Robust oracle estimation and uncertainty quantification for possibly sparse quantiles

Eduard Belitser, Paulo Serra and Alexandra Vegelien

Department of Mathematics, VU Amsterdam, Amsterdam, The Netherlands

ABSTRACT
A general many quantiles + noise model is studied in the robust formulation (allowing non-normal, non-independent observations), where the identifiability requirement for the noise is formulated in terms of quantiles rather than the traditional zero expectation assumption. We propose a penalisation method based on the quantile loss function with appropriately chosen penalty function making inference on possibly sparse high-dimensional quantile vector. We apply a local approach to address the optimality by comparing procedures to the oracle sparsity structure. We establish that the proposed procedure mimics the oracle in the problems of estimation and uncertainty quantification (under the so-called EBR condition). Adaptive minimax results over sparsity scale follow from our local results.

1. Introduction
Since the emergence of statistical science, the classical ‘signal+noise’ paradigm for observed data is the main testbed in theoretical statistics for a vast number of methods and techniques for various inference problems and in various optimality frameworks. Many other practically important situations can be reduced to or approximated by the ‘signal+noise’ setting, capturing the statistical essence of the original (typically, more complex) model and preserving its main features in a pure form. There is a huge literature on ‘signal+noise’ setting with big variety of combinations of the main ingredients in the study. We mention the following ingredients: assumptions on the observation model (moment conditions, independence, normality, etc.); applied methodology (LSE, penalisation, shrinkage, thresholding, (empirical) Bayesian approach, projection, FDR method); studied inference problem (estimation, detection, model selection, testing, posterior contraction, uncertainty quantification, structure recovery); structural assumptions (smoothness, sparsity, clustering, shape constraints, such as monotonicity, unimodality and convexity, loss functions (expectations of powers of $\ell_2$-norm, $\ell_q$-norm; Hamming loss, etc.); optimality framework and pursued criteria (minimax rate, oracle rate, control of FDR, FNR, expectation of the loss function, exponential bounds for the large deviations of the loss functions, etc.). A small sample of relevant literature includes (Donoho, Johnstone,
Suppose we observe $X = (X_1, \ldots, X_n) \sim P^n_\theta$, where $\theta = (\theta_1, \ldots, \theta_n) \in \mathbb{R}^n$ is the parameter of interest. We shall suppress $n$ from the notation to simply write $P_\theta$ for $P^n_\theta$.

A natural example is given by $X_i \text{ind} \sim N(\theta_i, 1), i = 1, \ldots, n$, but this specific distributional assumption on the data generating process will not be imposed. The main modelling assumption in this note is that, for some fixed level $\tau \in (0, 1)$,

$$P(X_i - \theta_i \leq 0) = \tau, \quad i = 1, \ldots, n.$$

In other words, we deal with the setting many quantiles + noise, as $\theta_i$’s are the $\tau$-quantiles of the observed $X_i$’s and the goal is to make an inference on the high-dimensional parameter $\theta$. One can study and make inference on several quantiles simultaneously across different levels $\tau \in (0, 1)$. While the mean is a commonly used measure of centrality, it is heavily influenced by outliers or extreme values in the dataset. In contrast, quantiles are robust to outliers and provide information about the spread of the data as well as the location of the central values, cf., Zou and Yuan (2008), Jiang, Wang, and Bondell (2013) and Jiang, Bondell, and Wang (2014). The setting ‘many quantiles + noise’ with sparse quantile vectors arises in statistical and machine learning applications where the goal is to estimate a large number of quantiles for a high-dimensional dataset. The number of quantiles to be estimated can be very large, while the data may be sparse, meaning that only a small fraction of the entries in the data are non-zero. In such applications, quantiles may be a more sensible object of study, especially when the distribution is skewed or contains outliers. In this setting, the interest of study focuses on the effect of the parameter values on the tails of a distribution of observations, in addition to, or instead of, the centre. For example, in finance, quantiles such as the value at risk (the value at risk in finance has the meaning of quantile) or expected shortfall are commonly used to estimate the potential losses of a portfolio of investments, rather than relying solely on the mean return. Estimating a large number of quantiles is important for risk management and portfolio optimisation, as it provides a more complete picture of the potential losses or gains of a portfolio. In genomics, estimating many quantiles can help identify genes or markers that are associated with disease. In image processing, estimating quantiles can help identify regions of interest or anomalies in images. Quantiles as objects of interest occur in diverse areas, including economics, biology, meteorology; cf. Santosa and Symes (1986), Abrevaya (2002), Koenker (2005), Belloni and Chernozhukov (2011), Hulmán et al. (2015), Gabriela Ciuperca (2018) and Belloni, Chernozhukov, and Kato (2019).

Besides quantile formulation, the next distinctive feature of our study is the robust formulation in the sense that we do not assume any particular form of the distribution of the observed data. Only a mild condition (Condition C1) is imposed. The observations do not need to be normal and, in fact no specific distribution is assumed, the distribution
of the ‘noise’ terms $X_i - \theta_i$, $i = 1, \ldots, n$, may depend on $\theta$ and do not even have to be independent. This makes the applicability scope of our results rather broad.

The third aspect of our study is the local approach. It is in general impossible to make sensible inference on a high-dimensional parameter without any additional structure, either in terms of assumptions on parameter, or on the observation model, or both. In this paper, we are concerned with sparsity structure, various version of which have been predominantly studied in the literature recently. Sparsity structure means that a relative majority of the parameter entries are all equal to some fixed value (typically zero). In the local approach, we address optimality by comparing procedures to the oracle sparsity structure (think of an ‘oracle observer’ that knows the best sparsity pattern of the actual $\theta$). When applying the local approach, the idea is not to rely on sparsity as such, but rather extract all the sparsity structure present in the data, and utilise it in the aimed inference problems. The claim that a procedure performs optimally in the local sense means that it attains the oracle quality, i.e. mimics the best sparsity structure pertinent to the true parameter, whichever it is. The local result imply minimax optimality over all scales (including the traditional sparsity class $\ell_0[s]$) that satisfy a certain condition.

The final feature of our study concerns the problem of uncertainty quantification. In recent years, focus in nonparametric statistics has shifted from point estimation to uncertainty quantification, a more challenging problem, much less results on this topic are available in the literature; cf. Szabó, van der Vaart, and van Zanten (2015), Belitser (2017), van der Pas, Szabó, and van der Vaart (2017) and Belitser and Nurushev (2020). Certain negative results (cf. Li 1989; Baraud 2004; Belitser and Nurushev 2019 and the references therein) show that in general a confidence set cannot simultaneously have coverage and the optimal size uniformly over all parameter values. This makes the construction of the so-called ‘honest’ confidence sets impossible, and a strategy recently pursued in the literature is to discard a set of ‘deceptive parameters’ to ensure coverage at the remaining parameter values, while maintaining the optimal size uniformly over the whole set. In an increasing order of generality, removing deceptive parameters is expressed by imposing the conditions of self-similarity, polished tail and excessive bias restriction (EBR), see Szabó et al. (2015), Belitser (2017), van der Pas et al. (2017) and Belitser and Nurushev (2020). All the above papers deal with $\ell_2$-loss framework for Gaussian observations in ‘many means + noise’ setting (a more general case is considered in Belitser and Nurushev 2020).

In this paper, we work with a new setting ‘many quantiles + general noise’, and use the quantile loss rather than the traditional $\ell_2$-loss. The quantile loss function has certain properties of usual loss functions, but it is not symmetric, the peculiarity essentially characterising the notion of quantile. We propose and exploit a new version of EBR condition expressed in terms of quantile loss. To the best of our knowledge, there are no results on uncertainty quantification neither for the quantile loss function, nor for the $\ell_1$-loss, also neither in local, nor global (minimax) formulation in this general robust setting. In this respect, our study presents the first step in this direction.

We finally summarise the main contributions of this paper.

