52nd IEEE Conference on Decision and Control
December 10-13, 2013. Florence, Italy

Least Squares Estimates and the Coverage of Least Squares Costs

Algo Carè, Simone Garatti, Marco C. Campi

***Abstract*— The least squares estimate $\hat{x}_N$ minimizes the sum of the squared residuals $\sum_{i=1}^{N} \|A_i x - b_i\|^2$ over a finite set of observations (A_i, b_i). At $x = \hat{x}_N$, the squared residuals $\|A_i \hat{x}_N - b_i\|^2$ are called the "empirical costs". Intuitively, the empirical costs carry information on the probability distribution of the cost $\|A\hat{x}_N - b\|^2$ that is paid for other, yet unseen, values of (A, b) taken from the same population as the observations (A_i, b_i). In this work, this intuition is set on solid theoretical grounds. We provide a precise characterization of the probabilities with which the cost does not exceed certain thresholds that are constructed from the empirical costs. These probabilities are called "coverages". All the results are derived in a setting where the observations are independent, while the framework is otherwise "agnostic" in that no a-priori assumptions about the underlying probability for (A, b) is made.**

I. INTRODUCTION AND PROBLEM SET-UP

Given the finite sample of data

$$\mathsf{D}^N = (A_1, b_1), (A_2, b_2), \ldots, (A_N, b_N),$$

where $A_i \in \mathbb{R}^{n\times d}$ and $b_i \in \mathbb{R}^n$, the least squares estimate $\hat{x}_N$ is defined as the minimizer of the sum of squared residuals

$$\sum_{i=1}^{N} \|A_i x - b_i\|^2,$$ [1]

where $\|\cdot\|$ denotes the Euclidean norm. The least squares method is relevant to many fields including statistics, systems and control, quantitative finance, econometrics and decision-making, to cite but a few.

In this paper, we assume that (A_i, b_i) are independent and identically distributed (i.i.d.) random elements with distribution F. The squared residuals evaluated at $\hat{x}_N$,

$$\mathbf{q}_i := \|A_i \hat{x}_N - b_i\|^2, \quad i = 1, \ldots, N,$$

are called the "empirical costs". Given a new observation (A, b) sampled from F independently of D^N, the cost of (A, b) evaluated at $\hat{x}_N$ is denoted with

$$\mathbf{q} := \|A\hat{x}_N - b\|^2.$$

The goal of this paper is to provide evaluations that $\mathbf{q}$ does not exceed certain thresholds constructed from the data.

A function $\mathbf{c}$ of the data D^N is called a *statistic*. For example, a statistic is $\mathbf{c} = \max_{i=1,\ldots,N} \mathbf{q}_i$. If an evaluation of the probability that $\mathbf{q}$ does not exceed $\mathbf{c}$ is provided, such en evaluation can be used as a descriptor of the performance of $\hat{x}_N$. The probability that $\mathbf{q}$ does not exceed $\mathbf{c}$ is called the "mean coverage" of $\mathbf{c}$ and is formally defined as follows.

Definition 1: Given a statistic $\mathbf{c}$ of the data D^N, the *mean coverage* of $\mathbf{c}$, is

$$\Pr\{\mathbf{q} \le \mathbf{c}\}.$$

⋆

In words, the mean coverage of $\mathbf{c}$ is the total probability of seeing a random sample D^N and that one more observation (A, b) independent of D^N carries a cost no higher than $\mathbf{c}(\mathsf{D}^N)$. The new instance (A, b) can be interpreted as the next instance (A_{N+1}, b_{N+1}) observed after the estimate $\hat{x}_N$ has been made.[2] The discussion is made more concrete through the following estimation problem in linear regression.

Example 1 (linear regression): Let u and y be two scalar random variables. We want to regress y against a polynomial of order $d-1$ in u. N independent observations $(u_1, y_1), \ldots, (u_N, y_N)$ are available. Letting

$$A_i = \left[1, u_i, {u_i}^2, \ldots, {u_i}^{d-1}\right] \in \mathbb{R}^{1\times d}$$

and

$$b_i = y_i, \quad \text{for } i = 1, \ldots, N,$$

the polynomial at the observed u_i writes $\hat{y}(u_i) = A_i x$ where x is the vector of parameters to be tuned, and the

This work was supported by the Ministero dellIstruzione, dellUniversità e della Ricerca (MIUR), and by the European Union under project FP7 257005 "MoVeS: Modeling, Verification, and Control of Complex Systems."

