilon(\epsilon_i + \gamma Z_i)/\sqrt{\sigma_\epsilon^2 + \gamma^2}$, where $Z_i \sim \mathcal{N}(0,1)$, independent of ϵ_i, and $\gamma = \mathbf{E}\left(\|\widehat{\beta}_\rho - \beta\|_2\right)$ - computed using the "correct" error distribution, i.e errors ϵ_i - to the average of $\|\widehat{\beta}_\rho - \beta\|_2$ under the "true" error distribution i.e, errors ϵ_i. This latter quantity is of course γ. **(b)** Added to the plot in (a) is the relative risk of $\widehat{\beta}(\rho)$ for errors $\eta_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$ vs the original errors ϵ_i (dotted line). 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. The true error distribution is the standard Laplacian distribution ($\sigma_\epsilon^2 = 2$). Each simulation uses the standard estimate of σ_ϵ^2 from the generated ϵ_i's. γ was computed using a first run of simulations using the "true" error distribution. The Huber loss in this plot is Huber_1 and not the default `R` Huber, which is $\text{Huber}_{1.345}$.

for these quantities, which could then be used in the `decon` package.

In Figure 3 we show the error rate of confidence intervals based on bootstrapping from the standardized predicted errors and from the deconvolution estimates of ϵ_i. We see that both methods control the Type I error, unlike bootstrapping from the residuals, and that both methods are conservative. There is little difference between the two methods with this sample size, though with $n = 100$, we observe the deconvolution performance to be worse in L_1 (data not shown). The deconvolution strategy, however, must depend on the distribution of the design matrix; for elliptical designs, 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



(a) L_1 loss (b) Huber loss

Figure 3: **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 (blue) and from deconvolution of predicted error (magenta). We demonstrate its improvement over the standard residual bootstrap (green) for (a) L_1 loss and (b) Huber loss. The error distribution is double exponential and the design matrix X is Gaussian, but otherwise the simulations parameters are as in Figure 1. The error rates on confidence intervals based on bootstrapping from a $N(0, \hat{\sigma}^2)$ (dashed curve) are as a lower bound on the problem. For the precise error rates see Table S3.

$p/n = 0.5$). This was not the case for the bootstrap using standardized predicted errors, where the Type I error for an elliptical design was only slightly higher than the target 0.05 (around 0.07, data not shown).

3 Bootstrapping the joint distribution of (y_i, X_i)

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 robust method of bootstrapping, allowing for the linear model to be incorrectly specified. It is also considered to be slightly more conservative compared 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 more closely 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, 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.

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 S1, S2). In L_1, even ratios as small as 0.1 are incredibly conservative, with the error rate dropping to less than 0.01. For Huber and L_2 loss, the severe loss of power in our simulations starts for ratios of 0.3 (see Tables S1,S5, S4).

A minimal requirement for the distribution of the bootstrapped data to give reasonable inferences is that the variance of the bootstrap estimator $\widehat{\beta}^*$ needs to be a good estimate of the variance of $\widehat{\beta}$. In the case of high dimensions, this is not the case. In Figure 5 we plot the ratio of the variance of $\widehat{\beta}^*$ to the variance of $\widehat{\beta}$ 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 bootstrap variance roughly overestimates the correct variance by a factor 1.3 in the case of L_2-regressions; for Huber and L_1 the bootstrap variance is roughly twice as large as it should be (Table S7).

In the simple case of L_2, we can further quantify this loss in power by comparing the size of the bootstrap confidence interval to the correct size 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 (see 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 is still 80% larger than needed; for the elliptical distributions we simulated, the width was as much as 3.5 times larger than the correct confidence interval. Furthermore, as

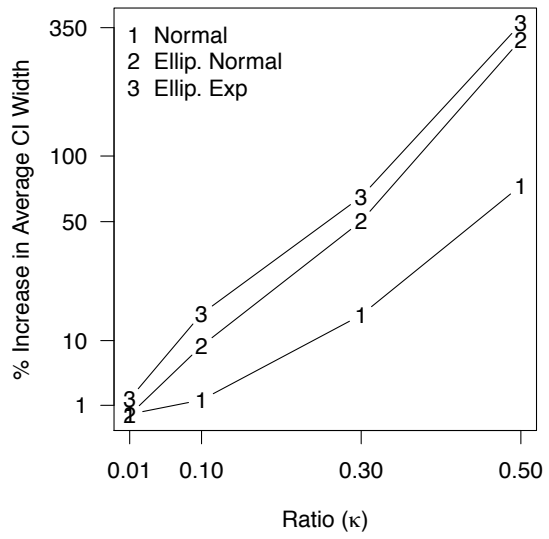


Figure 4: **Comparison of width of 95% confidence intervals of β_1 for L_2 loss:** Here we demonstrate the increase in the width of the confidence interval due to pairs bootstrapping. Shown on the y-axis is the percent increase of the average confidence interval width based on simulation ($n = 500$), 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 S2.

we can see in Figure 1, L_2 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 a limited but interesting setting, we can understand the properties of the (bootstrap) variance of $v'\widehat{\beta}^*$, and in particular its expectation. Define $\widehat{\beta}_w$ as the result of regressing y on X with random weight w_i for each observation (y_i, X_i). We call $\widehat{\beta}^*_w$ the resampled version of $\widehat{\beta}$ where we use random weights $\{w_i\}_{i=1}^n$ in the resampling of the pairs (y_i, X_i). Assuming the weights are independent of $\{y_i, X_i\}_{i=1}^n$, we have the equality in law $\widehat{\beta}^*_w = \widehat{\beta}_w | \{y_i, X_i\}_{i=1}^n$. For the standard pairs bootstrap, the estimate $\widehat{\beta}^*$ from a single resampling of the pairs is equivalent to $\widehat{\beta}^*_w$, where w is drawn from a multinomial distribution with expectation $1/n$ for each entry.

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 3.1. *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 $\mathcal{N}(0, \mathrm{Id}_p)$, 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 is, asymptotically as* $n \to \infty$

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

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

For a proof of this theorem and a consistent estimator of this expression, see Supplementary Text, Section S3. 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 expected bootstrap variance is non-negative.

In 3.1.2 below, we discuss possible extensions of this theorem, such as different design matrices or correlated predictors. Before doing so, we first will discuss the implications of this result to pairs bootstrapping.

3.1.1 Application to Pairs Bootstrapping

In the standard pairs bootstrap, the weights are actually 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 $w_i > 0$. We refer the reader to the corresponding discussion in El Karoui (2010) for explanations on how to handle these two problems in our context. Effectively, the previous result still holds in these conditions, provided $(1-1/e) > \kappa = \lim p/n$. For simplicity in the discussions that follow, we analyze the case $w_i \overset{iid}{\backsim}$ Poisson(1) which is asymptotically equivalent to the Multinomial$(n, 1/n)$ for the quantities of interest to us.