- For our new (robust) setting ‘many quantiles + general noise’ we propose a penalisation procedure for selecting the sparsity pattern, based on the quantile loss (as counterpart of the $\ell_2$-loss) and the corresponding penalty term.
• The estimated sparsity pattern is next used for the construction of an estimator of $\theta$ and a confidence set for $\theta$, again in terms of the quantile loss function.
• Two theorems establish the local (oracle) optimality of the estimator and confidence set (under the EBR condition), respectively.
• We provide an elegant and relatively short proofs of the main results; compare with the relatively laborious proofs of related results in Szabó et al. (2015), Belitser (2017), van der Pas et al. (2017) and Belitser and Nurushev (2020).
• The obtained results are robust and local in the sense explained above.
• The obtained local results imply adaptive minimax optimality for scales of classes that satisfy certain relation (in particular, for the traditional sparsity scale).

Organisation of the rest of the paper is as follows. In Section 2, we give a robust formulation of the observation model in the 'many quantiles + general noise' setting, and introduce some notations and preliminaries. In Section 3, we present the main results of the paper and discuss their consequences for the optimality in the minimax sense for the traditional sparsity scale. Section 4 contains a small simulation study. The proofs of the theorems are provided in Section 5.

2. Robust model formulation and preliminaries

We can formally rewrite the model stated in the introduction in the familiar ‘signal+noise’ form:

$$X_i = \theta_i + \xi_i, \quad \mathbb{P}_\theta(\xi_i \leq 0) = \tau, \quad i \in [n] = \{1, \ldots, n\}. \quad (1)$$

where $\xi_i = X_i - \theta_i$ can be thought of as ‘noise’. Note that all the quantities depend of course on $\tau$. Precisely, for $\tau \in (0, 1)$, in the model (1) we have $\theta = \theta(\tau), \xi = \xi(\tau)$ with $\mathbb{P}_\theta(\xi_i(\tau) \leq 0) = \tau, i = 1, \ldots, n$. Not to overload the notation, we suppress the dependence on $\tau$ in the sequel. As we mentioned in the Introduction, we study robust setting in the sense that we do not assume any particular distribution of $\xi = (\xi_1, \ldots, \xi_n)$. In fact, the $\xi_i$’s do not have to be identically distributed, their distribution may depend on $\theta$ and they do not even have to be independent.

Introduce the following asymmetric absolute deviation function:

$$\rho(x) = \rho_\tau(x) = \rho_{\tau,m}(x) = \sum_{i=1}^{m} x_i(\tau - 1\{x_i \leq 0\}), \quad x \in \mathbb{R}^{m}, \quad m \in \mathbb{N},$$

which we will also call quantile loss function, $1\{E\}$ denotes the indicator of event $E$.

Slightly abusing notation, we use the same notation $\rho(x)$ for any $m \in \mathbb{N}$, for example, most of the time it will be either $m = n$ or $m = 1$, depending on the dimensionality of the argument of the function. This will always be clear from the context. Often, we will suppress the dependence of $\rho$ on a fixed $\tau \in (0, 1)$, unless emphasising this dependence where it is relevant.

Remark 2.1: The origin of the quantile loss function $\rho_\tau(x), x \in \mathbb{R}$, is well explained in the book (Koenker 2005). The function $\rho_\tau(x)$ characterises the $\tau$-quantile of a random variable $Z$ in the sense that the $\tau$-quantile $\theta_\tau$ of $Z$ is known to minimise the criterion
function $E\rho_\tau(Z - \vartheta)$ with respect to $\vartheta$, i.e. arg min$_{\vartheta \in \mathbb{R}} E\rho_\tau(Z - \vartheta) = \theta_\tau$. Basically, this function plays the same role for characterising the quantiles as the quadratic function in characterising the expectation of a random variable.

The quantile loss function $\rho$ can be related to the $\ell_1$-criterion $|x| = \sum_{i=1}^m |x_i|$, $x = (x_1, \ldots, x_m) \in \mathbb{R}^m$: for any $\tau \in (0, 1)$,
\begin{equation}
(1 - c_\tau)|x| \leq \rho_\tau(x) \leq c_\tau|x|, \quad \text{with} \quad c_\tau = \max\{\tau, 1 - \tau\}.
\end{equation}

In particular, it follows that $\rho_{1/2}(x) = |x|/2$. Another useful property of the quantile loss function $\rho$ to be used later on is that, for any $x \in \mathbb{R}^m$,
\begin{equation}
\rho_\tau(-x) \leq C_\tau \rho_\tau(x), \quad \text{with} \quad C_\tau = \max\{1 - \tau, \tau\} = \frac{c_\tau}{1 - c_\tau}.
\end{equation}

**Remark 2.2:** The above quantile loss function evaluated at the difference $\rho(x - y)$, $x, y \in \mathbb{R}^m$, possesses all the properties of a metric $d(x, y)$ except for the symmetry. In particular, $\rho(0) = 0$ (zero at zero); if $x \neq y$, $\rho(x - y) > 0$ (positivity); and finally the triangle inequality holds
\begin{equation}
\rho(x + y) \leq \rho(x) + \rho(y), \quad x, y \in \mathbb{R}^m.
\end{equation}

Since we have as many parameters as observations, it is typically impossible to make inference on $\theta$ even in a weak sense unless the data possesses some structure. Here, we work with the structural assumption that $\theta$ is (possibly) a sparse vector. Specifically, we assume that $\theta \in L_1$, where $I \subseteq [n] = \{1, \ldots, n\}$ and $L_1 = \{x \in \mathbb{R}^n : x_i = 0, i \in I^c\}$, where $I^c = [n] \setminus I$. The structure studied here is the unknown ’true’ sparsity pattern $I_0 = I_0(\theta) \subseteq [n]$, that is, $\theta \in L_{I_0}$ and $I_0$ is the ’minimal’ sparsity structure in the sense that $I_0 = \arg \min_{I \subseteq [n] ; I \not\supseteq \theta} \sum_{i \in I} x_i$.

Introduce some further notation: let $\mathcal{I}$ denote the family of all subsets of $[n]$; for $x \in \mathbb{R}^n$, denote its $\ell_1$-norm by $|x| = \sum_{i=1}^n |x_i|$; $|S|$ denote the cardinality of a set $S$. Further let us introduce the so-called ’quantile projection’ $P_I$ onto $L_1$ (called just projection in what follows), with respect to the quantile loss function $\rho(x - y)$. For $x \in \mathbb{R}^n$, define $P_I x$ as follows:
\begin{equation}
\inf_{y \in L_1} \rho(x - y) = \rho(x - P_I x).
\end{equation}

In view of the properties of the quantile loss function $\rho$ given by Remark 2.2, the projection $P_I$ is readily found as the following linear operator:
\begin{equation}
P_I x = (x_i 1\{i \in I\}, \quad i = 1, \ldots, n) \in \mathbb{R}^n.
\end{equation}

We will use later a certain monotonicity property of the quantile projection operator.

If $I_1 \subseteq I_2$, then $\rho(P_{I_1} x) \leq \rho(P_{I_2} x), \quad x \in \mathbb{R}^n.$

For $s \geq 0$ and $I \in \mathcal{I}$, denote (with the convention $0 \log(a/0) = 0$ for $a > 0$)
\begin{equation}
\lambda(s) = s \log(en/s), \quad p(I) = (|I| \lambda(|I|))^{1/2} = |I| \left[\log(en/|I|)\right]^{1/2}.
\end{equation}

We have that $\lambda(s) \geq s$ for $s \in [0, n]$, and besides, since $\lambda(s)$ is increasing in $s \in (0, n]$, $\lambda(s) \geq 1 + \log(n)$ for all $s \in [n]$. The function $p(I)$ has a meaning of complexity of structure $I$. 
The following relation will be used later: for any $v > 1$,\[\sum_{I \in \mathcal{I}} e^{-v \lambda(|I|)} = 1 + \sum_{s \in [n]} \sum_{I: |I| = s} e^{-v \lambda(s)} \leq 1 + \sum_{s \in [n]} e^{-(v-1)s} \leq C_v, \tag{6}\]
with $C_v = 1 + (e^{v-1} - 1)^{-1} < \infty$.

The following conditions on $\xi = X - \theta$ is assumed throughout.