Algo Carè is with Dipartimento di Ingegneria dell'Informazione, Università di Brescia, via Branze 38, 25123 Brescia, Italia. E-mail: algo.care@ing.unibs.it, website: http://www.ing.unibs.it/algo.care/

Simone Garatti is with Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, piazza Leonardo da Vinci 32, 20133 Milano, Italia. E-mail: simone.garatti@polimi.it, website: http://home.deib.polimi.it/sgaratti/

Marco C. Campi is with Dipartimento di Ingegneria dell'Informazione, Università di Brescia, via Branze 38, 25123 Brescia, Italia. E-mail: marco.campi@ing.unibs.it, website: http://www.ing.unibs.it/campi/

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

[2] The term "mean coverage" is borrowed from the statistical literature. Given a D^N, the set $T(\mathsf{D}^N) := \{(A, b) : \mathbf{q} \le \mathbf{c}\}$ is a "tolerance region" in the space $\mathbb{R}^{n\times d} \times \mathbb{R}^n$ according to the statistical terminology, [1], [2]. A tolerance region depends on D^N. The probability of a tolerance region is commonly called the "coverage probability" of the tolerance region, and is written as $\Pr\{\mathbf{q} \le \mathbf{c}|\mathsf{D}^N\}$. The conditioning with respect to D^N emphasizes that such a probability depends on D^N since the tolerance region $T(\mathsf{D}^N)$ is a set that depends on D^N. By taking the expected value of the coverage of $T(\mathsf{D}^N)$ with respect to D^N, the "mean coverage" is obtained,

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

coefficients x are obtained by minimizing the sum of the squared residuals

$$\sum_{i=1}^{N} \|A_i x - b_i\|^2.$$

In this regression problem, the new instance corresponds to the next observed data point (u_{N+1}, y_{N+1}), and the mean coverage of $\mathbf{c}$ is the probability of the event that $(\hat{y}(u_{N+1}) - y_{N+1})^2 \leq \mathbf{c}(\mathsf{D}^N)$.

$\star$

The knowledge of the distribution F is normally required to compute the mean coverage of a statistic $\mathbf{c}$. However, as we shall see below, it is possible to construct specific statistics whose mean coverages are guaranteed "distribution-free", that is, they hold for *all* distributions F. In applications, these statistics can be used to predict the value of $\mathbf{q}$ even when F is unknown.

Before proceeding, a notation is introduced that will be in force throughout. Given a sample of scalar variables $r_1, r_2, \ldots, r_N$, we denote with $r_{(1)}, r_{(2)}, \ldots, r_{(N)}$ the order statistics of the r_i's, that is, the $r_{(i)}$'s are the r_i's in increasing order of value: $r_{(1)} \leq r_{(2)} \leq \cdots \leq r_{(N)}$. Also, we recall that the following classic result holds for any i.i.d. sample, see [3], [4].

Theorem 1: Let $r_1, r_2, \ldots, r_N$ be an i.i.d. sample from a distribution F_r on $\mathbb{R}$. For a new r sampled from F_r independently of $r_1, r_2 \ldots, r_N$, it holds that

$$\Pr\{r \leq r_{(i)}\} \geq \frac{i}{N+1}, \quad i = 1, \ldots, N. \tag{1}$$

$\star$

Theorem 1 states that the statistic $r_{(i)}$ is not exceeded with a probability at least of $\frac{i}{N+1}$, no matter what F_r is. Interestingly, this is a tight result, because if F_r is continuous, (1) holds with equality, $\Pr\{r \leq r_{(i)}\} = \frac{i}{N+1}$.

In our context of least squares estimation, the ordered empirical costs $\mathbf{q}_{(1)}, \ldots, \mathbf{q}_{(N)}$ can be used as statistics to bound $\mathbf{q}$. It is a fact, however, that Theorem 1 does not apply to $\mathbf{q}_{(1)}, \ldots, \mathbf{q}_{(N)}$. In fact, $\mathbf{q}_1, \ldots, \mathbf{q}_N, \mathbf{q}$ are not i.i.d., because they depend on all the data set D^N through $\hat{x}_N$. Moreover, $\hat{x}_N$ is chosen so as to minimize the sum of the squared residuals, so that the empirical costs are biased towards small values, and we expect that $\Pr\{\mathbf{q} \leq \mathbf{q}_{(i)}\} < \frac{i}{N+1}$. The following example illustrates that this intuition is indeed true.

