A Coverage Theory for Least Squares[1]

Algo Carè

Institute for Computer Science and Control (MTA SZTAKI), Budapest, Hungary.

(e-mail: algocare@gmail.com*)*

Simone Garatti

Politecnico di Milano, Italia.

(e-mail: simone.garatti@polimi.it*)*

Marco C. Campi

University of Brescia, Italy.

(e-mail: marco.campi@unibs.it*)*

Abstract. A sensible use of an estimation method requires that assessment criteria for the quality of the estimate be available. In this paper we present a coverage theory for the least squares estimate. By suitably modifying the empirical costs, one constructs statistics that are guaranteed to cover with known probability the cost associated with a next, still unseen, member of the population. All results of this paper are distribution-free and can be applied to least squares problems in use across a variety of fields.

Keywords: empirical distribution, coverage, statistics with distribution-free mean coverage, least squares, order statistics

[1] **This is the peer reviewed version of the following article: "A Coverage Theory for Least Squares, A. Car, S. Garatti, M.C. Campi, Journal of the Royal Statistical Society: Series B (Statistical Methodology)," which has been published in final form at [doi: 10.1111/rssb.12219]. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Self-Archiving.** The work of Algo Carè was supported by an Alain Bensoussan Fellowship of the European Research Consortium for Informatics and Mathematics (ERCIM), the Australian Research Council (ARC) under Discovery Grant DP130104028 and the University of Brescia and the Fondazione Cariplo and Regione Lombardia NOWERC project 2014-2256. The work of Simone Garatti was partly funded by The European Commission project UnCoVerCPS under grant number 643921 and by MIUR. The work of Marco C. Campi was partly funded by the University of Brescia H&W project CLAFITE and the Ministero dellIstruzione, dell'Università e della Ricerca.

1. Introduction

Given a sample of experimental observations (X_i, Y_i), $i = 1, \ldots, N$, where $X_i \in \mathbb{R}^{n \times d}$ and $Y_i \in \mathbb{R}^n$ (see the example given a few lines below to clarify the reasons why we consider a multidimensional Y_i and, correspondingly, a matrix structure for X_i), the least squares method consists in minimizing

$$\sum_{i=1}^{N} \|Y_i - X_i\beta\|^2, \tag{1}$$

with respect to the decision variable $\beta \in \mathbb{R}^d$, where $\|\cdot\|$ is Euclidean norm. The minimizer[2] is denoted by $\hat{\beta}_N$, and is called the *least squares estimate* or the *least squares solution*. Depending on the context, the least squares method has various interpretations that range from β being a parameter used to tune a descriptive model to β being a decision variable in a design process. An example of this second set-up is in order.

EXAMPLE 1 (SERVICE LOCATION). *Each member of a population is described by a two-dimensional vector p, which gives the position where the person lives, and a number $\rho \in [0,1]$ which assigns the person's rate of use of a given service (e.g. public laundry, post office, etc.). We are interested in determining a suitable location β to position the service. To this purpose, a sample of N members of the population is interviewed and their values (p_i, ρ_i), $i = 1, \ldots, N$, are recorded. The service location is then determined by minimizing $\sum_{i=1}^{N} \|\rho_i(p_i - \beta)\|^2$, which is the sum of squared home-service distances weighted by the rate of use of the service. This problem can be rewritten in the form* (1) *with $X_i = \rho_i \cdot I \in \mathbb{R}^{2\times 2}$, and $Y_i = \rho_i p_i \in \mathbb{R}^2$. Notice that a multidimensional Y_i and, correspondingly, a matrix structure for X_i turn up naturally in the formulation of this problem.* $\star$

The least squares method has become a standard in many applied fields that range from data-based and stochastic optimization, to robust filter design, system identification, and adaptive control. Irrespective of the application at hand, assessing the performance of $\hat{\beta}_N$ prior to its use is an important step to validate the solution, and the performance assessment of $\hat{\beta}_N$ is the subject of this paper.

Throughout, we assume that (X_i, Y_i), $i = 1, \ldots, N$, is an independent and identically distributed sample from a distribution F. For short, we shall denote the data set by D^N, namely, $D^N = \{(X_1, Y_1), \ldots, (X_N, Y_N)\}$. For a new pair (X, Y), define the *least squares cost* (or more briefly the *cost*) of (X, Y) as

$$\mathbf{q} := \|Y - X\hat{\beta}_N\|^2.$$

As (X, Y) varies according to F independently of D^N, the conditional distribution of $\mathbf{q}$ given $\hat{\beta}_N$ describes the cost paid by the population of (X, Y) corresponding to $\hat{\beta}_N$, and its knowledge may be used to support decisions of various type. For instance, in the service location problem of Example 1

[2] If the minimizer is not unique, the solution is determined by a tie-break rule.

knowing the distribution of $\mathbf{q}$ may support decisions on the service facility equipment that has to be acquired to dispatch goods, or even on whether one single service facility is not enough to serve the territory and two facilities should be built instead. However, computing the conditional distribution of $\mathbf{q}$ given $\hat{\beta}_N$ demands that one knows F, which is unrealistic in practice. Hence, for a practical performance assessment one aims at constructing descriptors of the conditional distribution of $\mathbf{q}$ that are based on the experimental data set D^N.[3]

One simple descriptor is the empirical mean $\frac{1}{N}\sum_{i=1}^{N}\|Y_i - X_i\hat{\beta}_N\|^2$. This is an estimator of $\mathbb{E}[\mathbf{q}|\hat{\beta}_N]$, the conditional mean of $\mathbf{q}$ given $\hat{\beta}_N$, and it has received much attention in the literature. Classic results characterize the deviation of $\frac{1}{N}\sum_{i=1}^{N}\|Y_i - X_i\hat{\beta}_N\|^2$ from $\mathbb{E}[\mathbf{q}|\hat{\beta}_N]$ when $N \to \infty$ (asymptotic results), Lehmann and Casella (1998), while more recent work based on the statistical learning theory extends these results to when N is finite, Vapnik and Chervonenkis (1971); Vapnik (1996).

1.1. *Goal of this paper: least squares cost coverages*

In this paper, we consider a more structured characterization of the least squares cost than its mean. The goal is to determine statistics $\mathbf{c}$ of the data set D^N that are threshold values for $\mathbf{q}$ with given probabilistic guarantees. In other words, referring to Fig. 1, the attention is shifted from quantifying the deviation of $\frac{1}{N}\sum_{i=1}^{N}\|Y_i - X_i\hat{\beta}_N\|^2$ from $\mathbb{E}[\mathbf{q}|\hat{\beta}_N]$, as in Fig. 1(a), to quantifying the probability that $\|Y - X\hat{\beta}_N\|^2$ falls in the bold segment below $\mathbf{c}$, as in Fig. 1(b). In this context, we want to establish rigorous results that hold for any finite N.

We start with the following definition of coverage and mean coverage.

Definition 1 (**coverage and mean coverage**). *Given a statistic* $\mathbf{c}$ *of the data set* D^N *and a pair* (X,Y) *distributed according to* F *and independent of* D^N*, the* coverage *of* $\mathbf{q} = \|Y - X\hat{\beta}_N\|^2$ *by* $(-\infty, \mathbf{c}]$ *is defined as*

$$\mathbb{P}\{\mathbf{q} \le \mathbf{c}|D^N\}; \tag{2}$$

the mean coverage *of* $\mathbf{q}$ *by* $(-\infty, \mathbf{c}]$ *is*

$$\mathbb{E}[\mathbb{P}\{\mathbf{q} \le \mathbf{c}|D^N\}] = \mathbb{P}\{\mathbf{q} \le \mathbf{c}\}.$$

⋆

The term "coverage" comes from the literature on tolerance/prediction regions, Wilks (1941); Scheffé and Tukey (1947); Fraser and Guttman (1956); Vardeman (1992); Di Bucchianico et al. (2001); Li and Liu (2008); Lei et al. (2013); Frey (2013), which can be explained as follows. For a given D^N, $T(D^N) := \{(X,Y) : \mathbf{q} \le \mathbf{c}\}$ is a region in the space $\mathbb{R}^{n\times d} \times \mathbb{R}^n$ and the coverage of $\mathbf{q}$ by

[3] In our study, the (X_i, Y_i) values are random draws from a population, and, hence, regression problems where the X_i are deterministic values are left out from our analysis.

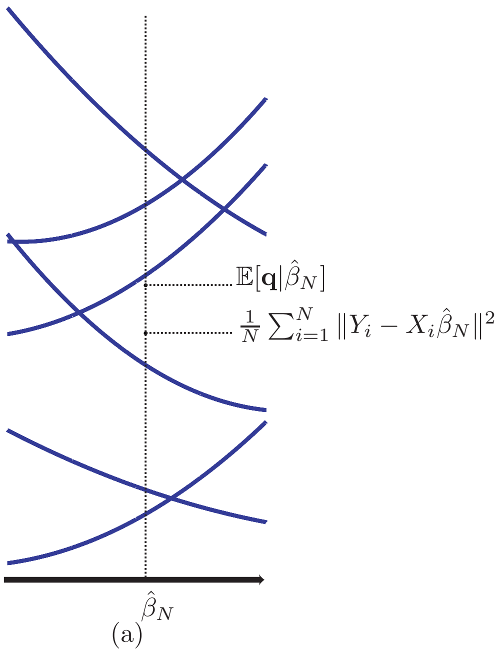


Figure 1. (a) empirical vs. conditional mean; (b) $\mathbf{c}$ statistic. The curved lines represent the squared residual functions $\|Y_i - X_i\hat{\beta}_N\|^2$.

$(-\infty, \mathbf{c}]$ is the coverage of (X, Y) by the region $T(D^N)$ in the sense of the above literature since $\mathbb{P}\{\mathbf{q} \leq \mathbf{c}|D^N\} = \int_{\{(X,Y)\}} \mathbb{1}\left(T(D^N)\right)\mathrm{d}F$ (here, $\mathbb{1}(\cdot)$ denotes indicator function).

For a given D^N, $\mathbf{c}$ is the quantile of the distribution of $\mathbf{q}$ corresponding to the probability value given by the coverage. That is, if e.g. the coverage is 90%, for the $\hat{\beta}_N$ and $\mathbf{c}$ given by the seen D^N, the probability mass of the (X, Y) pairs such that $\|Y - X\hat{\beta}_N\|^2 \leq \mathbf{c}$ is 90%.

The coverage of $\mathbf{q}$ by $(-\infty, \mathbf{c}]$ depends on D^N and is a random variable. The mean coverage is its expected value. The mean coverage is also equal to $\mathbb{P}\{\mathbf{q} \leq \mathbf{c}\}$, i.e., it is the total probability of seeing a random sample D^N, constructing $\mathbf{c}$, and then extracting one more instance of (X, Y) that incurs a cost smaller than or equal to $\mathbf{c}$. In an application with sequential observations, the mean coverage is the limit of the frequency with which the $(N + 1)$th observation incurs a cost less than or equal to the statistic $\mathbf{c}$ computed from the previous N observations when the observation window shifts along the time axis. See Section 3.1 for an example.

