CONDITIONAL PREDICTIVE INFERENCE FOR HIGH-DIMENSIONAL STABLE ALGORITHMS

By Lukas Steinberger[*] and Hannes Leeb[†]

University of Vienna

We investigate generically applicable and intuitively appealing prediction intervals based on leave-one-out residuals. The conditional coverage probability of the proposed intervals, given the observations in the training sample, is close to the nominal level, provided that the underlying algorithm used for computing point predictions is sufficiently stable under the omission of single feature/response pairs. Our results are based on a finite sample analysis of the empirical distribution function of the leave-one-out residuals and hold in non-parametric settings with only minimal assumptions on the error distribution. To illustrate our results, we also apply them to high-dimensional linear predictors, where we obtain uniform asymptotic conditional validity as both sample size and dimension tend to infinity at the same rate. These results show that despite the serious problems of resampling procedures for inference on the unknown parameters (cf. Bickel and Freedman, 1983; El Karoui and Purdom, 2015; Mammen, 1996), leave-one-out methods can be successfully applied to obtain reliable predictive inference even in high dimensions.

1. Introduction.

It is a fundamental task of statistical learning, when given an i.i.d. training sample of feature/response pairs (x_i, y_i) and an additional feature vector x_0, to provide a point prediction for the corresponding unobserved response variable y_0. In such a situation, a prediction interval that contains the unobserved response variable with a prescribed probability provides valuable additional information to the practitioner. In many applications, when measurements are costly, a training sample is obtained only once and is subsequently used to repeatedly construct point and interval predictions as new measurements of feature vectors become available. In

[*]Lukas Steinberger was supported by the Austrian Science Fund (FWF): P28233 and by the German Research Foundation (DFG): RO 3766/401.

[†]Hannes Leeb was supported by the Austrian Science Fund (FWF): P 28233-N32 and P 26354-N26.

MSC 2010 subject classifications: Primary 62G15, 62J02; secondary 62G20, 62J07

Keywords and phrases: prediction intervals, high-dimensional regression, algorithmic stability, jackknife

such a situation, it is desirable to control the conditional coverage probability of the prediction interval given the observations in the training sample, rather than the unconditional probability.

We study a very simple method based on leave-one-out residuals which is generic in the sense that it applies to a large class of possible point predictors, while providing asymptotically valid prediction intervals. For an i.i.d. training sample $T_n = (x_i, y_i)_{i=1}^n$ of size n, consisting of $\mathbb{R}^p \times \mathbb{R}$-valued feature/response pairs, and an additional feature vector x_0 in $\mathbb{R}^p$, suppose that we have decided to use a prediction algorithm $M_{n,p} : (\mathbb{R}^p \times \mathbb{R})^n \times \mathbb{R}^p \to \mathbb{R}$ to produce a point prediction $\hat{y}_0 = M_n(T_n, x_0)$ for the real unobserved response y_0. If $T_n^{[i]} = (x_j, y_j)_{j \neq i}$ is the sample without the i-th observation pair, compute leave-one-out residuals $\hat{u}_i = y_i - M_{n-1}(T_n^{[i]}, x_i)$, $1 \leq i \leq n$. Finally, to obtain a prediction interval for y_0, compute appropriate empirical quantiles $\hat{q}_{\alpha_1}$ and $\hat{q}_{\alpha_2}$ from the collection $\hat{u}_1, \dots, \hat{u}_n$ and report the leave-one-out prediction interval

$$PI_{\alpha_1,\alpha_2}^{(L1O)}(T_n, x_0) = (\hat{y}_0 + \hat{q}_{\alpha_1}, \hat{y}_0 + \hat{q}_{\alpha_2}].$$

The use of the half open interval is due to technical convenience and is inconsequential for practical purposes. In this paper we investigate the conditional coverage probability

$$P^{n+1}(y_0 \in PI_{\alpha_1,\alpha_2}^{(L1O)}(T_n, x_0) \| T_n),$$

first in finite samples, and then in more specific asymptotic settings where the dimension p of the feature vectors x_i increases at the same rate as sample size n. We find that even in these challenging scenarios where both n and p are large, the conditional coverage of $PI_{\alpha_1,\alpha_2}^{(L1O)}(T_n, x_0)$ is close to the nominal level $\alpha_2 - \alpha_1$. Note that the analogous procedure based on ordinary residuals $y_i - M_n(T_n, x_i)$ instead of leave-one-out residuals would, in general, not be valid in such a large-p scenario (cf. Bickel and Freedman, 1983).

Despite the remarkable simplicity of this method, and its apparent similarity to the jackknife, we are not aware of any rigorous analysis of its statistical properties. Our approach is very similar, in spirit, to the methods proposed in Butler and Rothman (1980), Stine (1985), Schmoyer (1992), Olive (2007) and Politis (2013), in the sense that it relies on resampling and leave-one-out ideas for predictive inference. But the methods from these references, like most resampling procedures in the literature, are investigated only in the classical large sample asymptotic regime where the number of available explanatory variables is fixed. Notable exceptions are Bickel and

Freedman (1983), Mammen (1996) and, recently, El Karoui and Purdom (2015). However, the latter articles draw mainly negative conclusions about resampling methods in high dimensions, arguing, for instance, that the famous residual bootstrap in linear regression, which relies on the consistent estimation of the true unknown error distribution, is unreliable when the number of variables in the model is not small compared to sample size. In contrast, we show that the leave-one-out prediction interval $PI_{\alpha_1,\alpha_2}^{(L1O)}$ does not suffer from these problems because it relies on estimation of the conditional distribution of the prediction error $P^{n+1}(y_0 - \hat{y}_0 \leq t \| T_n)$ instead of an estimator for the unconditional distribution of the error term $y_0 - \mathbb{E}[y_0 \| x_0]$. That the use of leave-one-out residuals leads to more reliable methods in high dimensions was also observed by El Karoui and Purdom (2015).

Our contribution is threefold. First, we show that the leave-one-out prediction interval is approximately conditionally valid given the training sample T_n, in the sense that

$$P^{n+1}\left(y_0 \in PI_{\alpha_1,\alpha_2}^{(L1O)}(T_n, x_0) \middle\| T_n\right) \approx \alpha_2 - \alpha_1.$$

The error term of the above approximation can be controlled in finite samples and asymptotically, provided that the employed prediction algorithm M_n is sufficiently stable under the omission of single feature/response pairs and that it has a bounded (in probability) estimation error as an estimator for the true unknown regression function. It is of paramount importance, however, to point out that we do not need to assume consistent estimation of the regression function and our leading examples are such that consistency fails.

Second, we show that the required stability and approximation properties are satisfied in many cases, including many linear predictors in high dimensional regression problems and even if the true model is not exactly linear. In particular, the proposed method is always valid if the employed predictor is consistent for the unknown regression function (or for an appropriate surrogate target), and is therefore applicable to complex data structures and methods such as non-parametric regression or LASSO prediction.

Third, we discuss issues of interval length and find that in typical situations predictors with smaller mean squared prediction error lead to shorter prediction intervals. For ordinary least squares prediction, we also investigate the impact of the dimensionality of the regression problem on the interval length and discuss the relationship between the leave-one-out method and an obvious sample splitting technique. All our results hold uniformly over large classes of data generating processes and under weak assumptions on the unknown error distribution (e.g., the errors may be heavy tailed and non-symmetric, and the standardized design vectors $\mathrm{Cov}[x_i]^{-1/2}x_i$ may have dependent components and a non-spherical distribution).

Our work is greatly inspired by El Karoui et al. (2013) and Bean et al. (2013) (see also El Karoui, 2013, 2018), who investigate efficiency of general M-estimators in linear regression when the number of regressors p is of the same order of magnitude as sample size n. In particular, the M-estimators studied in these references provide one leading example of a class of linear predictors for which our construction of prediction intervals leads to conditionally valid predictive inference even in high dimensions.

The remainder of the paper is organized as follows. In the following Subsection 1.1 we give a brief overview of alternative methods from the large body of literature on predictive inference in regression. Subsection 1.2 introduces the notation that is used throughout the paper. Sections 2 and 3 proceed along a general-to-specific scheme. We begin, in Subsection 2.1, by introducing the general leave-one-out method and the notion of conditional validity and we take a first step towards proving that the latter property is satisfied. In Subsection 2.2, we draw the connection between conditional validity and algorithmic stability and present our main results which provide generic sufficient conditions for conditional validity. In Section 3 we then show that these conditions can be verified in challenging statistical scenarios where regression function estimation and the bootstrap usually fail to be consistent. In particular, we consider linear predictors based on regularized M-estimators and based on James-Stein-type estimators in a situation where the number of regressors p is not small relative to sample size n. We also take a closer look at the ordinary least squares estimator, because its simplicity allows for a rigorous discussion of the resulting interval length. In Section 4, we then also discuss the important case where the employed predictor is consistent (possibly for some pseudo target rather than the true regression function) and we provide examples on non-parametric regression and high-dimensional LASSO. The case of consistency is an important test case for our method. Finally, in Section 5, we provide some further discussions and sketch possible extensions of our results. Most of the proofs are deferred to the supplementary material.

1.1. *Related work.* In a fully parametric setting, predictive inference is essentially a special case of parametric inference (see, e.g., Cox and Hinkley, 1974, Section 7.5). Constructing valid prediction sets becomes much more challenging, however, if one is interested in a non-parametric setting. By non-parametric, we do not only mean that the regression function can not be indexed by a finite dimensional Euclidean space, but also that the random fluctuations $y_i - \mathbb{E}[y_i \| x_i]$ about the conditional mean function can not be described by a parametric family of distributions.

1.1.1. *Tolerance regions.* A rather well researched and classical topic in the statistics literature is the construction of so called tolerance regions or tolerance limits, which are closely related to prediction regions. A tolerance region is a set valued estimate $TR_\alpha(T_n) \subseteq \mathbb{R}^m$ based on i.i.d. m-variate data $z_1, \dots, z_n$, $T_n = (z_1, \dots, z_n)$, such that the probability of covering an independent copy z_0 is close to a prescribed confidence level. More precisely, a $(1-\alpha, \rho)$ tolerance region TR is such that $P^n(P^{n+1}(z_0 \in TR \| T_n) \geq 1-\alpha) = \rho$, and TR is called a $(1-\alpha)$-expectation tolerance region, if $\mathbb{E}_{P^n}[P^{n+1}(z_0 \in TR \| T_n)] = P^{n+1}(z_0 \in TR) = 1 - \alpha$ (cf. Krishnamoorthy and Mathew, 2009). The study of non-parametric tolerance regions goes back at least to Wilks (1941, 1942), Wald (1943) and Tukey (1947) (see Krishnamoorthy and Mathew, 2009, for an overview and further references) and is traditionally based on the theory of order statistics of i.i.d. data. These researchers already obtained multivariate distribution-free methods, that is, tolerance regions that achieve a certain type of validity in finite samples without imposing parametric assumptions. The connection to prediction regions is apparent: If $z_i = (x_i, y_i)$, then a tolerance region $TR_\alpha(T_n)$ for $z_0 = (x_0, y_0)$ can be immediately used to obtain a prediction region for y_0 by setting $PR_\alpha(T_n, x_0) = \{y : (x_0, y) \in TR_\alpha(T_n)\}$. However, this is arguably not the most economical way of constructing a prediction region. In fact, the construction of a multivariate and possibly high-dimensional tolerance region appears to be a more ambitious goal than the construction of a prediction region for a univariate response variable. In particular, since estimation of the full density of z_0 – which could be used to compute an optimal highest density region – is usually not feasible if the dimension m is non-negligible compared to sample size n, one has to specify a shape for the tolerance region TR_α and it is not obvious which shapes are preferable in a non-parametric setting. For example, Bucchianico et al. (2001) provide results for smallest possible hyperrectangles and ellipsoids, but obtain only the classical large sample asymptotic results with fixed dimension. Chatterjee and Patra (1980) estimate the density non-parametrically, which fails in high dimensions. Li and Liu (2008) use a notion of data depth to avoid the specification of the shape, but the fully data driven method, again, is only shown to be valid asymptotically, with the dimension fixed. Finally, numerically computing the x_0-cut of TR_α to obtain PR_α is computationally demanding and the result is sensitive to the shape of TR_α.

1.1.2. *Conformal prediction.* A strand of literature which has emerged from the early ideas of non-parametric tolerance regions, but which is more prominent within the machine learning community than the statistics community, is called conformal prediction (Vovk et al., 1999, 2005, 2009). Conformal prediction is a very flexible general framework for construction of prediction regions that can be used in conjunction with any learning algorithm. The general idea is to construct a pivotal p-value $\pi(y_\star)$ to test $H_\star : y_0 = y_\star$ based on the sample T_n and x_0, for each possible value $y_\star$ of y_0, and to invert the test to obtain a prediction region for y_0, i.e., $PR_\alpha = \{y : \pi(y) \geq \alpha\}$. The method was primarily designed for an on-line learning setup (cf. Vovk et al., 2009), but has recently been popularized in the statistics community by Lei et al. (2017, 2013) and Lei and Wasserman (2014), who study it as a batch method. Aside from their flexibility, conformal prediction methods have the advantage that they are valid in finite samples, in the sense that the unconditional coverage probability $P^{n+1}(y_0 \in PR_\alpha)$ is no less than the nominal level $1-\alpha$, provided only that the feature/response pairs $(x_0, y_0), (x_1, y_1), \dots, (x_n, y_n)$ are exchangeable. On the other hand, their practical implementation is not so straight forward, because, for the test inversion, the p-value π has to be evaluated on a grid of possible y values, which is especially tricky if the conformal prediction region is not an interval (see Chen et al. (2017) and Lei (2017) for further discussion of these issues). Moreover, it is not clear if the classical conformal methods can also provide a form of conditional validity. In Vovk (2012), a version of conformal prediction was presented that achieves also a certain type of (approximate) conditional validity. However, the method relies on a sample splitting idea, which usually makes the prediction region unnecessarily wide (see Sections 3.4 and 5.2 for further discussion of sample splitting techniques). A different version of conditional validity (conditioning on x_0), is discussed in Barber et al. (2019a) (see also Remark 5.3 below).

1.1.3. *The jackknife+.* Barber et al. (2019b) recently proposed a modification of the leave-one-out method considered here. For the modified method, which they call jackknife+, they derived a finite-sample lower bound for the unconditional coverage probability, under the assumption that the feature/response pairs are exchangeable and without requiring that the prediction algorithm is stable. For a jackknife+ interval with nominal coverage probability $1-\alpha$, the lower bound is $1-2\alpha$; if the prediction algorithm is stable (under omission of a single feature/response pair), the lower bound moves closer towards $1-\alpha$ (provided that the interval is slightly modified further). In simulations and data-examples, Barber et al. (2019b) found that the jackknife+ performs essentially like the jackknife, i.e., like the method considered in this paper, unless the prediction algorithm is highly unstable; in the unstable case, jackknife+ outperforms jackknife. The use of an unstable prediction method is, of course, debatable, at least if the user is aware of the instability. The conditional performance of the jackknife+ interval, i.e., its coverage probability conditional on the training data, is yet to be analyzed.

1.2. *Preliminaries and notation.*

For $p \in \mathbb{N}$, let $\mathcal{Y} \subseteq \mathbb{R}$ and $\mathcal{X} \subseteq \mathbb{R}^p$ be Borel measurable sets and let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Moreover, let $\mathcal{P}$ be some class of Borel probability measures on $\mathcal{Z}$ and, for $n \in \mathbb{N}$, $n \geq 2$, let P^n denote the n-fold product measure of $P \in \mathcal{P}$. For $P \in \mathcal{P}$, we write $z_0 = (x_0, y_0)$ for a random vector distributed according to P and we write $T_n = (z_i)_{i=1}^n$, $z_i = (x_i, y_i)$, for a training sample, where $z_0, z_1, \dots, z_n$ are independent and identically distributed according to P. This means that (T_n, z_0) is distributed as P^{n+1}. By $m_P(x) := \mathbb{E}_P[y_0 \| x_0 = x]$, $m_P : \mathcal{X} \to \mathbb{R}$, we denote (a version of) the true unknown regression function, if it exists. We sometimes express the training data T_n as (X, Y), where $X = [x_1, \dots, x_n]'$ is of dimension $n \times p$ and $Y = (y_1, \dots, y_n)'$ is a random n-vector. Moreover, X' denotes the transpose of X, and we write $(X'X)^\dagger$ for the Moore-Penrose inverse of $X'X$. Similarly, we write $X_{[i]} = [x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n]'$ and $Y_{[i]} = (y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_n)'$.

Next, we formally define the notion of a (learning) algorithm and that of a predictor (or estimator) $\hat{m}_n$ and its leave-one-out equivalent $\hat{m}_n^{[i]}$. Consider a measurable function $M_{n,p} : \mathcal{Z}^n \times \mathcal{X} \to \mathbb{R}$. $M_{n,p}$ is also called a learning algorithm. For each vector $x \in \mathcal{X}$, we set $\hat{m}_n(x) = M_{n,p}(T_n, x)$ and $\hat{m}_n^{[i]}(x) = M_{n-1,p}(T_n^{[i]}, x)$, where $T_n^{[i]} = (z_j)_{j \neq i}$, $i = 1, \dots, n$, denotes the reduced training sample where the observation $z_i = (x_i, y_i)$ has been deleted. Thus whenever we are talking about a predictor, we implicitly talk about the pair of functions $(M_{n,p}, M_{n-1,p})$. A predictor $\hat{m}_n$ is called *symmetric* if for every choice of $z_1, \dots, z_n \in \mathcal{Z}$, every $x \in \mathcal{X}$ and every permutation π of n elements, $M_{n,p}((z_i)_{i=1}^n, x) = M_{n,p}((z_{\pi(i)})_{i=1}^n, x)$, and if the same holds true for $M_{n-1,p}$. Since the training data $T_n = (z_1, \dots, z_n)$ are assumed to be i.i.d., it is natural to consider symmetric predictors. Also note that, although computationally demanding, in principle any predictor $\hat{m}_n$ can be symmetrized by averaging over all possible permutations of the training data.

If $A(t) \in \mathcal{B}(\mathcal{Z})$, $t \in \mathcal{Z}^n$, is a collection of Borel subsets of $\mathcal{Z}$, then we define the conditional probability of $A(T_n)$ given the training sample $T_n = t$ by $P^{n+1}(z_0 \in A(T_n) \| T_n = t) := P(A(t))$. For example, if $PI(t, x)$ is an interval depending on $t \in \mathcal{Z}^n$ and $x \in \mathcal{X}$, then $P^{n+1}(y_0 \in PI(T_n, x_0) \| T_n = t) := P(\{(x, y) \in \mathcal{Z} : y \in PI(t, x)\})$, assuming measureability. If $f : D \to \mathbb{R}$ is a real function on some domain D, then $\|f\|_\infty = \sup_{s \in D} |f(s)|$. For $a, b \in \mathbb{R}$, we also write $a \vee b = \max(a, b)$, $a \wedge b = \min(a, b)$ and $a_+ = a \vee 0$, and let $\lceil \delta \rceil$ denote the smallest integer no less than $\delta \in \mathbb{R}$. We write $U \overset{\mathcal{L}}{=} V$, if the random quantities U and V are equal in distribution and the underlying probability space is clear from the context. By a slight abuse of notation, we also write $U \overset{\mathcal{L}}{=} \mathcal{L}_0$ if the random variable U is distributed according to the probability law $\mathcal{L}_0$ and, again, the underlying probability space is clear from the context.

For our asymptotic statements, we will also need the following conventions. Let $(p_n)_{n \in \mathbb{N}}$ be a sequence of positive integers. If, for each $n \in \mathbb{N}$, $\mathcal{P}_n$ is a collection of probability distributions on $\mathcal{Z}_n \subseteq \mathbb{R}^{p_n+1}$ and $\phi_n : \mathcal{Z}_n^n \times \mathcal{P}_n \to \mathbb{R}$ is a function such that for every $P \in \mathcal{P}_n$, $t \mapsto \phi_n(t, P)$ is measurable, then we say that ϕ_n is $\mathcal{P}_n$-uniformly bounded in probability if $\limsup_{n \to \infty} \sup_{P \in \mathcal{P}_n} P^n(|\phi_n(T_n, P)| > M) \to 0$, as $M \to \infty$, and write $\phi_n = O_{\mathcal{P}_n}(1)$. If $\sup_{P \in \mathcal{P}_n} P^n(|\phi_n(T_n, P)| > \varepsilon) \to 0$, as $n \to \infty$, for every $\varepsilon > 0$, then we say that ϕ_n converges $\mathcal{P}_n$-uniformly in probability to zero and write $\phi_n = o_{\mathcal{P}_n}(1)$. Similarly, we say that ϕ_n converges $\mathcal{P}_n$-uniformly in probability to $\psi_n : \mathcal{Z}_n^n \times \mathcal{P}_n \to \mathbb{R}$, which is also assumed to be measurable in its first argument, if $|\phi_n - \psi_n| = o_{\mathcal{P}_n}(1)$.

2. Main results.

2.1. *Leave-one-out prediction intervals and conditional validity.* For $\alpha \in (0, 1)$, we want to construct a prediction interval $PI_\alpha(T_n, x_0) = (\hat{m}_n(x_0) + L_\alpha(T_n), \hat{m}_n(x_0) + U_\alpha(T_n)]$ for y_0, where L_α and U_α are measurable functions on $\mathcal{Z}^n$, such that

$$(2.1) \qquad \sup_{P \in \mathcal{P}} \mathbb{E}_{P^n} \left[\left| P^{n+1} \left(y_0 \in PI_\alpha(T_n, x_0) \middle\| T_n \right) - (1 - \alpha) \right| \right]$$

is small. We can not expect the expression in (2.1) to be equal to zero for some fixed n and a reasonably large class $\mathcal{P}$ (see Remark 5.1 below). Therefore, we are content with (2.1) being close to zero as n, and possibly also p, is large. This notion of conditional validity is related to what Vovk (2013) calls *training conditional validity*, and which is itself closely related to the conventional notion of a $(1-\alpha, \rho)$ tolerance region for ρ close to 1 (cf. Krishnamoorthy and Mathew, 2009). However, these conventional definitions require only that the conditional coverage probability $P^{n+1}(y_0 \in PI_\alpha(T_n, x_0) \| T_n)$ is no less than the prescribed confidence level $1-\alpha$, with high probability, whereas the requirement that (2.1) is small also excludes overly conservative procedures. Note that if (2.1) is small, then also

$$\begin{aligned}
&\left|P^{n+1}\left(y_0 \in PI_\alpha(T_n, x_0)\right) - (1-\alpha)\right| \\
&\quad = \left|\mathbb{E}_{P^n}\left[P^{n+1}\left(y_0 \in PI_\alpha(T_n, x_0)\middle\| T_n\right) - (1-\alpha)\right]\right| \\
&\quad \le \mathbb{E}_{P^n}\left[\left|P^{n+1}\left(y_0 \in PI_\alpha(T_n, x_0)\middle\| T_n\right) - (1-\alpha)\right|\right]
\end{aligned}$$

will be small. Hence, the prediction interval is then also approximately unconditionally valid, uniformly over $P \in \mathcal{P}$.

If the conditional distribution function $s \mapsto \tilde{F}_n(s) := P^{n+1}(y_0 - \hat{m}_n(x_0) \le s \| T_n)$ is continuous, then, for $0 \le \alpha_1 < \alpha_2 \le 1$ fixed, there is an optimal shortest but infeasible interval

$$PI_{\alpha_1,\alpha_2}^{(OPT)} = (\hat{m}_n(x_0) + \tilde{q}_{\alpha_1}, \hat{m}_n(x_0) + \tilde{q}_{\alpha_2}] \tag{2.2}$$

in the set of all prediction intervals PI of the form $PI = PI(T_n, x_0) = (\hat{m}_n(x_0) + L(T_n), \hat{m}_n(x_0) + U(T_n)]$ that also satisfy

$$P^{n+1}\left(y_0 \le \inf PI \middle\| T_n\right) = \alpha_1, \quad \text{and} \tag{2.3}$$

$$P^{n+1}\left(y_0 \ge \sup PI \middle\| T_n\right) = 1 - \alpha_2 : \tag{2.4}$$

Simply choose $\tilde{q}_{\alpha_1}$ to be the largest α_1-quantile of $\tilde{F}$ and $\tilde{q}_{\alpha_2}$ to be the smallest α_2-quantile of $\tilde{F}_n$. This gives the user the flexibility to choose precisely what error probability of under and over-prediction she is willing to accept. Thus, for $PI_{\alpha_1,\alpha_2}^{(OPT)}$, (2.1) is actually equal to zero (for $\alpha_1 + 1 - \alpha_2 = \alpha$), at least if $\mathcal{P}$ contains only probability distributions on $\mathcal{Z}$ for which $\tilde{F}_n : \mathbb{R} \to [0,1]$ is almost surely continuous.