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

Relation to the performance of the bootstrap for inference The formula in Theorem 3.1 allows us to relate the expected variance of the bootstrap estimator and the variance of the least-squares estimator in the Gaussian design setting, which is asymptotically $\kappa/(1-\kappa)\sigma_\epsilon^2$ (relying on simple Wishart computations (Haff, 1979; Mardia et al., 1979) or random matrix theory). If our bootstrap worked for estimating the variance of the quantity of interest, we should have, at least

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

and hence should have

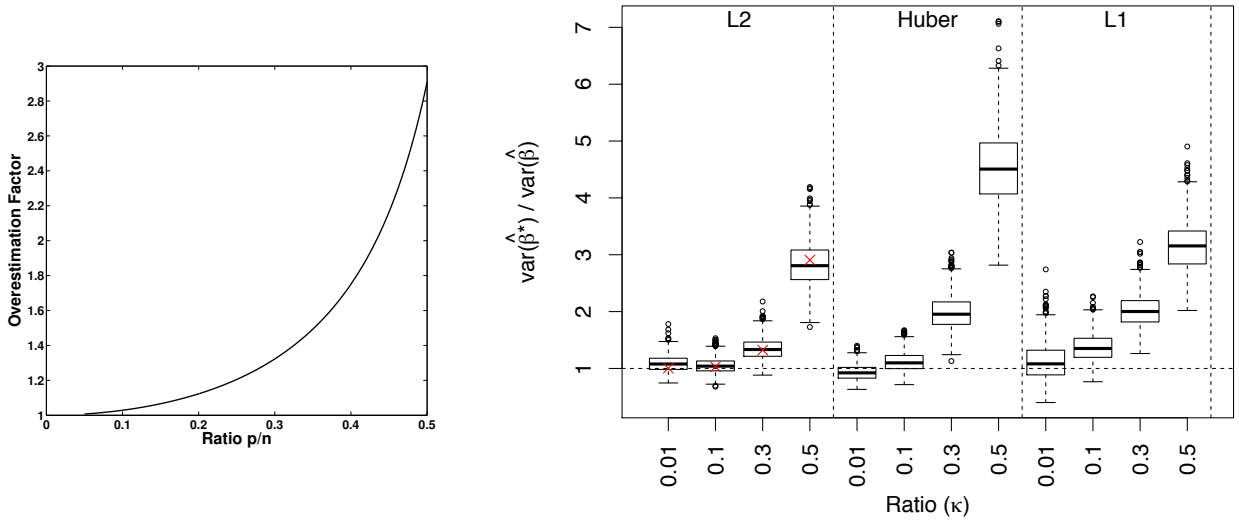
$$\mathbf{E}\left(\frac{1}{(1+cw_i)^2}\right) = \frac{1-\kappa}{1+\kappa} \, .$$

This is not the case for most weight distributions and in particular is not numerically verified for the Poisson(1) distribution that corresponds to the standard pairs bootstrap. See Figure 5a for a visual depiction of the problem, where we see that the theory 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.

3.1.2 Extensions of Theorem 3.1

Case of $\Sigma \neq \mathrm{Id}_p$ The first extension we might think of is to move from the case where $\Sigma = \mathrm{Id}_p$ to general Σ. In the case of Gaussian design, it is clear, when $p < n$, that all the results apply there, too, simply by replacing v by $\Sigma^{-1/2}v$ everywhere in the theoretical analysis. Since our corrections are multiplicative, they apply in this setting, too. So the formula involving only the weight distribution holds for $\Sigma \neq \mathrm{Id}_p$, provided Σ is positive definite, i.e all its eigenvalues are strictly positive. Based on the computations in the proof provided in the Supplementary Text, it is also possible to get a consistent estimator of this quantity when $\Sigma \neq \mathrm{Id}_p$ - however, this is a bit beside the point of our discussion here and we do not discuss the matter further.

Case of elliptical design In light of previous work on model robustness issues in high-dimensional statistics (see e.g El Karoui (2009), El Karoui (2010)), it is natural to ask whether the central results of Theorem 3.1 still apply when $X_i \overset{\mathcal{L}}{=} \lambda_i Z_i$, with λ_i a random variable independent of Z_i, and $Z_i \sim \mathcal{N}(0, \mathrm{Id}_p)$. We require $\mathbf{E}(\lambda_i^2) = 1$ so that $\mathrm{cov}(X_i) = \mathrm{Id}_p$, as in the assumptions



(a) L_2 (Theoretical) (b) All (Simulated)

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 3.1 using Poisson(1) weights to the asymptotic variance $\kappa/(1-\kappa)\sigma_\epsilon^2$. (b) boxplots of the ratio of the bootstrap variance of $\widehat{\beta}_1^*$ to the variance $\widehat{\beta}_1$, as calculated over 1000 simulations (i.e. $\mathrm{var}\left(\widehat{\beta}\right)$ is estimated across simulated design matrices X, and not conditional on X). The theoretical prediction for the mean of the distribution from Theorem 3.1 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 Supplementary Table S7.

of Theorem 3.1. The types of computations done in the proof can again be performed in that situation, though they become a bit more complicated. In light of the fact that even in the simpler case of the Gaussian design the corrections are very sensitive and hence not terribly robust (see discussion in Section 3.2), we do not present these computations here. We also note that a simple change of variables in the original formulation of the least squares problem show that understanding the elliptical situation is equivalent to understanding the situation of Gaussian design with heteroskedastic errors. (Effectively, the errors ϵ_i can be thought of as being replaced by $\tilde{\epsilon}_i = \epsilon_i/\lambda_i$, the weights w_i being replaced by $w_i\lambda_i^2$.)

Going beyond the Gaussian design As explained in several papers in random matrix theory, a number of the quantities appearing in our theorems will converge to the same limit when i.i.d Gaussian predictors are replaced by i.i.d predictors with mean 0 and variance 1 and "enough

moments". Hence, the results we present here should be fairly robust to changing normality assumptions to i.i.d-ness assumptions for the entries of the design matrix X. However, as has been explained in previous works - see e.g Diaconis and Freedman (1984); Hall et al. (2005); El Karoui (2009) - the assumption of Gaussian (or i.i.d) entries for the predictors is very limiting geometrically. It basically implies that the predictors X_i's are almost orthogonal to each other and that $\|X_i\|/\sqrt{p}$ is almost constant across i's, i.e the predictors live near a sphere. Ellipticity is a natural way to break this geometric constraint, but arguably does not capture an extremely wide variety of models.

3.2 Alternative weight distributions for resampling

The formula given in Theorem 3.1 suggests that resampling from a distribution $\hat{F}$ defined using weights other than i.i.d Poisson(1) (or, equivalently for our asymptotics, Multinominal(n,1/n)) should give us better bootstrap estimators. In fact, we should require, at least, that the bootstrap expected variance of these estimators match the correct variance $\kappa/(1-\kappa)\sigma_\epsilon^2$ (for the Gaussian design).