In this paper, our goal is to find statistics $\mathbf{c}$ that have a guaranteed mean coverage irrespective of the (unknown) distribution F. The statistics we will introduce have the additional property of being asymptotically tight in a precise sense specified later. Instead, we do not enter the theoretical study of the coverage, which exhibits difficulties that go beyond the analysis presented in this paper.

The following definition is in order.

DEFINITION 2 (**distribution-free mean coverage**). *Interval* $(-\infty, \mathbf{c}]$ *has a distribution-free* mean coverage τ *if*

$$\mathbb{P}\{\mathbf{q} \leq \mathbf{c}\} \geq \tau$$

holds for all distributions F. ⋆

One natural approach to follow when one seeks distribution-free statistics is to look at the squared residuals corresponding to $\hat{\beta}_N$

$$\mathbf{q}_i := \|Y_i - X_i \hat{\beta}_N\|^2, \quad i = 1, \ldots, N.$$

These $\mathbf{q}_i$'s are called *empirical costs*. Intuitively, the empirical costs carry information on how $\mathbf{q}$ distributes for the seen data set D^N. Note that the real line is split by the N empirical costs $\mathbf{q}_i$ in $N + 1$ intervals, and one might expect that each of these intervals carries on average a probability of $1/(N + 1)$ of containing $\mathbf{q}$. This is indeed what happens in a simplified context where N points are independently drawn on the real line and then ordered (order statistics). We briefly digress to describe this situation because it is useful for future comparison.

Consider a univariate independent random sample $r_i \in \mathbb{R}$, $i = 1, \ldots, N$, from a distribution F_r, and let $r_{(1)}, r_{(2)}, \ldots, r_{(N)}$ be the *order statistics* of the r_i's, that is, $r_{(1)} \leq r_{(2)} \leq \cdots \leq r_{(N)}$.[4] Then, the following well known result holds, see e.g. David and Nagaraja (2003).

PROPOSITION 1 (**order statistics**). *Let* r *be a new value sampled from* F_r *independently of* $r_1, r_2 \ldots, r_N$. *Then,*

$$\mathbb{P}\{r \leq r_{(i)}\} \geq \frac{i}{N+1}, \quad i = 1, \ldots, N,$$

i.e., $(-\infty, r_{(i)}]$ *has a distribution-free* mean coverage $\frac{i}{N+1}$. ⋆

This result holds with equality, i.e., $\mathbb{P}\{r \leq r_{(i)}\} = \frac{i}{N+1}$, for continuous distributions F_r.

In the context of least squares optimization of this paper, however, there is an extra element, which makes order statistics not applicable. This is that the empirical costs $\mathbf{q}_i$ are computed on a real line that originates from $\hat{\beta}_N$. Since $\hat{\beta}_N$ minimizes the squared residuals, this line is data-dependent and a bias arises so that the mean coverage of $(-\infty, \mathbf{q}_{(i)}]$ is in general less than $\frac{i}{N+1}$. A simple example in Appendix A illustrates this fact. This situation is similar to what happens in post-selection inference. It was noted as early as in the 1960s by Buehler and Fedderson (1963) and Brown (1967) that performing data-based model selection and deriving statistical inference from the selected model as

[4] Throughout this paper, for any collection of N real numbers, $a_1, a_2, \ldots, a_N$, notation $a_{(1)}, a_{(2)}, \ldots, a_{(N)}$ denotes the a_i's in ascending order.

though the model were deterministically assigned leads to invalid results, see also Pötscher (1991) and Benjamini and Yekutieli (2005) for more recent discussions. This problem has attracted much interest in recent years, in particular Berk et al. (2013) proposes to perform simultaneous inference to restore validity and Belloni et al. (2015), Tibshirani et al. (2016), Lee et al. (2016), and Belloni et al. (2014) derive valid statistical inference in various contexts that include Least Absolute Deviation regression (LAD), Forward Stepwise Regression (FS), Least Angle Regression (LAR), the Lasso, and quantile regression. In the present paper, the focus is different from that of the aforementioned contributions since we do not make variable selection and are interested in studying the mean coverage of the cost for a fixed structure. In this context, order statistics that are valid for a deterministic real line lose their validity and our goal is that of providing valid inference as explained in the following.

1.2. Main results of this paper

We construct statistics $\bar{\mathbf{q}}_1, \bar{\mathbf{q}}_2, \ldots, \bar{\mathbf{q}}_N$ such that each interval $(-\infty, \bar{\mathbf{q}}_{(i)}]$ has a distribution-free mean coverage $\frac{i}{N+1}$ (Theorem 1). The statistics $\bar{\mathbf{q}}_1, \bar{\mathbf{q}}_2, \ldots, \bar{\mathbf{q}}_N$ are obtained by adding a data-dependent margin to the empirical costs $\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_N$. Under mild assumptions, the margin is shown to tend to zero as N grows to infinity (Theorem 2), which shows that the statistics are asymptotically tight. Moreover, even with moderate data sets the margin turns out to be small enough to be practically useful. In the context of Proposition 1, the margin is zero and the result for order statistics is recovered as a particular case (Example 3). Moreover, the fact that the findings of this paper are distribution-free is key to their applicability since assuming that the distribution F is available is unrealistic in most applications. These results are theoretically proved, and demonstrated through simulation experiments in this paper.

The intuitive idea behind the derivation of the statistics $\bar{\mathbf{q}}_i$ is as follows. A prediction set $S(D^N)$ in the (X, Y) domain with distribution-free mean coverage $\mathbb{P}\{(X, Y) \in S(D^N)\} = \frac{i}{N+1}$ is first constructed. Then, $\bar{\mathbf{q}}_{(i)}$ is obtained by upper bounding the sup of the costs associated to the pairs (X, Y) that belong to $S(D^N)$, so that $(-\infty, \bar{\mathbf{q}}_{(i)}]$ carries a guaranteed mean coverage.

The prediction set $S(D^N)$ is constructed by resorting to so-called "conformal prediction", Vovk (2004); Vovk et al. (2005); Gammerman and Vovk (2007); Shafer and Vovk (2008); Lei et al. (2013); Lei and Wasserman (2014). The creators of this approach had the brilliant intuition that, in an exchangeable framework, discarding the pairs (X, Y) that turn out to be the less conformal in the set of N observations D^N augmented by (X, Y) generates regions with an a-priori known probability of containing a future observation. This approach was mainly motivated by prediction purposes, namely, the problem of finding a region in the (X, Y) domain for (X_{N+1}, Y_{N+1}). We construct a prediction set in exactly the same sense as that considered in the above references. In our paper, however, we make this construction as an intermediate step towards our final goal, which is different from that of predicting (X_{N+1}, Y_{N+1}). Our final goal is to derive rigorous and useful statistics to evaluate the performance of the least squares design for when the design is applied to a new member of the population. In this regard, it is important to note that the choice of the conformity measure has a large impact on the shape of the prediction set, which, in turn, affects the quality of the statistics on the cost. One first contribution of this paper is that of introducing a suitable conformity measure that is geared towards the achievement of tight statistics $\bar{\mathbf{q}}_{(i)}$. This is obtained by making the prediction set adhere to the part of the (X, Y) domain that has low cost corresponding to $\hat{\beta}_N$. This property is not met by other conformity measures available in the literature, and the interested reader is referred to the on-line supplementary material, Section 1, for a comparative example.

A second contribution of the paper is that of providing an explicit and easy-to-compute formula to evaluate the statistics $\bar{\mathbf{q}}_{(i)}$. This is important because an explicit computation of the sup of the cost over a highly complex prediction set is in general difficult to perform, and simply defining the statistics as the sup would leave the computational burden to the end user, resulting in an impractical approach. The easy-to-compute formula is carefully derived to reduce conservatism, as shown by asymptotic theorems and simulation examples.

1.3. Other related literature

An early study that is related to the subject matter of this paper is Saw et al. (1984, 1988). In these contributions, the authors have derived data-dependent Chebyshev inequalities that can be used in a scalar set-up corresponding in our notation to $X = 1$ and $Y \in \mathbb{R}$ to build statistics with distribution-free mean coverage. Applications of this result are found in various contexts among which UCBs (Upper Confidence Bounds) methods, Xu and Nelson (2013), neural curve tuning, Etzold and Eurich (2005), distance concentration, Kabán (2012), model reliability for train station parking errors, Chen and Gao (2012), and testing procedures, Beasley et al. (2004). However, the papers Saw et al. (1984, 1988) deal only with a scalar set-up and, most notably, due to their nature and scope, the statistics there obtained depend on the data sample only through the sample mean and variance. Hence, information that is valuable for our purpose of characterizing $\mathbf{q}$ remains unexploited.

Tight results on distribution-free mean coverages have been previously obtained by the authors of this paper in a different set-up, that of worst-case convex optimization, Calafiore and Campi (2005). Moreover, in Campi and Garatti (2008, 2011); Carè et al. (2015); Campi and Garatti (2016) the results of Calafiore and Campi (2005) have been strengthened by also computing the distribution of the coverage, which is important for determining confidence regions for the coverage values. All these studies hinge crucially upon the concept of support constraint, Calafiore and Campi (2005), a concept which does not carry over to the set-up of the present paper of least squares optimization. In actual facts, this is the very reason why the fundamental least squares method has so far not been object of consideration in our studies.

1.4. Structure of the paper

All main results of the paper are provided and discussed in Section 2. Section 3 contains numerical examples, while all the technical proofs are in the appendices.

The data and software code used in the numerical examples can be downloaded from the URL: `http://home.deib.polimi.it/sgaratti/coverageLS.htm`.

1.5. Notation

For a matrix M:

- M^T denotes the transpose of M;
- $M^\dagger$ denotes Moore-Penrose generalized inverse of M;
- $\|M\|$ is the spectral norm, i.e., $\|M\| = \sup_{\|x\|=1} \|Mx\|$, where the norm in the right-hand side is Euclidean norm;
- $\lambda_{\max}(M)$ denotes maximum eigenvalue of M;
- for a symmetric M, $M \succ 0$ ($M \succeq 0$) means that M is positive definite (semi-definite). For a pair of symmetric matrices M and N, $M \succ N$ ($M \succeq N$) means that $M - N$ is positive definite (semi-definite).

2. Statistics with Distribution-Free Mean Coverage

For convenience, the squared residuals are henceforth written as $\|Y_i - X_i\beta\|^2 = (\beta - v_i)^T K_i (\beta - v_i) + h_i$, where $K_i = X_i^T X_i$, $v_i = X_i^\dagger Y_i$, $h_i = \|Y_i - X_i v_i\|^2$. Note that $K_i \succeq 0$, but K_i can as well be singular, as is e.g. in regression problems with scalar Y where $K_i = X_i^T X_i$ has rank 1.