We propose the following simple jackknife-type idea to approximate the optimal infeasible procedure: For $\alpha \in [0,1]$, let $\hat{q}_\alpha$ denote an empirical α-quantile of the sample $\hat{u}_1, \dots, \hat{u}_n$ of leave-one-out residuals $\hat{u}_i = y_i - \hat{m}_n^{[i]}(x_i)$. To be more precise, we set $\hat{q}_\alpha = \hat{u}_{(\lceil n\alpha \rceil)}$ if $\alpha > 0$ and $\hat{q}_0 = \hat{u}_{(1)} - e^{-n}$ (any number strictly less than $\hat{u}_{(1)}$ would do), where $\hat{u}_{(1)} \le \hat{u}_{(2)} \le \dots \le \hat{u}_{(n)}$ are the order statistics of the leave-one-out residuals. Then the leave-one-out prediction interval is given by

$$PI_{\alpha_1,\alpha_2}^{(L1O)}(T_n, x_0) \;=\; \hat{m}_n(x_0) \;+\; \left(\hat{q}_{\alpha_1},\; \hat{q}_{\alpha_2}\right]. \tag{2.5}$$

Excluding the left endpoint turns out to be convenient for proving Lemma 2.1 below. The random distribution functions

$$\hat{F}_n(s) := \hat{F}_n(s;T_n) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(-\infty,s]}(\hat{u}_i) \tag{2.6}$$

and

$$\tilde{F}_n(s) := \tilde{F}_n(s;T_n) := P^{n+1}(y_0 - \hat{m}_n(x_0) \leq s \| T_n), \tag{2.7}$$

$s \in \mathbb{R}$, play a crucial role in the analysis of the leave-one-out prediction intervals.

The idea behind the leave-one-out procedure is remarkably simple. To estimate the conditional distribution $\tilde{F}_n$ of the prediction error $y_0 - \hat{m}_n(x_0)$ we simply use the empirical distribution $\hat{F}_n$ of the leave-one-out residuals $\hat{u}_i = y_i - \hat{m}_n^{[i]}(x_i)$. Notice that $\hat{m}_n$ is independent of (x_0, y_0), and $\hat{m}_n^{[i]}$ is independent of (x_i, y_i), and thus, $\hat{u}_i$ has almost the same distribution as the prediction error, except that $\hat{m}_n^{[i]}$ is calculated from one observation less than $\hat{m}_n$. In many cases this difference turns out to be negligible if n is large, even if p is relatively large too. Note, however, that the leave-one-out residuals $(\hat{u}_i)_{i=1}^n$ are not independent.

The following elementary result shows that, indeed, the main ingredient to establish conditional validity (2.1) of the leave-one-out prediction interval in (2.5) is consistent estimation of $\tilde{F}_n$ in Kolmogorov distance.

Lemma 2.1. *For $0 \leq \alpha_1 < \alpha_2 \leq 1$, and if the fixed (non-random) training sample $t_n \in \mathcal{Z}^n$ is such that the leave-one-out residuals $\hat{u}_i = \hat{u}_i(t_n)$, $i = 1, \ldots, n$, are all distinct, then the leave-one-out prediction interval defined in* (2.5) *satisfies*

$$\left| P\left(y_0 \in PI_{\alpha_1,\alpha_2}^{(L1O)}(t_n, x_0)\right) - \frac{\lceil n\alpha_2\rceil - \lceil n\alpha_1\rceil}{n} \right| \leq 2\|\hat{F}_n - \tilde{F}_n\|_\infty.$$

Remark 2.2. Note that the inequality of Lemma 2.1 is a purely algebraic statement for a fixed training set t_n. Also note that the coverage probability $P(y_0 \in PI_{\alpha_1,\alpha_2}^{(L1O)}(t_n, x_0))$ is a version of the conditional probability $P^{n+1}(y_0 \in PI_{\alpha_1,\alpha_2}^{(L1O)}(T_n, x_0) \| T_n = t_n)$.

Proof of Lemma 2.1. By definition,

$$\begin{aligned} P\left(y_0 \in PI_{\alpha_1,\alpha_2}^{(L1O)}(t_n, x_0)\right) &= \tilde{F}_n(\hat{q}_{\alpha_2}) - \tilde{F}_n(\hat{q}_{\alpha_1}) \\ &= \tilde{F}_n(\hat{q}_{\alpha_2}) - \hat{F}_n(\hat{q}_{\alpha_2}) + \hat{F}_n(\hat{q}_{\alpha_1}) - \tilde{F}_n(\hat{q}_{\alpha_1}) + \hat{F}_n(\hat{q}_{\alpha_2}) - \hat{F}_n(\hat{q}_{\alpha_1}). \end{aligned}$$

For $\alpha_1 > 0$,

$$\begin{aligned} n\hat{F}_n(\hat{q}_{\alpha_2}) - n\hat{F}_n(\hat{q}_{\alpha_1}) &= \left|\{i \le n : \hat{u}_{(i)} \le \hat{u}_{(\lceil n\alpha_2 \rceil)}\}\right| - \left|\{i \le n : \hat{u}_{(i)} \le \hat{u}_{(\lceil n\alpha_1 \rceil)}\}\right| \\ &= \lceil n\alpha_2 \rceil - \lceil n\alpha_1 \rceil, \end{aligned}$$

and $n\hat{F}_n(\hat{q}_{\alpha_2}) - n\hat{F}_n(\hat{q}_0) = \left|\{i \le n : \hat{u}_{(i)} \le \hat{u}_{(\lceil n\alpha_2 \rceil)}\}\right| - 0 = \lceil n\alpha_2 \rceil$. Thus,

$$\hat{F}_n(\hat{q}_{\alpha_2}) - \hat{F}_n(\hat{q}_{\alpha_1}) = \frac{\lceil n\alpha_2 \rceil - \lceil n\alpha_1 \rceil}{n},$$

which concludes the proof. □

By virtue of Lemma 2.1, most of what follows will be concerned with the analysis of $\|\hat{F}_n - \tilde{F}_n\|_\infty$. We are particularly interested in situations where, for a fixed $x \in \mathcal{X}$, $\hat{m}_n(x)$ does not concentrate around $m_P(x)$ with high probability but remains random (cf. Remark 5.2 below). In such cases, the unconditional distribution function of the prediction error $P^{n+1}(y_0 - \hat{m}_n(x_0) \le s) = \mathbb{E}_{P^n}[\tilde{F}_n(s)]$, the empirical distribution function of the ordinary residuals $\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{(-\infty,s]}(y_i - \hat{m}_n(x_i))$ and the true error distribution function $P(y_0 - m_P(x_0) \le s)$ need not be close to one another, because $\hat{m}_n$ may not contain enough information about the true regression function m_P (see, for instance, Bickel and Freedman (1983) and Bean et al. (2013) for a linear regression example where $m_P(x) = x'\beta_P$)[1]. Nevertheless, we will see that even in such a challenging scenario, it is often possible to estimate the conditional distribution $\tilde{F}_n$ of $y_0 - \hat{m}_n(x_0)$, given the training sample T_n, by the empirical distribution $\hat{F}_n$ of the leave-one-out residuals.

2.2. *The role of algorithmic stability.* In this section we present general results that relate the uniform estimation error $\|\hat{F}_n - \tilde{F}_n\|_\infty$ to a measure of stability of the estimator $\hat{m}_n$. For our first result, sample size $n \ge 2$ and dimension $p \ge 1$ are fixed. We only need the following condition on the class of distributions $\mathcal{P}$ on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$.

(C1) Under every $P \in \mathcal{P}$, the distribution of $z_0 = (x_0, y_0)$ has the following properties:[2] The regression function $x \mapsto m_P(x) = \mathbb{E}_P[y_0 \| x_0 = x]$

[1] It turns out, however, at least in the linear model $m_P(x) = x'\beta_P$ and for appropriate estimators of β_P, that the conditional distribution of the prediction error $\tilde{F}_n$ does stabilize at its mean, i.e., the unconditional distribution, even if n and p are of the same order of magnitude (cf. Section 3.3 and Lemma A.7 in the proof of Theorem 3.4).

[2] To be formally precise, one should interpret x_0 as the identity mapping of $\mathcal{X} \subseteq \mathbb{R}^p$ onto itself and y_0 as the identity mapping of $\mathcal{Y} \subseteq \mathbb{R}$ onto itself.

exists and the error term $u_0 := y_0 - m_P(x_0)$ is independent of the regressor vector x_0 and has a Lebesgue density $f_{u,P}$ with $\|f_{u,P}\|_\infty < \infty$.

Remark 2.3. The boundedness of the error density $f_{u,P}$ can be relaxed to a Hölder condition on the cdf of u_0 at the expense of a slightly more complicated theory.

Remark 2.4. Note that by continuity of the cdf of the error distribution u_0, for every $\alpha \in [0,1]$, there exists a quantile $q_{u,P}(\alpha)$ such that $P(u_0 \leq q_{u,P}(\alpha)) = \alpha$. However, $q_{u,P}(\alpha)$ may not be uniquely determined by this requirement.

Building on terminology from Bousquet and Elisseeff (2002) (see also Devroye and Wagner (1979)), we use the following notion of algorithmic stability.

Definition 1. *For $\eta > 0$ and $\mathcal{P}$ as in (C1), we say the predictor $\hat{m}_n$ is* η-stable *with respect to $\mathcal{P}$ if*

$$\sup_{P\in\mathcal{P}} \mathbb{E}_{P^{n+1}}\left[\left(\|f_{u,P}\|_\infty \left|\hat{m}_n(x_0) - \hat{m}_n^{[i]}(x_0)\right|\right) \wedge 1\right] \;\leq\; \eta, \quad \forall i = 1,\dots,n.$$

By exchangeability of $z_0, z_1, \dots, z_n$, it is easy to see that a symmetric predictor $\hat{m}_n$ is η-stable w.r.t. $\mathcal{P}$ if, and only if, $\mathbb{E}_{P^{n+1}}[(\|f_{u,P}\|_\infty|\hat{m}_n(x_0) - \hat{m}_n^{[1]}(x_0)|) \wedge 1] \leq \eta$ for all $P \in \mathcal{P}$. Also note that a 0-stable predictor can not depend on the training data in a non-trivial way (cf. Lemma B.4).

We are now in the position to state our main result on the estimation of $\tilde{F}_n(s) = P^{n+1}(y_0 - \hat{m}_n(x_0) \leq s\|T_n)$ by $\hat{F}_n(s) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{(-\infty,s]}(\hat{u}_i)$.

Theorem 2.5. *Suppose the class $\mathcal{P}$ satisfies Condition (C1) and the estimator $\hat{m}_n$ is symmetric and η-stable w.r.t. $\mathcal{P}$. Then, for every $P \in \mathcal{P}$, every $L \in [1,\infty)$ and every $\mu \in \mathbb{R}$, we have*

$$\begin{aligned}\mathbb{E}_{P^n}\left[\|\hat{F}_n - \tilde{F}_n\|_\infty\right] \;\leq\;& P(|y_0 - m_P(x_0)| > L)\\ &+ P^{n+1}(|m_P(x_0) - \hat{m}_n(x_0) - \mu| > L)\\ &+ 3\left(L\|f_{u,P}\|_\infty\left(\frac{1}{2n} + 3\eta\right)\right)^{1/3} + \sqrt{\frac{1}{n} + 6\eta}.\end{aligned}$$

For illustration and later use we also provide an asymptotic version of this result.

Corollary 2.6. *For $n \in \mathbb{N}$, let $p = p_n$ be a sequence of positive integers and let $\mathcal{P}_n$ be as in (C1) but with $\mathcal{X} = \mathcal{X}_n \subseteq \mathbb{R}^{p_n}$ depending on n. Suppose that for $P \in \mathcal{P}_n$, there exists $\sigma_P^2 = \sigma_{P,n}^2 \in (0,\infty)$ such that $\limsup_{n\to\infty} \sup_{P\in\mathcal{P}_n} \|f_{u/\sigma_P,P}\|_\infty < \infty$, where $f_{u/\sigma_P,P}(s) := \sigma_P f_{u,P}(s\sigma_P)$ is the scaled error density. Moreover, assume that the estimator $\hat{m}_n$ is symmetric and η_n-stable w.r.t. $\mathcal{P}_n$, such that $\eta_n \to 0$ as $n \to \infty$. If the scaled estimation errors $|m_P(x_0) - \hat{m}_n(x_0)|/\sigma_P$ and the scaled errors $|y_0 - m_P(x_0)|/\sigma_P$ both are $\mathcal{P}_n$-uniformly bounded, then*

$$\sup_{P\in\mathcal{P}_n} \mathbb{E}_{P^n}\left[\|\hat{F}_n - \tilde{F}_n\|_\infty\right] \xrightarrow[n\to\infty]{} 0.$$

Moreover, for $0 \le \alpha_1 < \alpha_2 \le 1$, the leave-one-out prediction interval is uniformly asymptotically conditionally valid, i.e.,

$$\sup_{P\in\mathcal{P}_n} \mathbb{E}_{P^n}\left[\left|P^{n+1}\left(y_0 \in PI_{\alpha_1,\alpha_2}^{(L1O)}(T_n, x_0)\middle\| T_n\right) - (\alpha_2 - \alpha_1)\right|\right] \xrightarrow[n\to\infty]{} 0.$$

Proof. Apply Theorem 2.5 with $L = l_n\sigma_P$, $\mu = 0$ and $l_n = o\left(\frac{1}{2n} + 3\eta_n\right)$, $l_n \to \infty$ as $n \to \infty$. For the second claim, note that under (C1), $P^n(\hat{u}_1 = \hat{u}_2) = 0$ and apply Lemma 2.1. □

Theorem 2.5 provides an upper bound on the risk of estimating the conditional prediction error distribution $\tilde{F}_n$ by the empirical distribution of the leave-one-out residuals $\hat{F}_n$. The upper bound crucially relies on the properties of the chosen estimator $\hat{m}_n$ for the true unknown regression function m_P. If the sample size is sufficiently large and if the estimator is sufficiently stable and has a moderate estimation error, then the parameter L can be chosen such that the upper bound is small. This is what we do in Corollary 2.6. It is important to note that Theorem 2.5 and Corollary 2.6 are informative also in case the estimator $\hat{m}_n$ is not consistent for m_P, as is often the case when $p/n \nrightarrow 0$. The bound of Theorem 2.5 also exhibits an interesting trade-off between the η-stability of $\hat{m}_n$ and the magnitude of its estimation error. More stable estimators are allowed to be less accurate whereas less stable estimators need to achieve higher accuracy in order to be as reliable for predictive inference purposes as a more stable algorithm.

The proof of Theorem 2.5 relies, among other things, on a result of Bousquet and Elisseeff (2002) which bounds the L^2-distance between the generalization error of a predictor $\hat{m}_n$ (i.e., $\mathbb{E}_{P^{n+1}}[(y_0 - \hat{m}_n(x_0))^2 \| T_n]$) and its estimate based on leave-one-out residuals, in terms of the stability properties of $\hat{m}_n$. See Section A.1 for details.

Theorem 2.5 and Corollary 2.6 show that the leave-one-out prediction interval in (2.5) is approximately uniformly conditionally valid, i.e., has the property that (2.1) is small at least for large n, provided that the underlying estimator $\hat{m}_n$ has two essential properties. First, the estimator must be η-stable with respect to the class $\mathcal{P}$ over which uniformity is desired, with an η value that is small if n is large. More precisely, we require

$$\eta_n = \sup_{P\in\mathcal{P}_n} \mathbb{E}_{P^{n+1}}\left[\left(\|f_{u,P}\|_\infty |\hat{m}_n(x_0) - \hat{m}_n^{[1]}(x_0)|\right) \wedge 1\right] \xrightarrow[n\to\infty]{} 0. \tag{2.8}$$

This is an intuitively appealing assumption since otherwise the leave-one-out residuals $\hat{u}_i = y_i - \hat{m}_n^{[i]}(x_i)$ may not be well suited to estimate the distribution of the prediction error $y_0 - \hat{m}_n(x_0)$. Second, the scaled estimation error $(m_P(x_0) - \hat{m}_n(x_0))/\sigma_P$ at the new observation x_0 must be bounded in probability, uniformly over the class $\mathcal{P}$. Formally,

$$\limsup_{n\to\infty} \sup_{P\in\mathcal{P}_n} P^{n+1}\left(\frac{|m_P(x_0) - \hat{m}_n(x_0)|}{\sigma_P} > M\right) \xrightarrow[M\to\infty]{} 0. \tag{2.9}$$

This is used to guarantee that the conditional distribution $\tilde{F}_n$ of the prediction error $y_0 - \hat{m}_n(x_0)$ given the training data is tight in an appropriate sense (cf. Lemma A.3(ii)), so that a pointwise bound on $|\hat{F}_n(t) - \tilde{F}_n(t)|$ can be turned into a uniform bound. The remainder of this paper is therefore mainly concerned with verifying these two conditions on the estimator $\hat{m}_n$ in several different contexts. From now on, as in Corollary 2.6, we will take on an asymptotic point of view.

3. Linear prediction with many variables.

In this section we investigate a scenario in which both consistent parameter estimation as well as bootstrap consistency fail (cf. Bickel and Freedman, 1983; El Karoui and Purdom, 2015), but the leave-one-out prediction interval is still asymptotically uniformly conditionally valid. See Section 4 for a discussion of scenarios where consistent parameter estimation is possible. For $\kappa \in [0, 1)$, we fix a sequence of positive integers (p_n), such that $p_n/n \to \kappa$ as $n \to \infty$ and $n > p_n + 1$ for all $n \in \mathbb{N}$. In case $\kappa > 0$, this type of 'large p, large n' asymptotics has the advantage that certain finite sample features of the problem are preserved in the limit, while offering a workable simplification. It turns out that conclusions drawn from this type of asymptotic analyses often provide remarkably accurate descriptions of finite sample phenomena.

When working with linear predictors $\hat{m}_n(x_0) = x_0'\hat{\beta}_n$, and if the feature vectors x_i have second moment matrix $\Sigma_P = \mathbb{E}_P[x_0x_0']$ under P, the conditions (2.8) and (2.9) can be verified as follows. For $\varepsilon > 0$,

$$\begin{aligned}
&\mathbb{E}_{P^{n+1}}\left[\left(\|f_{u,P}\|_\infty |\hat{m}_n(x_0) - \hat{m}_n^{[1]}(x_0)|\right) \wedge 1\right] \\
&\quad\le \left(1 \vee \|f_{u/\sigma_P,P}\|_\infty\right)\left(P^{n+1}\left(\frac{|x_0'\hat{\beta}_n - x_0'\hat{\beta}_n^{[1]}|}{\sigma_P} > \varepsilon\right) + \varepsilon\right) \\
&\quad\le \left(1 \vee \|f_{u/\sigma_P,P}\|_\infty\right)\left(\mathbb{E}_{P^n}\left[\left(\frac{1}{\varepsilon^2}\left\|\Sigma_P^{1/2}\left(\hat{\beta}_n - \hat{\beta}_n^{[1]}\right)/\sigma_P\right\|_2^2\right) \wedge 1\right] + \varepsilon\right),
\end{aligned}$$

where, for the second inequality, we have used the conditional Markov inequality along with independence of x_0 and T_n. Thus (2.8) follows if the scaled error densities $f_{u/\sigma_P,P}$, $P \in \mathcal{P}_n$, $n \in \mathbb{N}$, are uniformly bounded and

$$\sup_{P\in\mathcal{P}_n} P^n\left(\left\|\Sigma_P^{1/2}\left(\hat{\beta}_n - \hat{\beta}_n^{[1]}\right)\right\|_2 /\sigma_P > \varepsilon\right) \xrightarrow[n\to\infty]{} 0, \tag{3.1}$$

for each $\varepsilon > 0$. By a similar argument, we find that (2.9) follows if

$$\limsup_{n\to\infty} \sup_{P\in\mathcal{P}_n} P^n\left(\left\|\Sigma_P^{1/2}\left(\hat{\beta}_n - \beta_P\right)\right\|_2 /\sigma_P > M\right) \xrightarrow[M\to\infty]{} 0 \quad \text{and} \tag{3.2}$$

$$\limsup_{n\to\infty} \sup_{P\in\mathcal{P}_n} P\left(\frac{|m_P(x_0) - x_0'\beta_P|}{\sigma_P} > M\right) \xrightarrow[M\to\infty]{} 0, \tag{3.3}$$

for some vectors $\beta_P \in \mathbb{R}^{p_n}$, $P \in \mathcal{P}_n$, $n \ge 1$.

3.1. *Regularized M-estimators.* An important class of linear predictors for which our theory on the leave-one-out prediction interval applies are those based on regularized M-estimators investigated by El Karoui (2018) in the challenging scenario where p/n is not close to zero (see also Bean et al., 2013; El Karoui, 2013; El Karoui et al., 2013). For a given convex loss function $\rho : \mathbb{R} \to \mathbb{R}$ and a fixed tuning parameter $\gamma \in (0, \infty)$ (both not depending on n), consider the estimator

$$\hat{\beta}_n^{(\rho)} := \operatorname{argmin}_{b\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n \rho(y_i - x_i'b) + \frac{\gamma}{2}\|b\|_2^2. \tag{3.4}$$

In a remarkable tour de force, El Karoui (2018) studied the estimation error $\|\hat{\beta}_n^{(\rho)} - \beta\|_2$ as $p/n \to \kappa \in (0, \infty)$, in a linear model $y_i = x_i'\beta + u_i$, allowing for heavy tailed errors (including the Cauchy distribution) and non-spherical design (see Section 2.1 in El Karoui, 2018, for details on the technical assumptions). In particular, the author shows that $\|\hat{\beta}_n^{(\rho)} - \beta\|_2$

converges in probability to a deterministic positive and finite quantity $r_\rho(\kappa)$ and characterizes the limit through a system of non-linear equations. On the way to this result, El Karoui (2018, Theorem 3.9 together with Lemma 3.5 and the ensuing discussion) also establishes the stability property $\|\hat{\beta}_n^{(\rho)} - \hat{\beta}_{n,[1]}^{(\rho)}\|_2 \to 0$ in probability. Thus, under the assumptions maintained in that reference, (3.1), (3.2) and (3.3) hold, and the leave-one-out prediction interval (2.5) based on the linear predictor $\hat{m}_n(x_0) = x_0'\hat{\beta}_n^{(\rho)}$ is asymptotically conditionally valid, provided that also the boundedness condition $\limsup_{n\to\infty}\sup_{P\in\mathcal{P}_n}\|f_{u/\sigma_P,P}\|_\infty < \infty$ of Corollary 2.6 is satisfied. Finally, we note that a detailed assessment of the predictive performance of $\hat{\beta}_n^{(\rho)}$ in dependence on ρ requires a highly non-trivial analysis of $r_\rho(\kappa)$. For the asymptotic validity of the leave-one-out prediction interval, however, all the information needed on $r_\rho(\kappa)$ is that it is finite.

3.2. *James-Stein type estimators.* Another important example is the class of linear predictors $\hat{m}_n(x_0) = x_0'\hat{\beta}_n^{(JS)}$ based on James-Stein type estimators $\hat{\beta}_n^{(JS)}$ defined below. Here, we can allow for the following class of data generating processes.

(C2) Fix finite constants $C_0 > 0$ and $c_0 > 0$ and probability measures $\mathcal{L}_l$ and $\mathcal{L}_w$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, such that $\mathcal{L}_w$ has mean zero, unit variance and finite fourth moment, $\int s^2\mathcal{L}_l(ds) = 1$ and $\mathcal{L}_l((-c_0, c_0)) = 0$.

For every $n \in \mathbb{N}$, the class $\mathcal{P}_n = \mathcal{P}_n(\mathcal{L}_l, \mathcal{L}_w, C_0)$ consists of all probability measures on $\mathcal{Z}^n \subseteq \mathbb{R}^{p_n+1}$, such that the distribution of (x_0, y_0) under $P \in \mathcal{P}_n$ has the following properties: The x_0-marginal distribution of P is given by

$$x_0 \overset{\mathcal{L}}{=} l_0\Sigma_P^{1/2}(w_1, \dots, w_{p_n})',$$

where $w_1, \dots, w_{p_n}$ are i.i.d. according to $\mathcal{L}_w$, $l_0 \overset{\mathcal{L}}{=} \mathcal{L}_l$ is independent of the w_j and $\Sigma_P^{1/2}$ is the unique symmetric positive definite square root of a positive definite $p_n \times p_n$ covariance matrix Σ_P.

The response y_0 has mean zero and its conditional distribution given the regressors x_0 under P is

$$y_0\|x_0 \overset{\mathcal{L}}{=} m_P(x_0) + \sigma_P v_0,$$

where v_0 is independent of x_0 and has mean zero, unit variance and fourth moment bounded by C_0, where $m_P : \mathbb{R}^{p_n} \to \mathbb{R}$ is some measurable regression function with $\mathbb{E}_P[m_P(x_0)] = 0$ and $\sigma_P \in (0, \infty)$.

In words, under the distributions in $\mathcal{P}_n$, the feature/response pair (x_0, y_0) follows a non-Gaussian random design non-linear regression model with regression function m_P and error variance σ_P. Moreover, the feature vectors x_i are allowed to have a complex geometric structure, in the sense that the standardized design vector $\Sigma_P^{-1/2} x_1$ is not necessarily concentrated on a sphere of radius $\sqrt{p_n}$, as would be the case if $\mathcal{L}_l$ was supported on $\{-1, 1\}$ (see, e.g., El Karoui (2010, Section 3.2) and El Karoui (2018, Section 2.3.1) for further discussion of this point). The model $\mathcal{P}_n$ in (C2) is non-parametric, because the regression function m_P is unrestricted, up to being centered, and the error distribution is arbitrary, up to the requirements $\mathbb{E}_P[v_0] = 0$, $\mathbb{E}_P[v_0^2] = 1$ and $\mathbb{E}_P[v_0^4] \leq C_0$.