Example 2: Suppose that $N = 2$: $\mathsf{D}^2 = (A_1, b_1)$, (A_2, b_2), and assume that, with probability 1, $A_1 = A_2 = 1$ and $b_1 \neq b_2$. Based on D^2, the least squares estimate $\hat{x}_2$ and the empirical costs $\mathbf{q}_1, \mathbf{q}_2$ are computed. We will evaluate the probability that a new instance $(1, b)$ is such that $\mathbf{q} \leq \mathbf{q}_{(2)}$ and show that it is strictly less than $\frac{2}{3}$. First, notice that conditionally to any set of three instances, say $S = \{(1, b'), (1, b''), (1, b''')\}$, the probability of each

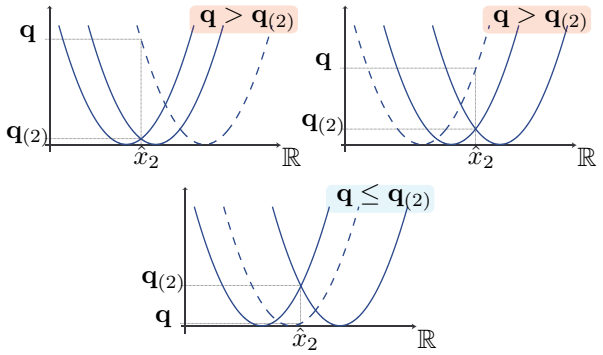


Fig. 1. The three parabolae of Example 2. The dashed parabola is $(x - b)^2$, while the other two correspond to the data D^2.

permutation of the elements in S is the same, that is, the role of the *new* instance $(1, b)$ is played by each element of S with probability $\frac{1}{3}$. As a consequence, for any set of three instances, the three situations represented in Fig. 1 are equally likely and, since $\mathbf{q} \leq \mathbf{q}_{(2)}$ holds true in one out of the three situations, integrating over all possible set of three instances yields $\Pr\{\mathbf{q} \leq \mathbf{q}_{(2)}\} = \frac{1}{3} < \frac{2}{3}$.

$\star$

The main achievement of this paper is to provide statistics $\bar{\mathbf{q}}_{(i)}$, $i = 1, \ldots, N$, such that

$$\Pr\{\mathbf{q} \leq \bar{\mathbf{q}}_{(i)}\} \geq \frac{i}{N+1}$$

holds true distribution-free, i.e., for every F. These statistics are obtained by adding a *margin* to the $\mathbf{q}_{(i)}$'s, according to a data-based rule that does not depend on F. This margin is small in normal cases and tends to zero as N grows to infinity.

Distribution-free results are of great interest since prior knowledge about F is often unrealistic to assume in practice. On the other hand, one may expect that a distribution-free result is conservative. In a sense, this paper contradicts this intuition by showing that a satisfactory and nonconservative characterization of the cost $\|A\hat{x}_N - b\|^2$ can be achieved by using $\bar{\mathbf{q}}_{(1)}, \ldots, \bar{\mathbf{q}}_{(N)}$, even for small number N.

A. Frequently used matrix notations

For a matrix M:

1) M^T = transpose matrix of M;
2) $M^\dagger$ = Moore-Penrose pseudoinverse of M;
3) $\|M\|$ = spectral norm = $\sup_{\|x\|=1} \|Mx\|$, where the norm in the right-hand side is the Euclidean norm;
4) $\lambda_{\max}(M)$ = maximum eigenvalue of M (M square matrix);
5) if M is symmetric, $M \succ 0$ ($M \succeq 0$) means M positive definite (semi-definite). $P \succ Q$ ($P \succeq Q$) means $P - Q$ positive definite (semi-definite).

For matrix concepts see e.g. [5], [6].

B. Bibliographic remarks

The least squares method dates back to Gauss and Legendre, see e.g. [7]. Ever since, it has been studied extensively and has found applications in an enormous variety of areas (e.g., linear regression theory [8], system identification [9], control [10], facility location [11], etc.). Much of the theoretical analysis focuses on bounding the deviation of $\mathbb{E}[\|A\hat{x}_N - b\|^2]$ from $\frac{1}{N}\sum_{i=1}^N \|A_i\hat{x}_N - b_i\|^2$. Classical results in this direction are asymptotic, e.g. [12], [9], while more recently results valid for finite N have started to appear, [13], based on the VC theory ([14], [15], [16]). The results of this paper are inherently different in that we do not aim at studying $\mathbb{E}[\|A\hat{x}_N - b\|^2]$; instead, we move towards characterizing the distribution of $\|A\hat{x}_N - b\|^2$ through the concept of coverage. This approach is in the spirit of [17], where sample-based min-max optimization is considered according to the approach of [18], [19], [20], [21]. Robust least squares have been considered in [22], [23].

C. Structure of the paper

The main theorem is provided in Section II followed by a discussion. A numerical example is given in Section III, and the outline of the proofs is given in Section IV.

II. MAIN RESULT

A. Main Theorem

To simplify the expression of our results, the cost $\|A_i x - b_i\|^2$ is rewritten as:

$$\|A_i x - b_i\|^2 = (x - v_i)^T K_i (x - v_i) + h_i,$$

with $K_i = A_i^T A_i$, $v_i = A_i^\dagger b_i$, $h_i = \|A_i v_i - b_i\|^2$. Observe that $K_i \succeq 0$ but not necessarily $K_i \succ 0$. For example, in the regression problem of Example 1, K_i is always a rank 1 matrix, so that $K_i \not\succ 0$ when $d > 1$.