When $\sum_{\substack{\ell=1 \\ \ell\neq i}}^N K_\ell \succ 0$, let $\bar{K}_i := K_i + 6K_i \left(\sum_{\substack{\ell=1 \\ \ell\neq i}}^N K_\ell \right)^{-1} K_i$. The modified empirical costs $\bar{\mathbf{q}}_i$ are then defined as follows

$$\bar{\mathbf{q}}_i := \begin{cases} (\hat{\beta}_N - v_i)^T \bar{K}_i (\hat{\beta}_N - v_i) + h_i, & \text{if } K_i \prec \frac{1}{6} \sum_{\substack{\ell=1 \\ \ell\neq i}}^N K_\ell \\ +\infty, & \text{otherwise.} \end{cases} \tag{3}$$

The next Theorem 1, which asserts that $(-\infty, \bar{\mathbf{q}}_{(i)}]$ has a distribution-free mean coverage $\frac{i}{N+1}$, is the main result of our study.

THEOREM 1 (**distribution-free mean coverage**). *Relation*

$$\mathbb{P}\{\mathbf{q} \leq \bar{\mathbf{q}}_{(i)}\} \geq \frac{i}{N+1}, \quad i = 1, \dots, N, \tag{4}$$

holds for any probability distribution F. ⋆

The technical proof of this theorem is differed to Appendix B. We now concentrate on discussing the meaning and importance of Theorem 1.

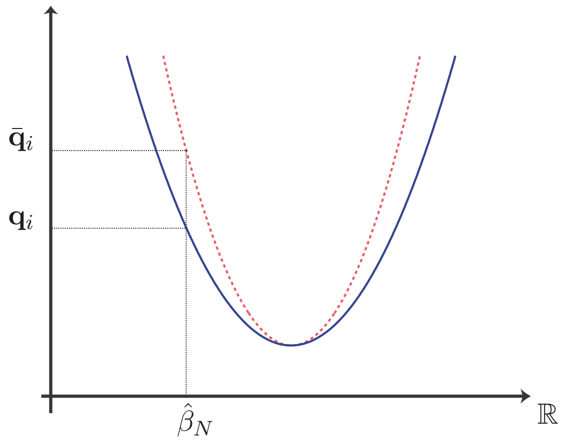


Figure 2. The paraboloid $(\beta - v_i)^T K_i(\beta - v_i) + h_i$ (continuous line) vs. the paraboloid $(\beta - v_i)^T \bar{K}_i(\beta - v_i) + h_i$ with increased curvature (dashed line). Their values at $\beta = \hat{\beta}_N$ are the empirical cost $\mathbf{q}_i$ and the modified empirical cost $\bar{\mathbf{q}}_i$, respectively.

a. [**geometric interpretation and intuitive explanation**]
$\bar{\mathbf{q}}_i$ has a nice geometric interpretation. The empirical cost $\mathbf{q}_i$ is the value of the paraboloid $(\beta - v_i)^T K_i(\beta - v_i) + h_i$ at $\beta = \hat{\beta}_N$. Instead, the modified empirical cost $\bar{\mathbf{q}}_i$ is obtained as the value at $\beta = \hat{\beta}_N$ of a paraboloid with increased curvature obtained by replacing K_i with $\bar{K}_i$, see Fig. 2.

The margin $\bar{\mathbf{q}}_i - \mathbf{q}_i$ is given by

$$(\hat{\beta}_N - v_i)^T \left(6K_i \left(\sum_{\substack{\ell=1 \\ \ell \neq i}}^{N} K_\ell \right)^{-1} K_i \right) (\hat{\beta}_N - v_i), \tag{5}$$

and it depends on the ratio of K_i to $\sum_{\substack{\ell=1 \\ \ell \neq i}}^{N} K_\ell$. If K_i is small compared to $\sum_{\substack{\ell=1 \\ \ell \neq i}}^{N} K_\ell$, then $\bar{\mathbf{q}}_i \approx \mathbf{q}_i$, which is normally the case except for moderate data sets (small N). In the opposite, when K_i is not small compared to $\sum_{\substack{\ell=1 \\ \ell \neq i}}^{N} K_\ell$, the margin can be larger. The intuitive reason for this is as follows. Corresponding to $\hat{\beta}_N$, the empirical costs are on average biased towards smaller values than the distribution of costs for the whole population. This is because the least squares estimate $\hat{\beta}_N$ is chosen in the point where the sum of the squared empirical costs is minimized. This biasing effect is larger for some empirical costs than for others. Suppose that one K_i is pretty large as compared to the others, to the point that K_i is even bigger than $\sum_{\substack{\ell=1\\ \ell\neq i}}^N K_\ell$. Then, the i-th data point plays an important role in determining the solution since the paraboloid $(\beta - v_i)^T K_i (\beta - v_i) + h_i$ has a strong "attraction effect" as compared to the attraction effect of other data points. As a consequence, the bias towards smaller values is more significant for the i-th paraboloid than for other paraboloids, which requires a bigger margin to compensate for this effect. In normal circumstances, one cannot expect that the situation is as sharp as in the hypothetical case used in the explanation above, and the margin given in (5) plays a subtle role in making the statistics valid in all conditions.

b. [the role of dimension *d*]

We have already observed that in regression problems with scalar Y, the matrix $K_i = X_i^T X_i$ has rank 1. This means that the paraboloid $(\beta - v_i)^T K_i (\beta - v_i) + h_i$ associated with a pair (X_i, Y_i) is flat in $d-1$ orthogonal directions, and (X_i, Y_i) does not influence the solution $\hat{\beta}_N$ but in one direction only. As a consequence, at least d observations are required for $\bar{\mathbf{q}}_i$ to be finite, and, moreover, the margin decreases roughly as d/N. This behavior has connections with the notion of overfitting in statistical learning. In other problems, on the other hand, the importance of d is toned down. This happens for example when all the K_i's are identity matrices, in which case a single pair (X_i, Y_i) impacts on all the d directions simultaneously, see Example 2 at the end of this section for one example of this situation.

c. [convergence of margin to zero]

In applications, it is almost the rule that $\sum_{\substack{\ell=1\\ \ell\neq i}}^N K_\ell$ grows faster than the largest of the K_i's. Then, the term $6K_i \left(\sum_{\substack{\ell=1\\ \ell\neq i}}^N K_\ell \right)^{-1} K_i$ in the definition of $\bar{K}_i$ vanishes as N grows, yielding $\bar{K}_i \to K_i$, and, hence, the margin tends to zero. This idea is formalized in the next Theorem 2, where it is shown that each margin $\bar{\mathbf{q}}_i - \mathbf{q}_i$ goes to zero as $N \to \infty$ provided that the distributions F is thin-tailed.

THEOREM 2 (**convergence**). *Assume that*

$$\mathbb{E}[K_i] \succ 0, \tag{6}$$

$$\exists \alpha, \bar{\chi} > 0 \textit{ such that } \forall \chi > \bar{\chi} \quad \mathbb{P}\{\|K_i\| > \chi\} \leq e^{-\alpha\chi}, \tag{7}$$

$$\exists \gamma, \bar{\nu} > 0 \textit{ such that } \forall \nu > \bar{\nu} \quad \mathbb{P}\{\|v_i\| > \nu\} \leq e^{-\gamma\nu}. \tag{8}$$

Then,

$$\max_{i=1,\ldots,N} (\bar{\mathbf{q}}_i - \mathbf{q}_i) \xrightarrow[N\to\infty]{} 0 \qquad \textit{almost surely.}$$

⋆

The proof is in Appendix C.

d. [distribution-free nature of the result]

Theorem 1 is universal, that is, no assumptions on F are made. Assumptions limit the applicability of a method in two distinct respects. First, the method is not applicable if the assumptions are not satisfied. Second, even if the assumptions are satisfied, the user may not know whether they are or are not. Thus, its distribution-free nature is a fundamental point of strength of the analytical instruments introduced in this paper. On the other hand, distribution-free results may be conservative. Theorem 3 below shows that if $\bar{\mathbf{q}}_{(i)}$ is replaced by $\mathbf{q}_{(i)}$ in the statement of Theorem 1, then the result in equation (4) holds with a reversed inequality for all F's satisfying a mild non-concentration condition. Thus, any possible conservatism is in the margin $\bar{\mathbf{q}}_i - \mathbf{q}_i$, and, since this margin converges to zero under natural conditions (see Theorem 2), and it is reasonably small in applications even for moderate data sets (see e.g. the examples at the end of this section and those in Section 3), the conclusion follows that the conservatism due to the distribution-free nature of the results is mild in the context of study of this paper.

THEOREM 3 (**upper bound on the mean coverage of** $(-\infty, \mathbf{q}_{(i)}]$). *Suppose that*

$$\mathbb{P}\{\|Y - X\beta\|^2 = \lambda\} = 0$$

holds for any $(\beta, \lambda) \in \mathbb{R}^d \times \mathbb{R}$. *Then,*

$$\mathbb{P}\{\mathbf{q} \leq \mathbf{q}_{(i)}\} \leq \frac{i}{N+1}, \quad i = 1, \dots, N.$$

⋆

The proof of Theorem 3 is in Appendix D.

e. [order selection in regression problems]

The findings of this paper are potentially useful for the problem of order selection in regression problems. A full development of a selection methodology, however, calls for extra knowledge that is not available at the present time, and we here briefly discuss this topic, which may serve as a stimulus for further research. Our Theorem 1 establishes a tight distribution-free mean coverage result. When, various model structures are considered, one can compare the modified empirical costs $\bar{\mathbf{q}}_{(i)}$ that, in different structures, attain the same mean coverage, and choose the structure with lowest $\bar{\mathbf{q}}_{(i)}$. This allows one to obtain a suitable trade-off between selecting a low-order model (which gives a large $\mathbf{q}_{(i)}$) and a high order model where the decrease of $\mathbf{q}_{(i)}$ is balanced by an increase of the margin $\bar{\mathbf{q}}_{(i)} - \mathbf{q}_{(i)}$ (see point b.). It is of interest to note that, for this procedure to stand on solid theoretical grounds, the coverage must also have a low variance for, otherwise, one runs into the risk of selecting a structure with guaranteed mean coverage but with significantly lower coverage for the sample at hand. While our empirical experience shows that the variance is indeed small in various contexts, see Section 3 for an example, at present no theoretical result on the variance is available. The authors of this paper believe that establishing a result in this direction would open important new avenues for model order selection.