**Condition C1.** For some $H_0, M_0 > 0$ and some positive function $\psi(u)$ monotonically increasing to infinity as $u \to \infty$, and all $I \in \mathcal{I}$, $I \neq \emptyset$, all $M \geq 0$,\[
\sup_{\theta \in \mathbb{R}^n} \mathbb{P}_\theta \left( \rho(P_I \xi) > M_p(I) \right) \leq H_0 e^{-(\psi(M) - M_0)\lambda(|I|)}. \tag{7}\]

Notice that Condition C1 is trivially fulfilled for $I = \emptyset$ with zero right-hand side. Let $\mathcal{P} = \mathcal{P}(H_0, M_0, \psi)$ stand for the family of distributions $\mathbb{P}_\theta$ satisfying Condition C1.

As $\lambda(s) \geq 1 + \log(n)$ for all $s \in [n]$, the right-hand side of (7) is further bounded by $H_0 e^{-(\psi(M) - M_0)\lambda(1)} \leq H_0 n^{-(\psi(M) - M_0)}$, which we will use in the proof.

**Remark 2.3:** Condition C1 is mild, for instance, it holds for independent sub-gaussian $\xi_i$’s. Recall one of the equivalent definitions of sub-gaussianity (see Vershynin 2018): a random variable $W$ is called $\sigma$-sub-gaussian for $\sigma > 0$ if for some $c_0 > 0$ $\mathbb{E}e^{c_0 W^2/\sigma^2} \leq 2$. In our case, $\sigma = 1$. For example, for $\xi_i \overset{\text{iid}}{\sim} \mathcal{N}(0, 1)$, $c_0 = 3/8$ (see Section 4). Recall that $\|x\|^2 = \sum_{i=1}^n x_i^2$ denotes the usual squared $\ell_2$-norm of $x \in \mathbb{R}^n$. Now, since $P_I \xi$ has at most $|I|$ non-zero entries, by (2) and the Cauchy–Schwarz inequality we have $\rho(P_I \xi) \leq c_\tau |P_I \xi| \leq c_\tau |I|^{1/2} \|P_I \xi\|$ (this also follows from the following inequality between two different $\ell_q$-norms: for $x \in \mathbb{R}^n$ and $0 < q_1 < q_2 \leq \infty$, $\|x\|_{q_1} \leq \|x\|_{q_2} \leq n^{1/q_1 - 1/q_2} \|x\|_{q_2}$). Recall also that $\lambda(s) \geq s$, $s \in [0, n]$. Using these and the Markov inequality, we obtain that, for $H_0 = 1$, $\psi(M) = c_0 c_\tau^2 M^2$ and $M_0 = \log 2$,
\[
\mathbb{P} \left( \rho(P_I \xi) \geq M_p(I) \right) \leq \mathbb{P} \left( c_\tau |I|^{1/2} \|P_I \xi\|_2 \geq M_p(I) \right) = \mathbb{P} \left( \|P_I \xi\|_2 \geq M e^{\{\lambda(|I|)\}^{1/2}} \right) \leq \mathbb{P} \left( c_0 \|P_I \xi\|_2^2 \geq c_0 c_\tau^{-2} M^2 \lambda(|I|) \right) \leq e^{\|I| \log 2 - c_0 c_\tau^{-2} M^2 \lambda(|I|)} \leq e^{-(\psi(M) - M_0)\lambda(|I|)}.
\]

One can extend the results to the case of the so-called sub-exponential errors, with adjusted complexity function $p(I)$; see Remark 3.3.

**Remark 2.4:** Condition C1 is clearly satisfied for bounded, arbitrarily dependent, $\xi_i$’s. This condition allows some interesting cases of dependent $\xi_i$’s. By arguing in the same way as in Belitser and Nurushev (2020), we can establish that Condition C1 holds also for $\xi_k$’s which follow an autoregressive model AR(1) with sub-gaussian white noise $\epsilon_k$’s (for appropriately chosen model parameters):
\[\xi_k = \alpha \xi_{k-1} + \epsilon_k, \quad k \in [n]; \quad \xi_0 = 0, \quad |\alpha| < 1.\]
In Belitser and Nurushev (2020), the normal white noise was used in the above model AR(1), but this can easily be extended to the sub-gaussian case by adjusting the constants involved.
Remark 2.5: We can extend our setting by including also a parameter \( \sigma > 0 \) by considering \( \sigma \xi_i \) instead of just \( \xi_i \) (and \( (X_i - \theta)/\sigma \) instead of \( X_i - \theta \)) in (1) and in Condition C1. In that case, the parameter \( \sigma > 0 \) is assumed to be known and fixed throughout. Together with \( n \), it reflects the information amount in the model in the sense that \( \sigma \to 0 \) means the flux of information.

We propose a penalised projection estimator. For \( \kappa > 0 \), define \( \hat{I} = \hat{I}(X, \kappa) \) to be a minimiser of the criterion \( C(I) = C(I, \kappa) = C(I, \kappa, X) \):

\[
C(I) = \rho(X - P_I X) + \kappa p(I) = \rho(P_{\hat{I}} X) + \kappa p(I),
\]

\[
\min_{I \in \mathcal{I}} C(I) = C(\hat{I}).
\] (8)

Remark 2.6: Since the proposed procedure \( \hat{I} \) for selecting sparsity pattern is based on the non-symmetric quantile loss function \( \rho \), the deviations from below and from above are treated differently. But the computation of the estimator \( \hat{I} \) of the sparsity pattern is not difficult. Indeed, it can be reduced to the search over \( n \) options: with \( b_k = \rho(X_k) \geq 0, k \in \{n\} \),

\[
\min_{I \in \mathcal{I}} C(I) = \min_{I \in \mathcal{I}} \left\{ \rho(P_I X) + \kappa p(I) \right\} = \min_{I \in \mathcal{I}} \left\{ \sum_{k \in I^c} \rho(X_k) + \kappa \log \log \left( \frac{en}{T} \right) \right\}
\]

\[
= \min_{i \in \{n\}} \left\{ \sum_{k \in I^c} b_k + \kappa \log \log \left( \frac{en}{T} \right) \right\} = \min_{i \in \{n\}} \left\{ B_i + \kappa \log \log \left( \frac{en}{T} \right) \right\},
\]

where \( b_{(1)} \leq \cdots \leq b_{(n)} \) is the ordered sequence of \( b_k \)'s and \( B_i = \sum_{k=1}^{n-i} b_{(k)} \).

Now, using the estimated sparsity structure \( \hat{I} \), define the estimator

\[
\hat{\theta} = \hat{\theta}(X, \kappa) = P_{\hat{I}} X.
\] (9)

From the triangle inequality (4), it follows that \( \rho(x - y) \geq \rho(x) - \rho(y) \). Using this, (3) and the definition (8), we derive that, for any \( I' \in \mathcal{I} \),

\[
\kappa (\rho(\hat{I}) - \rho(I')) \leq \rho(P_{I' l} X) - \rho(P_{I' \backslash l} X) = \rho(P_{I' l} X) - \rho(P_{I' \backslash l} X)
\]

\[
\leq \rho(P_{I' \backslash l} \theta) + \rho(\hat{P} X) - \rho(\hat{P} \backslash l \theta) + \rho(\hat{P} \backslash l (-\xi))
\]

\[
\leq \rho(\hat{P} \backslash l \theta) - \rho(\hat{P} \backslash l \theta) + \rho(\hat{P} \backslash l \xi) + C_T \rho(\hat{P} \xi) \]. (10)

For \( \varkappa > 0, \theta \in \mathbb{R}^n \) and \( I \in \mathcal{I} \), introduce the quantity

\[
r(\theta, I) = r_{\varkappa}(\theta, I) = \rho(\theta - P_I \theta) + \varkappa p(I) = \rho(P_{I \backslash I} \theta) + \varkappa p(I),
\]

which we call the quantile rate of the sparsity structure \( I \). The oracle sparsity structure \( I_0 = I_0(\theta) = I_0(\theta, \varkappa) \) is the one minimising \( r_{\varkappa}(\theta, I) \):

\[
\min_{I \in \mathcal{I}} r_{\varkappa}(\theta, I) = r_{\varkappa}(\theta, I_0) = r_{\varkappa}(\theta) = r(\theta). \] (11)

where its minimal value \( r(\theta) \) is called the oracle quantile rate, or just oracle rate.
3. Main results

In this section, we give the main results. All the constants in the below assertions depend on the fixed constants \( M_0, H_0 \) and function \( \psi \) appearing in Condition C1, the constant \( \kappa \) appearing in the oracle structure definition (8), and the quantile level \( \tau \).