We note that if we use $w_i = 1,\ \forall i$, the bootstrap variance will be 0, since we are not changing anything to the problem between each bootstrap repetition. On the other hand, we have seen that with $w_i \sim \text{Poisson}(1)$, the expected bootstrap variance was too large compared to $\kappa/(1-\kappa)\sigma_\epsilon^2$.

Because the formula in Theorem 3.1 appears fairly "continuous" in the weight distribution, we tried to find a parameter α such that if

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

the expected bootstrap variance would match the theoretical value of $\kappa/(1-\kappa)\sigma_\epsilon^2$. The rationale for this approach is that for $\alpha = 0$, the expected bootstrap variance is 0 and for $\alpha = 1$, the expected bootstrap variance is greater than the target value of $\kappa/(1-\kappa)\sigma_\epsilon^2$, as we saw working on the standard bootstrap.

We solved numerically this problem to find $\alpha(\kappa)$ (see Supplementary Table S6 and Supplementary Text, Subsection S3.1 for details of computation). We then used these values and performed bootstrap resampling using the weights defined in Equation (6). We evaluated bootstrap

	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 containing 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.

estimate of $var(\widehat{\beta}_1)$ as well as the confidence interval coverage of the true value. We find that this adjustment of the weights in estimating $\hat{F}$ restores the performance of the bootstrap estimates, resulting in accurate estimates of variance and appropriate levels of confidence interval coverage (Table 1).

Nonetheless, finding a good weight distribution to use in resampling requires knowing a great deal about the distribution of the design matrix or making many assumptions about the design matrix (see Subsubsection 3.1.2 for more on this). Another issue is the fact that small changes in the choice of α can result in fairly large changes in $\mathbf{E}\left(\text{var}\left(v'\widehat{\beta}_w|X,\epsilon\right)\right)$. For instance, for $\kappa = .5$, using $\widetilde{\alpha}(\kappa) = .95$ (fairly close to the correct value for $\kappa = .3$ and arguably pretty close to .92, the correct value for $\kappa = .5$), results in an expected bootstrap variance roughly 30% larger than $\kappa/(1-\kappa)\sigma_\epsilon^2$.

We think that with some more work, we could extend the results of Theorem 3.1 to the case of robust regression, with independent errors and most likely in the elliptical design case. If such results were obtained, then it would be in principle possible to find new weight distributions to use in bootstrapping. Those would of course be dependent on dimensionality, properties of the design and the error distributions as well as the loss function. In other words, they would be strongly dependent on assuming a particular statistical model - in which case asymptotic analysis is possible. This goes against the very premise of using resampling methods, which are advocated for situations where the statistician is not willing to assume a statistical model.

Hence the work we just presented on finding new weight distributions for bootstrapping gives

a proof of principle that this type of resampling ideas could be used 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 technique works for a wide variety of models, and hence is trustworthy in a broad variety of situations.

4 The Jackknife

In the context we are investigating, where we know that the distribution of $\widehat{\beta}_1$ is asymptotically normal (see arguments in El Karoui et al. (2013)), it is natural to ask whether we could simply use the jackknife to estimate the variance of $\widehat{\beta}_1$. The jackknife relies on leave-one-out procedures to estimate $var(\widehat{\beta}_1)$, only the estimate of variance is based on $\widehat{\beta}_{(i)}$: for a fixed vector v,

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

where $\tilde{\beta}$ is the mean of the $\widehat{\beta}_{(i)}$. The case of $\widehat{\beta}_1$ corresponds to picking $v = e_1$, i.e the first canonical basis vector.

Given the problems we just documented with the pairs bootstrap, it is natural to ask whether the jackknife fixes some of them, even though the jackknife is known to have its own problems (Efron (1982) or Koenker (2005), p.105). We note that at first glance, the results and expansions in El Karoui et al. (2013) and El Karoui (2013) might suggest that the jackknife is more robust to dimensionality issues than the standard pairs bootstrap for instance, since the leave-one-out ideas used in these papers are relatively robust to dimensionality. This is in sharp contrast with standard perturbation-analytic methods used to derive central limit theorems for standard M-estimators (see e.g van der Vaart (1998)) which can be used to show the validity of the bootstrap in lower dimensions that we consider here (see Mammen (1989), p. 385).

Perhaps surprisingly in light of these arguments, it turns out that the jackknife estimate of variance performs quite poorly. Again, the jackknife overestimates the variance of $v'\widehat{\beta}$ leading to extremely poor inference (Figure 1). 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 (Figure 6 and Supplementary table S7). In the case of L_1

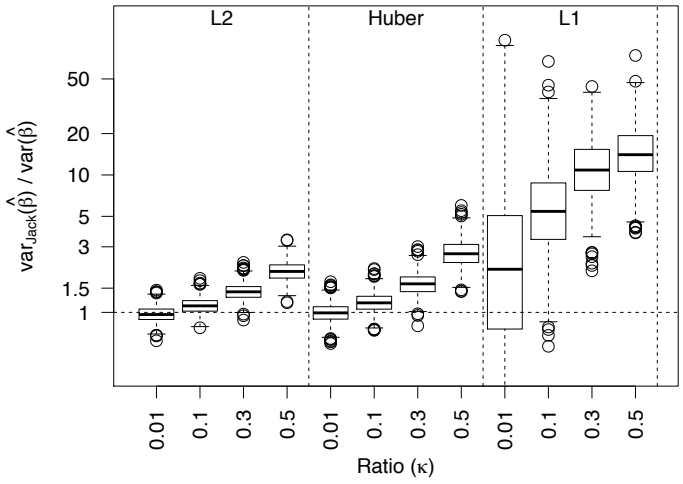


Figure 6: **Factor by which jackknife over-estimates the variance:** boxplots of the ratio of the jackknife estimate of the variance $\widehat{\beta}_1$ to the variance of $\widehat{\beta}_1$ as calculated over 1000 simulations. Simulations were with normal design matrix X and normal error ϵ_i with values 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 Supplementary Table S7.

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, 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.

Again, in the case of least-squares regression, we can theoretically evaluate the behavior of the jackknife. The proof of the following theorem is given in the supplementary material, Section S4.

Theorem 4.1. *Let us call varJACK the jackknife estimate of variance of $\widehat{\beta}_1$, the first coordinate of $\widehat{\beta}$. Suppose the design matrix X is such that $X_{i,j} \sim \mathcal{N}(0,1)$. Suppose $\widehat{\beta}$ is computed using least-squares and the errors ϵ have a variance. Then we have, as $n, p \to \infty$ and $p/n \to \kappa < 1$,*

$$\frac{\mathbf{E}\left(\textit{varJACK}\right)}{\mathrm{var}\left(\widehat{\beta}_1\right)} \to \frac{1}{1-\kappa}\,.$$

The same result is true for the jackknife estimate of variance of $v'\widehat{\beta}$, for any deterministic vector v with $\|v\|_2 = 1$.

