Supplementary material for the paper "A Coverage Theory for Least Squares"

Algo Carè

Institute for Computer Science and Control, Budapest, Hungary.

Simone Garatti

Politecnico di Milano, Italy.

Marco C. Campi

University of Brescia, Italy.

1. A comparison with other conformal prediction methods

Consider Example 1 in paper "A Coverage Theory for Least Squares" where, for simplicity, we take $p \in \mathbb{R}$, so that both X and Y are reals. p is uniformly distributed in $[0, 10]$ and $\rho = \exp(-p)$. $N = 99$ observations $\{(X_1, Y_1), \ldots, (X_{99}, Y_{99})\}$ are collected and displayed in Fig. s.1. We here compare two

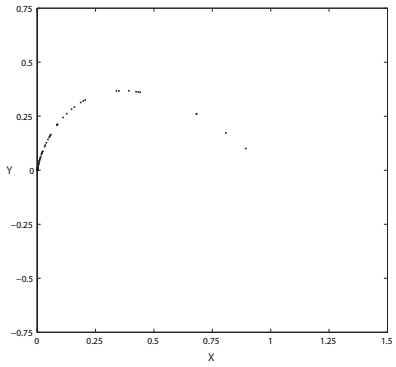
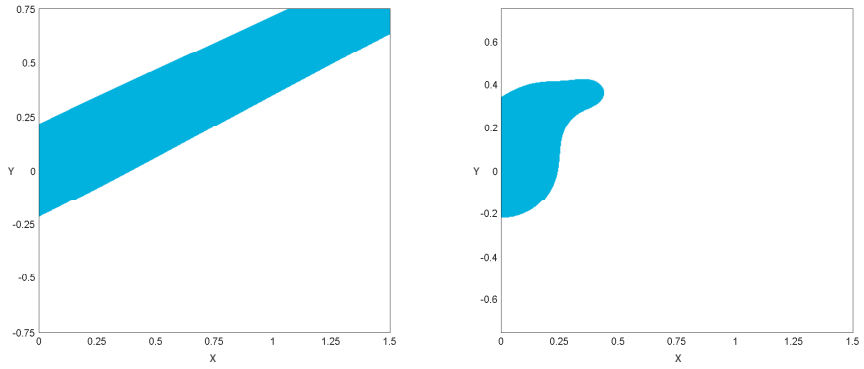


Figure s.1. The data set in the (X, Y) domain; pairs (X_i, Y_i) are represented by dots.

prediction sets in the (X, Y) domain with mean coverage 95%: that obtained with the new conformity measure introduced in paper "A Coverage Theory for Least Squares", Fig. s.2(a), and the prediction set that is constructed by the approach of Lei et al. (2013) with Gaussian kernel with covariance $10^{-2} \cdot I$, Fig. s.2(b).

In both cases, the sup of $\mathbf{q} = \|Y - X\hat{\beta}_N\|^2$ over the (X, Y) that belong to the prediction sets gives a threshold for $\mathbf{q}$ whose mean coverage is no smaller than 95%. With the 99 observations at hand, these sup are 0.0472 and 0.1184 in the two approaches, respectively.



(a) (b)

Figure s.2. (a) prediction set obtained with the new conformity measure introduced in this paper; (b) prediction set constructed by the kernel density approach of Lei et al. (2013).

One can see that the volume of our prediction set is bigger than that of the prediction set constructed using the method of Lei et al. (2013). However, the sup of the cost over our prediction set is smaller. The reason is that the conformity measure introduced in this paper has been specifically tailored to obtain small thresholds on the costs, while the conformity measure of Lei et al. (2013) aims to obtain sets with small Lebesgue volume. Hence, we see that our prediction set and the prediction set constructed using the method of Lei et al. (2013) have complementary features, and each is valuable in its own specific domain of application.

For problems in higher dimension, computing the sup of the cost over the prediction set is impractical. Theorem 1 in paper "A Coverage Theory for Least Squares" provides an easy-to-compute formula to upper bound this sup for our construction. In the present example, this formula gives a tight 0.0478.