The next result concerns the estimation problem.

**Theorem 3.1 (Estimation):** Let Condition C1 be fulfilled. Then for any \( \kappa > 0 \) and sufficiently large \( \kappa \), there exist positive \( M_1, H_1, m_1 \) such that

\[
P_\theta \left( \rho(\theta - \hat{\theta}) \geq M_1 r(\theta) \right) \leq H_1 e^{-m_1 \lambda(1)} \leq H_1 n^{-m_1}, \quad \theta \in \mathbb{R}^n,
\]

where \( \hat{\theta} \) is defined by (9).

Next we address the new problem of uncertainty quantification (UQ) for the parameter \( \theta \). A confidence set is defined in terms of quantile loss function \( \rho \) as follows:

\[
B(\tilde{\theta}, \tilde{r}) = \{ \theta \in \mathbb{R}^n : \rho(\theta - \tilde{\theta}) \leq \tilde{r} \},
\]

where the ‘centre’ \( \tilde{\theta} = \tilde{\theta}(X) : \mathbb{R}^n \mapsto \mathbb{R}^n \) and ‘radius’ \( \tilde{r} = \tilde{r}(X) : \mathbb{R}^n \mapsto \mathbb{R}_+ = [0, +\infty] \) are measurable functions of the data \( X \). The goal is to construct such a confidence set \( B(\tilde{\theta}, \tilde{r}) \) that for any \( \alpha_1, \alpha_2 \in (0, 1] \) and some function \( R(\theta), R : \mathbb{R}^n \to \mathbb{R}_+, \) there exist \( C, c > 0 \) such that

\[
\sup_{\theta \in \Theta_{cov}} P_\theta \left( \theta \notin B(\tilde{\theta}, \tilde{r}) \right) \leq \alpha_1, \quad \sup_{\theta \in \Theta_{size}} P_\theta \left( \tilde{r} \geq cR(\theta) \right) \leq \alpha_2,
\]

for some \( \Theta_{cov}, \Theta_{size} \subseteq \mathbb{R}^n \). The function \( R(\theta) \), called radial rate, is a benchmark for the effective radius of the confidence set \( B(\tilde{\theta}, \tilde{r}) \). The first expression in (14) is called coverage relation and the second size relation. It is desirable to find the smallest \( r(\theta) \), the biggest \( \Theta_{cov} \) and \( \Theta_{size} \) such that (14) holds and \( R(\theta) \asymp r(\Theta_{size}) \), where \( r(\Theta_{size}) \) is the optimal rate in estimation problem for \( \theta \). In our local approach, we pursue even more ambitious goal \( R(\theta) \asymp r(\theta) \), where \( r(\theta) \) is the oracle rate from the (local) estimation problem.

Typically, the so-called deceptiveness issue arises for the UQ problem in that the confidence set of the optimal size and high coverage can only be constructed for non-deceptive parameters (in particular, \( \Theta_{cov} \) cannot be the whole set \( \mathbb{R}^n \)). That is, coverage with an optimal sized radius is only possible if certain deceptive set of parameters is excluded from consideration, which is expressed by imposing some condition on the parameter. For example, the EBR (excessive bias restriction) condition in case of \( \ell_2 \)-norm is proposed in Belitser and Nurushev (2019), Belitser (2017) and Belitser and Ghosal (2020). Here we need an EBR-like condition (we keep the same term EBR), but now in terms of the quantile loss function \( \rho \).

**Condition EBR.** We say that parameter \( \theta \in \mathbb{R}^n \) satisfies the excessive bias restriction (EBR) condition with structural parameter \( t \geq 0 \) if \( \theta \in \Theta_{eb}(t) = \Theta_{eb}(t, \kappa) \) where

\[
\Theta_{eb}(t) = \left\{ \theta \in \mathbb{R}^n : \frac{\rho(\theta - P_{\hat{\theta}} \theta)}{P(\hat{\theta})} \leq t \right\},
\]

the oracle structure \( I_o = I_o(\theta) \) is defined by (11) (with the convention \( 0/0 = 0 \)).
The extent of the restriction \( \theta \in \Theta_{eb}(t) \) varies over different choices of constant \( t \), becoming more lenient as \( t \) increases, eventually covering the entire parameter space. In general, for any \( t \), a sequence of \( \theta \)'s can be found such that \( \theta \not\in \Theta_{eb}(t) \), when \( n \) varies. On the other hand, the set \( \Theta_{eb}(0) \) is not empty and consists of such \( \theta \)'s for which the oracle \( I_0 \) coincides with the true sparsity structure \( I_0 \). For a general discussion on EBR, we refer the reader to Belitser and Nurushev (2019). For sparsity structure in \( \ell_1 \)-sense, the EBR condition is satisfied if the minimal absolute value of the non-zero coordinates of \( \theta \) is larger than a certain lower bound, but this bound depends also on the number of the non-zero coordinates of \( \theta \). For example, for some sufficiently large \( C > 0 \),

\[
\bigcup_{I \in \mathcal{I}} \left\{ \theta \in L_I : |\theta_i| \geq C \log^{1/2} \left( \frac{|n|}{|I|} \right), \ i \in I \right\} \subseteq \Theta_{eb}(0).
\]

**Theorem 3.2 (Confidence ball):** Let the confidence set \( B(\cdot, \cdot) \) be defined by (13), \( r = p(\hat{I}) \), \( \hat{I} \) and \( \hat{\theta} \) be given by (8) and (9) respectively. Then for sufficiently large \( \kappa, \varkappa \) there exist constants \( M_2 = M_2(t), H_2, m_2, M_3, H_3, m_3 > 0 \) such that for any \( t \geq 0 \),

\[
\sup_{\theta \in \Theta_{eb}(t)} \mathbb{P}_\theta \left( \theta \not\in B(\hat{\theta}, M_2(t)\hat{r}) \right) \leq H_2 n^{-m_2}, \tag{16}
\]

\[
\sup_{\theta \in \mathbb{R}^n} \mathbb{P}_\theta \left( \hat{r} \geq M_3 r(\theta) \right) \leq H_3 n^{-m_3}. \tag{17}
\]

Constants \( H_i, M_i, m_i, i = 1, 2, 3 \), are all evaluated in the proofs of Theorems 3.1 and 3.2 in the form of several bounds, which depend on function \( \psi \), constants \( H_0, M_0 \) from Condition C1, the quantile level \( \tau \) and \( \varkappa \). There is no 'optimal' choice, in fact, many choices are possible, an improvement of one constant typically lead to worsening another one. It should also be noted that the constants (any choice satisfying the bounds in the proofs) are uniform over the whole family \( \mathcal{P} \) of distributions satisfying Condition C1, and can therefore be significantly improved for specific error distributions.

So far, all results are formulated in terms of the oracle. For estimation and the construction of a confidence set, this local approach delivers the most general results, as the convergence rate of the proposed estimator is directly linked with the oracle sparsity rather than the true structure, which may not be sparse (but very close to a sparse structure). However, if the parameter has a true sparse structure, the method will attain the quality pertinent to that true sparsity as well.