We note that the proof given in the supplementary material and previous results such as those

in El Karoui (2010) and El Karoui (2013) show that a similar analysis could be carried out in the case of elliptical X_i. This would result in a limiting result involving both the dimensionality factor κ and the distribution of the elliptical factors. It is also clear, since all these results 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).

We illustrate in Figure 7 the fact that correcting the jackknife estimate of variance by multiplying it by $1 - p/n$ yields correct confidence intervals for the setting of our theorem (normal X design matrix, L_2 loss). However, it is not a fix for all situations. In particular when the X matrix follows an elliptical distribution the correction of $1 - p/n$ from Theorem 4.1 gives little improvement even though the loss is still L_2, which demonstrates the sensitivity of the result on the assumptions on X (or more generally the geometric characteristics of the design matrix).

Corrections for more general settings For more general design distributions and loss functions, 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 the asymptotic limit, when $\Sigma = \mathrm{Id}_p$, preliminary heuristic calculations suggest to correct varJACK by dividing it by the factor

$$\text{correction Factor} = \frac{\text{trace}\left(\mathcal{S}^{-2}\right)/p}{\left[\text{trace}\left(\mathcal{S}^{-1}\right)/p\right]^2}\ . \tag{7}$$

Note that this conforms to our result in Theorem 4.1.

Equation (7) 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 $\psi'(x) = 1_{x=0}$.

We rescale the jackknife estimate of variance by numerically calculating the expected value

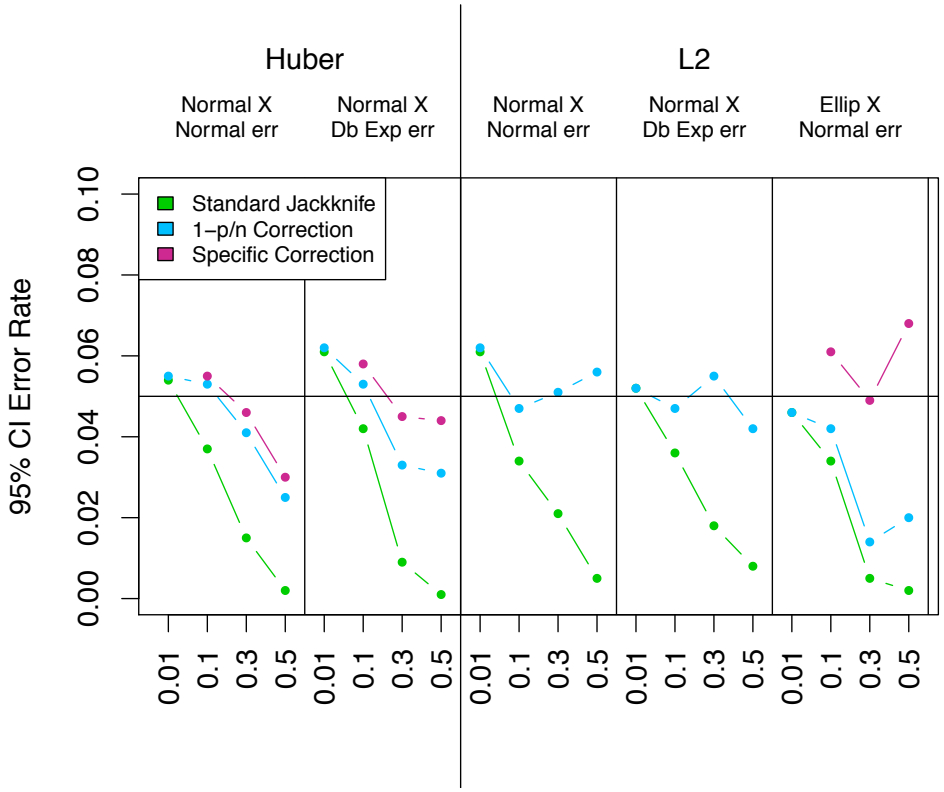


Figure 7: **Rescaling jackknife estimate of variance:** Shown are the error rates for confidence intervals for different re-scalings of the jackknife estimate of variance: the standard jackknife estimate (green); re-scaling using $1 - p/n$ as given in Theorem 4.1 for the L_2 case with normal design matrix X (blue); and re-scaling based on the heuristic in equation (7) for those settings not covered by the assumptions of Theorem 4.1 (magenta).

of the right hand side of Equation (7) and scaling the jackknife estimates of variance (Figure 7) (theoretical considerations suggest that for nice enough ρ's, the correction factor in Equation (7) is asymptotically deterministic and hence close to its mean).

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 confidence intervals. For the Huber loss (with $k = 1.345$, the default in R), it is less clear. We see a definite improvement as compared to the standard jackknife estimate, but the variance appears to be still too large leading to somewhat conservative inference. The improvement over the simpler correction of $1 - p/n$ is also not clear, perhaps owing to the fact that this Huber loss is not that different from the squared error loss.

It should be noted that the quality of the correction seem to depend on how smooth ψ is. In particular, even using the previous interpretations, the correction does not perform well for L_1 (at

least for $n = 1000$ and $\kappa = .1, .3, .5$, 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.

4.1 Extensions of Theorem 4.1 and case of $\Sigma \neq \mathrm{Id}_p$

As discussed in the pairs bootstrap case, some more technical work based on previously published techniques should permit extensions of the results to the case of $X_{i,j}$ being i.i.d with a "nice" mean 0, variance 1 distribution.

The case of $\Sigma \neq \mathrm{Id}_p$ is also tractable. For least squares, the correction factor is unchanged, since it is effectively enough to replace $v = e_1$ by $u_1 = \Sigma^{-1/2} v_1$ everywhere, and consider the problem of the Jackknife estimate of variance for $u_1' \widehat{\beta}$ in the null case $\Sigma = \mathrm{Id}_p$. The situation is more complicated for robust regression estimates. The value of the correction factor is the same regardless of Σ - however the expression we gave above depends on Σ being the identity and will not work in general when $\Sigma \neq \mathrm{Id}_p$. With a bit of technical work, it is fairly easy to modify the correction factor to estimate it from the data when $\Sigma \neq \mathrm{Id}_p$, but since this is quite tangential to the main thrust of our interests in this paper, we leave this for future work.

5 Conclusion

In this paper, we studied various resampling plans in the high-dimensional setting where p/n is not close to zero. One of our main findings is that different types of widely-used bootstraps will yield either conservative or anti-conservative confidence intervals. This is in sharp contrast to the low-dimensional setting where p is fixed and $n \to \infty$ or $p/n \to 0$. In effect, practitioners are left quite unsure of the statistical properties of these resampling methods, even in very simple settings.