Consider the following N statistics of the data D^N, for $i = 1, \ldots, N$,

$$\bar{\mathbf{q}}_i := \begin{cases} (\hat{x}_N - v_i)^T \bar{K}_i (\hat{x}_N - v_i) + h_i & \text{if } K_i \prec \frac{1}{6}\sum_{\substack{\ell=1 \\ \ell\neq i}}^N K_\ell \\ +\infty & \text{otherwise,} \end{cases} \tag{2}$$

where

$$\bar{K}_i := K_i + 6K_i \left(\sum_{\substack{\ell=1 \\ \ell\neq i}}^N K_\ell \right)^{-1} K_i,$$

and let $\bar{\mathbf{q}}_{(i)}$, $i = 1, \ldots, N$, the statistics obtained by ordering the $\bar{\mathbf{q}}_i$'s. The following theorem shows that $\bar{\mathbf{q}}_{(i)}$'s are statistics with guaranteed mean coverage.

Theorem 2: Irrespective of the probability distribution F, it holds that

$$\Pr\{\mathbf{q} \le \bar{\mathbf{q}}_{(i)}\} \ge \frac{i}{N+1}, \quad i = 1, \ldots, N. \tag{3}$$

⋆

For an outline of the proof see Section IV.

A couple of remarks are in order.

Remark 1 (geometric interpretation): Statistics $\bar{\mathbf{q}}_1, \ldots, \bar{\mathbf{q}}_N$, as well as their ordered versions $\bar{\mathbf{q}}_{(1)}, \ldots, \bar{\mathbf{q}}_{(N)}$, have a straight geometric interpretation. The empirical cost $\mathbf{q}_i$ is the value of the paraboloid $(x - v_i)^T K_i (x - v_i) + h_i$ at $x = \hat{x}_N$. According to Theorem 2, the corresponding $\bar{\mathbf{q}}_i$ is obtained by evaluating at $x = \hat{x}_N$ a modified version of the paraboloid, obtained by replacing the matrix K_i with $\bar{K}_i$, see Fig. 2. The modified $\bar{K}_i$ is given by the original K_i *plus* a term whose magnitude depends on the "ratio" of K_i and $\sum_{\substack{\ell=1 \\ \ell\neq i}}^N K_\ell$. If K_i is "small" with respect to $\sum_{\substack{\ell=1 \\ \ell\neq i}}^N K_\ell$, then $\bar{K}_i \approx K_i$, so that $\bar{\mathbf{q}}_i \approx \mathbf{q}_i$ (i.e. the margin is small), otherwise, $\bar{\mathbf{q}}_i$ may become large, or even infinite if $K_i \not\prec \frac{1}{6}\sum_{\substack{\ell=1 \\ \ell\neq i}}^N K_\ell$.

⋆

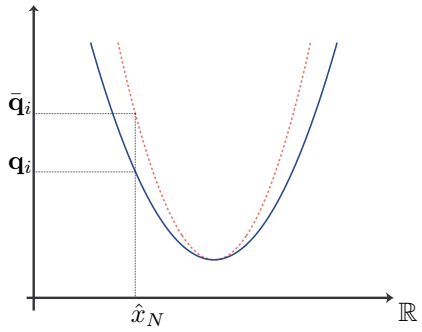


Fig. 2. The paraboloid $(x - v_i)^T K_i (x - v_i) + h_i$ associated with the i-th observation (continuous line) is compared with its modified version $(x - v_i)^T \bar{K}_i (x - v_i) + h_i$ (dashed line). The values at $x = \hat{x}_N$ are, respectively, the empirical cost $\mathbf{q}_i$ and $\bar{\mathbf{q}}_i$ as defined in (2).

Remark 2 (characterization of the margin): Under mild assumptions, as N increases the sum $\sum_{\substack{\ell=1 \\ \ell\neq i}}^N K_\ell$ becomes larger and larger compared to K_i, so that the term $K_i \left(\sum_{\substack{\ell=1 \\ \ell\neq i}}^N K_\ell \right)^{-1} K_i$ in the definition of $\bar{K}_i$ tends to zero yielding $\bar{K}_i \to K_i$ and $\bar{\mathbf{q}}_{(i)} \to \mathbf{q}_{(i)}$ for every i. Since, as it can be proven, the mean coverage of $\mathbf{q}_{(i)}$ is no more than $\frac{i}{N+1}$, this shows that the statistics $\bar{\mathbf{q}}_{(i)}$ are not conservative. The following Examples 3 and 4 illustrate this fact.