To illustrate this, consider any \( \theta \in \mathbb{R}^n \). The true structure \( I_0(\theta) \) and the oracle structure \( I_0(\hat{\theta}) \) in general do not coincide, but they are related by

\[
\rho(\theta) = r(\theta, I_0) \leq r(\theta, I_0) = \varkappa p(I_0) = \varkappa|I_0| \log^{1/2}(en/|I_0|). \tag{18}
\]

The first two relations hold by the definition of the oracle and the third because \( P_{I_0} \theta = \theta \) for the true structure. Since \( \rho(\theta) = \rho(\theta - P_{I_0} \theta) + \varkappa p(I_0) \), (18) implies that \( p(I_0) \leq p(I_0) \) and hence \( |I_0| \leq |I_0| \). Further, as \( |I_0(\theta)| \leq s \) for \( \theta \in \ell_0[s] \), (18) also yields

\[
\sup_{\theta \in \ell_0[s]} \rho(\theta) \leq \varkappa s \log^{1/2}(en/s).
\]

Besides, all the constants in Theorems 3.1 and 3.2 are uniform over \( \theta \in \ell_0[s] \) and \( \mathbb{P}_\theta \in \mathcal{P} \). The next result follows from these facts and Theorems 3.1 and 3.2.
Corollary 3.1: Under the conditions of Theorems 3.1 and 3.2, with the same choice of the constants,

\[
\sup_{\theta \in \ell_0[s]} \sup_{P_\theta \in P} \mathbb{P}_\theta \left( \rho(\theta - \hat{\theta}) \geq M_1 \xi s \log^{1/2}(en/s) \right) \leq H_1 n^{-m_1},
\]

\[
\sup_{\theta \in \ell_0[s]} \sup_{P_\theta \in P} \mathbb{P}_\theta \left( \hat{r} \geq M_3 \xi s \log^{1/2}(en/s) \right) \leq H_3 n^{-m_3},
\]

and (16) holds.

We claim that the obtained convergence rate \( s \log^{1/2}(en/s) \) in terms of quantile loss function is optimal over the class \( \ell_0[s] \) in the minimax sense. More precisely, the results of Johnstone and Silverman (2004) (in Johnstone and Silverman 2004, relations (17) and (18) with \( p = 0 \) and \( q = 1 \) and (2) imply that for the normal model \( P_\theta = \bigotimes_{i \in [n]} N(\theta_i, 1) \), there exist absolute \( C_1 \) such that

\[
\inf_{\tilde{\theta}} \sup_{\theta \in \ell_0[s]} \mathbb{E}_\theta |\theta - \tilde{\theta}| \geq C_1 s \log^{1/2}(en/s).
\]

This lower bound is not quite what we need to match with our upper bound because it is formulated in terms of expectation and the \( \ell_1 \)-loss. However, it is not difficult to establish the probability version of the lower bound: there exist absolute \( C_1, C_2 > 0 \) such that

\[
\inf_{\tilde{\theta}} \sup_{\theta \in \ell_0[s]} \mathbb{P}_\theta (|\theta - \tilde{\theta}| \geq C_1 s \log^{1/2}(en/s)) \geq C_2.
\]

The relation (2) connects the \( \ell_1 \)-loss with the quantile loss, so that the above relation implies that

\[
\inf_{\tilde{\theta}} \sup_{\theta \in \ell_0[s]} \mathbb{P}_\theta (\rho(\theta - \tilde{\theta}) \geq C_1 c^{-1} s \log^{1/2}(en/s)) \geq C_2.
\]

In this form, the lower bound matches our upper bound given by Corollary 3.1, basically showing that our procedure attains the optimal rate \( s \log^{1/2}(en/s) \) for the sparsity class \( \ell_0[s] \). This also means that the size of the constructed confidence ball \( B(\hat{\theta}, M_2(t)\hat{r}) \) is also optimal over the sparsity class \( \ell_0[s] \) in the minimax sense. However, the unavoidable price for the optimality in the size relation is that the coverage relation holds uniformly over \( \Theta_{eb}(t) \), not over \( \ell_0[s] \).

Remark 3.1: Interestingly, although there are some ‘deceptive’ \( \theta \) in \( \ell_0[s] \) that are not covered by \( \Theta_{eb}(t) \), there are also some \( \theta \)'s in \( \Theta_{eb}(t) \) which do not belong to the sparsity class \( \ell_0[s] \), but for which the coverage relation holds.

Remark 3.2: We derived the adaptive (sparsity \( s \) is also unknown) minimax results over the traditional sparsity scale \( \{\ell_0[s], s \in [n]\} \) as consequence of our local oracle results. The scope of our local result is even broader, the minimax results can be derived over any scale of classes \( \{\Theta_s, s \in \mathcal{S}\} \) with the corresponding minimax rates \( R(\Theta_s) \) as longs as, for some \( c > 0 \),

\[
\sup_{\theta \in \Theta_s} r(\theta) \leq c R(\Theta_s), \quad s \in \mathcal{S}.
\]
Indeed, if the above relation is fulfilled, we immediately obtain all the claims of Corollary 3.1 with \( \Theta_s \) instead of \( \ell_0[s] \), as consequences of Theorems 3.1 and 3.2. For example, it seems possible to derive the minimax results also for the scale of the \( \ell_s \)-balls: \( \ell_s[\eta] = \{ \theta \in \mathbb{R}^n : \frac{1}{n} \sum_{i=1}^n |\theta_i|^s \leq \eta^s \}, s \in [0, 2] \), with \( \eta = \eta_n \to 0 \) as \( n \to \infty \).

**Remark 3.3:** Interestingly, in relation to Remark 2.3, Condition C1 also holds for independent sub-exponential \( \xi_i \)'s, but with the adjusted function \( p(I) = p_{\exp}(I) = \lambda(|I|) \). Recall one of the equivalent definitions of sub-exponentiality (see Vershynin 2018): a random variable \( W \) is called \( \sigma \)-sub-exponential if \( \mathbb{E} e^{c_0|W|} \leq 2 \) for some \( c_0 > 0 \). For example, for the Laplace distribution \( \xi_i \sim f_\lambda(x) = \frac{\lambda}{2} e^{-\lambda|x|} \), \( 2 \geq \mathbb{E} e^{c_0|\xi|} = \frac{\lambda}{\lambda-c_0}, c_0 < \lambda \), we take \( c_0 = \frac{\lambda}{2} \). In particular, if \( \lambda = 1 \), \( c_0 = \frac{1}{2} \).

Now, since \( P:\tilde{\xi} \) has at most \( |I| \) non-zero entries, by (2) we have \( \rho(P:\tilde{\xi}) \leq c_\tau |P:\tilde{\xi}| = c_\tau \sum_{i \in I} |\xi_i| \). Recall also that \( \lambda(s) \geq s, s \in [0, n] \). Using these and the Markov inequality, we obtain that

\[
\mathbb{P} \left( \rho(P:\tilde{\xi}) \geq Mp_{\exp}(I) \right) \leq \mathbb{P} \left( c_\tau \sum_{i \in I} |\xi_i| \geq Mp_{\exp}(I) \right) 
= \mathbb{P} \left( c_0 \sum_{i \in I} |\xi_i| \geq \frac{M_0}{c_\tau} p_{\exp}(I) \right) 
\leq e^{\frac{M_0}{c_\tau} p_{\exp}(I)} \quad \leq e^{\frac{1}{|I|} \log 2 - c_0 c_\tau^{-1} M_{\exp}(I)} \quad \leq e^{\frac{1}{\lambda} \log 2 - c_0 c_\tau^{-1} M_{\exp}(I)},
\]

for \( \psi(M) = c_0 c_\tau^{-1} M \) and \( M_0 = \log 2 \). Similarly to the setting discussed in Remark 2.4, it is possible to generalise the case of independent sub-exponential errors \( \xi_i \)'s to the case of errors driven by an AR(1) model with sub-exponential white noise.

Thus, for sub-exponential \( \xi_i \)'s, Theorems 3.1 and 3.2 hold with the oracle rate \( r_{\exp}(\theta) = \min_{I \in \mathcal{I}} (\rho(P_I:\theta) + x\rho_{\exp}(I)) \). Notice a slight deterioration of the oracle rate \( r_{\exp}(\theta) \) as compared to the sub-gaussian case, as we now have the complexity \( p_{\exp}(I) = |I| \log(\text{en}/|I|) \) instead of \( p(I) = |I| \log(\text{en}/|I|)^{1/2} \). This will also lead to the corresponding slightly deteriorated global rate \( s \log(\text{en}/s) \) (instead of \( s \log^{1/2}(\text{en}/s) \)) in Corollary 3.1. Intuitively, a worse rate for the sub-exponential case is not surprising and should be considered as price for the ‘heavily tailedness’ of the errors \( \xi_i \)'s.