2. Proof of the key relationship (22) in paper "A Coverage Theory for Least Squares"

Start with noting that $\sum_{l\neq i}K_l \not\succ 0$ or $\gamma_i \geq \frac{1}{\sqrt{2}}$ implies that $\tilde{\mathbf{q}}_i = \infty$, see (15) in paper "A Coverage Theory for Least Squares", and $\tilde{\mathbf{q}}_i \geq \mu_i$ is therefore clearly true. Hence, throughout what follows we assume that

$$\sum_{l\neq i}K_l \succ 0 \text{ and } \gamma_i < \frac{1}{\sqrt{2}}.$$

By substituting in equation (21) of paper "A Coverage Theory for Least Squares" the expressions for $\mathbf{m}$ and $\mathbf{m}_i$ that are given in equations (19) and (20) of the same paper, we have

$$\begin{aligned}\mu_i = \quad & \sup_{K,v,h} \mathbf{Q}(\hat{\beta}_N) \\ & \text{subject to: } \mathbf{Q}(\hat{\beta}_N) \leq \mathbf{Q}(\hat{\beta}) - \mathbf{Q}(\hat{\beta}_N) + 2\mathbf{Q}_i(\hat{\beta}^{[i]}) - \mathbf{Q}_i(\hat{\beta}),\end{aligned}$$

i.e., μ_i is computed as the supremum of $\mathbf{Q}(\hat{\beta}_N)$ over the triples (K, v, h) such that $\mathbf{Q}(\hat{\beta}_N)$ does not exceed $\mathbf{Q}(\hat{\beta}) - \mathbf{Q}(\hat{\beta}_N) + 2\mathbf{Q}_i(\hat{\beta}^{[i]}) - \mathbf{Q}_i(\hat{\beta})$. Hence,

$$\mu_i \leq \sup_{K,v,h} \left\{ \mathbf{Q}(\hat{\beta}) - \mathbf{Q}(\hat{\beta}_N) + 2\mathbf{Q}_i(\hat{\beta}^{[i]}) - \mathbf{Q}_i(\hat{\beta}) \right\}. \tag{s.1}$$

Now we write $\mathbf{Q}$ as an explicit function of the optimization variables (K, v, h) with respect to which the sup in (s.1) is computed.

Note that:

$$\begin{aligned}
\hat{\beta} &= \Big(\sum K_l + K\Big)^{-1}\Big(\sum K_l v_l + Kv\Big), \\
\hat{\beta}_N &= \Big(\sum K_l\Big)^{-1}\sum K_l v_l, \\
\hat{\beta}^{[i]} &= \left(\sum_{l\neq i} K_l + K\right)^{-1}\left(\sum_{l\neq i} K_l v_l + Kv\right).
\end{aligned}$$

Let

$$w := \hat{\beta} - v = \Big(\sum K_l + K\Big)^{-1}\Big(\sum K_l v_l + Kv\Big) - v, \tag{s.2}$$

$$w_i := \hat{\beta} - v_i = \Big(\sum K_l + K\Big)^{-1}\Big(\sum K_l v_l + Kv\Big) - v_i, \tag{s.3}$$

and note that

$$\begin{aligned}
\hat{\beta}_N - v &= \Big(\sum K_l\Big)^{-1}\sum K_l v_l - v \\
&= \Big(\sum K_l\Big)^{-1}\Big(\sum K_l v_l - \sum K_l v\Big) \\
&= \Big(\sum K_l\Big)^{-1}\Big(\sum K_l + K\Big)\Big(\sum K_l + K\Big)^{-1}\Big[\sum K_l v_l + Kv - \Big(\sum K_l + K\Big)v\Big] \\
&= \left(I + \Big(\sum K_l\Big)^{-1}K\right)\left[\Big(\sum K_l + K\Big)^{-1}\Big(\sum K_l v_l + Kv\Big) - v\right] \\
&= \left(I + \Big(\sum K_l\Big)^{-1}K\right)w,
\end{aligned}$$

and similarly that

$$\hat{\beta}^{[i]} - v_i = \left(\sum_{l\neq i} K_l + K\right)^{-1}\left(\sum_{l\neq i} K_l v_l + Kv\right) - v_i = \left(I + \left(\sum_{l\neq i} K_l + K\right)^{-1} K_i\right) w_i.$$