⋆

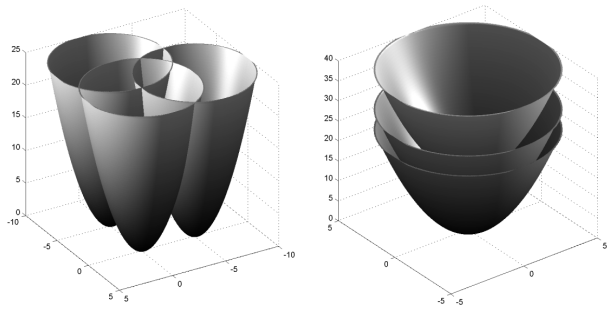
Example 3 (paraboloids with coplanar vertexes): Assume that $A_i = I$, $i = 1, \ldots, N$, yielding $K_i = I$, $v_i = b_i$, $h_i = 0$. See Fig. 3(a) for a visualization of the costs $\|A_i x - b_i\|^2$.

In this case $K_i \prec \frac{1}{6}\sum_{\ell\neq i} K_\ell \iff N \ge 8$, and

$$\bar{K}_i = \frac{N+5}{N-1} I,$$
$$\bar{\mathbf{q}}_{(i)} = \frac{N+5}{N-1} \mathbf{q}_{(i)},$$

for $N \ge 8$. Clearly, the margin $\bar{\mathbf{q}}_{(i)} - \mathbf{q}_{(i)} = \frac{6}{N-1}\mathbf{q}_{(i)}$ goes to zero as $1/N$.

⋆



(a) (b)

Fig. 3. (a) the cost functions $\|A_i x - b_i\|^2$ of Example 3; (b) the cost functions of Example 4.

Example 4 (stack of paraboloids): Assume that, for $i = 1, \ldots, N$,

$$A_i = \begin{bmatrix} I_{d\times d} \\ 0_{1\times d} \end{bmatrix} \quad \text{and} \quad b_i = \begin{bmatrix} 0_{d\times 1} \\ u_i \end{bmatrix},$$

where the subscripts denote the matrix dimensions (e.g. $0_{1\times d}$ is a row vector of zeros) and $u_1, \ldots, u_N$ are scalar values. In this case, $K_i = I_{d\times d}$, $v_i = 0$, $h_i = u_i^2$, and the cost functions $\|A_i x - b_i\|^2$ are as depicted in Fig. 3(b). As before, $\frac{1}{6}\sum_{\substack{\ell=1 \\ \ell\neq i}}^{N} K_\ell \succ K_i \iff N \geq 8$, while

$$\bar{K}_i = \frac{N+5}{N-1} I_{d\times d},$$
$$\bar{\mathbf{q}}_{(i)} = \mathbf{q}_{(i)},$$

for $N \geq 8$. Thus, for $N \geq 8$, it holds that

$$\Pr\left\{\mathbf{q} \leq \mathbf{q}_{(i)}\right\} \geq \frac{i}{N+1},$$

i.e., *there is no margin* between $\bar{\mathbf{q}}_{(i)}$ and $\mathbf{q}_{(i)}$ (compare with Theorem 1).

$\star$

III. NUMERICAL EXAMPLE

This example deals with the location of a facility in a given geographical area, see [11].

In the basic setting, the location of the facility has to be chosen so as to minimize the squared distance between the facility and the clients. To this purpose, N clients are randomly observed and their locations $p_1, \ldots, p_N$ are recorded. The facility location $\hat{x}_N \in \mathbb{R}^2$ is computed by minimizing $\sum_{i=1}^{N} \|x - p_i\|^2$, i.e., $\hat{x}_N$ is the geometric center of $p_1, \ldots, p_N$. This estimates the geometric center of the whole unknown client population. More in general, in order to take into account some factors other than distances (such as different importance of the clients, slope of the terrain, etc.), weighting matrices $A_1, \ldots, A_N$, each depending on the specific client observed, can be introduced, and $\hat{x}_N$ is then obtained as the minimizer of $\sum_{i=1}^{N} \|A_i(x - p_i)\|^2$. In order to obtain an evaluation of the performance of $\hat{x}_N$ with respect to the whole population of clients, we resort to the theory developed in Section II.