The question remains whether the resulting rate in this case \( \lambda(s) = s \log(\text{en}/s) \) is minimax over the sparsity scale \( \{\ell_0[s], s \in [n]\} \) for the \( \ell_1 \)-norm and sub-exponential errors. We did not find relevant results in the literature on this, but we conjecture that the minimax rate over sparsity scale for the \( \ell_q \)-norm with errors with density \( f_\xi(x) \propto e^{-|x|^\gamma} \), \( \gamma \in (0, 2] \), is expected to be \( s[\log(\text{en}/s)]^{q/\gamma} \).

### 4. A simulation study

In this section, we present a small simulation study. The main goal is to demonstrate the deceptiveness phenomenon for the UQ problem, which concerns the coverage relation of Theorem 3.2. Theorem 3.1 and the size relation of Theorem 3.2 will also be demonstrated in passing. Recall that the coverage relation in Theorem 3.2 holds only for the so-called non-deceptive parameters which are described by the EBR condition: \( \theta \in \Theta_{eb}(t) \). Below we
provide an example illustrating the failure of the coverage relation for a deceptive parameter \( \theta \). Exactly, we construct a sequence of ‘deceptive’ \( \theta_n \in \mathbb{R}^n, n \in \mathbb{N} \), such that, for any \( M_2, \mathbb{P}_{\theta_n}(\theta_n \notin B(\hat{\theta}, M_2 \hat{r})) \to 1 \) as \( n \to \infty \).

Consider the simplest setting in the model (1): \( \xi_i^{\text{ind}} \sim \mathcal{N}(0,1), i = 1, \ldots, n, \) and \( \tau = 1/2 \). In this case, according to Remark 3, Condition C1 is satisfied with \( H_0 = 1, \phi(M) = c_0 c_0^{-2} M^2 \) and \( M_0 = \log 2 \) where \( c_{0.5} = \frac{1}{2} \) and such \( c_0 \) that \( 2 \geq \mathbb{E} e^{\xi_0 W^2} = (1 - 2c_0)^{-1/2} \) for \( W \sim \mathcal{N}(0,1) \), leading to a choice \( c_0 = \frac{3}{8} \). Consider a parameter \( \theta_n = (d_n, 1, \ldots, 1) \in \mathbb{R}^n \) with any \( d_n > 2\exp(1) = 2\sqrt{\log 2/n} \), so that the oracle sparsity rate \( \mathbb{I}_{\theta} = \mathbb{I}_{\theta}(\theta_n) = \{1\} \) for all \( n \in \mathbb{N} \) and hence the ratio in the definition (15) of the EBR condition is \( \rho(\theta_n - \mathbb{I}_{\theta}(\theta_n)) = \frac{n^{-1}}{2 \log (en)^{1/2}} = t_n, n \in \mathbb{N} \). Since \( t_n \to \infty \) as \( n \to \infty \), this means that the sequence \( \{\theta_n\} \) can be seen as ‘deceptive’ in that for any \( t \geq 0 \) there exists \( n_0 \) such that \( \theta_n \notin \Theta_{e\theta}(t) \) for all \( n \geq n_0 \). In the below simulations, we took \( \kappa = 1.1, \nu = 1.4 \) and \( d_n = 10 \) which ensures the condition \( d_n > 2\sqrt{\log 2/n} \) for all considered \( n \)’s.

Introduce the following quantities: \( \tilde{D}_n = \frac{\rho(\theta_n - \hat{\theta})}{r(\theta_n)} \), \( \tilde{D}_n = \frac{\rho(\theta_n - \hat{\theta})}{\hat{r}} \), and \( \tilde{D}_n = \frac{\bar{r}}{r(\theta_n)} \). Now we perform the following simulations: for a chosen \( n \in \mathbb{N} \) and the above-described \( \theta_n \in \mathbb{R}^n \), generate \( B = 1000 \) data samples, then compute \( B \) values of \( \hat{\theta}(k), \bar{r}(k) \), and the corresponding \( \tilde{D}_n, \tilde{D}_n \) and \( \tilde{D}_n(k), k = 1, \ldots, B \). Since \( \{\rho(\theta_n - \hat{\theta}) \geq M_1 r(\theta_n)\} = \{\tilde{D}_n \geq M_1\} \), \( \{\theta_n \notin B(\hat{\theta}, M_2 \tilde{r})\} = \{\tilde{D}_n > M_2\} \) and \( \{\bar{r} \geq M_3 r(\theta_n)\} = \{\tilde{D}_n \geq M_3\} \), we can approximate the probabilities of the events from Theorems 3.1 and 3.2 as follows:

\[
\mathbb{P}_{\theta_n} \left( \rho(\theta_n - \hat{\theta}) \geq M_1 r(\theta_n) \right) \approx \frac{1}{B} \sum_{k=1}^{B} 1\{\tilde{D}_n(k) \geq M_1\},
\]
\[
\mathbb{P}_{\theta_n} \left( \theta_n \notin B(\hat{\theta}, M_2 \tilde{r}) \right) \approx \frac{1}{B} \sum_{k=1}^{B} 1\{\tilde{D}_n(k) > M_2\},
\]
\[
\mathbb{P}_{\theta_n} \left( \bar{r} \geq M_3 r(\theta_n) \right) \approx \frac{1}{B} \sum_{k=1}^{B} 1\{\tilde{D}_n(k) \geq M_3\}.
\]

In Figure 1, the boxplots of \( \tilde{D}_n(k) \)’s, \( \tilde{D}_n(k) \)’s and \( \tilde{D}_n(k) \)’s are plotted for several increasing \( n = 50, 100, 200, 400, 800, 1600, 3200 \). We see that the boxplots of \( \tilde{D}_n(k) \) (the plot on the left) stabilises around 1, this is because of the special choice of \( \theta_n \) for which \( \rho(\theta_n - \hat{\theta}) \) is asymptotically equivalent to the oracle rate \( r(\theta_n) \). Thus, the first probability in the above display will go to zero for any \( M_1 > 1 \), which shows that the claim of Theorem 3.1 holds also for this deceptive \( \{\theta_n\} \) as it should because the result of Theorem 3.1 is uniform over \( \theta \in \mathbb{R}^n \). The size relation of Theorem 3.2 holds as well (as it also should): the boxplots of \( \tilde{D}_n(k) \)’s (the plot on the right) even move to zero, showing that the third probability in the above display will converge to zero for any \( M_3 > 0 \).

The most interesting plot is in the middle of Figure 1, it shows that the boxplots of \( \tilde{D}_n(k) \)’s move up as \( n \) increases, ensuring that the fraction of data points such that \( \tilde{D}_n(k) > M_2 \) goes to 1 as \( n \to \infty \) for any constant \( M_2 \). Hence, \( \mathbb{P}_{\theta_n} (\theta_n \notin B(\hat{\theta}, M_2 \tilde{r})) \to 1 \) as \( n \to \infty \), for the chosen deceptive parameter sequence \( \{\theta_n\} \), which demonstrates the deceptiveness phenomenon.

Finally, we perform simulations for the sub-exponential case discussed in Remark 3.3 and a non-deceptive parameter \( \theta_n' = (d_n, \ldots, d_n, 0, \ldots, 0) \in \mathbb{R}^n \). Precisely, we consider the Laplace errors \( \xi_i^{\text{ind}} \sim \text{Laplace}(1,0), i = 1, \ldots, n \) (i.e. with density \( \frac{1}{2} e^{-|x|} \)), we set the first 20 entries of parameter \( \theta_n' \) equal to \( d_n \) and the remaining coordinates to zero. This means also
that we used complexity $p_{\exp}(I)$ instead of $p(I)$. All the other parameters of this simulation experiment remain the same as before.

Figure 2 shows that the claims of Theorems 3.1 and 3.2 hold for non-deceptive $\{\theta'_n\}$. We now see a downward trend in the boxplots of $\tilde{D}_{n}^{(k)}$'s as $n$ increases illustrating the good behaviour of the confidence set for non-deceptive $\{\theta'_n\}$ also in the sub-exponential case.

5. Proofs

In this section, we present the proofs of the theorems.