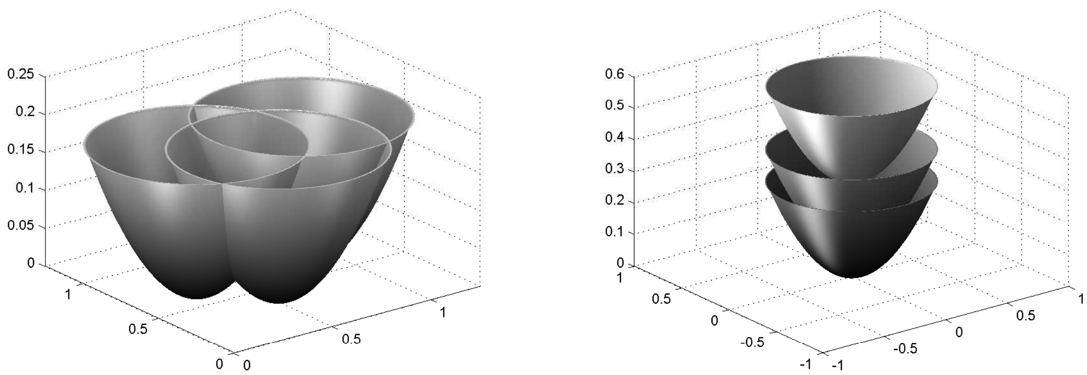
We end this section with two simple examples that further illustrate facts and results that we have discussed in this section. More complex numerical and empirical examples are given in the next Section 3.

EXAMPLE 2 (**paraboloids with coplanar vertexes**). *Suppose that* $n = d = 2$*,* $X = I$*, and* Y *is a random variable uniformly distributed in* $[0,1]^2$*. Then,* $K_i = I$*,* $v_i = Y_i$*,* $h_i = 0$*. Some of the cost functions* $\|Y_i - \beta\|^2$ *are shown in Fig. 3(a).*

In this case, $K_i \prec \frac{1}{6}\sum_{\substack{\ell=1\\ \ell\neq i}}^{N} K_\ell \Leftrightarrow N \geq 8$*, and*

$$\bar{K}_i = I + \frac{6}{N-1} I,$$
$$\bar{\mathbf{q}}_i = \mathbf{q}_i + \frac{6}{N-1} \mathbf{q}_i,$$

for $N \geq 8$*. Here* $\mathbf{q}_i$ *is upper bounded, in fact* $\mathbf{q}_i \leq \max_{\beta, Y_i \in [0,1]^2} \|Y_i - \beta\|^2 = 2$*, so that in this case the margin* $\bar{\mathbf{q}}_i - \mathbf{q}_i = \frac{6}{N-1}\mathbf{q}_i \leq \frac{12}{N-1}$ *goes to zero as* $1/N$*. See Fig. 4 for a graph depicting* $\max_{i=1,\ldots,N}(\bar{\mathbf{q}}_i - \mathbf{q}_i)$ *as a function of* N *in a simulated experiment.* ⋆



(a) (b)

Figure 3. The cost functions $\|Y_i - X_i\beta\|^2$ of Example 2 (left) and of Example 3 (right).

EXAMPLE 3 (**stack of paraboloids**). *Suppose that* X *and* Y *have the following structure*

$$X = \begin{bmatrix} I_{2\times 2} \\ 0_{1\times 2} \end{bmatrix}, \quad Y = \begin{bmatrix} 0_{2\times 1} \\ \alpha \end{bmatrix},$$

where subscript denotes the matrix dimension (e.g. $0_{1\times 2}$ *is a row vector with two zeros) and* α *is a random variable uniformly distributed in* $[0,1]$*. In this case, we have that* $K_i = I_{2\times 2}$*,* $v_i = 0_{2\times 1}$*,*

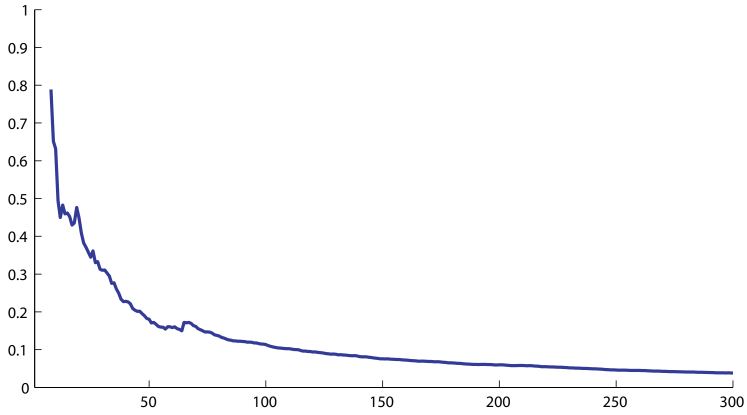


Figure 4. The largest margin $\max_{i=1,\dots,N}(\bar{\mathbf{q}}_i - \mathbf{q}_i)$ of Example 2 as a function of N.

$h_i = \alpha_i^2$. Some of the cost functions $\|\beta\|^2 + \alpha_i^2$ are depicted in Fig. 3(b).

Since all paraboloids have vertex in zero, it turns out that $\hat{\beta}_N = 0$ and $\bar{\mathbf{q}}_i = \mathbf{q}_i = \alpha_i^2$, $i = 1, \dots, N$, that is, the margin is zero in this case. As before, $K_i \prec \frac{1}{6}\sum_{\ell \neq i} K_\ell \Leftrightarrow N \geq 8$, and, for $N \geq 8$, Theorem 1 gives

$$\mathbb{P}\left\{\mathbf{q} \leq \bar{\mathbf{q}}_{(i)} = \mathbf{q}_{(i)}\right\} \geq \frac{i}{N+1}.$$

Interestingly enough, this is the same result that is obtained by applying Proposition 1 to $\mathbf{q}_i = \alpha_i^2$, that is, the order statistics result is recovered from the distribution-free Theorem 1. ⋆

3. Numerical Examples

Two examples are provided. The first example refers to stock prices. The second example aims at providing more intuition on certain concentration properties of the coverages.

3.1. An example in stock prices

We consider a data set of stock prices taken from the Bilkent University Function Approximation Repository (`http://funapp.cs.bilkent.edu.tr/DataSets/`). This is a public repository for "training and demonstration by machine learning and statistics community". The stock prices refer to ten aerospace companies, and were daily collected from January 1988 to October 1991. The whole data set can be represented by a 10×950 matrix $P = [P_{k,i}]$, whose column P_i contains the stock prices (in US dollars) for the ten companies at day i. Figure 5 profiles the trend of $P_{k,i}$ as a function of the day i, for $k = 1, 2, \dots, 10$.

Stock prices can be modeled as geometric Brownian motions, which, in discrete time, is written as

$$P_{k,i+1} = P_{k,i} + \mu_k P_{k,i} + \sigma_{k,i} P_{k,i},$$

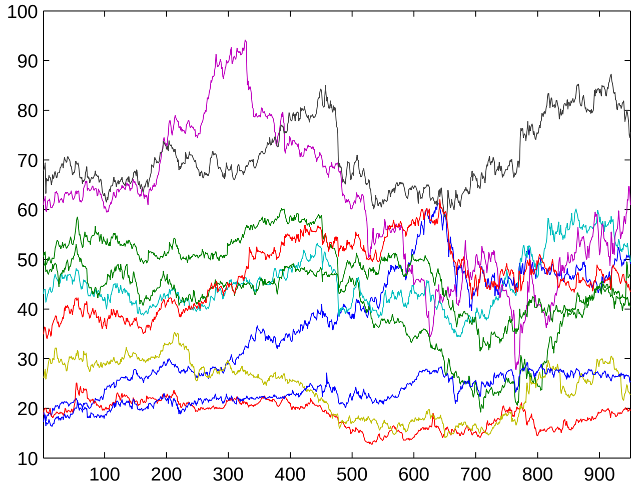


Figure 5. Stock prices from January 1988 to October 1991.

where μ_k is the percentage drift for the kth stock price, and $\sigma_{k,i}$ is a zero mean independent stochastic process that represents the percentage price volatility, see e.g. Hull (2009), Section 13.3. Letting

$$L_{k,i} = \frac{P_{k,i+1} - P_{k,i}}{P_{k,i}} = \mu_k + \sigma_{k,i}$$

be the rate-of-return of company k at day i, it is of interest to estimate $\mu = [\mu_1\ \mu_2\ \cdots\ \mu_{10}]^T$, the vector of percentage drift, but also to collect knowledge on the dispersion of the random variable $L_i = [L_{1,i}\ L_{2,i}\ \cdots\ L_{10,i}]^T$.

In practice, μ and the probability distribution of $\sigma_i = [\sigma_{1,i}\ \sigma_{2,i}\ \cdots\ \sigma_{10,i}]^T$ are time-varying. However, they can be considered constant over short time windows. In the following, estimation is performed over a moving window of 19 days, which is short enough for the approximation that μ and the probability distribution of σ_i are constant to approximately hold. The value μ_τ of μ at the τth time window is estimated by solving the least squares problem

$$\hat{\beta}_{19} = \arg\min_{\beta} \sum_{i=1}^{19} \|\beta - L_{\tau-1+i}\|^2.$$

In this context, $\mathbf{q} = \|\hat{\beta}_{19} - L_{\tau+19}\|^2$ is a synthetic scalar index – that is, the norm reduces the 10-dimensional vector of dispersions to a single real value – of how $L_{\tau+19}$ distributes around the estimate $\hat{\beta}_{19}$. It carries important information on the volatility of prices, and can be used by investors and governing bodies for decision making.

In the present setup, we have that the empirical costs

$$\mathbf{q}_i = \|\hat{\beta}_{19} - L_{\tau-1+i}\|^2, \quad i = 1, \ldots, 19$$

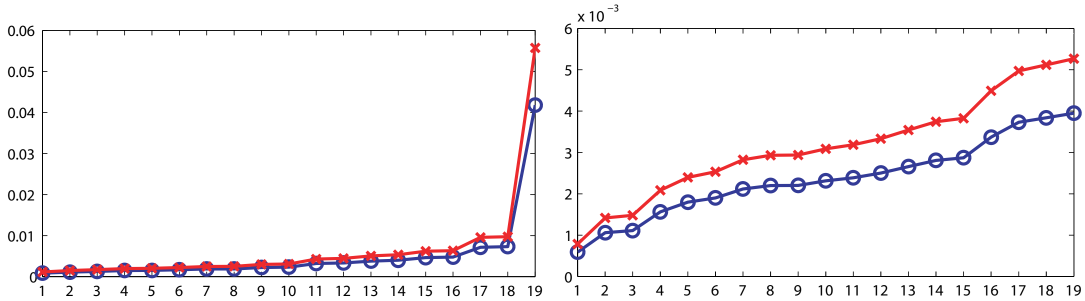


Figure 6. $\mathbf{q}_{(i)}$ (circles) vs. $\bar{\mathbf{q}}_{(i)}$ (crosses) in the first (left) and in the last (right) time window.