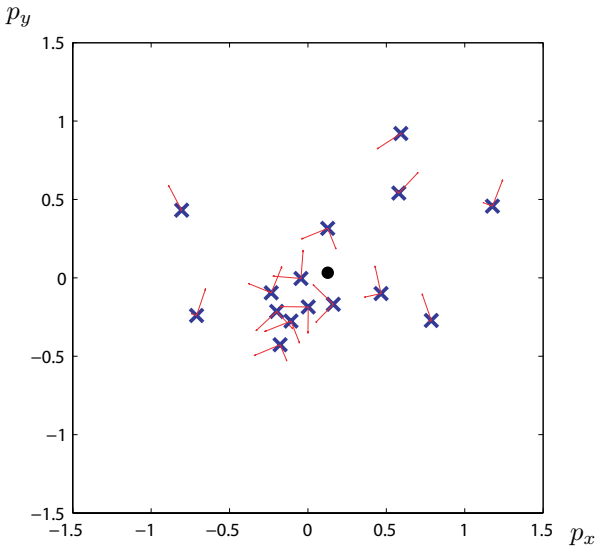


Fig. 4. Points $p_1, \ldots, p_{15} \in \mathbb{R}^2$ are represented by crosses. The eigenvectors of the matrix $K_i = A_i^T A_i$ associated with each point are also shown. The bold black dot is $\hat{x}_N$.

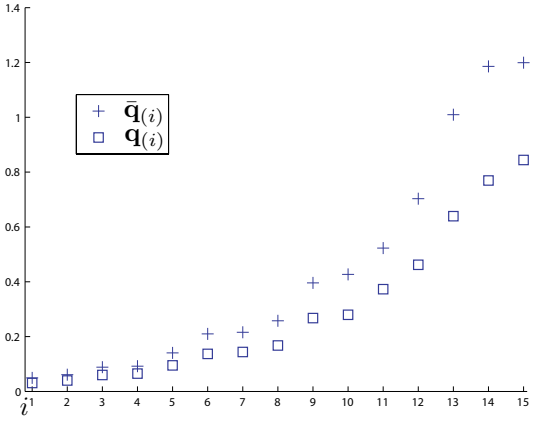


Fig. 5. $\mathbf{q}_{(i)}$ and $\bar{\mathbf{q}}_{(i)}$ for the facility location example.

A. Simulation setting

We generated a sample of $N = 15$ data from a population whose density function is a bivariate normal distribution with mean $(0, 0)$ and covariance matrix $\frac{1}{4}I$. The weighting matrix A associated with a client at point $p = (p_x, p_y)$ is as follows:

$$A = \begin{bmatrix} 1 - p_x^2 & p_x p_y \\ p_x p_y & 1 - p_y^2 \end{bmatrix}.$$

In Fig. 4, the obtained data sample and the computed estimate $\hat{x}_N$ are shown. Fig. 5 shows the ordered empirical costs $\mathbf{q}_{(1)}, \ldots, \mathbf{q}_{(N)}$ as well as $\bar{\mathbf{q}}_{(1)}, \ldots, \bar{\mathbf{q}}_{(N)}$ computed according to Theorem 2.

The *actual* mean coverages of $\bar{\mathbf{q}}_{(1)}, \ldots, \bar{\mathbf{q}}_{(N)}$ computed through a Monte-Carlo simulation are reported in Fig. 6. Note that $\Pr\{\mathbf{q} \leq \bar{\mathbf{q}}_{(i)}\} \geq \frac{i}{N+1}$, for any $i = 1, \ldots, 15$, in agreement with Theorem 2.

IV. PROOF OF THEOREM 2

Theorem 2 follows from Theorem 3 stated below.

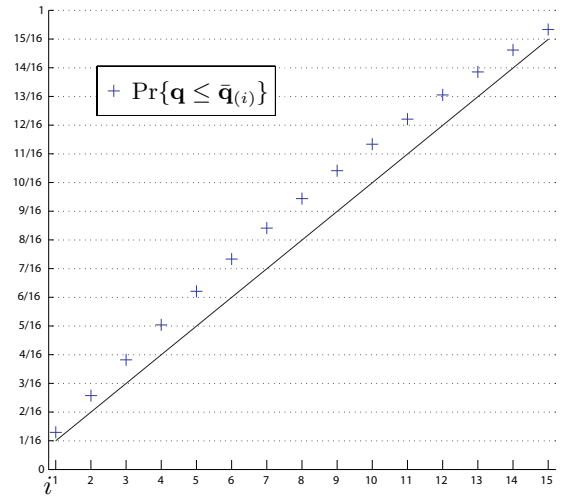


Fig. 6. The actual mean coverages of $\bar{\mathbf{q}}_{(i)}$.

Matrices K_i, $i = 1, \ldots, N$, are defined in Section II as $K_i = A_i^T A_i$. Thus, the K_i's are symmetric and positive semi-definite. Throughout, the simplified notations are in use

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

We start with a Lemma, whose proof follows from standard linear algebra.