To predict the value of y_0 from x_0 and a training sample $T_n = (x_i, y_i)_{i=1}^n$ with $n \geq p_n + 2$, generated from P^n, we consider linear predictors $\hat{m}_n(x_0) = x_0' \hat{\beta}_n(c)$, where $\hat{\beta}_n(c)$ is a James-Stein-type estimator given by

$$\hat{\beta}_n(c) \;=\; \begin{cases} \left(1 - \frac{c p_n \hat{\sigma}_n^2}{\hat{\beta}_n' X' X \hat{\beta}_n}\right)_+ \hat{\beta}_n, & \text{if } \hat{\beta}_n' X' X \hat{\beta}_n > 0, \\ 0, & \text{if } \hat{\beta}_n' X' X \hat{\beta}_n = 0, \end{cases}$$

for a tuning parameter $c \in [0, 1]$. Here $\hat{\beta}_n = (X'X)^\dagger X'Y$, $\hat{\sigma}_n^2 = \|Y - X\hat{\beta}_n\|_2^2/(n - p_n)$. The corresponding leave-one-out estimator $\hat{\beta}_n^{[i]}(c)$ is defined equivalently, but with X and Y replaced by $X_{[i]}$ and $Y_{[i]}$. Note that the leave-one-out equivalent of $\hat{\sigma}_n^2 = \hat{\sigma}_n^2(X, Y)$ is given by

$$\hat{\sigma}_{n,[i]}^2(X_{[i]}, Y_{[i]}) = \hat{\sigma}_{n-1}^2(X_{[i]}, Y_{[i]}) = \|Y_{[i]} - X_{[i]}\hat{\beta}_n^{[i]}\|_2^2/(n - 1 - p_n).$$

The ordinary least squares estimator $\hat{\beta}_n$ belongs to the class of James-Stein estimators. In particular, $\hat{\beta}_n(0) = \hat{\beta}_n$, because, with $P_X := X(X'X)^\dagger X'$, we have $\|P_X Y\|_2^2 = \hat{\beta}_n' X' X \hat{\beta}_n = 0$ if, and only if, $Y \in \operatorname{span}(P_X)^\perp = \operatorname{span}(X)^\perp$, and the latter clearly implies $\hat{\beta}_n = 0$.

Using James-Stein type estimators for prediction is motivated, e.g., by the optimality results of Dicker (2013) and the discussion in Huber and Leeb (2013). The next result shows that in the model (C2) with $p_n/n \to \kappa \in (0, 1)$ and if the deviation from a linear model is not too severe, the James-Stein-type estimators are sufficiently stable and their estimation errors are uniformly bounded in probability, just as required in (3.1) and (3.2).

Theorem 3.1. *For every $n \in \mathbb{N}$, let $\mathcal{P}_n = \mathcal{P}_n(\mathcal{L}_l, \mathcal{L}_w, C_0)$ be as in Condition (C2) and suppose that under every $P \in \mathcal{P}_n$, the error term v_0 in (C2) has a Lebesgue density. For $P \in \mathcal{P}_n$, define β_P to be the minimizer of*

$\beta \mapsto \mathbb{E}_P[(y_0 - \beta' x_0)^2]$ over $\mathbb{R}^{p_n}$. If $p_n/n \to \kappa \in [0,1)$, $0 \le c_n \le 1$ for all $n \in \mathbb{N}$, and

$$\limsup_{n\to\infty} \sup_{P\in\mathcal{P}_n} \mathbb{E}_P\left[\left(\frac{m_P(x_0) - x_0'\beta_P}{\sigma_P}\right)^2\right] < \infty, \tag{3.5}$$

then the positive part James-Stein estimator $\hat\beta_n(c_n)$ satisfies (3.2), i.e.,

$$\limsup_{n\to\infty} \sup_{P\in\mathcal{P}_n} P^n\left(\left\|\Sigma_P^{1/2}(\hat\beta_n(c_n) - \beta_P)/\sigma_P\right\|_2 > M\right) \xrightarrow[M\to\infty]{} 0.$$

If, in addition, $\kappa > 0$, then for every $\varepsilon > 0$, (3.1) is also satisfied, i.e.,

$$\sup_{P\in\mathcal{P}_n} P^n\left(\left\|\Sigma_P^{1/2}(\hat\beta_n(c_n) - \hat\beta_n^{[1]}(c_n))/\sigma_P\right\|_2 > \varepsilon\right) \xrightarrow[n\to\infty]{} 0.$$

Remark 3.2. Under the assumptions of Theorem 3.1, uniform asymptotic conditional validity of the leave-one-out prediction interval follows, provided that, in addition, the errors v_0 in Condition (C2) have uniformly bounded densities (cf. Corollary 2.6). To see this, note that (3.1) and (3.2) are conclusions of the theorem, that uniformly bounded fourth moment of the error implies $\mathcal{P}_n$-uniform boundedness and that (3.3) is a consequence of assumption (3.5).

Remark 3.3. The last statement of Theorem 3.1 can also be established for the case $\kappa = 0$ but would require a slightly different proof strategy. Since this case is statistically less interesting we omit it for the sake of brevity.

3.3. *Ordinary least squares and interval length.*

We investigate the special case of the ordinary least squares predictor $\hat m_n(x) = x'\hat\beta_n = x'(X'X)^\dagger X'Y$ in some more detail, because here also the length

$$\left|PI_{\alpha_1,\alpha_2}^{(L1O)}\right| = \hat q_{\alpha_2} - \hat q_{\alpha_1},$$

of the leave-one-out prediction interval (2.5) permits a reasonably simple asymptotic characterization. We consider a class $\mathcal{P}_n^{(lin)} = \mathcal{P}_n^{(lin)}(\mathcal{L}_l, \mathcal{L}_w, \mathcal{L}_v)$ which is a subset of the one of Condition (C2), with the additional assumption that the regression function m_P is linear and that the error distribution is fixed (up to arbitrary scaling).

(C3) Fix a finite constant $c_0 > 0$ and probability measures $\mathcal{L}_l$, $\mathcal{L}_w$ and $\mathcal{L}_v$ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, such that $\mathcal{L}_w$ and $\mathcal{L}_v$ have mean zero, unit variance and finite fourth moment, $\int s^2 \mathcal{L}_l(ds) = 1$ and $\mathcal{L}_l((-c_0, c_0)) = 0$.

For every $n \in \mathbb{N}$, the class $\mathcal{P}_n^{(lin)} = \mathcal{P}_n^{(lin)}(\mathcal{L}_l, \mathcal{L}_w, \mathcal{L}_v)$ consists of all probability measures on $\mathbb{R}^{p_n+1}$, such that the distribution of (x_0, y_0) under $P \in \mathcal{P}_n$ has the following properties: The x_0-marginal distribution of P is given by

$$x_0 \overset{\mathcal{L}}{=} l_0 \Sigma_P^{1/2}(w_1, \dots, w_{p_n})',$$

where $w_1, \dots, w_{p_n}$ are i.i.d. according to $\mathcal{L}_w$, $l_0 \overset{\mathcal{L}}{=} \mathcal{L}_l$ is independent of the w_j and $\Sigma_P^{1/2}$ is the unique symmetric positive definite square root of a positive definite $p_n \times p_n$ covariance matrix Σ_P.
The conditional distribution of the response y_0 given the regressors x_0 under P is

$$y_0 \| x_0 \overset{\mathcal{L}}{=} x_0'\beta_P + \sigma_P v_0,$$

where $v_0 \overset{\mathcal{L}}{=} \mathcal{L}_v$ is independent of x_0, and where $\beta_P \in \mathbb{R}^{p_n}$ and $\sigma_P \in (0, \infty)$.

Note that under (C3), the distributions $\mathcal{L}_l$, $\mathcal{L}_w$ and $\mathcal{L}_v$ are fixed, so that $\mathcal{P}_n^{(lin)}$ is a parametric model indexed by β_P, Σ_P and σ_P. However, these parameters may depend on sample size n, and the dimension p_n of β_P and Σ_P may increase with n. Subsequently, we aim at uniformity in these parameters.

Theorem 3.4. *Fix $\alpha \in [0,1]$. For every $n \in \mathbb{N}$, let $\mathcal{P}_n = \mathcal{P}_n^{(lin)}(\mathcal{L}_l, \mathcal{L}_w, \mathcal{L}_v)$ be as in (C3). If $p_n/n \to \kappa \in (0,1)$ then the scaled empirical α-quantile $\hat{q}_\alpha/\sigma_{P_n}$ of the leave-one-out residuals $\hat{u}_i = y_i - x_i'\hat{\beta}_n^{[i]}$ based on the OLS estimator $\hat{\beta}_n = (X'X)^\dagger X'Y$ converges $\mathcal{P}_n$-uniformly in probability to the corresponding α-quantile q_α of the distribution of*

$$lN\tau + v$$

and l, N, τ and v are defined as follows: $l \overset{\mathcal{L}}{=} \mathcal{L}_l$, $N \overset{\mathcal{L}}{=} \mathcal{N}(0,1)$, and $v \overset{\mathcal{L}}{=} \mathcal{L}_v$ are independent, and $\tau = \tau(\mathcal{L}_l, \kappa)$ is non-random.

The same statement holds also for $\kappa = 0$, provided that, in addition, $\mathcal{L}_v$ has a continuous and strictly increasing cdf and $p_n \to \infty$ as $n \to \infty$.

Here, the function $\kappa \mapsto \tau(\mathcal{L}_l, \kappa) \in [0, \infty)$ defined on $[0,1)$ has the following properties: For any $\mathcal{L}_l$ as in (C3), $\tau(\mathcal{L}_l, \kappa) = 0$ if, and only if, $\kappa = 0$. If $\mathcal{L}_l(\{-1,1\}) = 1$, then $\tau(\mathcal{L}_l, \kappa) = \sqrt{\kappa/(1-\kappa)}$.

Theorem 3.4 shows how the length $\hat{q}_{\alpha_2} - \hat{q}_{\alpha_1}$ of the leave-one-out prediction interval for the OLS predictor depends (asymptotically) on $\mathcal{L}_l$, $\mathcal{L}_v$ and $\kappa =$

$\lim_{n\to\infty} p_n/n$. For simplicity, let $\mathcal{L}_l(\{-1,1\}) = 1$ and consider an equal tailed interval, i.e., $\alpha_1 = \alpha/2 = 1-\alpha_2$. Figure 1 shows asymptotic interval lengths as functions of $\kappa \in [0,1]$ for different values of error level α in the cases $\mathcal{L}_v = \text{Unif}\{-1,1\}$ and $\mathcal{L}_v = \mathcal{N}(0,1)$. For a wide range of κ values ($\kappa \in [0,0.8]$), the interval length is almost constant. However, for high dimensional problems ($\kappa > 0.8$) the interval length increases dramatically, as expected, because here the asymptotic estimation error $\tau = \sqrt{\kappa/(1-\kappa)}$ explodes. We also get an idea about the impact of the error distribution, on which the practitioner has no handle. In particular, for large error levels ($\alpha = 0.6$) we even observe a non-monotonic dependence of the interval length on κ, which seems rather counterintuitive. This results from the non-monotonicity of $\tau^2 \mapsto IQR_\alpha(\mathcal{N}(0,\tau^2) * \mathcal{L}_v) = q_{1-\alpha/2} - q_{\alpha/2}$, where $*$ denotes convolution, which may only occur if the error distribution $\mathcal{L}_v$ is not log-concave (e.g., the blue curve for $\alpha = 0.6$ in Figure 1; cf. the discussion in Section 5.1). Finally, for large values of κ, and thus, for large values of τ, the error distribution has little effect on the interval length, because in that case the term $N\tau$ dominates the distribution of $N\tau + v$.

The result of Theorem 3.4 can be intuitively understood as follows. If the true model $\mathcal{P}_n^{(lin)}$ is linear and satisfies (C3) then the scaled prediction error under $P \in \mathcal{P}_n^{(lin)}$ is distributed as

$$\frac{y_0 - \hat{m}_n(x_0)}{\sigma_P} \overset{\mathcal{L}}{=} l_0(w_1,\dots,w_{p_n})\Sigma_P^{1/2}(\beta_P - \hat{\beta}_n)/\sigma_P + v_0,$$

and for n large, $\|\Sigma_P^{1/2}(\beta_P - \hat{\beta}_n)/\sigma_P\|_2 \approx \tau$ is approximately non-random, so that $(w_1,\dots,w_{p_n})\Sigma_P^{1/2}(\beta_P - \hat{\beta}_n)/\sigma_P \approx w_0' Z\tau$, where $Z := \Sigma_P^{1/2}(\beta_P - \hat{\beta}_n)/\|\Sigma_P^{1/2}(\beta_P - \hat{\beta}_n)\|_2$ is a random unit vector that is independent of $w_0 = (w_1,\dots,w_{p_n})'$. Thus, if p_n is large and Z satisfies the Lyapounov condition $\|Z\|_{2+\delta} \to 0$, then $w_0'Z \approx \mathcal{N}(0,1)$ (see Lemma A.7(ii)). This effect of additional Gaussian noise in the prediction error was also observed by El Karoui (2013, 2018); El Karoui et al. (2013); El Karoui and Purdom (2015). Note, however, that the conditions $\|\Sigma_P^{1/2}(\beta_P - \hat{\beta}_n)/\sigma_P\|_2 \approx \tau$ and $\|Z\|_{2+\delta} \to 0$ are not necessarily satisfied for any estimator $\hat{\beta}_n$. The former condition is indeed more generally satisfied by robust M-estimators of the form

$$\hat{\beta}_n^{(\rho)} = \text{argmin}_{b\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n \rho(y_i - x_i'b),$$

considered in El Karoui (2013, 2018) and under the model assumptions in that reference (cf. Section 3.1). Here, $\rho : \mathbb{R} \to \mathbb{R}$ is an appropriate convex loss function. If $\|\Sigma_P^{1/2}(\beta_P - \hat\beta_n^{(\rho)})/\sigma_P\|_2 \approx \tau < \infty$ holds, then the Lyapounov condition $\|Z\|_{2+\delta} \to 0$ is also satisfied by $\hat\beta_n^{(\rho)}$, provided that the standardized design vectors $\Sigma_P^{-1/2}x_i$ follow an orthogonally invariant distribution, because then one easily sees that

$$\hat\beta_n^{(\rho)} \;=\; \beta_P + \Sigma_P^{-1/2}\tilde\beta_n^{(\rho)} \;\overset{\mathcal{L}}{=}\; \beta_P + \|\tilde\beta_n^{(\rho)}\|_2\Sigma_P^{-1/2}U,$$

where $\tilde\beta_n^{(\rho)} = \operatorname{argmin}_{b\in\mathbb{R}^p}\frac{1}{n}\sum_{i=1}^n \rho(u_i - x_i'\Sigma_P^{-1/2}b)$ and U is uniformly distributed on the unit sphere and independent of $\|\tilde\beta_n^{(\rho)}\|_2 = \|\Sigma_P^{1/2}(\beta_P - \hat\beta_n^{(\rho)})/\sigma_P\|_2$, which is itself approximately constant equal to τ. However, this distributional invariance of the estimator, which is required for the Lyapounov property to hold, is not satisfied, e.g., by the James-Stein estimators (cf. Lemma B.3). If the mentioned conditions are not satisfied, much more complicated limiting distributions of the prediction error than the one of Theorem 3.4 may arise.

3.4. *Sample splitting.* An obvious alternative to the leave-one-out prediction interval (2.5) is to use a sample splitting method as follows. Decide on a fraction $\nu \in (0,1)$ and use only a number $n_1 = \lceil \nu n\rceil$ of observation pairs (x_i, y_i), $i \in S_\nu \subseteq \{1,\dots,n\}$, $|S_\nu| = n_1$, to compute an estimate $\hat m_{n_1}$. Now use the remaining $n - n_1$ observations to compute residuals $\hat u_i^{(\nu)} = y_i - \hat m_{n_1}(x_i)$, $i \in \{1,\dots,n\}\setminus S_\nu$. Since, conditionally on the observations corresponding to S_ν, these residuals are i.i.d. and distributed as $y_0 - \hat m_{n_1}(x_0)$, constructing a prediction interval of the form $[\hat m_{n_1}(x_0) + L, \hat m_{n_1}(x_0) + U]$ for y_0 is now equivalent to constructing a tolerance interval for $y_0 - \hat m_{n_1}(x_0)$ based on i.i.d. observations with the same distribution. One can now simply use appropriate empirical quantiles $L = \hat q_{\alpha_1}^{(\nu)}$ and $U = \hat q_{\alpha_2}^{(\nu)}$ from the sample splitting residuals $\hat u_i^{(\nu)}$ (see also Section 5.2). Such a procedure is suggested, e.g., by Vovk (2012) and Lei et al. (2017).

In order to formally study the length of this sample splitting interval we restrict to the case of OLS estimation, i.e., $\hat m_{n_1}(x) = x'\hat\beta_{n_1}$. Note that in this case, the estimator will not be unique if $n_1 < p_n$, so one usually requires $n_1 \geq p_n$. Now, by the same mechanism as discussed in Section 3.3, the empirical quantiles of the residuals $\hat u_i^{(\nu)}$, $i \in S_\nu^c$, converge (unconditionally) to the quantiles of $lN\tau' + u$, where now τ' is the non-random limit of $\|\Sigma_P^{1/2}(\beta_P - \hat\beta_{n_1})/\sigma_P\|_2$. In particular, if $\mathcal{L}_l$ degenerates to $\{-1,1\}$, then $\tau' = \sqrt{\kappa'/(1-\kappa')}$, where $\kappa' = \lim_{n\to\infty} p_n/n_1 = \kappa/\nu$. Thus, we can read off the asymptotic interval length of the sample splitting procedure from Figure 1 by simply adjusting the value of κ to κ/ν. For instance, in the

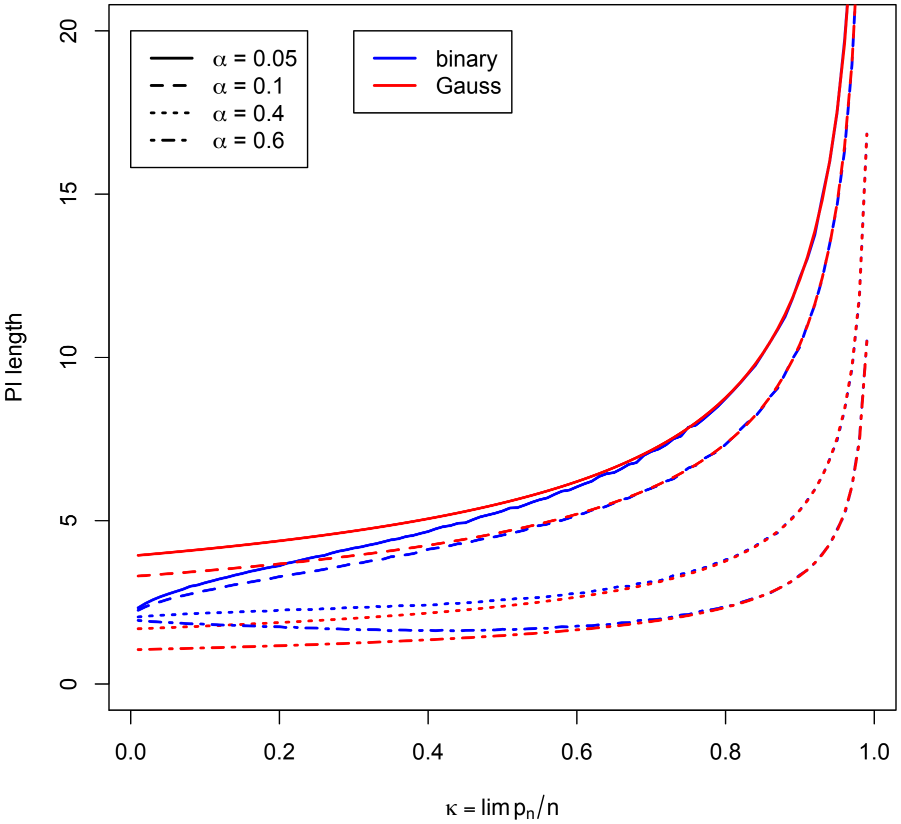

Fig 1. *Lengths of leave-one-out prediction intervals as a function of* $\kappa = \lim_{n\to\infty} p_n/n$ *for confidence level* $1-\alpha$ *and with Unif*$\{-1,1\}$ *(binary) and* $\mathcal{N}(0,1)$ *(Gauss) errors.*

binary error case with $\alpha = 0.05$, if $\kappa = 0.4$ and we use sample splitting with $\nu = 1/2$, then $\kappa' = 0.8$ and the asymptotic length of the leave-one-out prediction interval is about 4.7, while the asymptotic length of the sample splitting interval is about 9, so almost twice as wide.

4. Asymptotically degenerate (non-random) estimators.

Another important class of problems, where the conditions (2.8) and (2.9) of Subsection 2.2 are satisfied, are those where the estimator $\hat{m}_n$ asymptotically degenerates to some non-random function which need not be the true regression function $m_P : \mathcal{X} \to \mathbb{R}$. We point out that in the scenario considered in this section, the naive approach that tries to estimate the true unknown distribution of the errors u_i in the additive error model (C1) based on the ordinary residuals $y_i - \hat{m}_n(x_i)$ is often successful (asymptotically) for constructing conditionally valid prediction intervals, provided consistent estimation of m_P. This less challenging but more classical setting of asymptotically non-random predictors is an important test case for the leave-one-out method. We still consider asymptotic results where the number of explanatory variables $p = p_n$ can grow with sample size n. Thus, we consider a sequence $(p_n)_{n\in\mathbb{N}}$ and a sequence $(\mathcal{P}_n)_{n\in\mathbb{N}}$ of collections of probability measures on $\mathcal{Z}_n \subseteq \mathbb{R}^{p_n+1}$. Moreover, we have to slightly extend the usual definition of uniform consistency of an estimator sequence to cover also the leave-one-out estimate and the possibility of an asymptotically non-vanishing bias.

Definition 2 (Uniform Asymptotic Degeneracy (UAD)). *For every $n \in \mathbb{N}$, let $p_n \in \mathbb{N}$, let $\mathcal{P}_n$ be a collection of probability measures on $\mathcal{Z}_n$ and let $\sigma_n^2 : \mathcal{P}_n \to (0,\infty)$ be a positive functional on $\mathcal{P}_n$. We say that a sequence of symmetric predictors $\hat{m}_n(\cdot) = M_{n,p_n}(T_n, \cdot)$ is uniformly asymptotically degenerate (UAD) with respect to $(\mathcal{P}_n)_{n\in\mathbb{N}}$ and relative to $(\sigma_n^2)_{n\in\mathbb{N}}$, if there exists measurable functions $g_P : \mathbb{R}^{p_n} \to \mathbb{R}$, such that for every $\varepsilon > 0$,*

$$(4.1) \qquad \sup_{P\in\mathcal{P}_n} P^{n+1}\Big(|g_P(x_0) - M_{n,p_n}(T_n, x_0)| > \varepsilon\sigma_n(P)\Big) \xrightarrow[n\to\infty]{} 0 \quad \textit{and}$$

$$(4.2) \qquad \sup_{P\in\mathcal{P}_n} P^{n}\Big(|g_P(x_0) - M_{n-1,p_n}(T_n^{[1]}, x_0)| > \varepsilon\sigma_n(P)\Big) \xrightarrow[n\to\infty]{} 0.$$

The functional $\sigma_n^2(P)$ can be thought of, for instance, as the error variance $\sigma_n^2(P) = \mathrm{Var}_P[y_0 - m_P(x_0)]$, if it exists. Of course, conditions (4.1) and (4.2) coincide if the sequences (p_n), (σ_n^2) and $(\mathcal{P}_n)$ are constant. It is also easy to see that if $\hat{m}_n$ is UAD with respect to $(\mathcal{P}_n)$ and relative to $(\sigma_n^2)_{n\in\mathbb{N}}$, then the sequence of stability constants η_n satisfies (2.8), i.e.,

$$\eta_n = \sup_{P\in\mathcal{P}_n} \mathbb{E}_{P^{n+1}}\left[\left(\sigma_n(P)\|f_{u,P}\|_\infty \frac{|\hat{m}_n(x_0) - \hat{m}_n^{[1]}(x_0)|}{\sigma_n(P)}\right)\wedge 1\right] \xrightarrow[n\to\infty]{} 0,$$

provided that $\limsup_{n\to\infty}\sup_{P\in\mathcal{P}_n}\sigma_n(P)\|f_{u,P}\|_\infty < \infty$. Note that $f_{u/\sigma_n,P}(v) = \sigma_n f_{u,P}(\sigma_n v)$ is the density of the scaled error term $(y_0 - m_P(x_0))/\sigma_n$ under P, with $\sigma_n = \sigma_n(P)$. Furthermore, it is equally obvious that the UAD property of $\hat{m}_n$ together with $\limsup_{n\to\infty}\sup_{P\in\mathcal{P}_n} P(|m_P(x_0) - g_P(x_0)| > M\sigma_n(P)) \to 0$, as $M\to\infty$, implies (2.9).