Table 1. Empirical frequencies with which $\mathbf{q} \leq \bar{\mathbf{q}}_{(i)}$; $N = 19$.

i	4	8	12	16
Estimate of $\mathbb{P}\{\mathbf{q} \leq \bar{\mathbf{q}}_{(i)}\}$	0.28	0.51	0.69	0.85
$\frac{i}{20}$	0.2	0.4	0.6	0.8

correspond to the volatility observed over the time window, and the modified empirical costs are

$$\bar{\mathbf{q}}_i = \left(1 + \frac{1}{3}\right) \cdot \mathbf{q}_i, \quad i = 1, \ldots, 19.$$

Examples of the ordered $\mathbf{q}_{(i)}$ and $\bar{\mathbf{q}}_{(i)}$ are given in Fig. 6.

According to Theorem 1, the statistic $\bar{\mathbf{q}}_{(i)}$ has a mean coverage no smaller than $i/20$. Thus, as the time window slides along the time axis, relation $\mathbf{q} \leq \bar{\mathbf{q}}_{(i)}$ holds with a frequency at least of $i/20$. This property has been experimentally verified and the results for $i = 4, 8, 12, 16$ are reported in Table 3.1. Interestingly, $\mathbf{q}_{(i)}$ gave instead empirical frequencies that were systematically below $i/20$. With longer time windows the empirical results get closer to the theoretical evaluations, a fact that is in line with Theorem 2. As an example, Table 3.1 gives the results for a window of length 39.

3.2. A simulation example that describes the distribution of the coverages

In this second example, we investigate through simulation how the coverage of $(-\infty, \bar{\mathbf{q}}_{(i)}]$ distributes around its mean. When the distribution is peaked, the mean coverage approximates the coverage for the given data set.

Take $n = 1$, $d = 20$, and suppose that X is a random direction in $\mathbb{R}^{20}$ and Y is the scalar product

Table 2. Empirical frequencies with which $\mathbf{q} \leq \bar{\mathbf{q}}_{(i)}$; $N = 39$.

i	8	16	24	32
Estimate of $\mathbb{P}\{\mathbf{q} \leq \bar{\mathbf{q}}_{(i)}\}$	0.24	0.44	0.64	0.82
$\frac{i}{40}$	0.2	0.4	0.6	0.8

between X and a random Gaussian vector as follows

$$X = \frac{u^T}{\|u\|} \quad Y = Xv,$$

where $u, v \in \mathbb{R}^{20}$ are independent vectors both drawn according to a 20-variate normal density with identity covariance matrix and zero mean.[5] For a given data set $D^N = \{(X_1, Y_1), \dots, (X_N, Y_N)\}$, the coverage $\mathbb{P}\{\mathbf{q} \leq \bar{\mathbf{q}}_{(i)} | D^N\}$ is a number indicating the probability level of the quantile $\bar{\mathbf{q}}_{(i)}$. On the other hand, as the data set D^N varies, the coverage of $(-\infty, \bar{\mathbf{q}}_{(i)}]$ changes, and we are interested in recording its variability. For concreteness, we let $N = 199$ and computed via Montecarlo methods the coverage of the statistic $\bar{\mathbf{q}}_{(0.8\cdot(N+1))}$ with distribution-free mean coverage 0.8 in 10000 repeated experiments.

In dark in Fig. 7(a) is the histogram of the coverage of $(-\infty, \bar{\mathbf{q}}_{(0.8\cdot(N+1))}]$. This histogram is rather concentrated around its mean, and the coverage of $(-\infty, \bar{\mathbf{q}}_{(0.8\cdot(N+1))}]$ is above 0.8 in most cases. For completeness, the histogram of the coverage of $(-\infty, \mathbf{q}_{(0.8\cdot(N+1))}]$ is also depicted in Fig. 7(a). This histogram shows values almost systematically smaller than 0.8. The fact that the mean of the coverage of $(-\infty, \mathbf{q}_{(0.8\cdot(N+1))}]$ is smaller than 0.8 follows from Theorem 3.

Further, Figs. 7(b) and 7(c) depict the histograms of the coverages of $(-\infty, \bar{\mathbf{q}}_{(0.8\cdot(N+1))}]$ and $(-\infty, \mathbf{q}_{(0.8\cdot(N+1))}]$ for $N = 399$ and $N = 3999$, respectively. As N increases, the histograms become more and more concentrated. Moreover, in agreement with Theorem 2 where it is proved that $\bar{\mathbf{q}}_{(0.8\cdot(N+1))}$ tends to $\mathbf{q}_{(0.8\cdot(N+1))}$ almost surely, the two histograms approach each other.

A. A simple example showing the bias of $\mathbf{q}_{(i)}$

Suppose that $n = d = 1$, $X = 1$ and Y is random with continuous distribution. $N = 2$ observations are available. Based on $D^2 = \{(1, Y_1), (1, Y_2)\}$, the least squares estimate $\hat{\beta}_2$ and the empirical costs $\mathbf{q}_1, \mathbf{q}_2$ are computed. We will evaluate the probability that a new instance $(1, Y)$ is such that $\mathbf{q} \leq \mathbf{q}_{(2)}$ and show that this probability is strictly less than $\frac{2}{3}$. First, notice that, conditionally to any set of three instances, say $S = \{(1, Y'), (1, Y''), (1, Y''')\}$, the probability of each permutation of the elements in S is the same, that is, each element of S plays the role of *new* instance $(1, Y)$ with probability $\frac{1}{3}$. As a consequence, for any set of three instances, the three situations represented in Fig. 8 are equally likely and, since $\mathbf{q} \leq \mathbf{q}_{(2)}$ holds in one out of the three situations, integrating over all possible sets of three instances yields $\mathbb{P}\{\mathbf{q} \leq \mathbf{q}_{(2)}\} = \frac{1}{3} < \frac{2}{3}$.

[5]Note that $K = X^T X$ has rank 1; see the discussion following Theorem 1 for comments on how d affects $\bar{\mathbf{q}}_{(i)}$.

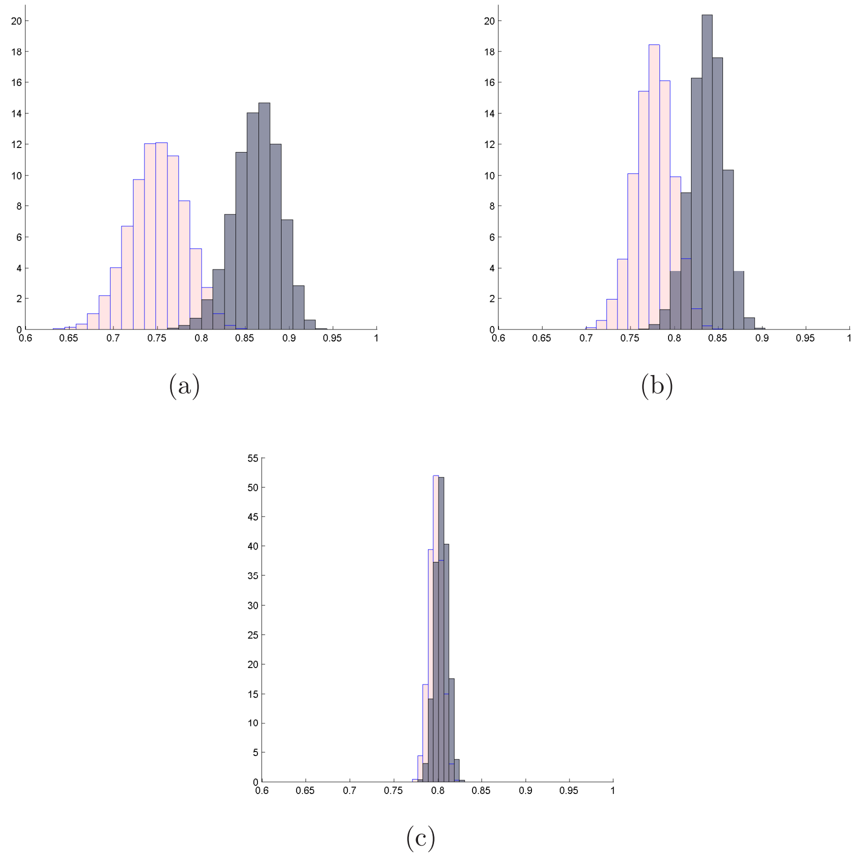


Figure 7. Histograms of the coverage of $(-\infty, \bar{\mathbf{q}}_{(0.8\cdot(N+1))}]$ (darker, on the right) and of $(-\infty, \mathbf{q}_{(0.8\cdot(N+1))}]$ (lighter, on the left) when (a) $N = 199$; (b) $N = 399$; (c) $N = 3999$.

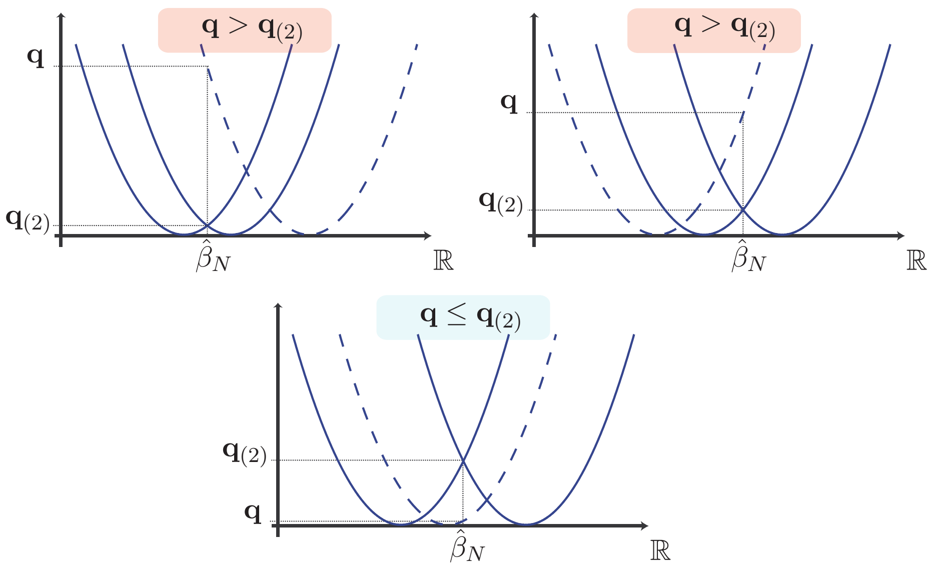


Figure 8. The three parabolas. The dashed parabola is $(Y - \beta)^2$, while the other two correspond to the data set D^2.

B. Proof of Theorem 1

We prove a slightly stronger result than Theorem 1. This stronger result is stated below as Theorem 4. In turn, we show that Theorem 1 follows from Theorem 4.