Under various assumptions underlying our simulations, we explained theoretically the phenomena we were observing in our numerical work. At a high-level, the failure of the residuals-bootstrap can be explained by the fact that the residuals tend to have a very different marginal distribution in the setting we consider than the true errors. For the pairs bootstrap, an explanation comes from the fact that spectral properties of high-dimensional weighted covariance matrices

are very different from those of unweighted covariance matrices. Once again this is not the case in low-dimension.

Under those assumptions, the various resampling plans can essentially be fixed to give confidence intervals with approximately correct coverage probability. We note however that under these simple assumptions, asymptotic theory has been developed that can be used to create confidence intervals without relying on resampling (El Karoui et al. (2013); Bean et al. (2013); El Karoui (2013)). We also note that these corrections tend to be based on certain non-trivial properties of the design matrix - hence they violate the tenets of resampling approaches which promise a simple and universal numerical method to get accurate solutions to a broad class of problems. A possible exception is that of resampling the standardized predicted errors, which continued to perform reasonably well for a variety of simulation settings.

We note that we have not done a completely exhaustive analysis of the many problem-specific bootstrap methods that have been proposed in the huge literature on resampling - though we have not seen corrections based on dimensionality before. We have nonetheless tried to use many popular methods which are widely cited and are commonly described in research monographs. Furthermore, we have tried more complicated ways to build confidence intervals than the simple ones (e.g. bias correction methods), but have found them to be erratic in high-dimension.

One critique of the practical value of our work on the bootstrap is that in high dimensions, many practitioners might prefer sparse techniques rather than the standard regression techniques we outline here. We would first note even for fairly low ratios of $p/n = 0.1$, we see some degradation in performance of the bootstrap, for example for L_1 regression or even in L_2 when the design matrix X is not Gaussian. Furthermore, the problematic characteristics of the bootstrap that we see are likely to extend into many other settings of high dimensional multivariate analysis.

Hence, the conclusion of our paper is quite unsettling in that it is very unclear how standard resampling approaches perform in even moderately high-dimension, especially in situations that are beyond the current reach of theoretical arguments. Those situations are of course the very setting where using resampling techniques makes strong practical sense, making the problem even more acute. The conclusion of our analyses is that resampling techniques tend to be unreliable and perform very poorly in even simple problems where we can check their performance against a benchmark.

SUPPLEMENTARY MATERIAL

Supplementary Text More detailed description of simulations and proofs of the theorems stated in main text (pdf)

Supplementary Figures Supplementary Figures referenced in the main text (pdf)

Supplementary Tables Supplementary Tables referenced in the main text (pdf)

References

BEAN, D., BICKEL, P. J., EL KAROUI, N., and YU, B. (2013). Optimal m-estimation in high-dimensional regression. *Proceedings of the National Academy of Sciences* **110**, 14563–14568. URL http://www.pnas.org/content/110/36/14563.abstract.

BERAN, R. and SRIVASTAVA, M. S. (1985). Bootstrap tests and confidence regions for functions of a covariance matrix. *Ann. Statist.* **13**, 95–115.

BICKEL, P. J. and FREEDMAN, D. A. (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.* **9**, 1196–1217. URL http://links.jstor.org/sici?sici=0090-5364(198111)9:6<1196:SATFTB>2.0.CO;2-R&origin=MSN.

BICKEL, P. J. and FREEDMAN, D. A. (1983). Bootstrapping regression models with many parameters. In *A Festschrift for Erich L. Lehmann*, Wadsworth Statist./Probab. Ser., pp. 28–48. Wadsworth, Belmont, Calif.

BICKEL, P. J., GÖTZE, F., and VAN ZWET, W. R. (1997). Resampling fewer than n observations: gains, losses, and remedies for losses. *Statist. Sinica* **7**, 1–31. Empirical Bayes, sequential analysis and related topics in statistics and probability (New Brunswick, NJ, 1995).

CHERNICK, M. R. (1999). *Bootstrap Methods: A Practitioner's Guide*. Wiley.

DAVISON, A. C. and HINKLEY, D. V. (1997). *Bootstrap methods and their application*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.

DIACONIS, P. and FREEDMAN, D. (1984). Asymptotics of graphical projection pursuit. *Ann. Statist.* **12**, 793–815. URL http://dx.doi.org/10.1214/aos/1176346703.

EATON, M. L. and TYLER, D. E. (1991). On Wielandt's inequality and its application to the asymptotic distribution of the eigenvalues of a random symmetric matrix. *Ann. Statist.* **19**, 260–271.

EFRON, B. (1979). Bootstrap methods: another look at the jackknife. *Ann. Statist.* **7**, 1–26. URL http://links.jstor.org/sici?sici=0090-5364(197901)7:1<1:BMALAT>2.0.CO;2-6&origin=MSN.

EFRON, B. (1982). *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.

EFRON, B. and TIBSHIRANI, R. J. (1993). *An introduction to the bootstrap*, volume 57 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, New York.

EL KAROUI, N. (2009). Concentration of measure and spectra of random matrices: Applications to correlation matrices, elliptical distributions and beyond. *The Annals of Applied Probability* **19**, 2362–2405.

EL KAROUI, N. (2010). High-dimensionality effects in the Markowitz problem and other quadratic programs with linear constraints: risk underestimation. *Ann. Statist.* **38**, 3487–3566. URL http://dx.doi.org/10.1214/10-AOS795.

EL KAROUI, N. (2013). Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: rigorous results. *arXiv:1311.2445* ArXiv:1311.2445.

EL KAROUI, N. (2013). On the realized risk of high-dimensional markowitz portfolios. *SIAM Journal in Financial Mathematics* **4(1)**, http://dx.doi.org/10.1137/090774926.

EL KAROUI, N., BEAN, D., BICKEL, P., LIM, C., and YU, B. (2011). On robust regression with high-dimensional predictors. Technical Report 811, UC, Berkeley, Department of Statistics. Originally submitted as manuscript AoS1111-009. Not under consideration anymore.

EL KAROUI, N., BEAN, D., BICKEL, P. J., LIM, C., and YU, B. (2013). On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences* URL http://www.pnas.org/content/early/2013/08/15/1307842110.abstract.

EL KAROUI, N. and KOESTERS, H. (2011). Geometric sensitivity of random matrix results: consequences for shrinkage estimators of covariance and related statistical methods. *Submitted to Bernoulli* Available at arXiv:1105.1404 (68 pages).

FAN, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *Ann. Statist.* **19**, 1257–1272. URL http://dx.doi.org/10.1214/aos/1176348248.

HAFF, L. R. (1979). An identity for the Wishart distribution with applications. *J. Multivariate Anal.* **9**, 531–544. URL http://dx.doi.org/10.1016/0047-259X(79)90056-3.

HALL, P. (1992). *The bootstrap and Edgeworth expansion*. Springer Series in Statistics. Springer-Verlag, New York. URL http://dx.doi.org/10.1007/978-1-4612-4384-7.