In the remainder of this subsection we list a number of examples where the UAD property of $\hat{m}_n$, and therefore (assuming (C1) and the mentioned boundedness conditions, including the one on the error $(y_0 - m_P(x_0))/\sigma_P$, c.f. Corollary 2.6) also asymptotic conditional validity of the leave-one-out prediction interval, holds. We emphasize that the conditions on the statistical model $\mathcal{P}$, that are imposed in the subsequent examples, are taken from the respective reference and we do not claim that they are minimal.

Example 4.1 (Non-parametric regression estimation). Consider a constant sequence of dimension parameters $p_n = p \in \mathbb{N}$. For positive finite constants L and C, let $\mathcal{P}(L,C)$ denote the class of probability distributions P on $\mathcal{Z} = \mathcal{X}\times\mathcal{Y} \subseteq \mathbb{R}^{p+1}$ such that $P(|y_0|\le L) = 1 = P(\|x_0\|_2 \le L)$ and whose corresponding regression function $m_P : \mathbb{R}^p \to \mathbb{R}$ is C-Lipschitz, i.e., $|m_P(x_1) - m_P(x_2)| \le C\|x_1 - x_2\|_2$ for all $x_1, x_2 \in \mathcal{X}$. Györfi et al. (2002, Chapter 7) show that if $\hat{m}_n$ is either an appropriate kernel estimate, a partitioning estimate or a nearest-neighbor estimate, all with fully data driven choice of tuning parameter, then

$$\sup_{P\in\mathcal{P}(L,C)} P^{n+1}(|\hat{m}_n(x_0) - m_P(x_0)| > \varepsilon) \xrightarrow[n\to\infty]{} 0$$

for every $\varepsilon > 0$. Because of the data driven choice of tuning parameter, which is usually done by a sample splitting procedure, the estimators in Györfi et al. (2002) are generally not symmetric in the input data. However, it is easy to see that symmetrized versions of those estimators are still UAD. Simply note that it is no restriction to assume $|\hat{m}_n(x_0) - m_P(x_0)| \le 2L$, such that convergence in probability and converges in L_1 are equivalent, and study the L_1 estimation error of the symmetrized estimator.

Example 4.2 (High-dimensional linear regression with the LASSO). Consider a non-decreasing sequence $(K_n)_{n\in\mathbb{N}}$ of positive real numbers and a sequence of dimension parameters $(p_n)_{n\in\mathbb{N}}$ such that $K_n^4 \log(p_n)/n \to 0$ as $n \to \infty$. For a positive finite constant M, let $\mathcal{P}_n(M)$ denote the class of probability distributions on $\mathbb{R}^{p_n+1}$, such that under $P \in \mathcal{P}_n(M)$, the pair (x_0, y_0) has the following properties:

- $\|x_0\|_\infty \leq M$, almost surely.
- Conditional on x_0, y_0 is distributed as $\mathcal{N}(x_0'\beta_P, \sigma_P^2)$, for some $\beta_P \in \mathbb{R}^{p_n}$ and $\sigma_P^2 \in (0, \infty)$.
- The parameters β_P and σ_P^2 satisfy $\max(\|\beta_P\|_1, \sigma_P) \leq K_n$.

In particular, we have $m_P(x_0) = x_0'\beta_P$. Chatterjee (2013, Theorem 1) shows that any estimator $\hat{\beta}_n^{(K_n)}$ which minimizes

$$\beta \quad \mapsto \quad \sum_{i=1}^n (y_i - \beta' x_i)^2 \quad \text{subject to} \quad \|\beta\|_1 \leq K_n$$

satisfies

$$\sup_{P\in\mathcal{P}_n(M)} P^{n+1}\left(\left|x_0'\hat{\beta}_n^{(K_n)} - m_P(x_0)\right| > \varepsilon\right) \xrightarrow[n\to\infty]{} 0$$

for every $\varepsilon > 0$. Clearly, here the leave-one-out estimate has the same asymptotic property, because $K_{n-1}^4 \log p_n/(n-1) \to 0$. Note that in this example, consistent estimation of the parameters β_P and σ_P^2 would require additional assumptions on the distribution of the feature vector x_0 (so called 'compatibility conditions', see Bühlmann and van de Geer (2011)), and therefore, it is not immediately clear whether the standard Gaussian prediction interval based on estimates $\hat{\beta}_n$ and $\hat{\sigma}_n^2$ and a Gaussian quantile is asymptotically valid in the present setting. Furthermore, the result of Chatterjee (2013) can be extended also to the non-Gaussian case, where the standard Gaussian prediction interval certainly fails.

Example 4.3 (Ridge regression with many variables). A qualitatively different parameter space is considered in Lopes (2015), who shows uniform consistency of ridge regularized estimators in a linear model under a boundedness assumption on the regression parameter β_P and a specific decay rate of eigenvalues of $\mathbb{E}_P[x_0 x_0']$.

Example 4.4 (Misspecified regression estimation). A classical strand of literature on the asymptotics of Maximum-Likelihood under misspecification has established various conditions under which the MLE is not consistent for the true unknown parameter, but for a pseudo parameter that corresponds to the projection of the true data generating distribution onto the maintained working model. See, for example, Huber (1967), White (1980a,b) or Fahrmeir (1990). A common pseudo target in random design regression is the minimizer of $\beta \mapsto \mathbb{E}_P[(y_0 - \beta' x_0)^2]$.

5. Discussion and further remarks.

In this section we collect several further thoughts on the leave-one-out prediction intervals. We discuss some properties of the proposed method that we have established above but which we believe hold in much higher generality. We also draw some further connections to other methods such as sample splitting, tolerance regions and prediction regions based on non-parametric density estimation, and we provide further intuition. Finally, we sketch possible extensions and open problems.

5.1. *Predictor efficiency and interval length.*

Recall that if $T_n \in \mathcal{Z}^n$ and P are such that

$$s \mapsto \tilde{F}_n(s; T_n) = P^{n+1}(y_0 - \hat{m}_n(x_0) \le s \| T_n),$$

is continuous, the optimal infeasible interval

$$PI^{(OPT)}_{\alpha_1,\alpha_2} = \hat{m}_n(x_0) + (\tilde{q}_{\alpha_1}, \tilde{q}_{\alpha_2}]$$

in (2.2) is the shortest interval of the form $\hat{m}_n(x_0) + (L(T_n), U(T_n)]$ such that (2.3) and (2.4) are satisfied. In this infeasible scenario, the only way in which the 'user' can influence the length of $PI^{(OPT)}_{\alpha_1,\alpha_2}$ is via the choice of predictor $\hat{m}_n$. This choice clearly affects the conditional distribution $\tilde{F}_n$ of the prediction error $y_0 - \hat{m}_n(x_0)$ and, thus, potentially its inter-quantile-range $\tilde{q}_{\alpha_2} - \tilde{q}_{\alpha_1}$. Since we only care about minimizing the inter-quantile-range of the conditional distribution $\tilde{F}_n$, for the rest of this subsection we consider the training data T_n to be fixed and non-random. Thus, the predictor $\hat{m}_n : \mathbb{R}^p \to \mathbb{R}$ is also non-random. Now we would like to use a predictor $\hat{m}_n$ such that the prediction error $y_0 - \hat{m}_n(x_0)$ has short inter-quantile-range. For simplicity, assume that $y_0 = m_P(x_0) + u_0$, where the error term u_0 has mean zero and is independent of the features x_0. Therefore, the prediction error is given by

$$y_0 - \hat{m}_n(x_0) \;=\; m_P(x_0) - \hat{m}_n(x_0) \;+\; u_0,$$

i.e., the sum of the estimation error $m_P(x_0) - \hat{m}_n(x_0)$ and the innovation u_0. Following Lewis and Thompson (1981), we say that a continuous univariate distribution P_1 is more dispersed than P_0 if, and only if, any two quantiles of P_1 are further apart than the corresponding quantiles of P_0. Now we note that minimizing the inter-quantile-rage of the prediction error $y_0 - \hat{m}_n(x_0)$ is, in general, not equivalent to minimizing the inter-quantile-rage of $m_P(x_0) - \hat{m}_n(x_0)$, because of the effect of the error term u_0. However, if the distribution of the error term u_0 has a log-concave density, then the distribution of $y_0 - \hat{m}_n^{(1)}(x_0)$ is more dispersed than that of $y_0 - \hat{m}_n^{(0)}(x_0)$, if, and only if, $m_P(x_0) - \hat{m}_n^{(1)}(x_0)$ is more dispersed than $m_P(x_0) - \hat{m}_n^{(0)}(x_0)$ (see Theorem 8 of Lewis and Thompson, 1981). Thus, under log-concave error distributions, interval length of $PI_{\alpha_1,\alpha_2}^{(OPT)}$ is directly related to prediction accuracy of the point predictor $\hat{m}_n$ in use. These considerations naturally carry over to the feasible analog $PI_{\alpha_1,\alpha_2}^{(L1O)}$ defined in (2.5). In Section 3.3, in the special case of a linear model and ordinary-least-squares prediction, we have discussed the issue of interval length in some more detail and provided a rigorous description of the asymptotic interval length in a high-dimensional regime. This sheds more light on the connection between the length of $PI_{\alpha_1,\alpha_2}^{(L1O)}$ and the estimation error $m_P(x_0) - \hat{m}_n(x_0)$. However, the lessons learned from the linear model appear to be valid in a much more general situation. In particular, we see that, at least for log-concave error distributions, the lengths of leave-one-out prediction intervals can be used to evaluate the relative efficiency of competing predictors.

5.2. *The case of a naive predictor and sample splitting.* Next, we discuss the important special case where we naively decide to work with a predictor $M_{n,p}(T_n, x_0) = m(x_0)$, $m : \mathcal{X} \to \mathbb{R}$, that does not depend on the training data T_n at all.[3] In this case, the predictor and its leave-one-out analog coincide and the (leave-one-out) residuals $\hat{u}_i = y_i - m(x_i)$ for $i = 1, \dots, n$, are actually independent and identically distributed according to the non-random distribution $\tilde{F}_n(s) = P^{n+1}(y_0 - m(x_0) \le s \| T_n) = P(y_0 - m(x_0) \le s)$ and $\hat{F}_n$ is their empirical distribution function. Therefore, by Lemma 2.1 and the Dvoretzky-Kiefer-Wolfowitz (DKW) inequality (Massart, 1990), if $\tilde{F}_n$ is continuous, we get for every $\varepsilon > 0$ that

$$P^n\left(\left| P\left(y_0 \in PI_{\alpha_1,\alpha_2}^{(L1O)}(T_n, x_0)\right) - \frac{\lceil n\alpha_2 \rceil - \lceil n\alpha_1 \rceil}{n} \right| > \varepsilon\right) \le 2\exp\left(-\frac{n\varepsilon^2}{2}\right).$$

[3] Note that this covers, in particular, the case where we do not even use, or do not have available, the feature vectors $x_0, \dots, x_n$, i.e., $m \equiv 0$. In this case, a *prediction interval* for y_0 that is only based on $y_1, \dots, y_n$ is more commonly referred to as a *tolerance interval*.

Integrating this tail probability also yields

$$\sup_{P\in\mathcal{P}} \mathbb{E}_{P^n}\left[\left|P\left(y_0 \in PI_{\alpha_1,\alpha_2}^{(L1O)}(T_n, x_0)\right) - \frac{\lceil n\alpha_2\rceil - \lceil n\alpha_1\rceil}{n}\right|\right] \le \sqrt{\frac{2\pi}{n}},$$

where $\mathcal{P}$ contains all probability measures on $\mathbb{R}^{p+1}$ for which $\tilde{F}_n$ is continuous. We also point out that in the present case where the predictor does not depend on T_n, the problem of constructing a prediction interval for y_0 can actually be reduced to finding a non-parametric univariate tolerance interval for $y_0 - m(x_0)$ based on the i.i.d. copies $(y_i - m(x_i))_{i=1}^n$. For this problem classical solutions are available, based on the theory of order statistics of i.i.d. data (cf. Krishnamoorthy and Mathew, 2009, Chapter 8). Unfortunately, the problem changes dramatically, once we try to learn the true regression function m_P from the training data T_n and use $\hat{m}_n(x_0) = M_{n,p}(T_n, x_0)$ to predict y_0, because then the leave-one-out residuals are no longer independent and the conditional distribution function $\tilde{F}$ of the prediction error $y_0 - \hat{m}_n(x_0)$ given T_n is random. Thus, in the general case we can not expect to obtain equally powerful and elegant results as above and we can not resort to the theory of order statistics of i.i.d. data. In particular, we note that the bound of Theorem 2.5 is still somewhat sub-optimal in this trivial case where the estimator does not depend on the training sample T_n. In that case, $\eta = 0$, but the derived bound still depends on the distribution of the estimation error $m_P(x_0) - m(x_0)$, even though in that case the alternative bound obtained above by the DKW inequality does no longer involve the estimation error. It is an open problem to establish a concentration inequality for $\|\hat{F}_n - \tilde{F}_n\|_\infty$ analogous to the DKW inequality but in the general case of dependent leave-one-out residuals and random $\tilde{F}_n$.

The discussion of the previous paragraph also applies to the case where the predictor m was obtained as an estimator for m_P, but from another independent training sample $S_k = (x_j^*, y_j^*)_{j=1}^k$ of k i.i.d. copies of (x_0, y_0). This situation can be seen as a sample splitting method, where k of the overall $n+k$ observations are used to compute the point predictor $m = \hat{m}_k$ and the remaining n observations in T_n are used as a validation set to estimate the conditional distribution of the prediction error $y_0 - \hat{m}_k(x_0)$ given S_k (and T_n), from the (conditionally on S_k) i.i.d. residuals $y_i - \hat{m}_k(x_i)$, $i = 1, \dots, n$. Such a procedure is discussed, for instance, by Lei et al. (2017) and Vovk (2012). Note that under the assumptions of the previous paragraph, such a method is asymptotically conditionally valid if the size n of the validation set diverges to infinity. However, this method uses only k of the $n+k$ available observation pairs for prediction, such that the point predictor $\hat{m}_k$ based on S_k is not as efficient as the analogous predictor based on the full sample $S_k \cup T_n$. This typically results in a larger prediction interval than necessary, because then the conditional distribution of the prediction error $y_0 - \hat{m}_k(x_0)$ is usually more dispersed than that of $y_0 - \hat{m}_{k+n}(x_0)$. See also the discussion in Subsections 3.4 and 5.1.

5.3. *Further remarks.*

Remark 5.1 (On exact conditional validity). Suppose that the class $\mathcal{P}$ contains at least the data generating distributions P_0 and P_1, where for $j \in \{0, 1\}$

$$P_j = \mathcal{N}_{p+1}(0, \sigma_j^2 I_{p+1}), \quad \sigma_j^2 > 0,\ \sigma_0^2 \neq \sigma_1^2,$$

and that we decide to predict y_0 by some linear predictor $\hat{m}_n(x_0) = x_0'\hat{\beta}_n$. We shall show that for every $\alpha \in (0, 1/2)$, it is impossible to construct a prediction interval of the form $PI_\alpha(T_n, x_0) = x_0'\hat{\beta}_n + [L_\alpha(T_n), U_\alpha(T_n)]$ based on a finite sample T_n and x_0, such that (2.1) is equal to zero.

Proof. If (2.1) is equal to zero, then for both $j = 0, 1$ and P_j^n-almost all samples T_n,

$$\begin{aligned} 1 - \alpha &= P_j^{n+1}(y_0 \in PI_\alpha(T_n, x_0) \| T_n) \\ &= P_j^{n+1}(L_\alpha(T_n) \leq y_0 - x_0'\hat{\beta}_n \leq U_\alpha(T_n) \| T_n) \\ &= \Phi\left(\frac{U_\alpha(T_n)}{\sigma_j\sqrt{\|\hat{\beta}_n\|_2^2 + 1}}\right) - \Phi\left(\frac{L_\alpha(T_n)}{\sigma_j\sqrt{\|\hat{\beta}_n\|_2^2 + 1}}\right). \end{aligned}$$

Since $1 - \alpha > 1/2$, we must have $L_\alpha < 0 < U_\alpha$, almost surely, and it is easy to see that the function

$$g_{l,u}(\nu) := \Phi\left(\frac{u}{\nu}\right) - \Phi\left(\frac{l}{\nu}\right), \quad g_{l,u} : (0, \infty) \to (0, 1),$$

is continuous and strictly decreasing, provided that $l < 0 < u$, and thus, for such l and u, $g_{l,u}$ is invertible. Therefore, for $j = 0, 1$ and for P_j^n-almost all samples T_n (equivalently, for Lebesgue almost all $T_n \in \mathbb{R}^{n(p+1)}$), we have

$$\tilde{\sigma}_n^2(T_n) := \left(\frac{g_{L_\alpha, U_\alpha}^{-1}(1 - \alpha)}{\sqrt{\|\hat{\beta}_n\|_2^2 + 1}}\right)^2 = \sigma_j^2.$$

In other words, there exists $T_n \in \mathcal{Z}^n$, such that $\sigma_0^2 = \tilde{\sigma}_n^2(T_n) = \sigma_1^2$, a contradiction. □

Remark 5.2. Consistent estimation of the true regression function $m_P : \mathcal{X} \to \mathbb{R}$ from an i.i.d. sample of size n is usually not possible if the dimension p of $\mathcal{X}$ is non-negligible compared to n. For example, in a Gaussian linear model where the only unknown parameter is the p-vector β of regression coefficients, it is impossible to consistently estimate the conditional mean $m_P(x_0) = \mathbb{E}_P[y_0 \| x_0] = \beta' x_0$, unless $p/n \to 0$ or strong assumptions are imposed on the parameter space (cf. Dicker, 2012).

Remark 5.3. A natural approach for constructing non-parametric prediction sets is to estimate the conditional density of y_0 given x_0 (if it exists), because, as can be easily shown, a highest density region of the conditional density of y_0 given x_0 provides the smallest (in terms of Lebesgue measure) prediction region $PR_\alpha(x_0)$ for y_0 that controls the conditional coverage probability given x_0, i.e., that satisfies

$$P(y_0 \in PR_\alpha(x) \| x_0 = x) \; \geq \; 1 - \alpha \quad \text{for } P\text{-almost all } x. \tag{5.1}$$

This condition has been called *object conditional validity* by Vovk (2013). However, object conditional validity is often too much to ask for. First of all, as shown by Barber et al. (2019a) (see also Lei and Wasserman, 2014; Vovk, 2013), for continuous distributions there are no non-trivial prediction sets based on a finite sample that satisfy (5.1). Moreover, even if we are content with *asymptotic* object conditional validity, learning the relevant properties of the conditional density of y_0 given x_0 is typically only possible if the dimension of the feature vector x_0 is much smaller than the available sample size (cf. Remark 5.2). Therefore, since our focus in the present paper is on high-dimensional problems, we do not aim at object conditional validity.

Remark 5.4 (On heteroskedasticity). The length of the leave-one-out prediction interval in (2.5), as it stands, does not depend on the value of x_0. An immediate way to account for heteroskedasticity is the following. Consider, in addition, an estimator $\hat{\sigma}_n^2(x) = S(T_n, x)$ of the conditional variance $\text{Var}[y_0 \| x_0 = x]$. Then a prediction interval can be computed as $\hat{m}_n(x_0) + (\hat{q}_{\alpha_1}, \hat{q}_{\alpha_2}]\hat{\sigma}_n(x_0)$, where now, $\hat{q}_\alpha$ is an empirical α-quantile of the leave-one-out residuals

$$\hat{u}_i = \frac{y_i - \hat{m}_n^{[i]}(x_i)}{\hat{\sigma}_{n,[i]}(x_i)}, \quad i = 1, \ldots, n.$$

Remark 5.5 (Computational simplifications). Computing the leave-one-out prediction interval may be computationally costly, because the model has to be re-fitted n-times on each of the possible reduced samples $T_n^{[i]}$, $i = 1, \dots, n$, in order to compute the leave-one-out residuals $\hat{u}_i = y_i - \hat{m}_n^{[i]}(x_i)$. Sometimes, it is possible to devise a shortcut for the computation of these residuals. For example, in case of ordinary least squares prediction $\hat{m}_n(x) = x'\hat{\beta}_n = x'(X'X)^\dagger X'Y$, if $X'_{[i]}X_{[i]}$ has full rank, we have the well known identity

$$\hat{u}_i = y_i - x_i'\hat{\beta}_n^{[i]} = \frac{y_i - x_i'\hat{\beta}_n}{1 - x_i'(X'X)^{-1}x_i},$$

such that the n-vector of leave-one-out residuals can be computed as

$$\left[\operatorname{diag}(I_n - X(X'X)^{-1}X')\right]^{-1}(I_n - X(X'X)^{-1}X')Y.$$

Hence, the model has to be fitted only once. If such a simplification is not possible, and the computation of all the residuals $\hat{u}_i$, $i = 1, \dots, n$, is too costly, then one will typically restrict to using only a smaller number of those residuals, e.g., $\hat{u}_i$, $i = 1, \dots, l$, with $l \ll n$.

Acknowledgements. The authors thank the participants of the “ISOR Research Seminar in Statistics and Econometrics” at the University of Vienna for discussion of an early version of the paper. In particular, we want to thank Benedikt M. Pötscher and David Preinerstorfer for valuable comments.

References.

Aven, T. (1985). Upper (lower) bounds on the mean of the maximum (minimum) of a number of random variables. *J. Appl. Probab. 22*(3), 723–728.

Bai, Z. and J. W. Silverstein (2010). *Spectral Analysis of Large Dimensional Random Matrices* (2nd ed.). Springer Series in Statistics. New York: Springer.

Bai, Z. D. and Y. Q. Yin (1993). Limit of the smallest eigenvalue of a large dimensional sample covariance matrix. *Ann. Probab. 21*(3), 1275–1294.

Barber, R. F., E. J. Candes, A. Ramdas, and R. J. Tibshirani (2019a). The limits of distribution-free conditional predictive inference. *arXiv preprint arXiv:1903.04684*.

Barber, R. F., E. J. Candes, A. Ramdas, and R. J. Tibshirani (2019b). Predictive inference with the jackknife+. *Ann. Statist.* (forthcoming).

Bean, D., P. J. Bickel, N. El Karoui, and B. Yu (2013). Optimal m-estimation in high-dimensional regression. *Proc. Natl. Acad. Sci. USA 110*(36), 14563–14568.

Bickel, P. J. and D. A. Freedman (1983). Bootstrapping regression models with many parameters. In P. Bickel, K. Doksum, and J. Hodges (Eds.), *A Festschrift for Erich L. Lehmann*, pp. 28–48. Wadsworth Inc.

Billingsley, P. (1995). *Probability and Measure* (3rd ed.). New York: Wiley.

Bousquet, O. and A. Elisseeff (2002). Stability and generalization. *J. Mach. Learn. Res. 2*, 499–526.

Bucchianico, A. D., J. H. J. Einmahl, and N. A. Mushkudiani (2001). Smallest nonparametric tolerance regions. *Ann. Statist. 29*(5), 1320–1343.

Bühlmann, P. and S. van de Geer (2011). *Statistics for High-dimensional Data.* Berlin: Springer.

Butler, R. and E. D. Rothman (1980). Predictive intervals based on reuse of the sample. *J. Amer. Statist. Assoc. 75*(372), 881–889.

Chatterjee, S. (2013). Assumptionless consistency of the lasso. *arXiv preprint arXiv:1303.5817*.

Chatterjee, S. K. and N. K. Patra (1980). Asymptotically minimal multivariate tolerance sets. *Calcutta Statist. Assoc. Bull. 29*(1-2), 73–94.

Chen, W., K.-J. Chun, and R. F. Barber (2017). Discretized conformal prediction for efficient distribution-free inference. *arXiv preprint arXiv:1709.06233*.

Cox, D. R. and D. V. Hinkley (1974). *Theoretical Statistics.* London: Chapman and Hall.

Devroye, L. and T. J. Wagner (1979). Distribution-free inequalities for the deleted and holdout error estimates. *IEEE Trans. Inform. Theory 25*(2), 202–207.

Dicker, L. (2012). Optimal estimation and prediction for dense signals in high-dimensional linear models. *arXiv:1203.4572*.

Dicker, L. H. (2013). Optimal equivariant prediction for high-dimensional linear models with arbitrary predictor covariance. *Electron. J. Stat. 7*, 1806–1834.

El Karoui, N. (2010). The spectrum of kernel random matrices. *Ann. Statist. 38*(1), 1–50.

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

El Karoui, N. (2018). On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators. *Probab. Theory Relat. Fields 170*(1-2), 95–175.

El Karoui, N., D. Bean, P. J. Bickel, C. Lim, and B. Yu (2013). On robust regression with high-dimensional predictors. *Proc. Natl. Acad. Sci. USA 110*(36), 14557–14562.

El Karoui, N. and E. Purdom (2015). Can we trust the bootstrap in high-dimension?

Fahrmeir, L. (1990). Maximum likelihood estimation in misspecified generalized linear models. *Statistics 21*(4), 487–502.

