Electronic Journal of Statistics
Vol. 12 (2018) 1822–1841
ISSN: 1935-7524
https://doi.org/10.1214/18-EJS1447

Solution of linear ill-posed problems by model selection and aggregation

Felix Abramovich

Department of Statistics and Operations Research
Tel Aviv University, Israel
e-mail: felix@post.tau.ac.il

Daniela De Canditiis*

Istituto per le Applicazioni del Calcolo "M. Picone",
CNR Rome, Italy
e-mail: d.decanditiis@iac.cnr.it

and

Marianna Pensky†

Department of Mathematics University of Central Florida
Orlando, USA
e-mail: Marianna.Pensky@ucf.edu

Abstract: We consider a general statistical linear inverse problem, where the solution is represented via a known (possibly overcomplete) dictionary that allows its sparse representation. We propose two different approaches. A model selection estimator selects a single model by minimizing the penalized empirical risk over all possible models. By contrast with direct problems, the penalty depends on the model itself rather than on its size only as for complexity penalties. A Q-aggregate estimator averages over the entire collection of estimators with properly chosen weights. Under mild conditions on the dictionary, we establish oracle inequalities both with high probability and in expectation for the two estimators. Moreover, for the latter estimator these inequalities are sharp. The proposed procedures are implemented numerically and their performance is assessed by a simulation study.

MSC 2010 subject classifications: Primary 62G05; secondary 62C10.
Keywords and phrases: Aggregation, ill-posed linear inverse problem, model selection, oracle inequality, overcomplete dictionary.

Received October 2017.

Contents

*Partially supported by INdAM-GNCS Project 2018.
†Partially supported by National Science Foundation (NSF), grants DMS-1407475 and DMS-1712977.

1. Introduction

Linear inverse problems, where the data is available not on the object of primary interest but only in the form of its linear transform, appear in a variety of fields: medical imaging (X-ray tomography, CT and MRI), astronomy (blurred images), finance (model calibration of volatility) to mention just a few. The main difficulty in solving inverse problems is due to the fact that most of practically interesting and relevant cases fall into the category of so-called ill-posed problems, where the solution cannot be obtained numerically by simple inversion of the transform. In statistical inverse problems the data is, in addition, corrupted by random noise that makes the solution even more challenging.

Statistical linear inverse problems have been intensively studied and there exists an enormous amount of literature devoted to various theoretical and applied aspects of their solutions. We refer a reader to [7] for review and references therein.

Let $\mathcal{G}$ and $\mathcal{H}$ be two separable Hilbert spaces and $A : \mathcal{G} \to \mathcal{H}$ be a bounded linear operator. Consider a general statistical linear inverse problem

$$y = Af + \varepsilon, \tag{1.1}$$

where y is the observation, $f \in G$ is the (unknown) object of interest, ε is a white noise with a (known) noise level σ. For ill-posed problems A^{-1} does not exist as a linear bounded operator.

Most of approaches for solving (1.1) essentially rely on reduction of the original problem to a sequence model using the following general scheme:

1. Choose some orthonormal basis $\{\phi_j\}$ on $\mathcal{G}$ and expand the unknown f in (1.1) as
$$f = \sum_j \langle f, \phi_j \rangle_G \ \phi_j \tag{1.2}$$
2. Define ψ_j as the solution of $A^* \psi_j = \phi_j$, where A^* is the adjoint operator, that is, $\psi_j = A(A^*A)^{-1}\phi_j$. Reduce (1.1) to the equivalent sequence model:
$$\langle y, \psi_j \rangle_H \ = \ \langle Af, \psi_j \rangle_H + \langle \varepsilon, \psi_j \rangle_H \ = \ \langle f, \phi_j \rangle_G + \langle \varepsilon, \psi_j \rangle_H, \tag{1.3}$$
where for ill-posed problems $\mathrm{Var}\,(\langle y, \psi_j \rangle_H) = \sigma^2 \|\psi_j\|_H^2$ increases with j. Following the common terminology, an inverse problem is called *mildly ill-posed*, if the variances increase polynomially and *severely ill-posed* if their growth is exponential (see, e.g., [7]).