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

$$K_i^{\frac{1}{2}} \left(\sum_{\ell \neq i} K_\ell \right)^{-1} K_i^{\frac{1}{2}} \prec \gamma I \iff K_i \prec \gamma \sum_{\ell \neq i} K_\ell, \tag{4}$$

and

$$K_i^{\frac{1}{2}} \left(\sum_{\ell \neq i} K_\ell \right)^{-1} K_i^{\frac{1}{2}} \preceq \gamma I \iff K_i \preceq \gamma \sum_{\ell \neq i} K_\ell. \tag{5}$$

$\star$

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

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

and

$$W_i := K_i + (4 + 2\gamma_i) K_i \left(\sum_{\ell \neq i} K_\ell \right)^{-1} K_i. \tag{6}$$

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

$$K_i^{\frac{1}{2}} \left(\sum_{\ell \neq i} K_\ell \right)^{-1} K_i^{\frac{1}{2}} \preceq \gamma_i I, \tag{7}$$

and, hence,

$$\begin{aligned} W_i &= K_i + (4 + 2\gamma_i) K_i^{\frac{1}{2}} \left(K_i^{\frac{1}{2}} \left(\sum_{\ell \neq i} K_\ell \right)^{-1} K_i^{\frac{1}{2}} \right) K_i^{\frac{1}{2}} \\ &\preceq K_i + (4 + 2\gamma_i) \gamma_i K_i = (1 + 4\gamma_i + 2\gamma_i^2) K_i. \end{aligned} \tag{8}$$

Applying Lemma 1 to (7) 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 (8) yields

$$\begin{aligned} W_i &\preceq (1 + 4\gamma_i + 2\gamma_i^2) \frac{\gamma_i}{1 + \gamma_i} \sum K_\ell \prec [\text{since } \gamma_i < \frac{1}{\sqrt{2}}] \\ &\prec 2 \sum K_\ell, \end{aligned} \tag{9}$$

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

$$\tilde{\mathbf{q}}_i := \begin{cases} (\hat{x}_N - v_i)^T \tilde{K}_i (\hat{x}_N - v_i) + h_i & \text{if } \sum_{\ell \neq i} K_\ell \succ 0 \\ & \text{and } \gamma_i < \frac{1}{\sqrt{2}} \\ +\infty & \text{otherwise.} \end{cases} \tag{10}$$

Theorem 3: Irrespective of the probability distribution F, it holds that

$$\Pr\{\mathbf{q} \leq \tilde{\mathbf{q}}_{(i)}\} \geq \frac{i}{N+1}.$$

$\star$

The proof of Theorem 3 is given in [24]. We show here that Theorem 2 follows from Theorem 3. To this end, it is enough to show that $\tilde{\mathbf{q}}_i \leq \bar{\mathbf{q}}_i$, $i = 1, \ldots, N$. 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 3. We now show that, if $\gamma_i < \frac{1}{6}$, then $\tilde{K}_i \preceq \bar{K}_i$, from which $\tilde{\mathbf{q}}_i \leq \bar{\mathbf{q}}_i$. Due to that $\gamma_i < \frac{1}{6}$, (8) 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,

$$\begin{aligned} \tilde{K}_i &= W_i + W_i \left(2\sum K_\ell - W_i \right)^{-1} W_i \\ &\preceq W_i + W_i \left(2\sum_{\ell \neq i} K_\ell \right)^{-1} W_i \\ &= [\text{substitute (6) for } W_i \text{ and let} \\ &\qquad \Phi := K_i^{\frac{1}{2}} \left(\sum_{\ell \neq i} K_\ell \right)^{-1} K_i^{\frac{1}{2}}] \end{aligned}$$

$$= K_i + K_i^{\frac{1}{2}} \left(\frac{9+4\gamma_i}{2}\Phi + (4+2\gamma_i)\Phi^2 + 2(2+\gamma_i)^2\Phi^3 \right) K_i^{\frac{1}{2}}$$

$$\preceq [\text{since } \Phi \preceq \gamma_i I]$$

$$\preceq K_i + K_i^{\frac{1}{2}} \left(\frac{9+4\gamma_i}{2}\Phi + (4+2\gamma_i)\gamma_i\Phi + 2(2+\gamma_i)^2\gamma_i^2\Phi \right) K_i^{\frac{1}{2}}$$

$$= K_i + (4.5+6\gamma_i+10\gamma_i^2+8\gamma_i^3+2\gamma_i^4) K_i \left(\sum_{\ell \neq i} K_\ell \right)^{-1} K_i$$

$$\preceq [\text{since } 4.5+6\gamma_i+10\gamma_i^2+8\gamma_i^3+2\gamma_i^4 < 6 \text{ for } \gamma_i < \frac{1}{6}]$$

$$\preceq \bar{K}_i.$$

Wrapping up, if $K_i \prec \frac{1}{6}\sum_{\ell \neq i} K_\ell$, then $\tilde{K}_i \preceq \bar{K}_i \implies \tilde{\mathbf{q}}_i \leq \bar{\mathbf{q}}_i \implies$ Theorem 2 follows from Theorem 3.