HALL, P., MARRON, J. S., and NEEMAN, A. (2005). Geometric representation of high dimension, low sample size data. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67**, 427–444.

HIRIART-URRUTY, J.-B. and LEMARÉCHAL, C. (2001). *Fundamentals of convex analysis*. Grundlehren Text Editions. Springer-Verlag, Berlin. Abridged version of ıt Convex analysis and minimization algorithms. I [Springer, Berlin, 1993; MR1261420 (95m:90001)] and ıt II [ibid.; MR1295240 (95m:90002)].

JOHNSTONE, I. (2001). On the distribution of the largest eigenvalue in principal component analysis. *Ann. Statist.* **29**, 295–327.

KOENKER, R. (2005). *Quantile regression*, volume 38 of *Econometric Society Monographs*. Cambridge University Press, Cambridge. URL http://dx.doi.org/10.1017/CBO9780511754098.

MAMMEN, E. (1989). Asymptotics with increasing dimension for robust regression with applications to the bootstrap. *Ann. Statist.* **17**, 382–400. URL http://dx.doi.org/10.1214/aos/1176347023.

MAMMEN, E. (1992). Bootstrap, wild bootstrap, and asymptotic normality. *Probab. Theory Related Fields* **93**, 439–455. URL http://dx.doi.org/10.1007/BF01192716.

MAMMEN, E. (1993). Bootstrap and wild bootstrap for high-dimensional linear models. *Ann. Statist.* **21**, 255–285. URL http://dx.doi.org/10.1214/aos/1176349025.

MARDIA, K. V., KENT, J. T., and BIBBY, J. M. (1979). *Multivariate analysis*. Academic Press [Harcourt Brace Jovanovich Publishers], London. Probability and Mathematical Statistics: A Series of Monographs and Textbooks.

MCKEAN, J. W., SHEATHER, S. J., and HETTMANSPERGER, T. P. (1993). The Use and Interpretation of Residuals Based on Robust Estimation. *Journal of the American Statistical Association* **88**, 1254–1263.

PAJOR, A. and PASTUR, L. (2009). On the limiting empirical measure of eigenvalues of the sum of rank one matrices with log-concave distribution. *Studia Math.* **195**, 11–29. URL http://dx.doi.org/10.4064/sm195-1-2.

PARZEN, M. I., WEI, L. J., and YING, Z. (1994). A resampling method based on pivotal estimating functions. *Biometrika* **81**, 341–350. URL http://dx.doi.org/10.1093/biomet/81.2.341.

POLITIS, D. N., ROMANO, J. P., and WOLF, M. (1999). *Subsampling*. Springer Series in Statistics. Springer-Verlag, New York. URL http://dx.doi.org/10.1007/978-1-4612-1554-7.

SHORACK, G. R. (1982). Bootstrapping robust regression. *Comm. Statist. A—Theory Methods* **11**, 961–972.

SILVERSTEIN, J. W. (1995). Strong convergence of the empirical distribution of eigenvalues of large-dimensional random matrices. *J. Multivariate Anal.* **55**, 331–339.

STROOCK, D. W. (1993). *Probability theory, an analytic view*. Cambridge University Press, Cambridge.

VAN DER VAART, A. W. (1998). *Asymptotic statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.

WACHTER, K. W. (1978). The strong limits of random matrix spectra for sample matrices of independent elements. *Annals of Probability* **6**, 1–18.

WANG, X. and WANG, B. (2011). Deconvolution estimation in measurement error models: The r package decon. *Journal of Statistical Software* **39**, 1–24.

WU, C.-F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *Ann. Statist.* **14**, 1261–1350. URL `http://dx.doi.org/10.1214/aos/1176350142`. With discussion and a rejoinder by the author.

APPENDIX

Notations : in this appendix, 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.

S1 Description of Simulations and other Numerics

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, $\beta = 0$, so $y_i = \epsilon_i$.

2. Estimate $\hat{\beta}$ 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). 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 $\hat{\beta}^*$ 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, we wrote an internal function that left out each observation in turn and recalculated $\hat{\beta}_{(i)}$. For testing the deconvolution method for residual resampling (Section 2.3), we did not repeatedly resample from the deconvolution estimate of the distribution of ϵ_i; instead we drew a single draw from that estimate and resampled from that single draw. This made testing the proof of concept of the deconvolution method easier in our existing structure using `boot`, but is inferior to redrawing repeatedly from the deconvolution estimate of the distribution.

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 $(\pm 1.96\sqrt{\widehat{var}_{Jack}(\hat{\beta}_1)})$

S1.1 Values of 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

- 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
  - $\lambda_i \sim Exp(\sqrt{2})$ (i.e. mean $1/\sqrt{2}$)
  - $\lambda_i \sim N(0,1)$
  - $\lambda_i \sim Unif(0.5, 1.5)$

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, 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 $0.01, 0.1, 0.3, 0.5$.

S1.2 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` (MOSEK, 2014) as our optimization engine. We used $n = 500$ and $1,000$ simulations to compute the mean of the quantities we were interested in.

S1.3 Plotting of Figure 2a

This figure was 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 $\mathbf{E}\left(\|\widehat{\beta} - \beta\|_2\right)$. 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.

S2 Residual bootstrap (p/n close to 1)

We analyze the problem when p/n is close to 1 and prove Theorem 2.1.

Proof of Theorem 2.1. Recall the system describing the asymptotic limit of $\|\widehat{\beta}_\rho - \beta\|$ when $p/n \to \kappa$ 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 El Karoui et al. (2013); El Karoui (2013)): $\|\widehat{\beta}_\rho - \beta\| \to r_\rho(\kappa)$ 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:

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

In this system, 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, prox$(c\rho)(x) \simeq \frac{x}{1+c}$. Hence, $x -$ 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 $\widehat{z}_\epsilon$ has extremely low probability of falling outside the interval $(-s_\kappa, s_\kappa)$ - recall that $1 - \kappa \to 0$. 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 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.) □

S3 On the expected Variance of the bootstrap estimator (Proof of Theorem 3.1)

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,

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

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).

For ease of exposition in what follows, we take $v = e_p$, the p-th canonical basis vector. (Invariance arguments mentioned below show that this choice is made without loss of generality in the setting we are studying.)

We consider the simple case where $X_i \overset{iid}{\backsim} \mathcal{N}(0, \mathrm{Id}_p)$. This allows us to work with results in El Karoui et al. (2011); El Karoui et al. (2013), El Karoui (2013). We note that by invariance - see the aforementioned papers - theoretical considerations can be studied without loss of generality in the case where $\beta = 0$.

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 that 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,

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

$e_{i,[p]}$ 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, $\hat{c} = \frac{1}{n}\mathrm{trace}\left(S_w^{-1}\right) + \mathrm{o}_{L_2}(1)$, 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 $\frac{1}{n}\mathrm{trace}\left(S_w^{-1}\right)$ is asymptotically deterministic in the situation under investigation with our assumptions on w and X, i.e $\frac{1}{n}\mathrm{trace}\left(S_w^{-1}\right) = c + \mathrm{o}_{L_2}(1)$, where $c = \mathbf{E}\left(\frac{1}{n}\mathrm{trace}\left(S_w^{-1}\right)\right)$.