Matrices K_i, $i = 1, \ldots, N$, are defined in Section 2 as $K_i = X_i^T X_i$. Thus, the K_i's are symmetric and positive semi-definite. Throughout, the following simplified notation is in use

$$\sum K_\ell \text{ stands for } \sum_{\ell=1}^{N} K_\ell, \quad \sum_{\ell \neq i} K_\ell \text{ stands for } \sum_{\substack{\ell=1 \\ \ell \neq i}}^{N} K_\ell.$$

The next lemma is frequently used in this section.

Lemma 1. *Assume that* $\sum_{\ell \neq i} K_\ell \succ 0$. *For any* $\gamma \geq 0$, *the following two equivalences hold:*

$$K_i^{\frac{1}{2}} \left(\sum_{\ell \neq i} K_\ell \right)^{-1} K_i^{\frac{1}{2}} \prec \gamma I \iff K_i \prec \gamma \sum_{\ell \neq i} K_\ell, \tag{9}$$

$$K_i^{\frac{1}{2}} \left(\sum_{\ell \neq i} K_\ell \right)^{-1} K_i^{\frac{1}{2}} \preceq \gamma I \iff K_i \preceq \gamma \sum_{\ell \neq i} K_\ell. \tag{10}$$

⋆

Proof. The case $\gamma = 0$ is easily verified by inspection. Suppose that $\gamma > 0$. For given matrices $A \in \mathbb{R}^{p \times p}$, $B \in \mathbb{R}^{p \times q}$, $C \in \mathbb{R}^{q \times q}$ with $A \succ 0$ and $C \succ 0$, the following relation among Schur complements holds: $C - B^T A^{-1} B \succ 0$ $[\succeq 0] \iff A - B C^{-1} B^T \succ 0$ $[\succeq 0]$. The lemma follows by taking $A = \gamma I$, $B = K_i^{\frac{1}{2}}$, and $C = \sum_{\ell \neq i} K_\ell$. □

We next introduce some definitions that are used later in the statement of Theorem 4.

If $\sum_{\ell \neq i} K_\ell \succ 0$, let

$$\gamma_i := \lambda_{\max} \left(K_i^{\frac{1}{2}} \left(\sum_{\ell \neq i} K_\ell \right)^{-1} K_i^{\frac{1}{2}} \right),$$

$$W_i := K_i + (4 + 2\gamma_i) K_i \left(\sum_{\ell \neq i} K_\ell \right)^{-1} K_i. \tag{11}$$

Suppose further that $\gamma_i < \frac{1}{\sqrt{2}}$, then matrix $2 \sum K_\ell - W_i$ is invertible. To show this, note that, γ_i being the maximum eigenvalue of $K_i^{\frac{1}{2}} \left(\sum_{\ell \neq i} K_\ell \right)^{-1} K_i^{\frac{1}{2}}$, we have that

$$K_i^{\frac{1}{2}} \left(\sum_{\ell \neq i} K_\ell \right)^{-1} K_i^{\frac{1}{2}} \preceq \gamma_i I, \tag{12}$$

and, hence,

$$\begin{aligned} W_i &= K_i + (4 + 2\gamma_i) K_i^{\frac{1}{2}} \left(K_i^{\frac{1}{2}} \left(\sum_{\ell \neq i} K_\ell \right)^{-1} K_i^{\frac{1}{2}} \right) K_i^{\frac{1}{2}} \\ &\preceq K_i + (4 + 2\gamma_i) \gamma_i K_i = (1 + 4\gamma_i + 2\gamma_i^2) K_i. \end{aligned} \tag{13}$$

Applying Lemma 1 to (12) gives $K_i \preceq \gamma_i \sum_{\ell \neq i} K_\ell$, from which $K_i \preceq \frac{\gamma_i}{1+\gamma_i} \sum K_\ell$. Substituting this result in (13), under the condition $\gamma_i < \frac{1}{\sqrt{2}}$, yields

$$W_i \preceq (1 + 4\gamma_i + 2\gamma_i^2) \frac{\gamma_i}{1+\gamma_i} \sum K_\ell \prec 2 \sum K_\ell, \tag{14}$$

which proves the invertibility of $2\sum K_\ell - W_i$.

If $\sum_{\ell \neq i} K_\ell \succ 0$ and $\gamma_i < \frac{1}{\sqrt{2}}$, define $\tilde{K}_i := W_i + W_i (2\sum K_\ell - W_i)^{-1} W_i$. Let

$$\tilde{\mathbf{q}}_i := \begin{cases} (\hat{\beta}_N - v_i)^T \tilde{K}_i (\hat{\beta}_N - v_i) + h_i, & \text{if } \sum_{\ell \neq i} K_\ell \succ 0 \text{ and } \gamma_i < \frac{1}{\sqrt{2}} \\ +\infty, & \text{otherwise.} \end{cases} \tag{15}$$

Theorem 4. *Relation*

$$\mathbb{P}\{\mathbf{q} \le \tilde{\mathbf{q}}_{(i)}\} \ge \frac{i}{N+1}.$$

holds for any probability distribution F. ⋆

Before proving Theorem 4, we show that Theorem 1 follows from Theorem 4. To this end, it is enough to show that $\tilde{\mathbf{q}}_i \le \bar{\mathbf{q}}_i$, $i = 1, \dots, N$. When $\bar{\mathbf{q}}_i = \infty$, this is trivially true, so we consider the case when $\bar{\mathbf{q}}_i$ is finite, which holds if $K_i \prec \frac{1}{6} \sum_{\ell \neq i} K_\ell$. In view of Lemma 1, condition $K_i \prec \frac{1}{6} \sum_{\ell \neq i} K_\ell$ implies that $\gamma_i < \frac{1}{6}$, which strengthens condition $\gamma_i < \frac{1}{\sqrt{2}}$ used in Theorem 4. We now show that, if $\gamma_i < \frac{1}{6}$, then $\tilde{K}_i \preceq \bar{K}_i$, from which $\tilde{\mathbf{q}}_i \le \bar{\mathbf{q}}_i$.
Due to that $\gamma_i < \frac{1}{6}$, (13) gives $W_i \preceq 2K_i$, so that

$$2\sum K_\ell - W_i \succeq 2\sum K_\ell - 2K_i = 2\sum_{\ell \neq i} K_\ell.$$

Thus,

$$\begin{aligned} \tilde{K}_i &= W_i + W_i \Big(2\sum K_\ell - W_i\Big)^{-1} W_i \\ &\preceq W_i + W_i \left(2\sum_{\ell \neq i} K_\ell\right)^{-1} W_i \\ &= K_i + K_i^{\frac{1}{2}} \left(\frac{9 + 4\gamma_i}{2} \Phi + (4 + 2\gamma_i)\Phi^2 + 2(2+\gamma_i)^2 \Phi^3\right) K_i^{\frac{1}{2}}, \end{aligned}$$

where we substituted (11) for W_i and let $\Phi = K_i^{\frac{1}{2}} \Big(\sum_{\ell \neq i} K_\ell\Big)^{-1} K_i^{\frac{1}{2}}$. Since $\Phi \preceq \gamma_i I$, we get

$$\begin{aligned} \tilde{K}_i &\preceq K_i + K_i^{\frac{1}{2}} \left(\frac{9 + 4\gamma_i}{2} \Phi + (4 + 2\gamma_i)\gamma_i \Phi + 2(2+\gamma_i)^2 \gamma_i^2 \Phi\right) K_i^{\frac{1}{2}} \\ &= K_i + (4.5 + 6\gamma_i + 10\gamma_i^2 + 8\gamma_i^3 + 2\gamma_i^4) K_i \left(\sum_{\ell \neq i} K_\ell\right)^{-1} K_i \\ &\preceq \bar{K}_i, \end{aligned}$$

where the last inequality follows from the fact that $4.5 + 6\gamma_i + 10\gamma_i^2 + 8\gamma_i^3 + 2\gamma_i^4 < 6$ for $\gamma_i < \frac{1}{6}$. Wrapping up, if $K_i \prec \frac{1}{6}\sum_{\ell \neq i} K_\ell$, then $\tilde{K}_i \preceq \bar{K}_i \Longrightarrow \tilde{\mathbf{q}}_i \leq \bar{\mathbf{q}}_i \Longrightarrow$ Theorem 1 follows from Theorem 4.

Proof of Theorem 4.

To ease the notation, let

$$\mathbf{Q}_i(\beta) := (\beta - v_i)^T K_i(\beta - v_i) + h_i = \|Y_i - X_i\beta\|^2, \quad (16)$$

$$\mathbf{Q}(\beta) := (\beta - v)^T K(\beta - v) + h = \|Y - X\beta\|^2. \quad (17)$$

With these definitions, we can write

$$\hat{\beta}_N = \arg\min_\beta \sum_{i=1}^N \mathbf{Q}_i(\beta), \quad \mathbf{q}_i = \mathbf{Q}_i(\hat{\beta}_N), \quad \mathbf{q} = \mathbf{Q}(\hat{\beta}_N).$$

It is also convenient to introduce the minimizer of the least squares cost augmented with $\mathbf{Q}(\beta)$, namely,[6]

$$\hat{\beta} := \arg\min_\beta \left\{ \sum_{i=1}^N \mathbf{Q}_i(\beta) + \mathbf{Q}(\beta) \right\},$$

and the minimizer of the augmented least squares cost without the ith term, that is,

$$\hat{\beta}^{[i]} := \arg\min_\beta \left\{ \sum_{\substack{\ell=1 \\ \ell \neq i}}^N \mathbf{Q}_\ell(\beta) + \mathbf{Q}(\beta) \right\}, \; i = 1, \ldots, N. \quad (18)$$

The following random variables $\mathbf{m}$ and $\mathbf{m}_i$, $i = 1, \ldots, N$, exhibit a precise ranking property indicated in Lemma 2. Define:

$$\mathbf{m} := \begin{cases} \mathbf{Q}(\hat{\beta}_N) + \left[\mathbf{Q}(\hat{\beta}_N) - \mathbf{Q}(\hat{\beta})\right], & \text{if } \sum K_\ell \succ 0 \\ +\infty, & \text{otherwise,} \end{cases} \quad (19)$$

$$\mathbf{m}_i := \begin{cases} \mathbf{Q}_i(\hat{\beta}^{[i]}) + \left[\mathbf{Q}_i(\hat{\beta}^{[i]}) - \mathbf{Q}_i(\hat{\beta})\right], & \text{if } \sum_{\ell \neq i} K_\ell + K \succ 0 \\ +\infty, & \text{otherwise.} \end{cases} \quad (20)$$

LEMMA 2.

$$\mathbb{P}\{\mathbf{m} \leq \mathbf{m}_{(i)}\} \geq \frac{i}{N+1}, \; i = 1, \ldots, N.$$

⋆

PROOF. Random variables $\mathbf{m}$ and $\mathbf{m}_i$ are all obtained by applying the same function to permutations of an i.i.d. (independent and identically distributed) sample of $N+1$ elements, namely $\{(X_1, Y_1), \ldots, (X_N, Y_N), (X, Y)\}$. Hence, $\mathbf{m}$ and $\mathbf{m}_i$ are exchangeable random variables. Conditionally on a set of $N+1$ fixed values taken by $\mathbf{m}$ and $\mathbf{m}_i$ in any order (that is, the first value is taken

[6] If the solution is not unique, the solution is determined by the same tie-break rule that is used to determine $\hat{\beta}_N$ when the minimizer of the least squares cost is not unique. The same applies to $\hat{\beta}^{[i]}$ defined below.

by any one of the variables $\mathbf{m}$ or $\mathbf{m}_i$, the second value by any one of the remaining variables, and so on), relation $\mathbf{m} \leq \mathbf{m}_{(i)}$ (which means that $\mathbf{m}$ has the lowest value, or the second lowest, ..., or the i-th lowest) holds with probability $\frac{i}{N+1}$ or more (more can occur because of ties). Integrating over all the possible values taken by $\mathbf{m}$ and $\mathbf{m}_i$, the result is obtained. □

According to the terminology of Vovk et al. (2005), equations (19) and (20) introduce a conformity measure, and $\mathbf{m}$, $\mathbf{m}_i$ are the corresponding conformity scores of (X, Y), (X_i, Y_i).