Györfi, L., M. Kohler, A. Krzyżak, and H. Walk (2002). *A Distribution-Free Theory of Nonparametric Regression.* Springer Series in Statistics. New York: Springer.

Huber, N. and H. Leeb (2013). Shrinkage estimators for prediction out-of-sample: Conditional performance. *Commun. Statist. - Theory Methods 42*(7), 1246–1264.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability 1*, 221–233.

Krishnamoorthy, K. and T. Mathew (2009). *Statistical Tolerance Regions: Theory, Applications, and Computation.* Wiley Series in Probability and Statistics. Hoboken, New Jersey: Wiley.

Lei, J. (2017). Fast exact conformalization of lasso using piecewise linear homotopy. *arXiv preprint arXiv:1708.00427*.

Lei, J., M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman (2017). Distribution-free predictive inference for regression. *J. Amer. Statist. Assoc.*.

Lei, J., J. Robins, and L. Wasserman (2013). Distribution-free prediction sets. *J. Amer. Statist. Assoc. 108*(501), 278–287.

Lei, J. and L. Wasserman (2014). Distribution-free prediction bands for non-parametric regression. *J. Roy. Statist. Soc. Ser. B 76*, 71–96.

Lewis, T. and J. W. Thompson (1981). Dispersive distributions, and the connection between dispersivity and strong unimodality. *J. Appl. Probab. 18*(1), 76–90.

Li, J. and R. Y. Liu (2008). Multivariate spacings based on data depth: I. construction of nonparametric multivariate tolerance regions. *Ann. Statist. 36*(3), 1299–1323.

Lopes, M. E. (2015). *Some Inference Problems in High-Dimensional Linear Models.* Ph. D. thesis, UC Berkeley.

Mammen, E. (1996). Empirical process of residuals for high-dimensional linear models. *Ann. Statist. 24*(1), 307–335.

Massart, P. (1990). The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Ann. Probab. 18*(3), 1269–1283.

Olive, D. J. (2007). Prediction intervals for regression models. *Comput. Statist. Data Anal. 51*(6), 3115–3122.

Politis, D. N. (2013). Model-free model-fitting and predictive distributions. *Test 22*(2), 183–221.

Schmoyer, R. L. (1992). Asymptotically valid prediction intervals for linear models. *Technometrics 34*(4), 399–408.

Stine, R. A. (1985). Bootstrap prediction intervals for regression. *J. Amer. Statist. Assoc. 80*(392), 1026–1031.

Tukey, J. W. (1947). Non-parametric estimation ii. statistically equivalent blocks and tolerance regions – the continuous case. *Ann. Math. Statist. 18*(4), 529–539.

van der Vaart, A. W. (2007). *Asymptotic Statistics* (8th ed.). Cambridge Series in Statistical and Probabilistic Mathematics. New York: Cambridge University Press.

Vovk, V. (2012). Conditional validity of inductive conformal predictors. In *Asian conference on machine learning*, pp. 475–490.

Vovk, V. (2013). Conditional validity of inductive conformal predictors. *Machine Learning 92*(2), 349–376.

Vovk, V., A. Gammerman, and C. Saunders (1999). Machine-learning applications of algorithmic randomness. In I. Bratko and S. Dzeroski (Eds.), *Proceedings of the Sixteenth International Conference on Machine Learning*, ICML '99, pp. 444–453. Morgan Kaufmann Publishers Inc.

Vovk, V., A. Gammerman, and G. Shafer (2005). *Algorithmic Learning in a Random World.* New York: Springer.

Vovk, V., I. Nouretdinov, and A. Gammerman (2009). On-line predictive linear regression. *Ann. Statist. 37*(3), 1566–1590.

Wald, A. (1943). An extension of Wilks' method for setting tolerance limits. *Ann. Math. Statist. 14*(1), 45–55.

White, H. (1980a). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica 48*(4), 817–838.

White, H. (1980b). Using least squares to approximate unknown regression functions. *Internat. Econom. Rev. 21*(1), 149–170.

Wilks, S. S. (1941). Determination of sample sizes for setting tolerance limits. *Ann. Math. Statist. 12*(1), 91–96.

Wilks, S. S. (1942). Statistical prediction with special reference to the problem of tolerance limits. *Ann. Math. Statist. 13*(4), 400–409.

APPENDIX A: PROOFS OF MAIN RESULTS

We repeatedly use the following argument: Suppose that A_n, $n \geq 1$, are real-valued random variables on $(\Omega, \mathcal{F}, P)$ and $\mathcal{F}_n$, $n \geq 1$, are sub-sigma fields of $\mathcal{F}$. If $\mathbb{E}[|A_n|\|\mathcal{F}_n]$ is well-defined and converges to zero in probability, then A_n converges to zero in probability. Indeed

$$P(|A_n| > \varepsilon) = \mathbb{E}[P(|A_n| > \varepsilon\|\mathcal{F}_n) \wedge 1] \leq \mathbb{E}\left[\left(\frac{1}{\varepsilon}\mathbb{E}[|A_n|\|\mathcal{F}_n]\right) \wedge 1\right].$$

In this upper bound, the integrand converges to zero in probability by assumption, and this integrand is bounded by 1 by construction. In view of the dominated convergence theorem, this upper bound converges to zero. Moreover, if $\mathbb{E}[A_n\|\mathcal{F}_n]$ is well-defined and converges to $c \in \mathbb{R}$ in probability and if, in addition, $\mathrm{Var}[A_n\|\mathcal{F}_n] \to 0$ in probability, then $A_n \to c$ in probability. Indeed,

$$\mathbb{E}[(A_n - c)^2\|\mathcal{F}_n] = \mathrm{Var}[A_n\|\mathcal{F}_n] + (\mathbb{E}[A_n\|\mathcal{F}_n] - c)^2$$

converges to zero in probability. From the preceding statement the claim follows.

A.1. Proof of Theorem 2.5. The proof relies on the following result, which is a special case of Lemma 9 (Equation (9)) in Bousquet and Elisseeff (2002) (see also Devroye and Wagner, 1979) applied with the loss function $\ell(f, z) = \mathbb{1}_{(-\infty,s]}(y - f(x))$, $f : \mathcal{X} \to \mathcal{Y}$, $z = (y, x) \in \mathcal{Z}$, $s \in \mathbb{R}$ and $M = 1$, in their notation.

Lemma A.1 (Bousquet and Elisseeff (2002)). *If the estimator $\hat{m}_n$ is symmetric, then*

$$\mathbb{E}_{P^n}\left[\left(\hat{F}_n(s) - \tilde{F}_n(s)\right)^2\right] \\ \leq \frac{1}{2n} + 3\mathbb{E}_{P^{n+1}}\left[\left|\mathbb{1}_{(-\infty,s]}(y_0 - \hat{m}_n(x_0)) - \mathbb{1}_{(-\infty,s]}(y_0 - \hat{m}_n^{[1]}(x_0))\right|\right],$$

for every $s \in \mathbb{R}$ and every probability distribution P on $\mathcal{Z}$.

Under Condition (C1), it is elementary to relate the upper bound of Lemma A.1 to the η-stability of $\hat{m}_n$. We defer the proof until the end of this section.

Lemma A.2. *Let $\mathcal{P}$ be a collection of probability measures on $\mathcal{Z} = \mathcal{Y} \times \mathcal{X}$ that satisfies Condition (C1). Then, for every $s \in \mathbb{R}$ and every $P \in \mathcal{P}$,*

$$\mathbb{E}_{P^{n+1}}\left[\left|\mathbb{1}_{(-\infty,s]}(y_0 - \hat{m}_n(x_0)) - \mathbb{1}_{(-\infty,s]}(y_0 - \hat{m}_n^{[1]}(x_0))\right|\right] \\ \leq \mathbb{E}_{P^{n+1}}\left[\left(\|f_{u,P}\|_\infty|\hat{m}_n(x_0) - \hat{m}_n^{[1]}(x_0)|\right) \wedge 1\right].$$

To turn the pointwise bound of Lemma A.1 into a uniform one, we need a certain continuity and tightness property of $\tilde{F}_n$.

Lemma A.3. *Let $\mathcal{P}$ be a collection of probability measures on $\mathcal{Z} = \mathcal{Y} \times \mathcal{X}$ that satisfies Condition (C1) and fix a training sample $t_n \in \mathcal{Z}^n$.*

(i) If $P \in \mathcal{P}$ and $s_1, s_2 \in \mathbb{R}$, then

$$\left|\tilde{F}_n(s_1; t_n) - \tilde{F}_n(s_2; t_n)\right| \leq \|f_{u,P}\|_\infty |s_1 - s_2|.$$

(ii) Let $P \in \mathcal{P}$, $\delta \in [0,1]$, $\mu \in \mathbb{R}$ and $c \in (0,\infty)$, and define $\overline{s} = \mu + c + q_{u,P}(\delta)$ and $\underline{s} = \mu - c + q_{u,P}(\delta)$, where $q_{u,P}(\delta) \in \bar{\mathbb{R}}$ is an arbitrary δ-quantile of $f_{u,P}$. Then,

$$\begin{aligned} 1 - \tilde{F}_n(\overline{s}; t_n) &\leq (1-\delta) + P\left(m_P(x_0) - M_{n,p}(t_n, x_0) - \mu > c\right), \\ \tilde{F}_n(\underline{s}; t_n) &\leq \delta + P\left(m_P(x_0) - M_{n,p}(t_n, x_0) - \mu < -c\right). \end{aligned}$$

We provide the proofs of Lemma A.2 and Lemma A.3 below, after the main argument is finished. The proof of Theorem 2.5 is now a finite sample version of the proof of Polya's theorem. Fix $P \in \mathcal{P}$, $t_n \in \mathcal{Z}^n$, $\mu \in \mathbb{R}$, $\varepsilon > 0$ and $c = L \in [1, \infty)$. Set $\delta_1 = F_{u,P}(-L)$ and $\delta_2 = F_{u,P}(L)$, where $F_{u,P}(s) := \int_{-\infty}^{s} f_{u,P}(v)\, dv$. Next, choose the (possibly non-unique) quantiles $q_{u,P}(\delta_1) = -L$ and $q_{u,P}(\delta_2) = L$ and consider $\overline{s}$ and $\underline{s}$ as in Lemma A.3(ii) with $\delta = \delta_2$ and $\delta = \delta_1$, respectively, i.e., $\overline{s} = \mu + 2L$ and $\underline{s} = \mu - 2L$. We split up the interval $[\underline{s}, \overline{s})$ into K intervals $[s_{j-1}, s_j)$, $j = 1, \dots, K$, with endpoints $\underline{s} =: s_0 < s_1 < \cdots < s_K := \overline{s}$, such that $s_j - s_{j-1} \leq \varepsilon$. We may thus take $K = \lceil (\overline{s} - \underline{s})/\varepsilon \rceil = \lceil 4L/\varepsilon \rceil$. If $s < s_0$, then

$$\begin{aligned} \hat{F}_n(s) - \tilde{F}_n(s) &\geq 0 - \tilde{F}_n(s_0) \geq -|\hat{F}_n(s_0) - \tilde{F}_n(s_0)| - \tilde{F}_n(s_0), \\ \hat{F}_n(s) - \tilde{F}_n(s) &\leq \hat{F}_n(s_0) \leq |\hat{F}_n(s_0) - \tilde{F}_n(s_0)| + \tilde{F}_n(s_0). \end{aligned}$$

Furthermore, if $s \geq s_K$, then

$$\begin{aligned} \hat{F}_n(s) - \tilde{F}_n(s) &\geq \hat{F}_n(s_K) - 1 \\ &\geq -|\hat{F}_n(s_K) - \tilde{F}_n(s_K)| - \left(1 - \tilde{F}_n(s_K)\right), \\ \hat{F}_n(s) - \tilde{F}_n(s) &\leq 1 - \tilde{F}_n(s_K) \\ &\leq |\hat{F}_n(s_K) - \tilde{F}_n(s_K)| + 1 - \tilde{F}_n(s_K). \end{aligned}$$

Finally, for $j \in \{1, \dots, K\}$ and $s \in [s_{j-1}, s_j)$,

$$\begin{aligned} \hat{F}_n(s) - \tilde{F}_n(s) &\geq -|\hat{F}_n(s_{j-1}) - \tilde{F}_n(s_{j-1})| - \left(\tilde{F}_n(s_j) - \tilde{F}_n(s_{j-1})\right), \\ \hat{F}_n(s) - \tilde{F}_n(s) &\leq |\hat{F}_n(s_j) - \tilde{F}_n(s_j)| + \left(\tilde{F}_n(s_j) - \tilde{F}_n(s_{j-1})\right). \end{aligned}$$

Thus, discretizing the supremum over $\mathbb{R}$, we get

$$
\begin{aligned}
&\sup_{s\in\mathbb{R}} |\hat{F}_n(s) - \tilde{F}_n(s)| \\
&= \sup_{s<s_0} |\hat{F}_n(s) - \tilde{F}_n(s)| \vee \sup_{s\geq t_K} |\hat{F}_n(s) - \tilde{F}_n(s)| \\
&\qquad \vee \max_{j=1,\dots,K} \sup_{s\in[s_{j-1},s_j)} |\hat{F}_n(s) - \tilde{F}_n(s)| \\
&\leq \quad \Big(|\hat{F}_n(s_0) - \tilde{F}_n(s_0)| + \tilde{F}_n(s_0)\Big) \vee \Big(|\hat{F}_n(s_K) - \tilde{F}_n(s_K)| + 1 - \tilde{F}_n(s_K)\Big) \\
&\qquad \vee \max_{j=1,\dots,K} \Big(\Big[|\hat{F}_n(s_{j-1}) - \tilde{F}_n(s_{j-1})| \vee |\hat{F}_n(s_j) - \tilde{F}_n(s_j)|\Big] + \tilde{F}_n(s_j) - \tilde{F}_n(s_{j-1})\Big).
\end{aligned}
$$

Next using the abbreviation $e_{n,P}(x_0) = m_P(x_0) - M_{n,p}(t_n, x_0)$ and both parts of Lemma A.3, we can further bound this as

$$
\begin{aligned}
&\max_{j=0,\dots,K} \Big(|\hat{F}_n(s_j) - \tilde{F}_n(s_j)|\Big) \\
&\qquad + \varepsilon\|f_{u,P}\|_\infty + \delta_1 + (1-\delta_2) + P(|e_{n,P} - \mu| > c) \\
&= \max_{j=0,\dots,K} \Big(|\hat{F}_n(s_j) - \tilde{F}_n(s_j)|\Big) \\
&\qquad + \varepsilon\|f_{u,P}\|_\infty + P(|u_0| > L) + P(|e_{n,P} - \mu| > L).
\end{aligned}
$$

Now, using Lemma 2.1 of Aven (1985), the expectation (w.r.t. the training data t_n) of the maximum can be bounded by

$$
\left(\sum_{j=0}^{K} \mathbb{E}_{P^n}\left[|\hat{F}_n(s_j) - \tilde{F}_n(s_j)|^2\right]\right)^{\frac{1}{2}}.
$$

Finally, applying Lemmas A.1 and A.2, the expression on the previous display is bounded by

$$
\left((K+1)\left[\frac{1}{2n} + 3\eta\right]\right)^{\frac{1}{2}}.
$$

We have established the bound

$$
\begin{aligned}
\mathbb{E}_{P^n}\left[\|\hat{F}_n - \tilde{F}_n\|_\infty\right] &\leq \varepsilon\|f_{u,P}\|_\infty + P(|u_0| > L) + P^{n+1}(|e_{n,P} - \mu| > L) \\
&\qquad + \left(\left(\frac{4L}{\varepsilon} + 2\right)\left[\frac{1}{2n} + 3\eta\right]\right)^{\frac{1}{2}}.
\end{aligned}
$$

Write $\nu_n := \frac{1}{2n} + 3\eta$. To simplify the minimization over ε, we use $\sqrt{4L/\varepsilon + 2} \leq \sqrt{4L/\varepsilon} + \sqrt{2}$ and minimize

$$
\varepsilon \mapsto \varepsilon\|f_{u,P}\|_\infty + \left(\frac{4L\nu_n}{\varepsilon}\right)^{\frac{1}{2}}.
$$

It is easy to see that this is minimized at $\varepsilon^* = \left(\frac{\sqrt{L\nu_n}}{\|f_{u,P}\|_\infty}\right)^{2/3}$. Plugging this back into the upper bound yields

$$\begin{aligned}\mathbb{E}_{P^n}\left[\|\hat{F}_n - \tilde{F}_n\|_\infty\right] \le P(|u_0| > L) + P(|e_{n,P} - \mu| > L)\\ + 3(L\|f_{u,P}\|_\infty \nu_n)^{1/3} + (2\nu_n)^{1/2}.\end{aligned}$$

□

Proof of Lemma A.2. The integrand on the left of the desired inequality is equal to

$$\mathbb{1}_{\left\{y_0 - \hat{m}_n(x_0) \le s < y_0 - \hat{m}_n^{[1]}(x_0)\right\}} + \mathbb{1}_{\left\{y_0 - \hat{m}_n^{[1]}(x_0) \le s < y_0 - \hat{m}_n(x_0)\right\}}.$$

Note that the two sets above are disjoint. Thus, using the abbreviations $e_{n,P} = m_P(x_0) - \hat{m}_n(x_0)$ and $e_{n,P}^{[1]} = m_P(x_0) - \hat{m}_n^{[1]}(x_0)$ together with the independence of x_0 and u_0, the conditional expectation of the sum in the previous display, given the training data and x_0, can be bounded as

$$\begin{aligned}&P^{n+1}(e_{n,P} \le s - u_0 < e_{n,P}^{[1]} \| T_n, x_0) + P^{n+1}(e_{n,P}^{[1]} \le s - u_0 < e_{n,P} \| T_n, x_0)\\ &= \int\limits_{s-(e_{n,P}^{[1]}(x_0)) \vee (e_{n,P}(x_0))}^{s-(e_{n,P}^{[1]}(x_0)) \wedge (e_{n,P}(x_0))} f_{u,P}(v)\, dv \wedge 1 \le \left(\|f_{u,P}\|_\infty |\hat{m}_n(x_0) - \hat{m}_n^{[1]}(x_0)|\right) \wedge 1.\end{aligned}$$

□

Proof of Lemma A.3. For $P \in \mathcal{P}$, $t_n \in \mathcal{Z}^n$ and $s_1 > s_2$, abbreviate $e_n(P) = m_P(x_0) - \hat{m}_n(x_0)$ and note

$$\begin{aligned}\tilde{F}_n(s_1) - \tilde{F}_n(s_2) &= P\left(s_2 < y_0 - \hat{m}_n(x_0) \le s_1\right)\\ &= P\left(s_2 - e_n(P) < u_0 \le s_1 - e_n(P)\right)\\ &= \mathbb{E}_P\left[\int\limits_{s_2 - e_n(P)}^{s_1 - e_n(P)} f_{u,P}(v)\, dv\right] \quad \le \quad \|f_{u,P}\|_\infty (s_1 - s_2),\end{aligned}$$

in view of independence between x_0 and u_0 imposed by Condition (C1). So the first claim follows upon reversing the roles of s_1 and s_2. For the second claim, take $\overline{s}$ and $\underline{s}$ as in the lemma to obtain

$$\begin{aligned}\tilde{F}_n(\overline{s}) &= P\left(u_0 \le \overline{s} - e_n(P)\right)\\ &\ge P\left(u_0 \le \mu + c + q_{u,P}(\delta) - e_n(P), e_n(P) - \mu \le c\right)\\ &\ge P\left(u_0 \le q_{u,P}(\delta), e_n(P) - \mu \le c\right)\\ &= P\left(u_0 \le q_{u,P}(\delta)\right) \cdot P\left(e_n(P) - \mu \le c\right)\\ &= \delta \cdot (1 - P\left(m_P(x_0) - \hat{m}_n(x_0) - \mu > c\right)).\end{aligned}$$

This implies the first bound. The second one is obtained analogously by

$$\begin{aligned}\tilde{F}_n(\underline{s}) &= P\left(u_0 \le \underline{s} - e_n(P)\right)\\ &\le P\left(u_0 \le \underline{s} - e_n(P), e_n(P) - \mu \ge -c\right) + P(e_n(P) - \mu < -c)\\ &\le P\left(u_0 \le q_{u,P}(\delta), e_n(P) - \mu \ge -c\right) + P(e_n(P) - \mu < -c)\\ &\le \delta + P(e_n(P) - \mu < -c).\end{aligned}$$

This finishes the proof. □

A.2. Proof of Theorem 3.1. We begin by showing that under the assumptions of Theorem 3.1, both of its conclusions hold with $c_n = 0$ (OLS) and irrespective of $\kappa \in [0, 1)$. The proof of the following result is deferred to the end of the subsection.

Lemma A.4. *For every $n \in \mathbb{N}$, let $\mathcal{P}_n = \mathcal{P}_n(\mathcal{L}_l, \mathcal{L}_v, C_0)$ be as in Condition (C2). For $P \in \mathcal{P}_n$, define β_P to be the minimizer of $\beta \mapsto \mathbb{E}_P[(y_0 - \beta' x_0)^2]$ over $\mathbb{R}^{p_n}$. If $p_n/n \to \kappa \in [0, 1)$ and*

$$\limsup_{n\to\infty} \sup_{P\in\mathcal{P}_n} \mathbb{E}_P\left[\left(\frac{m_P(x_0) - x_0'\beta_P}{\sigma_P}\right)^2\right] < \infty, \tag{A.1}$$

then the ordinary least squares estimator $\hat{\beta}_n = (X'X)^\dagger X'Y$ satisfies

$$\limsup_{n\to\infty} \sup_{P\in\mathcal{P}_n} P^n\left(\left\|\Sigma_P^{1/2}(\hat{\beta}_n - \beta_P)/\sigma_P\right\|_2^2 > M\right) \xrightarrow[M\to\infty]{} 0,$$

and for every $\varepsilon > 0$,

$$\sup_{P\in\mathcal{P}_n} P^n\left(\left\|\Sigma_P^{1/2}(\hat{\beta}_n - \hat{\beta}_n^{[1]})/\sigma_P\right\|_2^2 > \varepsilon\right) \xrightarrow[n\to\infty]{} 0.$$

We proceed with the proof of Theorem 3.1. In order to achieve uniformity over $\mathcal{P}_n$, we consider an arbitrary sequence $P_n \in \mathcal{P}_n$ and abbreviate $m_n =$

m_{P_n}, $\beta_n = \beta_{P_n}$, $\Sigma_n = \Sigma_{P_n}$ and $\sigma_n = \sigma_{P_n}$ and we write $\mathbb{E}_n = \mathbb{E}_{P_n^n}$, $\mathrm{Var}_n = \mathrm{Var}_{P_n^n}$, etc. We have to show that $\limsup_{n\to\infty} P_n^n(\|\Sigma_n^{1/2}(\hat\beta_n(c_n)-\beta_n)/\sigma_n\|_2^2 > M) \to 0$ as $M\to\infty$, and, provided that $\kappa>0$, that $P_n^n(\|\Sigma_n^{1/2}(\hat\beta_n(c_n)-\hat\beta_n^{[1]}(c_n))/\sigma_n\|_2^2 > \varepsilon) \to 0$, as $n\to\infty$, for every $\varepsilon > 0$.

Define $\delta_n^2 = \beta_n'\Sigma_n\beta_n/\sigma_n^2$, $t_n^2 = \hat\beta_n' X'X\hat\beta_n/(n\sigma_n^2)$ and

$$s_n = \begin{cases} \left(1 - \frac{p_n}{n}\frac{c_n}{t_n^2}\frac{\hat\sigma_n^2}{\sigma_n^2}\right)_+, & \text{if } t_n^2 > 0,\\ 1, & \text{if } t_n^2 = 0,\end{cases}$$

such that $0\le s_n \le 1$, and $\hat\beta_n(c_n) = s_n\hat\beta_n$, because $t_n^2 = 0$ if, and only if, $\hat\beta_n = 0$. We abbreviate $D := \limsup_{n\to\infty}\sup_{P\in\mathcal{P}_n}\mathbb{E}_P[(m_P(x_0)-\beta_P'x_0)^2]/\sigma_P^2$. The following properties are useful and will be verified after the main argument is finished.