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 $c_i = c + \mathrm{o}_P(1)$, 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

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

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 $\hat{c}$ 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,

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

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

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

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),

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

Hence,

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

Now, in expectation, using e.g El Karoui (2013), $\mathbf{E}\left(\|\widehat{g}_{(i)}\|_2^2\right)(1+\mathrm{o}(1)) = \mathbf{E}\left(\|\widehat{b}\|_2^2\right) = p\mathbf{E}\left(\widehat{b}_p^2\right)$. 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)

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

since the previous expression effectively relates trace $\left(D_w X(X'D_wX)^{-1}X'\right)$ 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

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

Hence, for the bagged estimate, we have the equation

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

We conclude that

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

Note that $\frac{\kappa}{1-\kappa}\sigma^2 = \mathbf{E}\left(\|\widehat{\beta}_{sLS}\|_2^2\right)$, 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,

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

So we conclude that

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

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

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

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

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

we see that

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

(Note that under mild conditions on ϵ, X and w, we also have $\|\widehat{\beta}_w-\beta\|_2^2 = \mathbf{E}\left(\|\widehat{\beta}_w-\beta\|_2^2\right) + \mathrm{o}_{L_2}(1)$ - 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 $\left((\boldsymbol{X'D_wX})^{-2}\boldsymbol{X'D_{w^2}X}\right)$ Of course,

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

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

$$\mathrm{trace}\left((X'D_wX)^{-2}X'D_{w^2}X\right) = \frac{1}{n}\sum_{i=1}^n w_i^2\frac{X_i'\widehat{\Sigma}^{-2}X_i}{n} .$$

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

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

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

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

We call, for $z\in\mathbb{C}$, $c(z) = \frac{1}{n}\mathrm{trace}\left((\widehat{\Sigma}-z\mathrm{Id}_p)^{-1}\right)$ and $c_i(z) = X_i'(\widehat{\Sigma}_{(i)}-z\mathrm{Id}_p)^{-1}X_i$, 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 $c_i(z) = c(z)(1 + \mathrm{o}_P(1))$ 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

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

and using the previously discussed identity

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

we have

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

In other words,

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

Now,

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

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

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

Hence,

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

The fact that we can take expectations on both sides of this equation and that $\mathrm{o}_P(1)$ is in fact $\mathrm{o}_{L_2}(1)$ 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

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

Using the fact that

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

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, $c(0) \to c$, 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

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

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

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

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 $\frac{1}{p}\text{trace}\left((\widehat{\Sigma}_w + \tau \text{Id}_p)^{-k}\right)$, 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 $\tau \to 0$ 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.

S3.1 On acceptable weight distributions

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, i.e the least-squares one in the case we are considering. Here, this variance is asymptotically $\kappa/(1-\kappa)\sigma_\epsilon^2$.

Recall that in the main text, we proposed to use

$$w_i \overset{iid}{\backsim} 1 - \alpha + \alpha\text{Poisson}(1)$$

To determine α numerically so that

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

we performed 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 S6. This means that we stopped the algorithm when the ratio of the two terms in the previous display was within 1% of 1. We used a sample size of 10^6 to estimate all the expectations.

S3.2 Numerics for Figure 5a

This figure, related to the current discussion 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 $\mathbf{E}\left(1/(1+Wc)\right) = 1 - p/n$. Once c was found, we approximated the expectation in Theorem 3.1 in the same fashion as we just described.

Once we had computed the quantity appearing in Theorem 3.1, 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.

S4 Jackknife Variance (Proof of Theorem 4.1)

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

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

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

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

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 $X_i \sim \mathcal{N}(0, \mathrm{Id}_p)$. 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

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

We conclude that

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

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 $n, p \to \infty$,

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

Hence, in our asymptotics,

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

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,

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

Therefore,

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

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

When $v = e_1$, we therefore have

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

Therefore, in that situation,

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

In other words,

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

S4.1 Dealing with the centering issue

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

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

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 $\frac{1}{n}\sum_{i=1}^n v'\widehat{\Sigma}_{(i)}^{-1}X_i\epsilon_i \to 0$ 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 $p/n/(1-p/n) + \mathrm{o}_P(1)$. Let us call $\mathsf{c} = p/n/(1-p/n)$.

We conclude that

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

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

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

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

S4.2 Putting everything together

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

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

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

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

This completes the proof of Theorem 4.1

S4.3 Extension to more involved design 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 misestimated 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.

In the case of robust regression, the correction factors presented here will also be problematic and not fix the problems.

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 in El Karoui (2013). Checking all the approximations we made in this Jackknife computation would requires 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.

References

DAVISON, A. C. and HINKLEY, D. V. (1997). *Bootstrap methods and their application*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.

EL KAROUI, N. (2009). Concentration of measure and spectra of random matrices: Applications to correlation matrices, elliptical distributions and beyond. *The Annals of Applied Probability* **19**, 2362–2405.

EL KAROUI, N. (2010). High-dimensionality effects in the Markowitz problem and other quadratic programs with linear constraints: risk underestimation. *Ann. Statist.* **38**, 3487–3566. URL `http://dx.doi.org/10.1214/10-AOS795`.

EL KAROUI, N. (2013). Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: rigorous results. *arXiv:1311.2445* ArXiv:1311.2445.

EL KAROUI, N. (2013). On the realized risk of high-dimensional markowitz portfolios. *SIAM Journal in Financial Mathematics* **4(1)**, http://dx.doi.org/10.1137/090774926.

EL KAROUI, N., BEAN, D., BICKEL, P., LIM, C., and YU, B. (2011). On robust regression with high-dimensional predictors. Technical Report 811, UC, Berkeley, Department of Statistics. Originally submitted as manuscript AoS1111-009. Not under consideration anymore.

EL KAROUI, N., BEAN, D., BICKEL, P. J., LIM, C., and YU, B. (2013). On robust regression with high-dimensional predictors. *Proceedings of the National Academy of Sciences* URL `http://www.pnas.org/content/early/2013/08/15/1307842110.abstract`.

EL KAROUI, N. and KOESTERS, H. (2011). Geometric sensitivity of random matrix results: consequences for shrinkage estimators of covariance and related statistical methods. *Submitted to Bernoulli* Available at arXiv:1105.1404 (68 pages).

GRANT, M. and BOYD, S. (2008). Graph implementations for nonsmooth convex programs. In *Recent Advances in Learning and Control* (BLONDEL, V., BOYD, S., and KIMURA, H., editors), Lecture Notes in Control and Information Sciences, pp. 95–110. Springer-Verlag Limited. `http://stanford.edu/~boyd/graph_dcp.html`.