**Proof:** First we introduce some notation. Define the events

$$E_1 = \{\rho(\theta - \hat{\theta}) \geq M_1 r(\theta)\}, \quad E_2 = \{r(\theta, \hat{I}) \leq A_1 r(\theta)\}, \quad E_3 = \{\rho(\frac{1}{\sqrt{n}}) \leq A_2 p(\tilde{I})\},$$

where the constants $M_1,A_1,A_2 > 0$ are to be chosen later. We evaluate the probability of interest $P_\theta(E_1) = P_\theta(\rho(\theta - \hat{\theta}) \geq M_1 r(\theta))$ as follows:

$$P_\theta(E_1) = P_\theta(E_1 \cap E_2 \cap E_3) + P_\theta(E_1 \cap E_2^c \cap E_3) + P_\theta(E_1 \cap E_2^c \cap E_3^c) \leq T_1 + T_2 + T_3.$$  

(19)

We bound these three probabilities separately.
First we bound $T_1$. Using the properties (3) and (4) of the quantile loss function $\rho$, we derive that the event $E_1 \cap E_2 \cap E_3$ implies that

$$M_1 r(\theta) \leq \rho(\theta - \hat{\theta}) = \rho(\theta - P_{ij}X) \leq \rho(\theta - P_{ij}\theta) + \rho(-P_{ij}\hat{\theta}) \leq \rho(\theta - P_{ij}\theta) + A_2 C_r p(\hat{I}) \leq \max\{1, \frac{A_2 C_r}{\xi}\} r(\theta, \hat{I}) \leq \max\{1, \frac{A_2 C_r}{\xi}\} A_1 r(\theta).$$

Hence, if $M_1 > A_1 \max\{1, \frac{A_2 C_r}{\xi}\}$, then

$$T_1 = \mathbb{P}_\theta(E_1 \cap E_2 \cap E_3) \leq \mathbb{P}_\theta\left(M_1 r(\theta) \leq \max\{1, \frac{A_2 C_r}{\xi}\} A_1 r(\theta)\right) = 0. \quad (20)$$

To bound $T_2$, write

$$T_2 = \mathbb{P}_\theta(E_1 \cap E_2^c \cap E_3^c) \leq \mathbb{P}_\theta(E_3^c) \leq \sum_{I \in \mathcal{I}} \mathbb{P}\left(\rho(P_{ij}\hat{\theta}) > A_2 p(\hat{I})\right).$$

If $A_2 > 0$ is such that $\psi(A_2) > M_0 + 1$, then using Condition C1 (7), we obtain that

$$\mathbb{P}\left(\rho(P_{ij}\hat{\theta}) > A_2 p(\hat{I})\right) \leq H_0 e^{-A_3 \lambda(|I|)}, \quad I \in \mathcal{I}, I \neq \emptyset,$$

where we set $A_3 = \psi(A_2) - M_0$. Recall that $\lambda(s)$ is increasing for $s \in (0, n]$. This, (6) and the two previous displays entail that, since $A_3 > 1$ (so that $A_4 = \frac{(A_3 - 1)}{2} > 0$),

$$T_2 \leq \mathbb{P}_\theta(E_3^c) \leq \sum_{I \in \mathcal{I}} \mathbb{P}\left(\rho(P_{ij}\hat{\theta}) > A_2 p(\hat{I})\right) \leq H_\xi \sum_{I \in \mathcal{I}: |I| \geq 1} e^{-A_3 \lambda(|I|)} \leq H_\xi e^{-A_4 \lambda(1)} \sum_{I \in \mathcal{I}: |I| \geq 1} e^{-(1+4A_4) \lambda(|I|)} \leq H_1 e^{-A_4 \lambda(1)}. \quad (21)$$

Finally, we bound $T_3 = \mathbb{P}_\theta(E_1 \cap E_2^c \cap E_3)$, using (10) with $I' = I_0$ and $\kappa > A_2$, we obtain that under $E_2^c \cap E_3$,

$$C_r \rho(P_{ij}\hat{\theta}) \geq \rho(P_{ij}\hat{\theta}) + \kappa p(\hat{I}) - \rho(P_{ij}\theta) - \kappa p(I_0) - \rho(P_{ij}\hat{\theta}) \geq \rho(P_{ij}\hat{\theta}) + \kappa p(\hat{I}) - \rho(P_{ij}\theta) - \kappa p(I_0) - A_2 p(\hat{I}) \geq \min\{\frac{\kappa - A_2}{\xi}, 1\} r(\theta, \hat{I}) - \max\{\frac{\kappa}{\xi}, 1\} r(\theta) \geq \min\{\frac{\kappa - A_2}{\xi}, 1\} A_1 r(\theta) - \max\{\frac{\kappa}{\xi}, 1\} r(\theta) = A_5 r(\theta) \geq A_5 \kappa p(\theta),$$

as long as $A_5 = \min\{\frac{\kappa - A_2}{\xi}, 1\} A_1 - \max\{\frac{\kappa}{\xi}, 1\} > 0$. We conclude that the event $E_2^c \cap E_3$ implies the event $\{\rho(P_{ij}\hat{\theta}) \geq A_5 \kappa p(I_0)\}$, so that, by Condition C1 (7),

$$T_3 \leq \mathbb{P}_\theta(E_2^c \cap E_3) \leq \mathbb{P}_\theta\left(\rho(P_{ij}\hat{\theta}) \geq A_5 \kappa p(I_0)\right) \leq H_\xi e^{-A_6 \lambda(1)}, \quad (22)$$

where $A_6 = \psi\left(\frac{A_5 \kappa}{C_r}\right) - M_0 > 0$ if $A_5 > 0$ is chosen so large that $\psi\left(\frac{A_5 \kappa}{C_r}\right) > M_0$.

To summarise the choices of the constants, we need to take such $\kappa, M_1 > 0$ (in the claim of the theorem) and such constants $A_1, A_2 > 0$ (in the proof of the theorem) that
By using the above relation, we obtain that, under the event $C_\kappa > 1$

From (10), Condition $C_1$ (7) and $A_8 = \psi(A_7) - M_0 > 0$. Then, under $E_3 \cap E_4$, from (10) with $I' = I \cup I_0$, it follows that,

Using the last display, (21), Condition $C_1$ (7) and $A_8 = \psi(A_7) - M_0 > 0$, we bound

Now we establish the coverage property. The constants $M_1$, $H_1$, and $m_1$ are defined in Theorem 3.1. Take $M_2 = \frac{M_1(t + \kappa)}{\delta}$. where fixed $\delta \in (0,1)$ is from the definition of the
event $E_t$. If $\theta \in \Theta_{eb}(t)$, then, in view of (11),

$$r(\theta) = \rho(P_{t\theta}^\theta) + \varepsilon p(I_0) \leq (t + \varepsilon)p(I_0).$$

So, $p(I_0) \geq (t + \varepsilon)^{-1}r(\theta)$ for all $\theta \in \Theta_{eb}(t)$. Combining this with Theorem 3.1 and (24) yields that, uniformly in $\theta \in \Theta_{eb}(t)$,

$$\mathbb{P}_\theta\left( \theta \notin B(\hat{\theta}, M_2\hat{r}) \right) \leq \mathbb{P}_\theta\left( \rho(\theta - \hat{\theta}) > M_2\hat{r}, \hat{r} \geq \delta p(I_0) \right) + \mathbb{P}_\theta\left( \hat{r} < \delta p(I_0) \right)$$

$$\leq \mathbb{P}_\theta\left( \rho(\theta - \hat{\theta}) > M_1(t + \varepsilon)p(I_0) \right) + \mathbb{P}_\theta\left( p(\hat{I}) < \delta p(I_0) \right)$$

$$\leq \mathbb{P}_\theta\left( \rho(\theta - \hat{\theta}) > M_1r(\theta) \right) + \mathbb{P}_\theta\left( p(\hat{I}) < \delta p(I_0) \right)$$

$$\leq H_1e^{-m_1\lambda(1)} + H_1e^{-A_4\lambda(1)} + H_0e^{-A_8\lambda(1)}.$$

The coverage relation follows.