Lemma A.5. *Under the assumptions of Theorem 3.1 we have: $\hat\sigma_n^2/\sigma_n^2$ and $\sigma_n^2/\hat\sigma_n^2$ are P_n-uniformly bounded in probability, $P_n^n(\hat\sigma_n^2 = 0) = 0$ (provided that $n \ge p_n + 2$) and $P_n^n(t_n^2 = 0)\to 0$. Furthermore, we have $P_n^n(t_n^2 \ge \kappa/2) \to 1$ if $\delta_n \to \delta\in[0,\infty)$. All the statements of the lemma continue to hold also for the leave-one-out analogs $t_{n,[1]}^2 := \hat\beta_n^{[1]\prime} X_{[1]}' X_{[1]}\hat\beta_n^{[1]}/(n\sigma_n^2)$ and $\hat\sigma_{n,[1]}^2 = \|Y_{[1]} - X_{[1]}\hat\beta_n^{[1]}\|_2^2/(n-1-p_n)$ of t_n^2 and $\hat\sigma_n^2$.*

The quantity of interest in the first claim of the theorem can be bounded as

$$\begin{aligned}\left\|\Sigma_n^{1/2}\left(\hat\beta_n(c_n)-\beta_n\right)/\sigma_n\right\|_2 &= \left\|\Sigma_n^{1/2}s_n\left(\hat\beta_n-\beta_n\right)/\sigma_n + \Sigma_n^{1/2}(s_n-1)\beta_n/\sigma_n\right\|_2\\ &\le \left\|\Sigma_n^{1/2}\left(\hat\beta_n-\beta_n\right)/\sigma_n\right\|_2 + (1-s_n)\delta_n. \qquad \text{(A.2)}\end{aligned}$$

Therefore, by Lemma A.4, it remains to show that $\limsup_{n\to\infty} Q_n(M)\to 0$ as $M\to\infty$, where $Q_n(M) = P_n^n((1-s_n)\delta_n > M)$. For fixed $M\in(1,\infty)$ and fixed $n\in\mathbb{N}$, we distinguish the cases $\delta_n < M^{1/2}$ and $\delta_n \ge M^{1/2}$. In the former case, $Q_n(M) = 0$. In the latter case, we proceed as follows. First, notice that

$$\begin{aligned} Q_n(M) = P_n^n\left((1-s_n)\delta_n > M, t_n^2 > 0\right) &\le P_n^n\left(\frac{p_n}{n}\frac{c_n}{t_n^2}\frac{\hat\sigma_n^2}{\sigma_n^2}\delta_n > M, t_n^2 > 0\right)\\ &= P_n^n\left(\frac{p_n}{n}\frac{c_n}{t_n^2/\delta_n^2}\frac{\hat\sigma_n^2}{\sigma_n^2} > M, t_n^2 > 0\right). \qquad \text{(A.3)}\end{aligned}$$

Furthermore, we trivially have $Y = X\beta_n + \sigma_n \tilde{v}$, where $\tilde{v} := (Y - X\beta_n)/\sigma_n$ has components $\tilde{v}_i = (m_n(x_i) - \beta_n' x_i)/\sigma_n + (y_i - m_n(x_i))/\sigma_n$, and, using the reverse triangle inequality and the notation $\|a\|_{P_X} = \sqrt{a' P_X a}$, we have

$$\begin{aligned} t_n = \sqrt{\frac{1}{n}\frac{Y'P_XY}{\sigma_n^2}} &= \|X\beta_n + \sigma_n\tilde{v}\|_{P_X}(n\sigma_n^2)^{-1/2} \\ &\geq \left|\|X\beta_n\|_{P_X} - \|\sigma_n\tilde{v}\|_{P_X}\right|(n\sigma_n^2)^{-1/2} \\ &= \left|\sqrt{\frac{\beta_n'\Sigma_n^{1/2}(\tilde{X}'\tilde{X}/n)\Sigma_n^{1/2}\beta_n}{\sigma_n^2}} - \sqrt{\frac{\tilde{v}'P_X\tilde{v}}{n}}\right|, \end{aligned}$$

where $\tilde{X} := (\tilde{x}_1, \ldots, \tilde{x}_n)' := X\Sigma_n^{-1/2}$ and $P_X := X(X'X)^\dagger X'$. Therefore, on the event

$$A_n(M) = \{\|\tilde{v}/\sqrt{n}\|_2^2 \leq M^{1/2}, \lambda_{\min}(\tilde{X}'\tilde{X}/n) > c_0^2(1-\sqrt{\kappa})^2/2 > M^{-1/2}\},$$

we have $\tilde{v}'P_X\tilde{v}(n\delta_n^2)^{-1} \leq M^{-1/2}$ (because $\delta_n \geq \sqrt{M}$) and

$$\beta_n'\Sigma_n^{1/2}(\tilde{X}'\tilde{X}/n)\Sigma_n^{1/2}\beta_n(\sigma_n^2\delta_n^2)^{-1} > c_0^2(1-\sqrt{\kappa})^2/2 > M^{-1/2},$$

so that on this event $t_n/\delta_n \geq c_0(1-\sqrt{\kappa})/\sqrt{2} - M^{-1/4} \geq 0$. Thus, turning back to (A.3) and using Markov's inequality, we obtain

$$\begin{aligned} &P_n^n\left(\frac{p_n}{n}\frac{c_n}{t_n^2/\delta_n^2}\frac{\hat{\sigma}_n^2}{\sigma_n^2} > M, t_n^2 > 0\right) \leq P_n^n\left(\frac{\hat{\sigma}_n^2}{\sigma_n^2} > Mt_n^2/\delta_n^2, t_n^2 > 0\right) \\ &\quad\leq P_n^n\left(\frac{\hat{\sigma}_n^2}{\sigma_n^2} > Mt_n^2/\delta_n^2, A_n(M)\right) + P_n^n\left(A_n(M)^c\right) \\ &\quad\leq P_n^n\left(\frac{\hat{\sigma}_n^2}{\sigma_n^2} > M\left(c_0(1-\sqrt{\kappa})/\sqrt{2} - M^{-1/4}\right)^2\right) + \frac{2D+1}{M^{1/2}} \\ &\qquad + P_n^n\left(\lambda_{\min}(\tilde{X}'\tilde{X}/n) \leq c_0^2(1-\sqrt{\kappa})^2/2\right) + P_n^n(c_0^2(1-\sqrt{\kappa})^2/2 \leq M^{-1/2}), \end{aligned}$$

for sufficiently large n. In view of Lemma B.1(i) in Appendix B and P_n-boundedness of $\hat{\sigma}_n^2/\sigma_n^2$ (Lemma A.5), the limit superior of the upper bound is equal to a function $Q(M) \geq 0$ that vanishes as $M \to \infty$. Therefore, we have shown that $\limsup_{n\to\infty} Q_n(M) \leq Q(M) \to 0$ as $M \to \infty$.

To establish the claim about the stability of $\hat{\beta}_n(c_n)$ we proceed in a similar way. First, note that

$$\begin{aligned} &\|\Sigma_n^{1/2}(\hat{\beta}_n(c_n) - \hat{\beta}_n^{[1]}(c_n))\|_2/\sigma_n = \|(s_n - s_n^{[1]})\Sigma_n^{1/2}\hat{\beta}_n + s_n^{[1]}\Sigma_n^{1/2}(\hat{\beta}_n - \hat{\beta}_n^{[1]})\|_2/\sigma_n \\ &\quad\leq |s_n - s_n^{[1]}|\|\Sigma_n^{1/2}\hat{\beta}_n/\sigma_n\|_2 + |s_n^{[1]}|\|\Sigma_n^{1/2}(\hat{\beta}_n - \hat{\beta}_n^{[1]})/\sigma_n\|_2 \\ &\quad\leq |s_n - s_n^{[1]}|\|\Sigma_n^{1/2}(\hat{\beta}_n - \beta_n)/\sigma_n\|_2 + |s_n - s_n^{[1]}|\delta_n + |s_n^{[1]}|\|\Sigma_n^{1/2}(\hat{\beta}_n - \hat{\beta}_n^{[1]})/\sigma_n\|_2, \end{aligned}$$

where $s_n^{[1]}$ is defined like s_n, but using $t_{n,[1]}$ and $\hat{\sigma}^2_{n,[1]}$. In view of Lemma A.4, it is easy to see that it remains to show that $|s_n - s_n^{[1]}|(1+\delta_n) = o_{P_n}(1)$. We argue along subsequences. Let n' be an arbitrary subsequence of n. Then by compactness of the extended real line, there exists a further subsequence n'' of n', such that $\delta_{n''} \to \delta \in [0,\infty]$. If we can show that for every $\varepsilon > 0$

$$P_{n''}^{n''}(|s_{n''} - s_{n''}^{[1]}|(1+\delta_{n''}) > \varepsilon) \xrightarrow[n''\to\infty]{} 0,$$

then the claim follows. For simplicity, we write n instead of n'' and we distinguish the cases $\delta = \infty$ and $\delta \in [0,\infty)$.

If $\delta = \infty$, then it suffices to show that $(s_n - s_n^{[1]})\delta_n$ converges to zero in P_n^n-probability. By Lemma A.5 we have $P_n^n(t_n^2 = 0) \to 0$ and the same holds for $t^2_{n,[1]}$, such that it suffices to show that

$$P_n^n(|s_n - s_n^{[1]}|\delta_n > \varepsilon, t_n > 0, t_{n,[1]} > 0) \to 0.$$

If $t_n > 0$, set $r_n = \frac{p_n}{n}\frac{c_n}{t_n^2}\frac{\hat{\sigma}_n^2}{\sigma_n^2}$, such that $s_n = (1-r_n)_+$ on this event, and define $r_n^{[1]} = \frac{p_n}{n}\frac{c_n}{t^2_{n,[1]}}\frac{\hat{\sigma}^2_{n,[1]}}{\sigma_n^2}$, provided that $t_{n,[1]} > 0$. Thus, if both t_n and $t_{n,[1]}$ are positive, we have

$$|s_n - s_n^{[1]}|\delta_n \le |r_n - r_n^{[1]}|\delta_n \le \left| \frac{\delta_n^2}{t_n^2}\frac{\hat{\sigma}_n^2}{\sigma_n^2} - \frac{\delta_n^2}{t^2_{n,[1]}}\frac{\hat{\sigma}^2_{n,[1]}}{\sigma_n^2} \right| \frac{1}{\delta_n}.$$

But in the first part of the proof we have already established that t_n^2/δ_n^2 is lower bounded by $c_0^2(1-\sqrt{\kappa})^2/4$ with asymptotic probability one, provided that $\delta_n^2 \to \infty$ (recall the case $\delta_n \ge M^{1/2}$ and the set $A_n(M)$, and let $M = \delta_n^2 \to \infty$), and an analogous argument applies to $t^2_{n,[1]}/\delta_n^2$. Thus, it follows from the P_n-boundedness of $\hat{\sigma}_n^2/\sigma_n^2$ and $\hat{\sigma}^2_{n,[1]}/\sigma_n^2$ that the upper bound in the previous display converges to zero in P_n^n-probability.

If $\delta \in [0,\infty)$, it suffices to show that $|s_n - s_n^{[1]}|$ converges to zero in P_n^n-probability. As before, we restrict to the event $\{t_n > 0, t_{n,[1]} > 0\}$. Note that due to the positive part mapping in the definition of s_n, the absolute difference $|s_n - s_n^{[1]}|$ vanishes if both r_n and $r_n^{[1]}$ are greater than or equal to 1, and is otherwise bounded by $|r_n - r_n^{[1]}| \le \max(|r_n/r_n^{[1]} - 1|, |r_n^{[1]}/r_n - 1|)$, provided that r_n and $r_n^{[1]}$ are positive. Thus, it remains to verify that $r_n^{[1]}/r_n$ converges to 1 in P_n^n-probability and that both $P_n^n(r_n = 0)$ and $P_n^n(r_n^{[1]} = 0)$ converge to zero. The latter statement follows from Lemma A.5, in fact it shows that $P_n^n(r_n = 0) = 0 = P_n^n(r_n^{[1]} = 0)$ provided that $n \geq p_n + 2$. Finally, to show that $r_n^{[1]}/r_n \to 1$ in P_n^n-probability, define $S_{[1]} := \tilde{X}_{[1]}' \tilde{X}_{[1]} = \sum_{i=2}^n \tilde{x}_i \tilde{x}_i'$ and note that by the Sherman-Morrison formula (see also the proof of Lemma A.4 below) we have

$$\begin{aligned}\hat{\beta}_n' X' X \hat{\beta}_n &= \hat{\beta}_n^{[1]\prime} X_{[1]}' X_{[1]} \hat{\beta}_n^{[1]} + (x_1' \hat{\beta}_n^{[1]})^2 + 2 x_1' \hat{\beta}_n^{[1]} (y_1 - x_1' \hat{\beta}_n^{[1]}) \\ &\quad + (y_1 - x_1' \hat{\beta}_n^{[1]})^2 \frac{\tilde{x}_1' S_{[1]}^{-1} \tilde{x}_1}{1 + \tilde{x}_1' S_{[1]}^{-1} \tilde{x}_1} \\ &\leq \hat{\beta}_n^{[1]\prime} X_{[1]}' X_{[1]} \hat{\beta}_n^{[1]} + y_1^2,\end{aligned}$$

at least on the event $B_n := \{\lambda_{\min}(S_{[1]}) > 0\}$, which has asymptotic P_n^n-probability one by Lemma B.1. Thus, on B_n, $t_n^2 / t_{n,[1]}^2 = 1 + g_n$, where

$$|g_n| \leq \frac{y_1^2}{\hat{\beta}_n^{[1]\prime} X_{[1]}' X_{[1]} \hat{\beta}_n^{[1]}}. \tag{A.4}$$

By Lemma A.5, and since $\kappa > 0$, $\hat{\beta}_n^{[1]\prime} X_{[1]}' X_{[1]} \hat{\beta}_n^{[1]} / (n\sigma_n^2) = t_{n,[1]}^2$ is bounded away from zero with asymptotic probability one. Thus, for the desired convergence of $t_n^2/t_{n,[1]}^2$ to 1, it remains to show that the numerator in (A.4) divided by $n\sigma_n^2$ converges to zero in P_n^n-probability. But this now follows from Condition (C2), assumption (3.5) and the fact that $\delta < \infty$. The proof is finished if we can also show that $\hat{\sigma}_n^2 / \hat{\sigma}_{n,[1]}^2$ converges to 1, in P_n^n-probability. To this end, we apply the Sherman-Morrison formula once more to get

$$I_n - P_X = I_n - P_{\tilde{X}} = \begin{pmatrix} \frac{1}{1+\tilde{x}_1' S_{[1]}^{-1} \tilde{x}_1}, & -\frac{\tilde{x}_1' S_{[1]}^{-1} \tilde{X}_{[1]}'}{1+\tilde{x}_1' S_{[1]}^{-1} \tilde{x}_1} \\ -\frac{\tilde{X}_{[1]} S_{[1]}^{-1} \tilde{x}_1}{1+\tilde{x}_1' S_{[1]}^{-1} \tilde{x}_1}, & I_{n-1} - P_{\tilde{X}_{[1]}} + \frac{\tilde{X}_{[1]} S_{[1]}^{-1} \tilde{x}_1 \tilde{x}_1' S_{[1]}^{-1} \tilde{X}_{[1]}'}{1+\tilde{x}_1' S_{[1]}^{-1} \tilde{x}_1} \end{pmatrix},$$

on the event B_n. Thus, on this event,

$$\begin{aligned}\hat{\sigma}_n^2 (n - p_n) = Y'(I_n - P_X) Y &= Y_{[1]}' (I_{n-1} - P_{X_{[1]}}) Y_{[1]} \\ &\quad + \frac{(y_1 - x_1' \hat{\beta}_n^{[1]})^2}{1 + \tilde{x}_1' S_{[1]}^{-1} \tilde{x}_1},\end{aligned}$$

such that $\frac{\hat{\sigma}_n^2}{\hat{\sigma}_{n,[1]}^2} \frac{n-p_n}{n-1-p_n} =: 1 + h_n$, where

$$|h_n| \leq \frac{(y_1 - x_1' \hat{\beta}_n^{[1]})^2}{(n - 1 - p_n)\sigma_n^2} \frac{\sigma_n^2}{\hat{\sigma}_{n,[1]}^2}.$$

But it is easy to see that the upper bound converges to zero in P_n^n-probability by Condition (C2), assumption (3.5), Lemmas A.4 and A.5, and because $n - p_n \to \infty$ and $\delta < \infty$. □

Proof of Lemma A.5. We use the notation $e_i := \frac{m(x_i) - x_i'\beta_n}{\sigma_n}$, $v_i := \frac{y_i - m(x_i)}{\sigma_n}$, $\check{v} := e + v$, where $e = (e_1, \dots, e_n)'$ and $v = (v_1, \dots, v_n)'$, such that $Y = X\beta_n + \sigma_n \check{v} = X\beta_n + \sigma_n e + \sigma_n v$. For the first claim simply observe that

$$\frac{\hat{\sigma}_n^2}{\sigma_n^2} = \frac{Y'(I_n - P_X)Y}{(n-p_n)\sigma_n^2} = \frac{n}{n-p_n}\frac{\check{v}'(I_n - P_X)\check{v}}{n} \le \frac{n}{n-p_n}\left\|\frac{\check{v}}{\sqrt{n}}\right\|_2^2,$$

and that $\mathbb{E}_n\|\check{v}\|_2^2 = n\mathbb{E}_n[\check{v}_1^2] \le n(2D+1)$, for sufficiently large n. For boundedness of the reciprocal we first note that $P_n^n(\hat{\sigma}_n^2 = 0) = \mathbb{E}_n[P_n^n(Y'(I_n - P_X)Y = 0\|X)] = P_n^n(I_n - P_X = 0) = 0$, because the conditional distribution of Y given X under P_n^n is absolutely continuous with respect to Lebesgue measure and $n \ge p_n + 2$. Similarly, $P_n^n(t_n = 0) = P_n^n(\hat{\beta}_n = 0) = \mathbb{E}_n[P_n^n((X'X)^\dagger X'Y = 0\|X)] = P_n^n(X(X'X)^\dagger = 0) = P_n^n(X = 0) = (\mathcal{L}_w(\{0\}))^{np_n} \to 0$. Next we show that $\hat{\sigma}_n^2/\sigma_n^2$ is bounded from below by $(1-\kappa)/2$ with asymptotic probability one. To this end, note that

$$\frac{\hat{\sigma}_n^2}{\sigma_n^2} = \frac{Y'(I_n - P_X)Y}{(n-p_n)\sigma_n^2} \ge 2\frac{e'(I_n - P_X)v}{n} + \left\|\frac{(I_n - P_X)v}{\sqrt{n}}\right\|_2^2,$$

where the conditional expectation of the mixed term given X is equal to zero and its conditional variance converges to zero in P_n^n-probability because of $\mathbb{E}_n[e_i^2] \le 2D$, for sufficiently large n. The conditional expectation of the last term in the previous display is $\operatorname{trace}(I_n - P_X)/n = \operatorname{trace}(I_n - P_{\tilde{X}})/n = 1 - p_n/n$, with asymptotic probability one in view of Lemma B.1(i). Using independence of the v_i and a little algebra, its conditional variance can be computed as

$$\begin{aligned}\operatorname{Var}_n\left[\frac{v'(I_n - P_X)v}{n}\middle\| X\right] &= \frac{2\operatorname{trace}((I_n - P_X)^2)}{n^2} + \frac{(\mathbb{E}_n[v_1^4] - 3)}{n^2}\sum_{i=1}^n (I_n - P_X)_{ii}^2 \\ &\le 2\frac{n}{n^2} + \frac{(C_0 + 3)n}{n^2} \to 0. \qquad \text{(A.5)}\end{aligned}$$

This establishes the boundedness of $\sigma_n^2/\hat{\sigma}_n^2$. For the remaining statement about t_n^2, suppose that $\delta_n^2 \to \delta \in [0,\infty)$ and note that

$$t_n^2 = \frac{Y'P_XY}{n\sigma_n^2} = \left\|\frac{\tilde{X}\Sigma_n^{1/2}\beta_n}{\sqrt{n\sigma_n^2}} + \frac{P_Xe}{\sqrt{n}} + \frac{P_Xv}{\sqrt{n}}\right\|_2^2.$$

Abbreviate $W_n := \frac{\tilde{X}\Sigma_n^{1/2}\beta_n}{\sqrt{n\sigma_n^2}} + \frac{P_X e}{\sqrt{n}}$ and observe $t_n^2 \geq 2W_n' P_X v/\sqrt{n} + \|P_X v/\sqrt{n}\|_2^2$. The conditional expectation of the mixed term $W_n' P_X v/\sqrt{n}$ given X is equal to zero, and its conditional variance is bounded by $\|W_n/\sqrt{n}\|_2^2$. But $\|W_n\|_2^2$ is bounded in P_n^n-probability, in view of the facts that $\delta < \infty$, $\mathbb{E}_n[\tilde{X}'\tilde{X}/n] = I_n$ and $\mathbb{E}_n[e_i^2] \leq 2D$, for sufficiently large n. Thus, the mixed term is $o_{P_n^n}(1)$. For $\|P_X v/\sqrt{n}\|_2^2$ one easily verifies that its conditional expectation given X is $\operatorname{trace}(P_X)/n = \operatorname{trace}(P_{\tilde{X}})/n$, which converges to $\kappa \in [0,1)$ in P_n^n-probability, because $P_n^n(\lambda_{\min}(\tilde{X}'\tilde{X}) = 0) \to 0$ by Lemma B.1. Furthermore, as above, its conditional variance can easily be computed as

$$\operatorname{Var}_n\left[\frac{v' P_X v}{n}\middle\| X\right] = \frac{2\operatorname{trace}(P_X^2)}{n^2} + \frac{(\mathbb{E}_n[v_1^4] - 3)}{n^2}\sum_{i=1}^n (P_X)_{ii}^2$$
$$\leq \frac{2p_n}{n^2} + \frac{(C_0+3)p_n}{n^2} \to 0.$$

Thus, $\|P_X v/\sqrt{n}\|_2^2$ converges to κ, in P_n^n-probability, which establishes the asymptotic lower bound on t_n^2. The results about the leave-one-out quantities can be established analogously. □

Proof of Lemma A.4. Fix $n \in \mathbb{N}$ and $P \in \mathcal{P}_n$. For simplicity, we write $m = m_P$, $\Sigma = \Sigma_P$, $\beta = \beta_P$ and $\sigma^2 = \sigma_P^2$ and abbreviate $\tilde{X} := X\Sigma^{-1/2}$. For $\xi > 0$, consider the event $A_n := A_n(\xi) := \{\lambda_{\min}(\tilde{X}'\tilde{X}/n) > \xi\}$. On this event, we observe that $\Sigma^{1/2}(\hat{\beta}_n - \beta)/\sigma = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\check{v}$, where $\check{v} = (\check{v}_1, \dots, \check{v}_n)'$, $\check{v}_i = (m(x_i) - \beta' x_i)/\sigma + v_i$ and $v_i = (y_i - m(x_i))/\sigma$, for $i = 1, \dots, n$. Thus, on A_n, $\|\Sigma^{1/2}(\hat{\beta}_n - \beta)/\sigma\|_2^2 = \check{v}'\tilde{X}(\tilde{X}'\tilde{X})^{-2}\tilde{X}'\check{v} = \check{v}'\tilde{X}(\tilde{X}'\tilde{X})^{-1/2}(\tilde{X}'\tilde{X})^{-1}(\tilde{X}'\tilde{X})^{-1/2}\tilde{X}'\check{v} \leq \|\check{v}/\sqrt{n}\|_2^2 \|(\tilde{X}'\tilde{X}/n)^{-1}\|_2$. Hence, using Condition (C2), we obtain

$$\begin{aligned}
&P^n(\|\Sigma^{1/2}(\hat{\beta}_n - \beta)/\sigma\|_2^2 > M)\\
&\quad\leq P^n(\|\check{v}/\sqrt{n}\|_2^2 \|(\tilde{X}'\tilde{X}/n)^{-1}\|_2 > M, A_n(\xi)) + P^n(A_n(\xi)^c)\\
&\quad\leq P^n(\|\check{v}/\sqrt{n}\|_2^2/\xi > M) + P^n(A_n(\xi)^c)\\
&\quad\leq \frac{\mathbb{E}_P[\check{v}_1^2]}{M\xi} + P^n(\lambda_{\min}(\tilde{X}'\tilde{X}/n) \leq \xi).
\end{aligned}$$

Since $\mathbb{E}_P[\check{v}_1^2] = \mathbb{E}_P[(m(x_0) - \beta' x_0)^2/\sigma^2] + 1$, in view of (C2), and because $P^n(\lambda_{\min}(\tilde{X}'\tilde{X}/n) \leq \xi)$ does not depend on the parameters β, Σ and σ^2, Lemma B.1(ii) implies the first claim if we set $\xi = c_0^2(1-\sqrt{\kappa})^2/2 > 0$.