GRANT, M. and BOYD, S. (2014). CVX: Matlab software for disciplined convex programming, version 2.1. `http://cvxr.com/cvx`.

HAFF, L. R. (1979). An identity for the Wishart distribution with applications. *J. Multivariate Anal.* **9**, 531–544. URL `http://dx.doi.org/10.1016/0047-259X(79)90056-3`.

HIRIART-URRUTY, J.-B. and LEMARÉCHAL, C. (2001). *Fundamentals of convex analysis*. Grundlehren Text Editions. Springer-Verlag, Berlin. Abridged version of ıt Convex analysis and minimization algorithms. I [Springer, Berlin, 1993; MR1261420 (95m:90001)] and ıt II [ibid.; MR1295240 (95m:90002)].

HORN, R. A. and JOHNSON, C. R. (1990). *Matrix analysis*. Cambridge University Press, Cambridge. Corrected reprint of the 1985 original.

KATO, T. (1995). *Perturbation theory for linear operators*. Classics in Mathematics. Springer-Verlag, Berlin. Reprint of the 1980 edition.

KOENKER, R. (2013). *quantreg: Quantile Regression*. URL http://CRAN.R-project.org/package=quantreg. R package version 5.05.

LEDOUX, M. (2001). *The concentration of measure phenomenon*, volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI.

MARDIA, K. V., KENT, J. T., and BIBBY, J. M. (1979). *Multivariate analysis*. Academic Press [Harcourt Brace Jovanovich Publishers], London. Probability and Mathematical Statistics: A Series of Monographs and Textbooks.

MILLER, P. D. (2006). *Applied asymptotic analysis*, volume 75 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI.

MOREAU, J.-J. (1965). Proximité et dualité dans un espace hilbertien. *Bull. Soc. Math. France* **93**, 273–299.

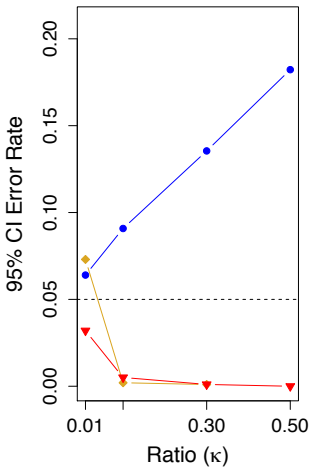
MOSEK (). *Rmosek: The R to MOSEK Optimization Interface*. URL http://rmosek.r-forge.r-project.org/,http://www.mosek.com/. R package version 7.0.5.

MOSEK (2014). MOSEK Optimization Toolbox. Available at www.mosek.com.

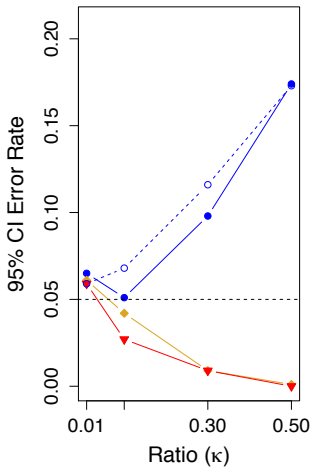
SILVERSTEIN, J. W. (1995). Strong convergence of the empirical distribution of eigenvalues of large-dimensional random matrices. *J. Multivariate Anal.* **55**, 331–339.

VENABLES, W. N. and RIPLEY, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. URL http://www.stats.ox.ac.uk/pub/MASS4. ISBN 0-387-95457-0.

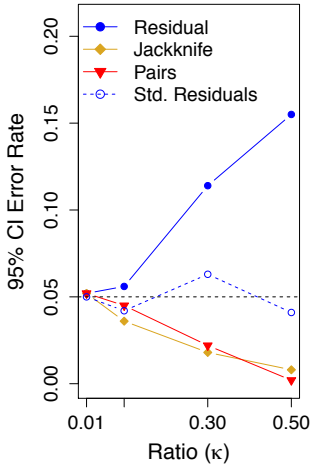
WEISBERG, S. (2014). *Applied linear regression*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, fourth edition.



(a) L_1 loss

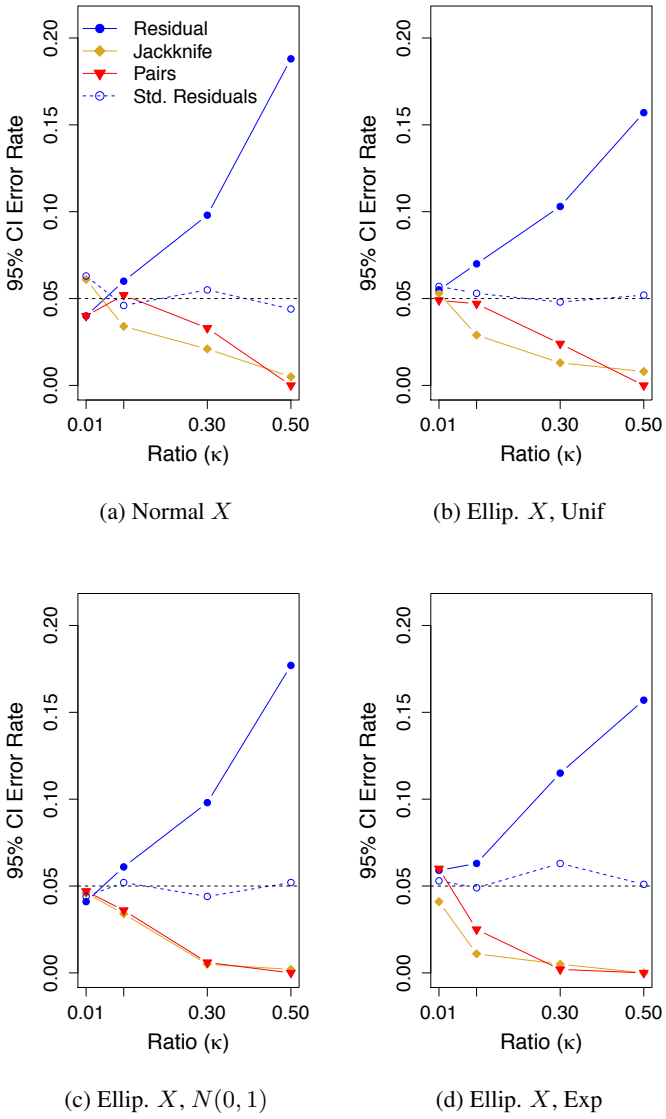


(b) Huber loss



(c) L_2 loss

Figure S1: **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.



(a) Normal X (b) Ellip. X, Unif

(c) Ellip. X, $N(0,1)$ (d) Ellip. X, Exp

Figure S2: **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 $\lambda_i \sim U(.5, 1.5)$, (c) elliptical with $\lambda_i \sim N(0,1)$, and (d) elliptical with $Exp(\sqrt{2})$. 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.