Using these expressions in the definitions (16) and (17) of $\mathbf{Q}_i$ and $\mathbf{Q}$ in paper "A Coverage Theory for Least Squares" yields

$$\begin{aligned}
\mathbf{Q}(\hat{\beta}) &= w^T K w + h, \\
\mathbf{Q}(\hat{\beta}_N) &= w^T\left(I + \Big(\sum K_l\Big)^{-1}K\right)^T K\left(I + \Big(\sum K_l\Big)^{-1}K\right)w + h, \\
\mathbf{Q}_i(\hat{\beta}^{[i]}) &= w_i^T\left(I + \left(\sum_{l\neq i} K_l + K\right)^{-1} K_i\right)^T K_i\left(I + \left(\sum_{l\neq i} K_l + K\right)^{-1} K_i\right)w_i + h_i, \\
\mathbf{Q}_i(\hat{\beta}) &= w_i^T K_i w_i + h_i.
\end{aligned}$$

Now, substituting in (s.1) and noting that h is canceled when taking the difference between $\mathbf{Q}(\hat{\beta})$ and $\mathbf{Q}(\hat{\beta}_N)$, we have that (recall that K and K_i are symmetric)

$$\begin{aligned}\mu_i &\leq \sup_{K,v}\left\{ w^T\left(K-\left(I+\left(\sum K_l\right)^{-1}K\right)^T K\left(I+\left(\sum K_l\right)^{-1}K\right)\right)w \right.\\ &\quad \left. +w_i^T\left(2\left(I+\left(\sum_{l\neq i}K_l+K\right)^{-1}K_i\right)^T K_i\left(I+\left(\sum_{l\neq i}K_l+K\right)^{-1}K_i\right)-K_i\right)w_i+h_i\right\}\\ &= \sup_{K,v}\left\{-w^TK\left(\sum K_l\right)^{-1}Z(K)\left(\sum K_l\right)^{-1}Kw+w_i^TV(K)w_i+h_i\right\},\end{aligned} \tag{s.4}$$

where

$$Z(K) := 2\sum K_l + K, \tag{s.5}$$

$$V(K) := 2\left(I+\left(\sum_{l\neq i}K_l+K\right)^{-1}K_i\right)^T K_i\left(I+\left(\sum_{l\neq i}K_l+K\right)^{-1}K_i\right)-K_i, \tag{s.6}$$

In (s.4), the dependence on K and v shows up explicitly and also implicitly through the definition of w and w_i in (s.2) and (s.3).

A simplification in the study of (s.4) is obtained by noting that (s.2) defines a one-to-one correspondence between (K,v) and (K,w). In fact, given (K,v), (s.2) permits one to compute w. On the other hand, given (K,w), v is computed through relationship

$$v = \left(\sum K_l\right)^{-1}\sum K_lv_l - \left(I+\left(\sum K_l\right)^{-1}K\right)w. \tag{s.7}$$

Therefore, the supremum with respect to K,v in (s.4) is equivalent to the supremum with respect to K,w provided that w_i is also written as a function of K,w. This is easily done by substituting in (s.3) the expression (s.7) for v, so obtaining

$$\begin{aligned}w_i &= \left(\sum K_l+K\right)^{-1}\left(\sum K_lv_l+K\left[\left(\sum K_l\right)^{-1}\sum K_lv_l-\left(I+\left(\sum K_l\right)^{-1}K\right)w\right]\right)-v_i\\ &= \left(\sum K_l+K\right)^{-1}\left(I+K\left(\sum K_l\right)^{-1}\right)\sum K_lv_l\\ &\quad -\left(\sum K_l+K\right)^{-1}\left(K+K\left(\sum K_l\right)^{-1}K\right)w-v_i\\ &= \left(\sum K_l+K\right)^{-1}\left(\sum K_l+K\right)\left(\sum K_l\right)^{-1}\sum K_lv_l\\ &\quad -\left(\sum K_l+K\right)^{-1}\left(\sum K_l+K\right)\left(\sum K_l\right)^{-1}Kw-v_i\\ &= \left(\sum K_l\right)^{-1}\sum K_lv_l-v_i-\left(\sum K_l\right)^{-1}Kw\\ &= \hat{\beta}_N-v_i-\left(\sum K_l\right)^{-1}Kw.\end{aligned}$$