For the stability property, we abbreviate $S_{[1]} = \tilde{X}_{[1]}'\tilde{X}_{[1]}$, $\tilde{\beta}_n := \Sigma^{1/2}(\hat{\beta}_n - \beta)/\sigma$ and $\tilde{\beta}_n^{[1]} = \Sigma^{1/2}(\hat{\beta}_n^{[1]} - \beta)/\sigma$, and consider the event $B_n = \{\lambda_{\min}(S_{[1]}) >$

$0\}$. On this event, also $\lambda_{\min}(\tilde{X}'\tilde{X}) = \lambda_{\min}(S_{[1]} + \tilde{x}_1\tilde{x}_1') > 0$, where $\tilde{X} = [\tilde{x}_1, \dots, \tilde{x}_n]'$ and $\tilde{X}_{[1]} = [\tilde{x}_2, \dots, \tilde{x}_n]'$, and the Sherman-Morrison formula yields

$$\begin{aligned}
\tilde{\beta}_n \;&=\; (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\check{v} \;=\; (S_{[1]} + \tilde{x}_1\tilde{x}_1')^{-1}(\tilde{X}_{[1]}'\check{v}_{[1]} + \tilde{x}_1\check{v}_1)\\
&= \left(S_{[1]}^{-1} - \frac{S_{[1]}^{-1}\tilde{x}_1\tilde{x}_1' S_{[1]}^{-1}}{1 + \tilde{x}_1' S_{[1]}^{-1}\tilde{x}_1}\right)(\tilde{X}_{[1]}'\check{v}_{[1]} + \tilde{x}_1\check{v}_1)\\
&= \tilde{\beta}_n^{[1]} - \frac{S_{[1]}^{-1}\tilde{x}_1\tilde{x}_1'\tilde{\beta}_n^{[1]}}{1 + \tilde{x}_1' S_{[1]}^{-1}\tilde{x}_1} + S_{[1]}^{-1}\tilde{x}_1\check{v}_1 - S_{[1]}^{-1}\tilde{x}_1\check{v}_1\frac{\tilde{x}_1' S_{[1]}^{-1}\tilde{x}_1}{1 + \tilde{x}_1' S_{[1]}^{-1}\tilde{x}_1}\\
&= \tilde{\beta}_n^{[1]} + \frac{S_{[1]}^{-1}\tilde{x}_1(\check{v}_1 - \tilde{x}_1'\tilde{\beta}_n^{[1]})}{1 + \tilde{x}_1' S_{[1]}^{-1}\tilde{x}_1},
\end{aligned}$$

and thus, $\|\Sigma^{1/2}(\hat{\beta}_n - \hat{\beta}_n^{[1]})/\sigma\|_2^2 = (1 + \tilde{x}_1' S_{[1]}^{-1}\tilde{x}_1)^{-2}\tilde{x}_1' S_{[1]}^{-2}\tilde{x}_1(\check{v}_1 - \tilde{x}_1'\tilde{\beta}_n^{[1]})^2 \leq 2(\check{v}_1^2 + (\tilde{x}_1'\tilde{\beta}_n^{[1]})^2)\tilde{x}_1' S_{[1]}^{-2}\tilde{x}_1$. Clearly, the squared error term $\check{v}_1^2$ is $\mathcal{P}_n$-uniformly bounded in probability because $\mathbb{E}_P[\check{v}_1^2] = \mathbb{E}_P[(m(x_0) - \beta' x_0)^2/\sigma^2] + 1$, as above; $\mathbb{E}[(\tilde{x}_1'\tilde{\beta}_n^{[1]})^2 \| \tilde{\beta}_n^{[1]}] = \|\tilde{\beta}_n^{[1]}\|_2^2$ is also $\mathcal{P}_n$-uniformly bounded in probability, by the same argument as in the first paragraph, which implies that $(\tilde{x}_1'\tilde{\beta}_n^{[1]})^2$ is $\mathcal{P}_n$-uniformly bounded in probability; and $\mathbb{E}[\tilde{x}_1' S_{[1]}^{\dagger 2}\tilde{x}_1 \| S_{[1]}] = \operatorname{trace} S_{[1]}^{\dagger 2} \to 0$, $\mathcal{P}_n$-uniformly in probability, by Lemma B.1. Therefore, we have $P^n(\|\Sigma^{1/2}(\hat{\beta}_n - \hat{\beta}_n^{[1]})/\sigma\|_2^2 > \varepsilon, B_n) \leq P^n(2O_{\mathcal{P}_n}(1)o_{\mathcal{P}_n}(1) > \varepsilon, B_n) \to 0$. Moreover, $P^n(B_n^c) = P^{n-1}(\lambda_{\min}(S_{[1]}) = 0) \to 0$, uniformly over $\mathcal{P}_n$, in view of Lemma B.1. □

A.3. Proof of Theorem 3.4. We begin by stating a few more results on the OLS estimator that hold in the linear model (C3). The proof is deferred to the end of the subsection.

Lemma A.6. *Under the assumptions of Theorem 3.4, the OLS estimator $\hat{\beta}_n = (X'X)^{\dagger}X'Y$, satisfies*

$$\sup_{P\in\mathcal{P}_n} P^n\left(\left|\left\|\Sigma_P^{1/2}(\hat{\beta}_n - \beta_P)/\sigma_P\right\|_2 - \tau\right| > \varepsilon\right) \xrightarrow[n\to\infty]{} 0, \quad \textit{and}$$

$$\sup_{P\in\mathcal{P}_n} P^n\left(\left\|\Sigma_P^{1/2}(\hat{\beta}_n - \beta_P)/\sigma_P\right\|_4 > \varepsilon\right) \xrightarrow[n\to\infty]{} 0,$$

for every $\varepsilon > 0$. Here, $\tau = \tau(\mathcal{L}_l, \kappa) \in [0, \infty)$ depends only on $\mathcal{L}_l$ and $\kappa \in [0, 1)$ and has the following properties: For any $\mathcal{L}_l$ as in (C3), $\tau(\mathcal{L}_l, \kappa) = 0$ if, and only if, $\kappa = 0$. If $\mathcal{L}_l(\{-1, 1\}) = 1$, then $\tau(\mathcal{L}_l, \kappa) = \sqrt{\kappa/(1-\kappa)}$.

The next result will be instrumental to establish convergence of the conditional law $P(y_0 - x_0'\hat{\beta}_n \leq t \| T_n)$ to the distribution of $lN\tau + v$, for l, N, τ, v as in the statement of the theorem. Its proof is also deferred until after the main argument is finished.

Lemma A.7. *Fix arbitrary positive constants $\tau \in [0, \infty)$, $\delta \in (0, 2]$ and $c \in (0, \infty)$ and let $(p_n)_{n\in\mathbb{N}}$ be a sequence of positive integers. On some probability space $(\Omega, \mathcal{A}, \mathbb{P})$, let v_0 and l_0 be real random variables and let $W_0 = (w_{0j})_{j=1}^{\infty}$ be a sequence of i.i.d. real random variables such that W_0, v_0 and l_0 are jointly independent, $|l_0| \geq c$, $\mathbb{E}[l_0^2] = 1$, $\mathbb{E}[w_{01}] = 0$, $\mathbb{E}[w_{01}^2] = 1$ and $\mathbb{E}[|w_{01}|^{2+\delta}] < \infty$. For $n \in \mathbb{N}$ and $b \in \mathbb{R}^{p_n}$, define $w_n = (w_{01}, \dots, w_{0p_n})'$, $G(t, b) = \mathbb{P}(l_0 w_n' b + v_0 \leq t)$ and $F(t) = \mathbb{P}(l_0 N\tau + v_0 \leq t)$, where $N \overset{\mathcal{L}}{=} \mathcal{N}(0, 1)$ is independent of (l_0, v_0). Consider positive sequences $g_1, g_2 : \mathbb{N} \to (0, 1)$, such that $g_j(n) \to 0$, as $n \to \infty$, $j = 1, 2$. Suppose that one of the following cases applies.*

(i) $\tau = 0$ and $t \mapsto \mathbb{P}(v_0 \leq t)$ is continuous. In this case, set

$$B_n = \{b \in \mathbb{R}^{p_n} : \|b\|_2 \leq g_1(n)\}.$$

(ii) $\tau > 0$ and $p_n \to \infty$ as $n \to \infty$. In this case, set

$$B_n = \{b \in \mathbb{R}^{p_n} : |\|b\|_2 - \tau| \leq g_1(n), b \neq 0, \|b\|_{2+\delta}/\|b\|_2 \leq g_2(n)\}.$$

(iii) $\tau > 0$ and $w_{01} \overset{\mathcal{L}}{=} \mathcal{N}(0, 1)$. In this case, set

$$B_n = \{b \in \mathbb{R}^{p_n} : |\|b\|_2 - \tau| \leq g_1(n)\}.$$

Then, using the convention that $\sup \varnothing = 0$,

$$\sup_{b\in B_n} \sup_{t\in\mathbb{R}} |G(t, b) - F(t)| \xrightarrow[n\to\infty]{} 0. \tag{A.6}$$

We now turn to the proof of Theorem 3.4. In order to achieve uniformity in $P_n \in \mathcal{P}_n^{lin}$, we consider sequences of parameters $\beta_n \in \mathbb{R}^{p_n}$, $\sigma_n^2 \in (0, \infty)$ and $\Sigma_n \in S_{p_n}$ (where S_{p_n} is the set of all symmetric, positive definite $p_n \times p_n$ matrices). All the operators $\mathbb{E}$, Var and Cov are to be understood with respect to P_n^n.

We have to show that, for arbitrary but fixed $\alpha \in [0, 1]$, $\hat{q}_\alpha/\sigma_n$ converges in P_n^n-probability to q_α, the α quantile of the distribution of $lN\tau + v$, the cdf of which we denote by F. In either case of Theorem 3.4, it is easy to see that the quantile q_α is unique. Note further that for $\alpha \in (0, 1]$, $\hat{q}_\alpha =$

$\hat{F}_n^\dagger(\alpha) := \inf\{t \in \mathbb{R} : \hat{F}_n(t) \geq \alpha\}$. We treat the case $\alpha \in \{0,1\}$ separately at the end of the proof, because $q_1 = -q_0 = \infty$. To deal with the empirical quantiles we use a standard argument. For $\alpha \in (0,1)$ and $\varepsilon > 0$, consider

$$P_n^n(|\hat{q}_\alpha/\sigma_n - q_\alpha| > \varepsilon) = P_n^n(\hat{q}_\alpha/\sigma_n > q_\alpha + \varepsilon) + P_n^n(\hat{q}_\alpha/\sigma_n < q_\alpha - \varepsilon).$$

To bound the first probability on the right, abbreviate $J_i := \mathbb{1}_{\{\hat{u}_i/\sigma_n > q_\alpha + \varepsilon\}}$ and note that by definition of the OLS predictor, the leave-one-out residuals $\hat{u}_i = y_i - x_i'(X_{[i]}'X_{[i]})^\dagger X_{[i]}'Y_{[i]}$, $i = 1, \dots, n$, and thus also the J_i, $i = 1, \dots, n$, are exchangeable under P_n^n. A basic property of the quantile function $\hat{F}_n^\dagger$ (cf. van der Vaart, 2007, Lemma 21.1) yields

$$\begin{aligned} P_n^n(\hat{q}_\alpha/\sigma_n > q_\alpha + \varepsilon) &= P_n^n\left(\alpha > \hat{F}_n(\sigma_n(q_\alpha + \varepsilon))\right) \\ &= P_n^n\left(1 - \hat{F}_n(\sigma_n(q_\alpha + \varepsilon)) > 1 - \alpha\right) \\ &= P_n^n\left(\frac{1}{n}\sum_{i=1}^n (J_i - \mathbb{E}[J_1]) > 1 - \alpha - \mathbb{E}[J_1]\right) \\ &= P_n^n\left(\frac{1}{n}\sum_{i=1}^n (J_i - \mathbb{E}[J_i]) > F_n(q_\alpha + \varepsilon) - \alpha\right), \end{aligned}$$

where $F_n(t) := P_n^n(\hat{u}_1/\sigma_n \leq t)$ is the marginal cdf of the scaled leave-one-out residuals. If we can show that

$$F_n(t) \to F(t), \qquad \forall t \in \mathbb{R}, \tag{A.7}$$

as $n \to \infty$, then $F_n(q_\alpha + \varepsilon) \to F(q_\alpha + \varepsilon) > \alpha$, because q_α is unique, and thus the probability in the second to last display can be bounded, at least for n sufficiently large, using Markov's inequality, by

$$\begin{aligned} &(F_n(q_\alpha + \varepsilon) - \alpha)^{-2}\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n (J_i - \mathbb{E}[J_i])\right|^2\right] \\ &\quad = (F_n(q_\alpha + \varepsilon) - \alpha)^{-2}\left(\frac{1}{n}\operatorname{Var}[J_1] + \frac{n(n-1)}{n^2}\operatorname{Cov}(J_1, J_2)\right), \end{aligned}$$

where the equality holds in view of the exchangeability of the J_i. An analogous argument yields a similar upper bound for the probability $P_n^n(\hat{q}_\alpha/\sigma_n \leq q_\alpha - \varepsilon)$ but with $(F_n(q_\alpha + \varepsilon) - \alpha)^{-2}$ replaced by $(\alpha - F_n(q_\alpha - \varepsilon))^{-2}$, and J_i replaced by $K_i = \mathbb{1}_{\{\hat{u}_i/\sigma_n \leq q_\alpha - \varepsilon\}}$. The proof will thus be finished if we can establish (A.7) and show that $\operatorname{Cov}(J_1, J_2)$ and $\operatorname{Cov}(K_1, K_2)$ converge to zero as $n \to \infty$. We only consider $\mathrm{Cov}(J_1, J_2) = \mathrm{Cov}(1 - J_1, 1 - J_2)$, as the argument for $\mathrm{Cov}(K_1, K_2)$ is analogous. Write $\delta = q_\alpha + \varepsilon$ and

$$\mathrm{Cov}(1-J_1, 1-J_2) = P_n^n(\hat{u}_1/\sigma_n \leq \delta, \hat{u}_2/\sigma_n \leq \delta) - P_n^n(\hat{u}_1/\sigma_n \leq \delta) P_n^n(\hat{u}_2/\sigma_n \leq \delta).$$

Now,

$$\begin{pmatrix} \hat{u}_1/\sigma_n \\ \hat{u}_2/\sigma_n \end{pmatrix} = \begin{pmatrix} \hat{u}_{1[2[}/\sigma_n \\ \hat{u}_{2[1]}/\sigma_n \end{pmatrix} + \begin{pmatrix} \hat{e}_1 \\ \hat{e}_2 \end{pmatrix},$$

where $\hat{u}_{i[j]} = y_i - x_i' \hat{\beta}_n^{[ij]}$, $\hat{\beta}_n^{[ij]} = (X_{[ij]}' X_{[ij]})^\dagger X_{[ij]}' Y_{[ij]}$, and $\hat{e}_i = (\hat{u}_i - \hat{u}_{i[j]})/\sigma_n = x_i'(\hat{\beta}_n^{[ij]} - \hat{\beta}_n^{[i]})/\sigma_n$, for $\{i, j\} = \{1, 2\}$. Therefore, $\mathbb{E}[\hat{e}_i \| Y_{[i]}, X_{[i]}] = 0$, because $\mathbb{E}[x_i] = 0$, and $\mathbb{E}[\hat{e}_i^2 \| Y_{[i]}, X_{[i]}] = \|\Sigma^{1/2}(\hat{\beta}_n^{[i]} - \hat{\beta}_n^{[ij]})/\sigma_n\|_2^2$, which converges to zero in P_n^n-probability, by Lemma A.4 (for a sample of size $n-1$ instead of n), which applies here because (C2) is satisfied under (C3). Hence, $\hat{e}_1$ and $\hat{e}_2$ converge to zero in probability. The joint distribution function of $\hat{u}_{1[2]}/\sigma_n$ and $\hat{u}_{2[1]}/\sigma_n$ can be written as

(A.8)

$$\begin{aligned} &P_n^n(\hat{u}_{1[2]}/\sigma_n \leq s, \hat{u}_{2[1]}/\sigma_n \leq t) \\ &\quad = \mathbb{E}\left[P_n^n\left(x_1'(\beta_n - \hat{\beta}_n^{[12]})/\sigma_n + v_1 \leq s, x_2'(\beta_n - \hat{\beta}_n^{[12]})/\sigma_n + v_2 \leq t \middle\| Y_{[12]}, X_{[12]}\right)\right] \\ &\quad = \mathbb{E}\left[G_n\left(s, \Sigma^{1/2}(\beta_n - \hat{\beta}_n^{[12]})/\sigma_n\right) G_n\left(t, \Sigma^{1/2}(\beta_n - \hat{\beta}_n^{[12]})/\sigma_n\right)\right], \end{aligned}$$

where, for $t \in \mathbb{R}$ and $b \in \mathbb{R}^{p_n}$, G_n is defined as $G_n(t, b) = P_n(b'\Sigma^{-1/2}x_0 + v_0 \leq t)$. Note that G_n depends only on $\mathcal{L}_l$, $\mathcal{L}_w$, $\mathcal{L}_v$ and on n, through p_n. If we abbreviate $\tilde{\beta}_n^{[12]} = \Sigma^{1/2}(\beta_n - \hat{\beta}_n^{[12]})/\sigma_n$ and $\tilde{\beta}_n^{[1]} = \Sigma^{1/2}(\beta_n - \hat{\beta}_n^{[1]})/\sigma_n$, we arrive at

$$\mathrm{Cov}(1 - J_1, 1 - J_2) = \mathbb{E}\left[G_n\left(\delta, \tilde{\beta}_n^{[12]}\right)^2\right] - \mathbb{E}\left[G_n\left(\delta, \tilde{\beta}_n^{[1]}\right)\right]^2 + o(1),$$

provided the bivariate distribution function in (A.8) converges pointwise to a continuous limit. We finish the proof by showing that for all $t \in \mathbb{R}$, the bounded random variables $G_n(t, \tilde{\beta}_n^{[12]})$ and $G_n(t, \tilde{\beta}_n^{[1]})$ both converge to $F(t)$, in P_n^n-probability, and hence, (A.8) converges to $F(s)F(t)$, which is continuous. Note that this also implies (A.7), because $F_n(t) = \mathbb{E}[G_n(t, \tilde{\beta}_n^{[1]})]$.

To this end, we note that for an arbitrary measureable set $B_n \subseteq \mathbb{R}^{p_n}$ and for any $\varepsilon > 0$,

$$\begin{aligned} P_n^n\left(\sup_{t\in\mathbb{R}}\left|G_n(t,\tilde{\beta}_n^{[1]}) - F(t)\right| > \varepsilon\right) &\le P_n^n\left(\sup_{t\in\mathbb{R}}\left|G_n(t,\tilde{\beta}_n^{[1]}) - F(t)\right| > \varepsilon, \tilde{\beta}_n^{[1]} \in B_n\right) \\ &\quad + P_n^n\left(\tilde{\beta}_n^{[1]} \notin B_n\right) \\ &\le a_n(\varepsilon) \;+\; P_n^n\left(\tilde{\beta}_n^{[1]} \notin B_n\right), \end{aligned}$$

where $a_n(\varepsilon) = 1$ if $\sup_{b\in B_n}\sup_{t\in\mathbb{R}} |G_n(t,b) - F(t)| > \varepsilon$, and $a_n(\varepsilon) = 0$, else. Now, we first consider the case $\kappa = 0$. Thus, Lemma A.6, which also applies to $\hat{\beta}_n^{[1]}$, yields $\|\tilde{\beta}_n^{[1]}\|_2 \to \tau = 0$, as $n \to \infty$, in P_n^n-probability. Therefore, the probability in the last line of the previous display converges to zero if we take $B_n = \{b \in \mathbb{R}^{p_n} : \|b\|_2 \le g_1(n)\}$ and $g_1(n) \to 0$ sufficiently slowly, as $n \to \infty$. Hence, Lemma A.7(i) applies and shows that also $a_n(\varepsilon) \to 0$ as $n \to \infty$, for every $\varepsilon > 0$. If $\kappa > 0$, Lemma A.6 yields $\|\tilde{\beta}_n^{[1]}\|_2 \to \tau > 0$ and $\|\tilde{\beta}_n^{[1]}\|_4 \to 0$, in P_n^n-probability, as $n \to \infty$. Thus, the probability in the last line of the previous display converges to zero if we take $B_n = \{b \in \mathbb{R}^{p_n} : b \ne 0, |\|b\|_2 - \tau| \le g_1(n), \|b\|_4/\|b\|_2 \le g_2(n)\}$ and sequences g_1 and g_2 that converge to zero sufficiently slowly. Now Lemma A.7(ii) shows that also $a_n(\varepsilon) \to 0$ as $n \to \infty$, for every $\varepsilon > 0$. The same argument applies to $\tilde{\beta}_n^{[12]}$ instead of $\tilde{\beta}_n^{[1]}$, which finishes the proof in the case $\alpha \in (0,1)$.

Next, we treat the case $\alpha = 0$. In either case of the theorem, we have $\lim_{\gamma\to 0} q_\gamma = q_0 = -\infty$. By definition, $\hat{q}_0 \le \hat{q}_\gamma$, for any $\gamma \in (0,1)$. Thus, for any $M > 0$, there exists a $\gamma \in (0,1)$, such that $q_\gamma < -2M$ and $P_n^n(\hat{q}_0/\sigma_{P_n} < -M) \ge P_n^n(\hat{q}_\gamma/\sigma_{P_n} < -M) \to 1$, as $n \to \infty$, in view of the first part. In other words, $\hat{q}_0/\sigma_{P_n}$ converges to $-\infty = q_0$ in P_n^n-probability. A similar argument can be used to treat the case $\alpha = 1$. □

Proof of Lemma A.6. On the event $\{\lambda_{\min}(\tilde{X}'\tilde{X}) > 0\}$, which has asymptotic probability one in view of Lemma B.1(ii), notice the identity

$$\Sigma_P^{1/2}(\hat{\beta}_n - \beta_P)/\sigma_P = (\tilde{X}'\tilde{X})^\dagger \tilde{X}'v,$$

where $v = (v_1, \dots, v_n)' = (Y - X\beta_P)/\sigma_P$. Thus, the distribution under $P \in \mathcal{P}_n$ of the quantity of interest does not depend on the parameters β_P, σ_P^2 and Σ_P. Hence, without loss of generality, we assume for the rest of this proof that $\beta_P = 0$, $\sigma_P^2 = 1$ and $\Sigma_P = I_{p_n}$. First, we have to show that $\|\hat{\beta}_n\|_2 \to \tau \in [0,\infty)$, in probability, for a $\tau = \tau(\mathcal{L}_l, \kappa)$ with the properties mentioned in the lemma. To this end, consider the conditional mean

$$\mathbb{E}\left[\|\hat{\beta}_n\|_2^2\middle\| X\right] = \operatorname{trace}(X'X)^\dagger X'X(X'X)^\dagger = \operatorname{trace}(X'X)^\dagger \xrightarrow{a.s.} \tau^2,$$

by Lemma B.1(iv) and for τ as desired (cf. Remark B.2). From the same lemma we get convergence of the conditional variance

$$
\begin{aligned}
\operatorname{Var}\left[\|\hat{\beta}_n\|_2^2\Big\|X\right] &= \operatorname{Var}\left[v'X(X'X)^{\dagger 2}X'v\Big\|X\right] =: \operatorname{Var}[v'Kv\|X] \\
&= 2\operatorname{trace} K^2 + (\mathbb{E}[v_1^4]-3)\sum_{i=1}^{n} K_{ii}^2 \\
&\leq 2\operatorname{trace} K^2 + (\mathbb{E}[v_1^4]+3)\sum_{i,j=1}^{n} K_{ij}^2 = (\mathbb{E}[v_1^4]+5)\operatorname{trace} K^2 \\
&= (\mathbb{E}[v_1^4]+5)\operatorname{trace} X(X'X)^{\dagger 2}X'X(X'X)^{\dagger 2}X' \\
&= (\mathbb{E}[v_1^4]+5)\operatorname{trace}(X'X)^{\dagger 2} \xrightarrow{a.s.} 0.
\end{aligned}
$$

For the second claim it suffices to show that $\|\hat{\beta}_n\|_4^4 \to 0$, in probability. Notice that for $M := (m_1,\dots,m_{p_n})' := (X'X)^{\dagger}X'$, we have

$$
\|\hat{\beta}_n\|_4^4 = \|Mv\|_4^4 = \sum_{j=1}^{p_n}(m_j'v)^4 = \sum_{j=1}^{p_n}\sum_{i_1,i_2,i_3,i_4=1}^{n} m_{ji_1}m_{ji_2}m_{ji_3}m_{ji_4}v_{i_1}v_{i_2}v_{i_3}v_{i_4}.
$$

After taking conditional expectation given X, only terms with paired indices remain and we get

$$
\begin{aligned}
\mathbb{E}\left[\|\hat{\beta}_n\|_4^4\Big\|X\right] &= \sum_{j=1}^{p_n}\left(\mathbb{E}[v_1^4]\sum_{i=1}^{n} m_{ji}^4 + 3\sum_{i\neq k}^{n} m_{ji}^2 m_{jk}^2\right) \\
&\leq \sum_{j=1}^{p_n}\left(\mathbb{E}[v_1^4]\sum_{i,k=1}^{n} m_{ji}^2 m_{jk}^2 + 3\sum_{i,k=1}^{n} m_{ji}^2 m_{jk}^2\right) \\
&= (\mathbb{E}[v_1^4]+3)\sum_{j=1}^{p_n}(m_j'm_j)^2 \leq (\mathbb{E}[v_1^4]+3)\operatorname{trace}\sum_{i,j=1}^{p_n} m_i m_i' m_j m_j' \\
&= (\mathbb{E}[v_1^4]+3)\operatorname{trace}(M'M)^2 = (\mathbb{E}[v_1^4]+3)\operatorname{trace}(X'X)^{\dagger 2} \xrightarrow{a.s.} 0,
\end{aligned}
$$

by Lemma B.1(iii). □

Proof of Lemma A.7. First, in the case (i), for every $n \in \mathbb{N}$, take $b_n \in B_n = \{b \in \mathbb{R}^{p_n} : \|b\|_2 \leq g_1(n)\}$ and simply note that $l_0 b_n' w_n \to 0$, in probability, and thus $G(t, b_n) \to F(t)$ weakly. Since the limit is continuous, Polya's theorem yields uniform convergence in $t \in \mathbb{R}$. Since the sequence $b_n \in B_n$ was arbitrary, we also get uniform convergence over B_n.