Now, for a given D^N, $\mathbf{Q}(\hat{\beta}_N)$ is a function of (X, Y), or, equivalently, of (K, v, h). For $i = 1, \ldots, N$, consider the following maximization problem

$$\mu_i := \sup_{K,v,h} \mathbf{Q}(\hat{\beta}_N) \quad \text{subject to: } \mathbf{m} \leq \mathbf{m}_i. \tag{21}$$

In the on-line supplementary material, Section 2, the validity of the following key relation is proven (the proof of equation (22) is rather technical, and it has been moved to the on-line supplementary material due to page limits)

$$\mu_i \leq \tilde{\mathbf{q}}_i, \quad i = 1, \ldots, N. \tag{22}$$

Theorem 4 easily follows from (22). Indeed, note that (22) implies that

$$\mu_{(i)} \leq \tilde{\mathbf{q}}_{(i)}, \quad i = 1, \ldots, N. \tag{23}$$

On the other hand, with the definition

$$\nu_i := \sup_{K,v,h} \mathbf{Q}(\hat{\beta}_N) \quad \text{subject to: } \mathbf{m} \leq \mathbf{m}_{(i)}, \tag{24}$$

we also have that

$$\nu_i \leq \mu_{(i)}, \quad i = 1, \ldots, N, \tag{25}$$

as it can be argued by the following simple reasoning. Fix a value of i, say $i = \bar{i}$. Assume for simplicity that sup in (24) is actually a max (if not, the proof follows by a limiting argument), and let (K^*, v^*, h^*) be the maximizer. Corresponding to (K^*, v^*, h^*), $\mathbf{m} \leq \mathbf{m}_{(\bar{i})}$, which entails that (K^*, v^*, h^*) is feasible for at least $N - \bar{i} + 1$ values of i in (21). Hence, since μ_i in (21) is obtained by a sup operation, $\nu_{\bar{i}} = \mathbf{Q}^*(\hat{\beta}_N) \leq \mu_i$ for at least $N - \bar{i} + 1$ values of i, which implies that $\nu_{\bar{i}} \leq \mu_{(\bar{i})}$, that is, (25).

Since $\nu_i \leq \mu_{(i)}$ (relation (25)) and $\mu_{(i)} \leq \tilde{\mathbf{q}}_{(i)}$ (relation (23)), we obtain $\nu_i \leq \tilde{\mathbf{q}}_{(i)}$, and Theorem 4 remains proven as follows

$$\begin{aligned}\mathbb{P}\{\mathbf{q} \leq \tilde{\mathbf{q}}_{(i)}\} &= \mathbb{P}\{\mathbf{Q}(\hat{\beta}_N) \leq \tilde{\mathbf{q}}_{(i)}\} \\ &\geq \mathbb{P}\{\mathbf{Q}(\hat{\beta}_N) \leq \nu_i\} \\ &\geq \mathbb{P}\{\mathbf{m} \leq \mathbf{m}_{(i)}\} \\ &\geq \frac{i}{N+1},\end{aligned}$$

where the second last inequality follows because ν_i is the sup of $\mathbf{Q}(\hat{\beta}_N)$ when $\mathbf{m} \leq \mathbf{m}_{(i)}$, while the last inequality is Lemma 2. □

C. Proof of Theorem 2

Consider a function $f(N) > 0$ such that $\frac{\ln N}{f(N)} \to 0$. Thus, $\frac{\ln N^3}{\alpha f(N)} = \frac{3}{\alpha}\frac{\ln N}{f(N)} \to 0$ and we have that

$$\begin{aligned}\mathbb{P}\left\{\frac{1}{f(N)}\max_{i=1,\dots,N}\|K_i\| > \frac{\ln N^3}{\alpha f(N)}\right\} &= \mathbb{P}\left\{\max_{i=1,\dots,N}\|K_i\| > \frac{\ln N^3}{\alpha}\right\} \\ &\leq N\mathbb{P}\left\{\|K_i\| > \frac{\ln N^3}{\alpha}\right\} \\ &\leq Ne^{-\alpha\frac{\ln N^3}{\alpha}} = \frac{1}{N^2},\end{aligned}$$

where the last inequality follows from condition (7). Since $\sum_{N=1}^{\infty}\frac{1}{N^2} < \infty$, from the Borel-Cantelli Lemma (see e.g. Shiryaev (1995)) we conclude that

$$\lim_{N\to\infty}\frac{1}{f(N)}\max_{i=1,\dots,N}\|K_i\| = 0 \quad \text{almost surely.} \tag{26}$$

Similarly, using (8) in place of (7), it can be proved that

$$\lim_{N\to\infty}\frac{1}{f(N)}\max_{i=1,\dots,N}\|v_i\| = 0 \quad \text{almost surely.} \tag{27}$$

On the other hand, condition (7) also guarantees that the strong law of large numbers applies, so that

$$\lim_{N\to\infty}\frac{1}{N}\sum_{\ell=1}^{N}K_\ell = \mathbb{E}[K_i] \quad \text{almost surely.} \tag{28}$$

Since $\mathbb{E}[K_i] \succ 0$ (condition (6)), (26) and (28) with $f(N) = N$ yield

$$\begin{aligned}\lim_{N\to\infty}\min_{i=1,\dots,N}\left\|\frac{1}{N}\sum_{\substack{\ell=1\\ \ell\neq i}}^{N}K_\ell\right\| &\geq \lim_{N\to\infty}\left[\left\|\frac{1}{N}\sum_{\ell=1}^{N}K_\ell\right\| - \frac{1}{N}\max_{i=1,\dots,N}\|K_i\|\right] \\ &= \|\mathbb{E}[K_i]\| \\ &> 0 \quad \text{almost surely.}\end{aligned} \tag{29}$$

Moreover, again using (26) and (28) with $f(N) = N$, we also see that relation

$$\frac{1}{N} K_i \prec \frac{1}{7} \frac{1}{N} \sum_{\ell=1}^{N} K_\ell, \ i = 1, \dots, N,$$

or, equivalently, relation

$$K_i \prec \frac{1}{6} \sum_{\substack{\ell=1 \\ \ell \neq i}}^{N} K_\ell, \ i = 1, \dots, N,$$

holds for N large enough almost surely. Thus, almost surely, the $\bar{\mathbf{q}}_i$ defined in (3) are finite and equal to $(\hat{\beta}_N - v_i)^T \bar{K}_i (\hat{\beta}_N - v_i) + h_i$ for N large enough. Hence, using $\bar{K}_i = K_i + 6K_i \left(\sum_{\substack{\ell=1 \\ \ell \neq i}}^{N} K_\ell \right)^{-1} K_i$, it holds that

$$\begin{aligned}
&\max_{i=1,\dots,N} (\bar{\mathbf{q}}_i - \mathbf{q}_i) \\
&\quad = \max_{i=1,\dots,N} \left((\hat{\beta}_N - v_i)^T \bar{K}_i (\hat{\beta}_N - v_i) + h_i - (\hat{\beta}_N - v_i)^T K_i (\hat{\beta}_N - v_i) - h_i \right) \\
&\quad = \max_{i=1,\dots,N} (\hat{\beta}_N - v_i)^T (\bar{K}_i - K_i)(\hat{\beta}_N - v_i) \\
&\quad \leq \max_{i=1,\dots,N} 6 \left\| \left(\sum_{\substack{\ell=1 \\ \ell \neq i}}^{N} K_\ell \right)^{-1} \right\| \|K_i\|^2 \|\hat{\beta}_N - v_i\|^2 \\
&\quad \leq \max_{i=1,\dots,N} 6 \left\| \left(\frac{1}{N} \sum_{\substack{\ell=1 \\ \ell \neq i}}^{N} K_\ell \right)^{-1} \right\| \frac{\|K_i\|^2}{N^{\frac{1}{2}}} \frac{\|\hat{\beta}_N\|^2 + \|v_i\|^2}{N^{\frac{1}{2}}},
\end{aligned}$$

for N large enough. The last expression tends to zero almost surely in view of (29), (26) and (27) with $f(N) = N^{\frac{1}{4}}$, and the fact that $\hat{\beta}_N = \left(\frac{1}{N} \sum_{\ell=1}^{N} K_\ell \right)^{-1} \left(\frac{1}{N} \sum_{\ell=1}^{N} K_\ell v_\ell \right)$ converges almost surely. This concludes the proof. □

D. Proof of Theorem 3

To ease the presentation, the following notation is in order

$$\mathbf{q}_i^{[k]} = \mathbf{Q}_i(\hat{\beta}^{[k]}), \ i = 0, \dots, N; \ k = 0, \dots, N,$$

where $\mathbf{Q}_i$, $i = 1, \dots, N$, is defined in (16) and $\mathbf{Q}_0 := \mathbf{Q}$, where $\mathbf{Q}$ is defined in (17); $\hat{\beta}^{[k]}$, $k = 1, \dots, N$, is defined in (18), and $\hat{\beta}^{[0]} := \hat{\beta}_N$. In this new notation $\mathbf{q}$ is written as $\mathbf{q}_0^{[0]}$, and $\mathbf{q}_i$ as $\mathbf{q}_i^{[0]}$, $i = 1, \dots, N$. Also, for $k = 0, \dots, N$, define $\mathbf{q}_{(i)}^{[k]} = \mathrm{ord}_{(i)} \left[\mathbf{q}_0^{[k]}, \dots, \mathbf{q}_{k-1}^{[k]}, \mathbf{q}_{k+1}^{[k]}, \dots, \mathbf{q}_N^{[k]} \right]$, where $\mathrm{ord}_{(\mathrm{i})}$ is the ith order statistic of the listed elements.