Substituting this expression for w_i in (s.4) we obtain (recall that K and K_i are symmetric)

$$\mu_i \leq \sup_{K,w} \left\{ w^T K \Big(\sum K_l\Big)^{-1} (V(K) - Z(K)) \Big(\sum K_l\Big)^{-1} Kw \right.$$
$$\left. - 2\Big(\hat{\beta}_N - v_i\Big)^T V(K) \Big(\sum K_l\Big)^{-1} Kw + \Big(\hat{\beta}_N - v_i\Big)^T V(K) \Big(\hat{\beta}_N - v_i\Big) + h_i \right\}.$$

Now, letting

$$A(K) := V(K) - Z(K), \tag{s.8}$$
$$B(K) := -2\Big(\hat{\beta}_N - v_i\Big)^T V(K), \tag{s.9}$$
$$C(K) := \Big(\hat{\beta}_N - v_i\Big)^T V(K) \Big(\hat{\beta}_N - v_i\Big) + h_i, \tag{s.10}$$

we have that

$$\mu_i \leq \sup_{K,w} \left\{ w^T K \Big(\sum K_l\Big)^{-1} A(K) \Big(\sum K_l\Big)^{-1} Kw + B(K) \Big(\sum K_l\Big)^{-1} Kw + C(K) \right\},$$

which, with the notation $y := (\sum K_l)^{-1} Kw$, implies

$$\begin{aligned} \mu_i &\leq \sup_{K,w} \left\{ y^T A(K) y + B(K) y + C(K) \right\} \\ &\leq \sup_{K,y} \left\{ y^T A(K) y + B(K) y + C(K) \right\}. \end{aligned} \tag{s.11}$$

In the next Fact 1 it is proven that $A(K) \prec 0$. As a consequence, the quadratic form $y^T A(K) y + B(K) y + C(K)$ admits a unique maximizer as a function of y.

FACT 1. *If* $\sum_{l \neq i} K_l \succ 0$ *and* $\gamma_i < \frac{1}{\sqrt{2}}$, *then* $A(K) \prec 0$, $\forall K \succeq 0$.

$\star$

PROOF. From (s.8), (s.5), and (s.6) we have that

$$\begin{aligned} A(K) &= 2\left(I + \left(\sum_{l \neq i} K_l + K \right)^{-1} K_i \right)^T K_i \left(I + \left(\sum_{l \neq i} K_l + K \right)^{-1} K_i \right) - K_i - 2\sum K_l - K \\ &= 2K_i^{\frac{1}{2}} \left(I + K_i^{\frac{1}{2}} \left(\sum_{l \neq i} K_l + K \right)^{-1} K_i^{\frac{1}{2}} \right)^2 K_i^{\frac{1}{2}} - K_i - 2\sum K_l - K \\ &\preceq 2K_i^{\frac{1}{2}} \left(I + K_i^{\frac{1}{2}} \left(\sum_{l \neq i} K_l + K \right)^{-1} K_i^{\frac{1}{2}} \right)^2 K_i^{\frac{1}{2}} - K_i - 2\sum K_l. \end{aligned} \tag{s.12}$$

Observe that

$$I + K_i^{\frac{1}{2}} \left(\sum_{l \neq i} K_l + K \right)^{-1} K_i^{\frac{1}{2}} \preceq I + K_i^{\frac{1}{2}} \left(\sum_{l \neq i} K_l \right)^{-1} K_i^{\frac{1}{2}}$$
$$\prec \left(1 + \frac{1}{\sqrt{2}} \right) I,$$

where the last inequality follows from (12) in paper "A Coverage Theory for Least Squares" using the assumption that $\gamma_i < \frac{1}{\sqrt{2}}$. Hence,[1]

$$\left(I + K_i^{\frac{1}{2}} \left(\sum_{l \neq i} K_l + K\right)^{-1} K_i^{\frac{1}{2}}\right)^2 \prec \left(1 + \frac{1}{\sqrt{2}}\right)^2 I.$$