Next, we consider the Gaussian case (iii), so $B_n = \{b \in \mathbb{R}^p : |\|b\|_2 - \tau| \leq g_1(n)\}$. For every $n \in \mathbb{N}$, let $b_n \in B_n$ be arbitrary, and note that $t \mapsto G(t, b_n)$ is the distribution function of $l_0 b_n' w_n + v_0$, where $w_n \overset{\mathcal{L}}{=} \mathcal{N}(0, I_{p_n})$, and l_0, w_n, v_0 are independent. Clearly, $l_0 b_n' w_n + v_0 \overset{\mathcal{L}}{=} l_0 N \|b_n\|_2 + v_0 \to l_0 N\tau + v_0$, weakly, and this limit has continuous distribution function F. Hence, by Polya's theorem, $\sup_t |\mathbb{P}(l_0 b_n' w_n + v_0 \leq t) - F(t)| \to 0$, as $n \to \infty$. And since the sequence $b_n \in B_n$ was arbitrary, the result follows.

In the general case (ii) first note that B_n may be empty. By our convention that $\sup \varnothing = 0$ it suffices to restrict to the subsequence n' for which $B_{n'} \neq \varnothing$. If this is only a finite sequence, then the result is trivial. For convenience, we write $n = n'$. So let $b_n \in B_n$ and define the triangular array $z_{nj} := b_{nj} w_{0j}$, $j = 1, \ldots, p_n$, which satisfies $\mathbb{E}[z_{nj}] = 0$ and $s_n^2 := \sum_{j=1}^{p} \mathbb{E}[z_{nj}^2] = \|b_n\|_2^2 \neq 0$. The Lyapounov condition is verified by

$$\begin{aligned}\sum_{j=1}^{p_n} s_n^{-(2+\delta)} \mathbb{E}[|z_{nj}|^{2+\delta}] &= \mathbb{E}\left[|w_{01}|^{2+\delta}\right] \left(\frac{\|b_n\|_{2+\delta}}{\|b_n\|_2}\right)^{2+\delta} \\ &\leq \mathbb{E}\left[|w_{01}|^{2+\delta}\right] [g_2(n)]^{2+\delta} \xrightarrow[n\to\infty]{} 0.\end{aligned}$$

Therefore, by Lyapounov's CLT (Billingsley, 1995, Theorem 27.3), we have

$$b_n' w_n / \|b_n\|_2 = \sum_{j=1}^{p_n} z_{nj}/s_n \xrightarrow[n\to\infty]{w} \mathcal{N}(0,1).$$

Since $b_n \in B_n$, we must have $\|b_n\|_2 \to \tau$ as $n \to \infty$, and thus, $b_n' w_n = \|b_n\|_2 b_n' w_n / \|b_n\|_2 \xrightarrow{w} N\tau$, where $N \overset{\mathcal{L}}{=} \mathcal{N}(0,1)$, as $n \to \infty$, and, by independence, $l_0 b_n' w_n + v_0 \xrightarrow{w} l_0 N\tau + v_0$. Since the distribution function of this limit is continuous, Polya's theorem yields $\sup_t |G(t, b_n) - F(t)| \to 0$, as $n \to \infty$. Now the proof is finished because this convergence holds for arbitrary sequences $b_n \in B_n$. □

APPENDIX B: AUXILIARY RESULTS

Lemma B.1. *On a common probability space $(\Omega, \mathcal{F}, \mathbb{P})$, consider an i.i.d. sequence $L_0 = \{l_i : i = 1, 2, \ldots\}$ of random variables satisfying $|l_1| \geq c > 0$, and a double infinite array $W_0 = \{w_{ij} : i, j = 1, 2, \ldots\}$ of i.i.d. random variables with mean zero, unit variance and $\mathbb{E}[w_{11}^4] < \infty$, such that L_0 and W_0 are independent. For a sequence of positive integers (p_n) with $p_n \leq n$, consider the $n \times p_n$ random matrix $\tilde{X} = \Lambda W$, where $\Lambda = \operatorname{diag}(l_1, \ldots, l_n)$ is* *diagonal and* $W = \{w_{ij} : i = 1, \dots, n; j = 1, \dots, p_n\}$*. Let* $(\tilde{X}'\tilde{X})^\dagger$ *denote the Moore-Penrose pseudo inverse of* $\tilde{X}'\tilde{X}$*. If* $p_n/n \to \kappa \in [0, 1)$ *then the following holds:*

(i) $\liminf_{n\to\infty} \lambda_{min}(\tilde{X}'\tilde{X}/n) \geq c^2(1 - \sqrt{\kappa})^2$*, almost surely.*
(ii) $\lim_{n\to\infty} \mathbb{P}(\lambda_{min}(\tilde{X}'\tilde{X}/n) \leq \varepsilon) = 0$ *for all* $\varepsilon < c^2(1 - \sqrt{\kappa})^2$*.*
(iii) If $m > 1$*, then* $\operatorname{trace}(\tilde{X}'\tilde{X})^{\dagger m} \to 0$*, almost surely, as* $n \to \infty$*.*
(iv) $\operatorname{trace}(\tilde{X}'\tilde{X})^\dagger \to \tau^2$ *almost surely, as* $n \to \infty$*, for some constant* $\tau \in [0, \infty)$ *that depends only on* κ *and on the distribution of* l_1^2 *and satisfies* $\tau = 0$ *if, and only if,* $\kappa = 0$*.*

Proof. Let $\lambda_1^{(n)} \leq \cdots \leq \lambda_{p_n}^{(n)}$ and $\mu_1^{(n)} \leq \cdots \leq \mu_{p_n}^{(n)}$ denote the ordered eigenvalues of $\tilde{X}'\tilde{X}/n$ and $W'W/n$, respectively. Then,

$$\lambda_1^{(n)} = \inf_{\|t\|=1} t'W'\Lambda^2 Wt/n \geq \left(\min_{i=1,\dots,n} l_i^2\right) \inf_{\|t\|=1} t'W'Wt/n \geq c^2\mu_1^{(n)},$$

and from the Bai-Yin Theorem (Bai and Yin, 1993) it follows that $\mu_1^{(n)} \to (1 - \sqrt{\kappa})^2 > 0$, almost surely, as $p_n/n \to \kappa \in [0, 1)$ (cf. Huber and Leeb, 2013, for the case $\kappa = 0$). This finishes the proof of part (i). Part (ii) is now a textbook argument: Simply note that for $k \leq n$, we have $\mathbb{P}(\lambda_1^{(n)} \leq \varepsilon) \leq \mathbb{P}(\inf_{r\geq k} \lambda_1^{(r)} \leq \varepsilon)$ and that $\inf_{r\geq k} \lambda_1^{(r)} \leq \inf_{r\geq k+1} \lambda_1^{(r)}$ for all $k \in \mathbb{N}$. Thus

$$\begin{aligned}\limsup_{n\to\infty} \mathbb{P}(\lambda_1^{(n)} \leq \varepsilon) &= \inf_{k\in\mathbb{N}} \sup_{n\geq k} \mathbb{P}(\lambda_1^{(n)} \leq \varepsilon)\\ &\leq \inf_{k\in\mathbb{N}} \mathbb{P}\left(\inf_{r\geq k} \lambda_1^{(r)} \leq \varepsilon\right) = \lim_{k\to\infty} \mathbb{P}\left(\inf_{r\geq k} \lambda_1^{(r)} \leq \varepsilon\right)\\ &= \mathbb{P}\left(\forall k \in \mathbb{N} : \inf_{r\geq k} \lambda_1^{(r)} \leq \varepsilon\right) = \mathbb{P}\left(\liminf_{n\to\infty} \lambda_1^{(n)} \leq \varepsilon\right) = 0.\end{aligned}$$

Next, abbreviate $\lambda_j = \lambda_j^{(n)}$, $\mu_j = \mu_j^{(n)}$, for $m \geq 1$ set $\alpha_m := c^{2m}(1-\sqrt{\kappa})^{2m}$ and, for $\alpha > 0$, define the functions h_0 and h_α by $h_0(y) = 1/|y|$ if $y \neq 0$ and $h_0(0) = 0$, and by $h_\alpha(y) = 1/|y|$, if $|y| > \alpha/2$ and $h_\alpha(y) = 2/\alpha$, if $|y| \leq \alpha/2$. With this notation, and from the previous considerations, we see that the difference between

$$\operatorname{trace}(\tilde{X}'\tilde{X})^{\dagger m} = n^{-m} \operatorname{trace}(\tilde{X}'\tilde{X}/n)^{\dagger m} = \frac{p_n}{n^m}\frac{1}{p_n}\sum_{j=1}^{p_n} h_0(\lambda_j^m),$$

and

$$\frac{p_n}{n^m}\frac{1}{p_n}\sum_{j=1}^{p_n} h_{\alpha_m}(\lambda_j^m)$$

converges to zero, almost surely, because $\lambda_j^m \geq \lambda_1^m \geq c^{2m}\mu_1^m \to \alpha_m > \alpha_m/2 > 0$, almost surely. But we have $n^{-m}\sum_{j=1}^{p_n} h_{\alpha_m}(\lambda_j^m) \leq (p_n/n^m)(2/\alpha_m) \to 0$, if $m > 1$, or if $m = 1$ and $\kappa = 0$. This finishes (iii) and the case $\kappa = 0$ of part (iv).

For the remainder of part (iv), let $m = 1$ and $\kappa > 0$, and first note that the empirical spectral distribution function $F_n^{\Lambda^2}$ of Λ^2 is simply given by the empirical distribution function of $l_1^2, \ldots, l_n^2$, and this converges weakly (even uniformly) to the distribution function of l_1^2, almost surely. Hence, from Theorem 4.3 in Bai and Silverstein (2010), it follows that, almost surely, the empirical spectral distribution function $F_n^{\tilde{X}'\tilde{X}/n}$ of $\tilde{X}'\tilde{X}/n$ converges vaguely, as $p_n/n \to \kappa \in (0, 1)$, to a non-random distribution function F that depends only on κ and on the distribution of l_1^2. From the argument in the previous paragraph we know that $\lambda_1 \geq c^2\mu_1 \to c^2(1-\sqrt{\kappa})^2 = \alpha_1 > 0$, almost surely, and thus the support of F must be lower bounded by α_1. Since h_{α_1} is continuous and vanishes at infinity, by vague convergence, we have (cf. Billingsley, 1995, relation (28.2))

$$\frac{1}{p_n}\sum_{j=1}^{p_n} h_{\alpha_1}(\lambda_j) = \int_{-\infty}^{\infty} h_{\alpha_1}(y)dF_n^{\tilde{X}'\tilde{X}/n}(y)$$

$$\xrightarrow{a.s.} \int_{-\infty}^{\infty} h_{\alpha_1}(y)dF(y) = \int_{-\infty}^{\infty} \frac{1}{y}\,dF(y) =: \tau_0^2 \in (0, 1/\alpha_1).$$

Thus

$$\frac{p_n}{n}\frac{1}{p_n}\sum_{j=1}^{p_n} h_{\alpha_1}(\lambda_j) \quad \xrightarrow{a.s.} \quad \kappa\tau_0^2 \ =: \ \tau^2 > 0.$$

□

Remark B.2. If the l_i in Lemma B.1 satisfy $|l_i| = 1$, almost surely, then τ in part (iv) is given by $\tau(\kappa) = \sqrt{\kappa/(1-\kappa)}$ (cf. Huber and Leeb, 2013, Lemma B.2).

Lemma B.3. *Suppose that for every $n \in \mathbb{N}$, the class $\mathcal{P}_n = \mathcal{P}_n^{(lin)}(\mathcal{L}_l, \mathcal{L}_w, \mathcal{L}_v)$ is as in Condition (C3) and $\mathcal{L}_l$ has a finite fourth moment. Furthermore, let $p_n/n \to \kappa > 0$ and $n > p_n$ for all $n \in \mathbb{N}$. Then, for every $c \in [0, 1]$, every $\eta \in (0, \infty]$ and every $\varepsilon \in (0, 1)$, the James-Stein-type estimator $\hat{\beta}_n(c)$ satisfies*

$$\sup_{P\in\mathcal{P}_n} P^n\left(\left\|\Sigma_P^{1/2}\left(\hat{\beta}_n(c) - \beta_P\right)/\sigma_P\right\|_{2+\eta} \geq \varepsilon c\sqrt{\kappa}/2\right) \quad \xrightarrow[n\to\infty]{} \quad 1.$$

Proof. Consider a sequence $P_n \in \mathcal{P}_n$, such that $\beta_{P_n} = \sigma_{P_n}\Sigma_{P_n}^{-1/2}(\sqrt{\kappa}, 0, \dots, 0)'$, so that $b := \Sigma_{P_n}^{1/2}\beta_{P_n}/\sigma_{P_n} = (\sqrt{\kappa}, 0, \dots, 0)' \in \mathbb{R}^{p_n}$ and $\|b\|_2 = \|b\|_q = \sqrt{\kappa}$, for every $q \in (0, \infty]$. Simple relations of ℓ^q-norms yield

$$\begin{aligned}
\left\|\Sigma_{P_n}^{1/2}\left(\hat{\beta}_n(c) - \beta_{P_n}\right)/\sigma_{P_n}\right\|_{2+\eta} &\geq \left\|\Sigma_{P_n}^{1/2}\left(\hat{\beta}_n(c) - \beta_{P_n}\right)/\sigma_{P_n}\right\|_{(2+\eta)\vee 4} \\
&= \left\|s_n\Sigma_{P_n}^{1/2}(\hat{\beta}_n - \beta_{P_n})/\sigma_{P_n} - (1-s_n)b\right\|_{(2+\eta)\vee 4} \\
&\geq \left||s_n|\left\|\Sigma_{P_n}^{1/2}(\hat{\beta}_n - \beta_{P_n})/\sigma_{P_n}\right\|_{(2+\eta)\vee 4} - |s_n - 1|\sqrt{\kappa}\right| \\
&\geq |s_n - 1|\sqrt{\kappa} - |s_n|\left\|\Sigma_{P_n}^{1/2}(\hat{\beta}_n - \beta_{P_n})/\sigma_{P_n}\right\|_{(2+\eta)\vee 4},
\end{aligned}$$

where s_n is defined as before, i.e.,

$$s_n = \begin{cases} \left(1 - \frac{p_n}{n}\frac{c}{t_n^2}\frac{\hat{\sigma}_n^2}{\sigma_n^2}\right)_+, & \text{if } t_n^2 > 0, \\ 1, & \text{else,} \end{cases}$$

and $t_n^2 = \hat{\beta}_n' X' X\hat{\beta}_n/(n\sigma_n^2)$, so that $\hat{\beta}_n(c) = s_n\hat{\beta}_n$. Clearly, we have $|s_n| \leq 1$ and $\left\|\Sigma_{P_n}^{1/2}(\hat{\beta}_n - \beta_{P_n})/\sigma_{P_n}\right\|_{(2+\eta)\vee 4} \leq \left\|\Sigma_{P_n}^{1/2}(\hat{\beta}_n - \beta_{P_n})/\sigma_{P_n}\right\|_4 \to 0$, in P_n^n-probability, by Lemma A.6. Therefore, we see that

$$\begin{aligned}
\text{(B.1)} \qquad & P_n^n\left(|s_n - 1|\sqrt{\kappa} - o_{P_n^n}(1) \geq \varepsilon c\sqrt{\kappa}/2\right) \\
&\leq P_n^n\left(\left\|\Sigma_{P_n}^{1/2}\left(\hat{\beta}_n(c) - \beta_{P_n}\right)/\sigma_{P_n}\right\|_{2+\eta} \geq \varepsilon c\sqrt{\kappa}/2\right).
\end{aligned}$$

Now, as in the proof of Lemma A.5 but with $e = 0$ (now the linear model is assumed correct),

$$t_n^2 = \left\|\frac{\tilde{X}\Sigma_{P_n}^{1/2}\beta_{P_n}}{\sqrt{n\sigma_{P_n}^2}} + \frac{P_{\tilde{X}}v}{\sqrt{n}}\right\|_2^2$$

and we showed already in that proof that the mixed term vanishes and $\|P_{\tilde{X}}v/\sqrt{n}\|_2^2 \to \kappa$ in P_n^n-probability. For the remaining quadratic form note that $\tilde{x}_i \overset{P_n^n}{=} l_i(w_{i1}, \dots, w_{ip_n})'$, where $l_i \sim \mathcal{L}_l$ and $w_{ij} \sim \mathcal{L}_w$ are all independent, for $i = 1, \dots, n$, $j = 1, \dots, p_n$, in view of Condition (C3). Thus

$A := \|\tilde{X}\Sigma_{P_n}^{1/2}\beta_{P_n}/\sqrt{n\sigma_{P_n}^2}\|_2^2 = \frac{1}{n}\sum_{i=1}^n (b'\tilde{x}_i)^2$. Clearly, $\mathbb{E}[A] = \|b\|_2^2 = \kappa$ and

$$\begin{aligned}
\operatorname{Var}[A] &= \frac{1}{n}\operatorname{Var}[(b'\tilde{x}_1)^2] = \frac{1}{n}\operatorname{Var}\left[l_1^2 \sum_{j_1,j_2}^{p_n} b_{j_1}b_{j_2}w_{1j_1}w_{2j_2}\right] \\
&\le \frac{1}{n}\mathbb{E}\left[l_1^4 \sum_{j_1,j_2,j_3,j_4}^{p_n} b_{j_1}b_{j_2}b_{j_3}b_{j_4}w_{1j_1}w_{2j_2}w_{1j_3}w_{2j_4}\right] \\
&= \frac{1}{n}\mathbb{E}[l_1^4]\left[\mathbb{E}[w_{11}^4]\sum_{j=1}^{p_n} b_j^4 + 3\sum_{j_1\neq j_2}^{p_n} b_{j_1}^2 b_{j_2}^2\right] \\
&\le \frac{1}{n}\mathbb{E}[l_1^4](\mathbb{E}[w_{11}^4]+3)\|b\|_2^4 \xrightarrow{n\to\infty} 0.
\end{aligned}$$

Thus, $t_n^2 \to \kappa + \kappa = 2\kappa$, in P_n^n-probability. Moreover, $\hat{\sigma}_n^2/\sigma_n^2 \to 1$, in P_n^n-probability, because its conditional mean given X converges to 1 and its conditional variance converges to zero (see the arguments in (A.5) and in the lines immediately before that display). Thus $|s_n - 1|\sqrt{\kappa} \to c\sqrt{\kappa}/2 > \varepsilon c\sqrt{\kappa}/2$, such that the probability in (B.1) converges to 1 and the proof is finished. □

Lemma B.4. *If $\hat{m}_n$ is a 0-stable predictor w.r.t. some class $\mathcal{P}$ of distributions on $\mathcal{Z}$, then there exists a collection $\{g_P : P \in \mathcal{P}\}$ of measurable functions $g_P : \mathcal{X} \to \mathbb{R}$, such that for all $P \in \mathcal{P}$,*

$$\begin{aligned}
&P^{n+1}\left(\{(T_n, z_0) \in \mathcal{Z}^{n+1} : M_{n,p}(T_n, x_0) = g_P(x_0)\}\right) = 1, \quad \textit{and} \\
&P^{n}\left(\{(T_{n-1}, z_0) \in \mathcal{Z}^{n} : M_{n-1,p}(T_{n-1}, x_0) = g_P(x_0)\}\right) = 1.
\end{aligned}$$

Remark B.5. Note that the dependence of g_P on P in Lemma B.4 can not be avoided. For example, suppose that $\mathcal{P} = \{P_0, P_1\}$ with disjoint supports $S_0 \cap S_1 = \varnothing$ and consider the naive algorithm

$$M_{n,p}(T_n, x_0) := \begin{cases} 0, & \text{if } T_n \in S_0^n, \\ 1, & \text{if } T_n \in S_1^n. \end{cases}$$

Clearly, this algorithm is 0-stable, but it does depend on T_n. This is not in contradiction with Lemma B.4, because $M_{n,p}(T_n, x_0) = g_P(x_0)$, where

$$g_P(x_0) := \begin{cases} 0, & \text{if } P = P_0, \\ 1, & \text{if } P = P_1. \end{cases}$$

The paradox is resolved by noticing that since the supports S_0 and S_1 are disjoint we can perfectly discriminate between P_0 and P_1.

Proof of Lemma B.4. Fix $P \in \mathcal{P}$. For $i = 1, \dots, n$, let $z_i, z_i' \in \mathcal{Z}$ and $x_0 \in \mathcal{X}$, and note that

$$\begin{aligned}
&\left|M_{n,p}(z_1, \dots, z_n, x_0) - M_{n,p}(z_1', \dots, z_n', x_0)\right| \\
&= \left|\sum_{i=1}^n \left[M_{n,p}(z_1', \dots, z_{i-1}', z_i, \dots, z_n, x_0) - M_{n,p}(z_1', \dots, z_i', z_{i+1}, \dots, z_n, x_0)\right]\right| \\
&\leq \sum_{i=1}^n \Big[\left|M_{n,p}(z_1', \dots, z_{i-1}', z_i, \dots, z_n, x_0) - M_{n-1,p}(z_1', \dots, z_{i-1}', z_{i+1}, \dots, z_n, x_0)\right| + \\
&\qquad \left|M_{n-1,p}(z_1', \dots, z_{i-1}', z_{i+1}, \dots, z_n, x_0) - M_{n,p}(z_1', \dots, z_i', z_{i+1}, \dots, z_n, x_0)\right| \Big].
\end{aligned}$$

By 0-stability, the integral of this upper bound with respect to P^{2n+1} is equal to zero. Therefore, applying Lemma B.6 with $f = M_{n,p}$, $S = \mathcal{Z}^n$, $P_S = P^n$, $T = \mathcal{X}$ and P_T equal to the x-marginal distribution of P, the first claim follows. The second claim is now a simple consequence of 0-stability. □

Lemma B.6. *Let $(S, \mathcal{S}, P_S)$ and $(T, \mathcal{T}, P_T)$ be two probability spaces, and let $f : S \times T \to \mathbb{R}$ be measurable w.r.t. the product sigma algebra $\mathcal{S} \otimes \mathcal{T}$ and the Borel sigma algebra on $\mathbb{R}$. If*

$$\int_{S^2 \times T} |f(s_1, t) - f(s_2, t)| \, dP_S \otimes P_S \otimes P_T(s_1, s_2, t) = 0,$$

then there exists a measurable function $g : T \to \mathbb{R}$, such that

$$P_S \otimes P_T\Big((s, t) : f(s, t) = g(t)\Big) = 1.$$

Proof. By Tonelli's theorem we have

$$\int_T |f(s_1, t) - f(s_2, t)| \, dP_T(t) = 0,$$

for $P_S \otimes P_S$-almost all (s_1, s_2), i.e., for all $(s_1, s_2) \in N^c \in \mathcal{S} \otimes \mathcal{S}$, where $P_S \otimes P_S(N) = 0$. Furthermore, whenever $(s_1, s_2) \in N^c$, then $f(s_1, t) = f(s_2, t)$, for P_T-almost all t, i.e., for all $t \in M(s_1, s_2)^c \in \mathcal{T}$, with $P_T(M(s_1, s_2)) = 0$. For $s_1 \in S$, consider $N_{s_1} := \{s \in S : (s_1, s) \in N\}$, i.e., the s_1-section of N, and use Tonelli again, to see that there exists a P_S-null set $L \in \mathcal{S}$, such that $P_S(N_{s_1}) = 0$, for all $s_1 \in L^c$.

Next, fix $s_1 \in L^c$ and define the set

$$A := A(s_1) := \{(s, t) \in S \times T : s \in N_{s_1}^c, t \in M(s_1, s)^c\},$$

as well as the function $g(t) := f(s_1, t)$, for $t \in T$.[4] We therefore have $A \subseteq \{(s,t) : f(s_1,t) = f(s,t)\} = \{(s,t) : g(t) = f(s,t)\}$ and, for $s \in N_{s_1}^c$, $A_s = M(s_1, s)^c$ has P_T-probability one. To conclude, we use Tonelli again, to obtain

$$P_S \otimes P_T(A) = \int_S P_T(A_s)\, dP_S(s) = \int_{N_{s_1}^c} P_T(A_s)\, dP_S(s) = P_S(N_{s_1}^c) = 1.$$

□

Lukas Steinberger
Department of Statistics and OR
Data Science @ Uni Vienna
University of Vienna
Oskar-Morgenstern-Platz 1
1090 Vienna, Austria
E-mail: lukas.steinberger@univie.ac.at

Hannes Leeb
Department of Statistics and OR
Data Science @ Uni Vienna
University of Vienna
Oskar-Morgenstern-Platz 1
1090 Vienna, Austria
E-mail: hannes.leeb@univie.ac.at

[4] Note that by construction, the function g depends not only on f, but also on the null set L, and thus on both the probability spaces $(S, \mathcal{S}, P_S)$ and $(T, \mathcal{T}, P_T)$.