Let us show the size property. Recall again the event $E_3 = \{\rho(P_l^\xi) \leq A_2p(\hat{I})\}$ from the proof of the previous theorem. By using (23) with $I' = I_0$, we derive that the event

$$\{p(\hat{I}) \geq M_3r(\theta)\} \cap E_3$$

implies the event

$$\{C_\tau \rho(P_l^\xi) \geq (\kappa - A_2)p(\hat{I}) - r(\theta) \leq A_9p(\hat{I}) \geq A_9M_3r(\theta) \geq A_9M_3\varepsilon p(I_0)\}$$

with

$$A_9 = \kappa - A_2 - M_3^{-1} > 0,$$

where $A_2$ is defined in the proof of Theorem 3.1. We thus have that, for any $\theta \in \mathbb{R}^n$,

$$\mathbb{P}_\theta(\hat{r} \geq M_3r(\theta)) \leq \mathbb{P}_\theta(e_3^T) + \mathbb{P}_\theta(p(\hat{I}) \geq M_3r(\theta), E_3)$$

$$\leq \mathbb{P}_\theta(e_3^T) + \mathbb{P}_\theta\left( p(\hat{I}) \geq M_3r(\theta), C_\tau \rho(P_l^\xi) \geq A_9p(\hat{I}) \right)$$

$$\leq \mathbb{P}_\theta(e_3^T) + \mathbb{P}_\theta\left( C_\tau \rho(P_l^\xi) \geq A_9p(\hat{I}) \geq A_9M_3\varepsilon p(I_0) \right)$$

$$\leq H_1e^{-A_4\lambda(1)} + H_0e^{-A_8\lambda(1)},$$

where the last inequality in the above display is obtained by (21) and Condition C1 (7), and $M_3$ is chosen to be so large that $A_{10} = \psi(A_9M_3\varepsilon/C_\tau) - M_0 > 0$. The size relation follows.

**Disclosure statement**

No potential conflict of interest was reported by the author(s).

**References**

Abramovich, F., Benjamini, Y., Donoho, D.L., and Johnstone, I.M. (2006), 'Adapting to Unknown Sparsity by Controlling the False Discovery Rate', *The Annals of Statistics*, 34, 584–653.

Abrevaya, J. (2002), 'The Effects of Demographics and Maternal Behavior on the Distribution of Birth Outcomes', in *Economic Applications of Quantile Regression*, eds. B. Fitzenberger, R. Koenker, and J. A. F. Machado, Berlin Heidelberg: Springer-Verlag, pp. 247–257.

Babenko, A., and Belitser, E. (2010), 'Oracle Convergence Rate of Posterior Under Projection Prior and Bayesian Model Selection', *Mathematical Methods of Statistics*, 19, 219–245.

Baraud, Y. (2004), 'Confidence Balls in Gaussian Regression', *The Annals of Statistics*, 32, 528–551.

Belitser, E. (2017), 'On Coverage and Local Radial Rates of Credible Sets', *The Annals of Statistics*, 45, 1124–1151.
Belitser, E., and Ghosal, S. (2020), ‘Empirical Bayes Oracle Uncertainty Quantification for Regression’, The Annals of Statistics, 31, 536–559.

Belitser, E., and Nurushev, N. (2019), General Framework for Projection Structures. ArXiv: 1904.01003.

Belitser, E., and Nurushev, N. (2020), ‘Needles and Straw in a Haystack: Robust Empirical Bayes Confidence for Possibly Sparse Sequences’, Bernoulli, 26, 191–225.

Bellec, P.C. (2018), ‘Sharp Oracle Inequalities for Least Squares Estimators in Shape Restricted Regression’, The Annals of Statistics, 46, 745–780.

Belloni, A., and Chernozhukov, V. (2011), ‘L1-Penalized Quantile Regression in High-Dimensional Sparse Models’, The Annals of Statistics, 39, 82–130.

Belloni, A., Chernozhukov, V., and Kato, K. (2019), ‘Valid Post-Selection Inference in High-Dimensional Approximately Sparse Quantile Regression Models’, Journal of the American Statistical Association, 114, 749–758.

Belloni, A., Chernozhukov, V., and Wang, L. (2011), ‘Square-root Lasso: Pivotal Recovery of Sparse Signals via Conic Programming’, Biometrika, 98, 791–806.

Benjamini, Y., and Hochberg, Y. (1995), ‘Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing’, Journal of the Royal Statistical Society: Series B, 57, 289–300.

Birgé, L., and Massart, P. (2001), ‘Gaussian Model Selection’, Journal of the European Mathematical Society, 3, 203–268.

Butucea, C., Ndaoud, M., Stepanova, N., and Tsybakov, A. (2018), ‘Variable Selection with Hamming Loss’, The Annals of Statistics, 46, 1837–1875.

Castillo, I., and van der Vaart, A. (2012), ‘Needles and Straw in a Haystack: Posterior Concentration for Possibly Sparse Sequences’, The Annals of Statistics, 40, 2069–2101.

Donoho, D.L., and Johnstone, I.M. (1994), ‘Minimax Risk Over $\ell_p$-Balls for $\ell_q$-Error’, Probability Theory and Related Fields, 99, 277–303.

Donoho, D.L., Johnstone, I.M., Hoch, J.C., and Stern, A.S. (1992), ‘Maximum Entropy and the Nearly Black Object (with Discussion)’, Journal of the Royal Statistical Society: Series B, 54, 41–81.

Efron, B. (2008), ‘Microarrays, Empirical Bayes and the Two-Groups Model’, Statistical Science, 23, 1–22.

Gabriela Ciuperca, G. (2018), ‘Test by Adaptive Lasso Quantile Method for Real-Time Detection of a Change-Point’, Metrika, 81, 689–720.

Hulmán, A., Witte, D., Kerényi, Z., Madarász, E., Tánczer, T., Bosnyák, Z., Szabó, É., Ferencz, V., Péterfalvi, A., Tabák, Á., and Nyári, T.A. (2015), ‘Heterogeneous Effect of Gestational Weight Gain on Birth Weight: Quantile Regression Analysis From a Population-Based Screening’, Annals of Epidemiology, 25, 133–137.

Jiang, L., Bondell, H.D., and Wang, H.J. (2014), ‘Interquantile Shrinkage and Variable Selection in Quantile Regression’, Computational Statistics & Data Analysis, 69, 208–219.

Jiang, L., Wang, H.J., and Bondell, H.D. (2013), ‘Interquantile Shrinkage in Regression Models’, Journal of Computational and Graphical Statistics, 22, 970–986.

Johnstone, I.M. (2017), Gaussian Estimation: Sequence and Wavelet Models, Book Draft.

Johnstone, I.M., and Silverman, B.W. (2004), ‘Needles and Straw in Haystacks: Empirical Bayes Estimates of Possibly Sparse Sequences’, The Annals of Statistics, 32, 1594–1649.

Koenker, R. (2005), Quantile Regression, Cambridge: Cambridge University Press.

Li, K.-C. (1989), ‘Honest Confidence Regions for Nonparametric Regression’, The Annals of Statistics, 17, 1001–1008.

Martin, R., and Walker, S.G. (2014), ‘Asymptotically Minimax Empirical Bayes Estimation of a Sparse Normal Mean Vector’, Electronic Journal of Statistics, 8, 2188–2206.

Santosa, F., and Symes, W.W. (1986), ‘Linear Inversion of Band-Limited Reflection Seismograms’, SIAM Journal on Scientific and Statistical Computing, 7, 1307–1330.

Szabó, B.T., van der Vaart, A.W., and van Zanten, J.H. (2015), ‘Frequentist Coverage of Adaptive Nonparametric Bayesian Credible Sets’, The Annals of Statistics, 43, 1391–1428.

derv der Pas, S.L., Szabó, B.T., and van der Vaart, A.W. (2017), ‘Uncertainty Quantification for the Horseshoe (with Discussion)’, Bayesian Analysis, 12, 1221–1274.
Vershynin, R. (2018), *High-Dimensional Probability. An Introduction with Applications in Data Science*, Cambridge: Cambridge University Press.

Zou, H., and Yuan, M. (2008), 'Composite Quantile Regression and the Oracle Model Selection Theory', *The Annals of Statistics*, 36, 1108–1126.