Substituting in (s.12), the conclusion is drawn that

$$A(K) \preceq 2\left(1 + \frac{1}{\sqrt{2}}\right)^2 K_i - K_i - 2\sum K_l = 2(\sqrt{2}+1)K_i - 2\sum K_l \prec 0,$$

where the last inequality follows since $\gamma_i < \frac{1}{\sqrt{2}}$ implies that $K_i^{\frac{1}{2}} \left(\sum_{l \neq i} K_l\right)^{-1} K_i^{\frac{1}{2}} \prec \frac{1}{\sqrt{2}} I$, from which, in view of Lemma 1 in paper "A Coverage Theory for Least Squares", $K_i \prec \frac{1}{\sqrt{2}} \sum_{l \neq i} K_l$, which in turn implies that $2(\sqrt{2}+1)K_i \prec 2\sum K_l$. □

The maximizer of $y^T A(K) y + B(K) y + C(K)$ is

$$y_{\max} = -\frac{1}{2} A(K)^{-1} B(K)^T,$$

so that, for any given K,

$$\begin{aligned} &\sup_y \left\{ y^T A(K) y + B(K) y + C(K) \right\} \\ &= -\frac{1}{4} B(K) A(K)^{-1} B(K)^T + C(K) \\ &= \left(\hat{\beta}_N - v_i\right)^T \left(V(K) - V(K)(V(K) - Z(K))^{-1} V(K)\right) \left(\hat{\beta}_N - v_i\right) + h_i, \end{aligned}$$

in virtue of (s.8)-(s.10). Thus, (s.11) gives

$$\mu_i \leq \sup_K \left(\hat{\beta}_N - v_i\right)^T \left(V(K) - V(K)(V(K) - Z(K))^{-1} V(K)\right) \left(\hat{\beta}_N - v_i\right) + h_i.$$

The final step of the proof of the key relationship (22) in paper "A Coverage Theory for Least Squares" consists in showing that

$$\left(\hat{\beta}_N - v_i\right)^T \left(V(K) - V(K)(V(K) - Z(K))^{-1} V(K)\right) \left(\hat{\beta}_N - v_i\right) + h_i \leq \tilde{\mathbf{q}}_i, \quad \forall K \succeq 0. \tag{s.13}$$

We start with the following fact.

FACT 2. *Assume that* $\sum_{l \neq i} K_l \succ 0$ *and* $\gamma_i < \frac{1}{\sqrt{2}}$. *Let* $W = W^T \succeq 0$ *such that*
(i) $V(K) \preceq W$, $\forall K \succeq 0$.
(ii) $W \prec 2\sum K_l$.
Then,

$$V(K) - V(K)(V(K) - Z(K))^{-1} V(K) \preceq W + W\left(2\sum K_l - W\right)^{-1} W, \quad \forall K \succeq 0.$$

⋆

[1] Though in general $A \prec B \not\Rightarrow A^2 \prec B^2$, it holds that $A \prec I \Rightarrow A^2 \prec I$.

PROOF. Rewrite (s.6) as

$$V(K) = K_i + 2K_i \left[\left(\sum_{l \neq i} K_l + K \right)^{-1} K_i \left(\sum_{l \neq i} K_l + K \right)^{-1} + 2 \left(\sum_{l \neq i} K_l + K \right)^{-1} \right] K_i$$

to see that $V(k) \succ 0$ if $K_i \succ 0$ while $V(K) \succeq 0$ if $K_i \succeq 0$.