Fix a value of $i \in \{1, \dots, N\}$. By the i.i.d. nature of $\mathbf{Q}_0, \mathbf{Q}_1 \dots, \mathbf{Q}_N$, we have that

$$\mathbb{P}\{\mathbf{q}_0^{[0]} \leq \mathbf{q}_{(i)}^{[0]}\} = \mathbb{P}\{\mathbf{q}_k^{[k]} \leq \mathbf{q}_{(i)}^{[k]}\}, \quad k = 1, \dots, N.$$

Hence, denoting by $\mathbb{1}(\cdot)$ the indicator function, we get

$$\begin{aligned}\mathbb{P}\{\mathbf{q} \leq \mathbf{q}_{(i)}\} &= \mathbb{P}\{\mathbf{q}_0^{[0]} \leq \mathbf{q}_{(i)}^{[0]}\} \\ &= \frac{1}{N+1}\sum_{k=0}^{N} \mathbb{P}\{\mathbf{q}_k^{[k]} \leq \mathbf{q}_{(i)}^{[k]}\} \\ &= \frac{1}{N+1}\sum_{k=0}^{N} \mathbb{E}\left[\mathbb{1}\left(\mathbf{q}_k^{[k]} \leq \mathbf{q}_{(i)}^{[k]}\right)\right] \\ &= \frac{1}{N+1}\mathbb{E}\left[\sum_{k=0}^{N} \mathbb{1}\left(\mathbf{q}_k^{[k]} \leq \mathbf{q}_{(i)}^{[k]}\right)\right]. \end{aligned} \tag{30}$$

The proof will be now completed by showing that

$$\sum_{k=0}^{N} \mathbb{1}\left(\mathbf{q}_k^{[k]} \leq \mathbf{q}_{(i)}^{[k]}\right) \leq i \tag{31}$$

holds almost surely, so that the right-hand side of (30) is bounded by $\frac{i}{N+1}$, which is the conclusion of the theorem. To show (31), define $S_k = \sum_{\substack{\ell=0 \\ \ell\neq k}}^{N} \mathbf{q}_\ell^{[k]}$, $k = 0, \ldots, N$, and consider a $S_{\bar{k}}$ such that

$$S_{\bar{k}} \leq S_k \quad \text{holds for at least } i \text{ indexes } k \text{ different from } \bar{k}. \tag{32}$$

The number of these indexes $\bar{k}$ is at least $N + 1 - i$. We show that $\mathbf{q}_{\bar{k}}^{[\bar{k}]} \geq \mathbf{q}_{(i)}^{[\bar{k}]}$. By contradiction, suppose instead that $\mathbf{q}_{\bar{k}}^{[\bar{k}]} < \mathbf{q}_{(i)}^{[\bar{k}]}$. Then, for any index k such that $\mathbf{q}_{(i)}^{[\bar{k}]} \leq \mathbf{q}_k^{[\bar{k}]}$, we have $S_{\bar{k}} = \sum_{\substack{\ell=0 \\ \ell\neq \bar{k}}}^{N} \mathbf{q}_\ell^{[\bar{k}]} > \sum_{\substack{\ell=0 \\ \ell\neq k}}^{N} \mathbf{q}_\ell^{[\bar{k}]}$. Since $\mathbf{q}_\ell^{[\bar{k}]} = \mathbf{Q}_\ell(\hat{\beta}^{[\bar{k}]})$ and $\hat{\beta}^{[k]}$ is the minimizer of $\sum_{\substack{\ell=0 \\ \ell\neq k}}^{N} \mathbf{Q}_\ell(\beta)$, we conclude that

$$S_{\bar{k}} > \sum_{\substack{\ell=0 \\ \ell\neq k}}^{N} \mathbf{Q}_\ell(\hat{\beta}^{[\bar{k}]}) \geq \sum_{\substack{\ell=0 \\ \ell\neq k}}^{N} \mathbf{Q}_\ell(\hat{\beta}^{[k]}) = \sum_{\substack{\ell=0 \\ \ell\neq k}}^{N} \mathbf{q}_\ell^{[k]} = S_k. \tag{33}$$

There are at least $N + 1 - i$ values of k such that $\mathbf{q}_{(i)}^{[\bar{k}]} \leq \mathbf{q}_k^{[\bar{k}]}$, so that, by (33), there are at least $N + 1 - i$ values of k such that $S_{\bar{k}} > S_k$. This contradicts (32). Thus, the conclusion is drawn that $\mathbf{q}_{\bar{k}}^{[\bar{k}]} \geq \mathbf{q}_{(i)}^{[\bar{k}]}$ is verified for all the indexes $\bar{k}$, which, as seen, are at least $N + 1 - i$. Since equality $\mathbf{q}_{\bar{k}}^{[\bar{k}]} = \mathbf{q}_{(i)}^{[\bar{k}]}$ holds only with probability 0 by the theorem assumption, it follows that $\mathbf{q}_{\bar{k}}^{[\bar{k}]} > \mathbf{q}_{(i)}^{[\bar{k}]}$ holds almost surely for the indexes $\bar{k}$, and (31) holds almost surely. This concludes the proof. □

References

Beasley, T. M., Page, G. P., Brand, J. P. L., Gadbury, G. L., Mountz, J. D. and Allison, D. B. (2004) Chebyshev's inequality for nonparametric testing with small N and alpha in microarray research. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **53**, 95–108.

Belloni, A., Chernozhukov, V. and Kato, K. (2014) Valid post-selection inference in high-dimensional approximately sparse quantile regression models. ArXiv preprint, http://arxiv.org/abs/1312.7186.

— (2015) Uniform post-selection inference for least absolute deviation regression and other Z-estimation problems. *Biometrika*, **102**, 77–94.

Benjamini, Y. and Yekutieli, D. (2005) False discovery rate-adjusted multiple confidence intervals for seleted parameters. *Journal of the American Statistical Association*, **100**, 71–81.

Berk, R., Brown, L., Buja, A., Zhang, K. and Zhao, L. (2013) Valid post-selection inference. *The Annals of Statistics*, **41**, 802–837.

Brown, L. (1967) The conditional level of Student's t-test. *The Annals of Mathematical Statistics*, **38**, 1068–1071.

Buehler, R. J. and Fedderson, A. P. (1963) Note on a conditional property of Student's t. *The Annals of Mathematical Statistics*, **34**, 1098–1100.

Calafiore, G. C. and Campi, M. C. (2005) Uncertain convex programs: randomized solutions and confidence levels. *Mathematical Programming*, **102**, 25–46.

Campi, M. C. and Garatti, S. (2008) The exact feasibility of randomized solutions of uncertain convex programs. *SIAM Journal on Optimization*, **19**, 1211–1230.

— (2011) A sampling-and-discarding approach to chance-constrained optimization: feasibility and optimality. *Journal of Optimization Theory and Applications*, **148**, 257–280.

— (2016) Wait-and-judge scenario optimization. *Mathematical Programming*. Published on-line DOI: 10.1007/s10107-016-1056-9.

Carè, A., Garatti, S. and Campi, M. C. (2015) Scenario min-max optimization and the risk of empirical costs. *SIAM Journal on Optimization*, **25**, 2061–2080.

Chen, D. and Gao, C. (2012) Soft computing methods applied to train station parking in urban rail transit. *Applied Soft Computing*, **12**, 759–767.

David, H. A. and Nagaraja, H. N. (2003) *Order Statistics (3rd Edition)*. New York, USA: Wiley.

Di Bucchianico, A., Einmahl, J. H. J. and Mushkudiani, N. A. (2001) Smallest nonparametric tolerance regions. *The Annals of Statistics*, **29**, 1320–1343.

Etzold, A. and Eurich, C. W. (2005) A direct, interval-based method for reconstructing stimuli from noise-robust tuning curves. *Neurocomputing*, **6566**, 103–109.

Fraser, D. A. S. and Guttman, I. (1956) Tolerance regions. *The Annals of Mathematical Statistics*, **27**, 162–179.

Frey, J. (2013) Data-driven nonparametric prediction intervals. *Journal of Statistical Planning and Inference*, **143**, 1039–1048.

Gammerman, A. and Vovk, V. (2007) Hedging predictions in machine learning. *The Computer Journal*, **50**, 151–163.

Hull, J. (2009) *Options, Futures and Other Derivatives (8th Edition)*. Pearson/Prentice Hall.

Kabán, A. (2012) Non-parametric detection of meaningless distances in high dimensional data. *Statistics and Computing*, **22**, 375–385.

Lee, J. D., Sun, D. L., Sun, Y. and Taylor, J. E. (2016) Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, **44**, 907–927.

Lehmann, E. L. and Casella, G. (1998) *Theory of point estimation (2nd Edition)*. Springer.

Lei, J., Robins, J. and Wasserman, L. (2013) Distribution-free prediction sets. *Journal of the American Statistical Association*, **108**, 278–287.

Lei, J. and Wasserman, L. (2014) Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society, Series B*, **76**, 71–96.

Li, J. and Liu, R. Y. (2008) Multivariate spacings based on data depth: I. Construction of nonparametric multivariate tolerance regions. *The Annals of Statistics*, **36**, 1299–1323.

Pötscher, B. M. (1991) Effects of model selection on inference. *Econometric Theory*, **7**, 163–185.

Saw, J. G., Yang, M. C. K. and Mo, T. C. (1984) Chebyshev inequality with estimated mean and variance. *The American Statistician*, **38**, 130–132.

— (1988) Corrections: Chebyshev inequality with estimated mean and variance. *The American Statistician*, **42**, 166.

Scheffé, H. and Tukey, J. W. (1947) Non-parametric estimation. II. Statistical equivalent blocks and tolerance regions - the continuous case. *The Annals of Mathematical Statistics*, **18**, 529–539.

Shafer, G. and Vovk, V. (2008) A tutorial on conformal prediction. *Journal of Machine Learning Research*, **9**, 371–421.

Shiryaev, A. N. (1995) *Probability (2nd Edition)*. New York, USA: Springer-Verlag.

Tibshirani, R. J., Taylor, J., Lockhart, R. and Tibshirani, R. (2016) Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, **111**, 600–620.

Vapnik, V. N. (1996) *Statistical Learning Theory*. New York, USA: Wiley.

Vapnik, V. N. and Chervonenkis, A. Y. (1971) On the uniform convergence of relative frequencies to their probabilities. *Theory of Probability and Its Applications*, **16**, 264–280.

Vardeman, S. B. (1992) What about the other intervals? *The American Statistician*, **46**, 193–197.

Vovk, V. (2004) A universal well-calibrated algorithm for on-line classification. *Journal of Machine Learning Research*, **5**, 575–604.

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

Wilks, S. S. (1941) Determination of sample sizes for setting tolerance limits. *The Annals of Mathematical Statistics*, **12**, 91–96.

Xu, W. L. and Nelson, B. L. (2013) Empirical stochastic branch-and-bound for optimization via simulation. *IIE Transactions*, **45**, 685–698.