REFERENCES

[1] S. S. Wilks, "Order statistics," *Bulletin of the American Mathematical Society*, vol. 54, no. 1, pp. 6–50, 1948.

[2] D. A. S. Fraser and I. Guttman, "Tolerance regions," *The Annals of Mathematical Statistics*, vol. 27, no. 1, pp. 162–179, 1956.

[3] S. S. Wilks, "Determination of sample sizes for setting tolerance limits," *The Annals of Mathematical Statistics*, vol. 12, no. 1, pp. 91–96, 1941.

[4] H. Scheffé and J. W. Tukey, "Non-parametric estimation. I. Validation of order statistics," *The Annals of Mathematical Statistics*, vol. 16, no. 2, pp. 187–192, 1945.

[5] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 1985.

[6] S. L. Campbell and C. D. Meyer, "The Moore-Penrose or generalized inverse," in *Generalized Inverses of Linear Transformations*. Philadelphia, Pennsylvania, USA: SIAM, 2009.

[7] R. L. Plackett, "Studies in the history of probability and statistics. XXIX: The discovery of the method of least squares," *Biometrika*, vol. 59, no. 2, pp. 239–251, 1972.

[8] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. New York, USA: Springer-Verlag New York, LLC, 2009.

[9] L. Ljung, *System Identification - Theory For the User*. Upper Saddle River, New Jersey, USA: Prentice Hall, 1999.

[10] B. D. O. Anderson and J. B. Moore, *Optimal Control: Linear Quadratic Methods*. Englewood Cliffs, New Jersey, USA: Prentice Hall, 1990.

[11] Z. Drezner, K. Klamroth, A. Schbel, and G. O. Wesolowsky, "The Weber problem," in *Facility location - applications and theory*, Z. Drezner and H. Hamacher, Eds. New York, USA: Springer-Verlag New York, LLC, 2002.

[12] E. L. Lehmann and G. Casella, *Theory of point estimation*, 2nd ed. Springer, 1998.

[13] G. C. Calafiore and F. Dabbene, "Near optimal solutions to least-squares problems with stochastic uncertainty," *Systems and Control Letters*, vol. 54, no. 12, pp. 1219–1232, 2005.

[14] M. Vidyasagar, *A Theory of Learning and Generalization*. London, UK: Springer-Verlag, 1997.

[15] T. Alamo, R. Tempo, and E. F. Camacho, "Randomized strategies for probabilistic solutions of uncertain feasibility and optimization problems," *IEEE Transactions on Automatic Control*, vol. 54, no. 11, pp. 2545–2559, 2009.

[16] R. Tempo, G. Calafiore, and F. Dabbene, *Randomized Algorithms for Analysis and Control of Uncertain Systems, with Applications*, 2nd ed. London, UK: Springer-Verlag, 2013.

[17] A. Carè, S. Garatti, and M. C. Campi, "Randomized min-max optimization: The exact risk of multiple cost levels," in *Proceedings of the IEEE Conference on Decision and Control*, Orlando, Florida, USA, 2011.

[18] G. C. Calafiore and M. C. Campi, "The scenario approach to robust control design," *IEEE Transactions on Automatic Control*, vol. 51, no. 5, pp. 742–753, 2006.

[19] M. C. Campi and S. Garatti, "The exact feasibility of randomized solutions of uncertain convex programs," *SIAM Journal on Optimization*, vol. 19, no. 3, pp. 1211–1230, 2008.

[20] ——, "A sampling-and-discarding approach to chance-constrained optimization: Feasibility and optimality," *Journal of Optimization Theory and Applications*, vol. 148, no. 2, pp. 257–280, 2011.

[21] S. Garatti and M. C. Campi, "Modulating robustness in control design: Principles and algorithms," *Control Systems, IEEE*, vol. 33, no. 2, pp. 36–51, 2013.

[22] L. El Ghaoui and H. Lebret, "Robust solutions to least-squares problems with uncertain data," *SIAM Journal on Matrix Analysis and Applications*, vol. 18, no. 4, pp. 1035–1064, 1997.

[23] H. Hindi and S. Boyd, "Robust solutions to l1, l2, and l-infinity uncertain linear approximation problems using convex optimization," in *Proceedings of the American Control Conference*, Philadelphia, Pennsylvania, USA, 1998.

[24] A. Carè, S. Garatti, and M. C. Campi, "Least squares estimates and the coverage of least squares costs," Brescia, Italy, Tech. Rep., 2013.