Suppose first that $K_i \succ 0$, so that $V(K) \succ 0$. Using (i), we have ($Z(K)$ was defined in (s.5))

$$\begin{aligned} V(K)^{-1} - Z(K)^{-1} &\succeq W^{-1} - Z(K)^{-1} \\ &\succeq W^{-1} - \Big(2 \sum K_l\Big)^{-1}. \end{aligned}$$

Because of (ii), $W^{-1} - (2\sum K_l)^{-1} \succ 0$, so that $W^{-1} - (2\sum K_l)^{-1}$ is invertible and we have

$$\big(V(K)^{-1} - Z(K)^{-1}\big)^{-1} \preceq \left(W^{-1} - \Big(2 \sum K_l\Big)^{-1} \right)^{-1}.$$

An application of the Matrix Inversion Lemma (see e.g. Hager (1989)) now gives

$$\begin{aligned} V(K) - V(K)(V(K) - Z(K))^{-1} V(K) &\preceq W - W\Big(W - 2\sum K_l\Big)^{-1} W \\ &= W + W\Big(2\sum K_l - W\Big)^{-1} W, \end{aligned}$$

which is the statement of Fact 2.

Suppose instead that $K_i \succeq 0$. Since (i) and (ii) give $V(K) \preceq W \prec 2\sum K_l$, for any $\epsilon > 0$ small enough it holds that

$$0 \prec V(K) + \epsilon I \preceq W + \epsilon I \prec 2\sum K_l.$$

Repeating the proof for the case $K_i \succ 0$ applied to $V(K) + \epsilon I$ and $W + \epsilon I$ in place of $V(K)$ and W yields

$$\begin{aligned} &V(K) + \epsilon I - (V(K) + \epsilon I)(V(K) + \epsilon I - Z(K))^{-1}(V(K) + \epsilon I) \\ &\preceq W + \epsilon I + (W + \epsilon I)\Big(2\sum K_l - W - \epsilon I\Big)^{-1}(W + \epsilon I), \end{aligned}$$

and the result is obtained by letting $\epsilon \to 0$. □

Fact 2 is now applied with $W = W_i$, where W_i is defined in equation (11) of paper "A Coverage Theory for Least Squares".

Rewrite $V(K)$ as

$$V(K) = K_i + 4K_i \left(\sum_{l \neq i} K_l + K \right)^{-1} K_i + 2K_i^{\frac{1}{2}} \left(K_i^{\frac{1}{2}} \left(\sum_{l \neq i} K_l + K \right)^{-1} K_i^{\frac{1}{2}} \right)^2 K_i^{\frac{1}{2}}.$$

Since using (12) in paper "A Coverage Theory for Least Squares" gives

$$K_i^{\frac{1}{2}}\left(\sum_{l\neq i}K_l+K\right)^{-1}K_i^{\frac{1}{2}}\preceq K_i^{\frac{1}{2}}\left(\sum_{l\neq i}K_l\right)^{-1}K_i^{\frac{1}{2}}\preceq\gamma_i I,$$

it holds that

$$V(K)\preceq K_i+4K_i\left(\sum_{l\neq i}K_l\right)^{-1}K_i+2\gamma_i K_i\left(\sum_{l\neq i}K_l\right)^{-1}K_i=W_i,\quad\forall K\succeq 0,$$

which is (i) in Fact 2. Relation (ii) in Fact 2 is also true because of equation (14) in paper "A Coverage Theory for Least Squares". Thus, Fact 2 gives

$$V(K)-V(K)(V(K)-Z(K))^{-1}V(K)\preceq W_i+W_i\Big(2\sum K_l-W_i\Big)^{-1}W_i,\quad\forall K\succeq 0,$$

and we conclude that

$$\begin{aligned}&\Big(\hat\beta_N-v_i\Big)^T\Big(V(K)-V(K)(V(K)-Z(K))^{-1}V(K)\Big)\Big(\hat\beta_N-v_i\Big)+h_i\\ \leq\ &\Big(\hat\beta_N-v_i\Big)^T\Big(W_i+W_i\Big(2\sum K_l-W_i\Big)^{-1}W_i\Big)\Big(\hat\beta_N-v_i\Big)+h_i=\tilde{\mathbf{q}}_i,\end{aligned}$$

which is (s.13). This concludes the proof of the key relationship (22) in paper "A Coverage Theory for Least Squares". □

References

Hager, W. W. (1989) Updating the inverse of a matrix. *SIAM Review*, **31**, 221–239.

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