3. Estimate the unknown coefficients $\langle f, \phi_j \rangle_G$ from empirical coefficients $\langle y, \psi_j \rangle_H$: $\widehat{\langle f, \phi_j \rangle}_G = \delta\{\langle y, \psi_j \rangle_H\}$, where $\delta(\cdot)$ is some truncating/shrinking/thresholding procedure (see, e.g., [7], Section 1.2.2.2 for a survey), and reconstruct f as
$$\hat{f} = \sum_j \widehat{\langle f, \phi_j \rangle}_G \ \phi_j$$

Efficient representation of f in a chosen basis $\{\phi_j\}$ in (1.2) is essential. In the widely-used singular value decomposition (SVD), ϕ_j's are the orthogonal eigenfunctions of the self-adjoint operator A^*A and $\psi_j = \lambda_j^{-1} A\phi_j$, where λ_j is the corresponding eigenvalue. SVD estimators are known to be optimal in various minimax settings over certain classes of functions (e.g., [20]; [10]; [9]). A serious drawback of SVD is that the basis is defined entirely by the operator A and ignores the specific properties of the object of interest $f \in \mathcal{G}$. Thus, for a given A, the same basis will be used regardless of the nature of a scientific problem at hand. While the SVD-basis could be very efficient for representing f in one area, it might yield poor approximation in the other. The use of SVD, therefore, restricts one within certain classes $\mathcal{G}$ depending on a specific operator A. See [15] for further discussion.

In wavelet-vaguelette decomposition (WVD), ϕ_j's are orthonormal wavelet series. Unlike SVD-basis, wavelets allows sparse representation for various classes of functions and the resulting WVD estimators have been studied in [15], [2], [21], [19]. However, WVD imposes relatively stringent conditions on A that are satisfied only for specific types of operators, mainly of convolution type.

A general shortcoming of orthonormal bases is due to the fact that they may be "too coarse" for efficient representation of unknown f. Since 90s, there was a growing interest in the atomic decomposition of functions over *overcomplete* dictionaries (see, for example, [23],[11], [16]). Every basis is essentially a *minimal* dictionary that allows only a unique (not necessarily sparse) representation. Such scarceness usually causes poor adaptivity [23]. Application of overcomplete dictionaries improves adaptivity of the representation, because one can choose now the most efficient (sparse) one among many available. One can see here an interesting analogy with colors. Theoretically, every other color can be generated by combining three basic colors (green, red and blue) in corresponding proportions. However, a painter would definitely prefer to use the whole available palette (overcomplete dictionary) to get the hues he needs! Selection of appropriate subset of atoms (model selection) that allows a sparse representation of a solution is a core element in such an approach. It becomes even more important for ill-posed inverse problems, where for large models the variance component in the risk of the estimator increases drastically. Pensky in [24] was probably the first to use overcomplete dictionaries for solving inverse problems. See the discussion on advantages of overcomplete dictionaries in applications to ill-posed problems in her paper.

Pensky in [24] utilized the Lasso techniques for model selection within the overcomplete dictionary, established oracle inequalities with high probability and applied the proposed procedure to several examples of linear inverse problems. However, as usual with Lasso, it required restrictive compatibility conditions on the design matrix Φ.

In this paper we propose two alternative approaches for overcomplete dictionaries based estimation in linear ill-posed problems. The first estimator is obtained by minimizing penalized empirical risk with a penalty on model M proportional to $\sum_{j\in M}\|\psi_j^2\|$. The second one is based on a Q-aggregation type procedure that is specifically designed for solution of linear ill-posed problems. We establish oracle inequalities for both estimators that hold with high probabilities and in expectation. Moreover, for the Q-aggregation estimator, the inequalities are sharp. Simulation study shows that the new techniques produce more accurate estimators than Lasso.

The rest of the paper is organized as follows. Section 2 introduces the notations and some preliminary results. The model selection and aggregation-type procedures are studied respectively in Section 3 and Section 4. The simulation study is described in Section 5. All proofs are given in the Appendix.

2. Setup and notations

Consider a discrete analog of a general statistical linear inverse problem (1.1):

$$y = Af + \varepsilon, \tag{2.1}$$

where $y \in \mathbb{R}^n$ is the vector of observations, $f \in \mathbb{R}^m$ is the unknown vector to be recovered, A is a known (ill-posed) $n \times m$ $(n \geq m)$ matrix with $rank(A) = m$, and $\varepsilon \sim N(0, \sigma^2 I_n)$.

In what follows $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$ denote respectively Euclidean norms and inner products. Let $\phi_j \in \mathbb{R}^m,\ j = 1, \cdots, p$ with $\|\phi_j\| = 1$ be a set of normalized vectors (dictionary), where typically $p > m$ (overcomplete dictionary). Let $\Phi_{m\times p}$ be the complete dictionary matrix with the columns $\phi_j,\ j = 1, \ldots, p$, and $\Psi_{n\times p}$ is such that $A^T\Psi = \Phi$, that is, $\Psi = A(A^TA)^{-1}\Phi$ and $\psi_j = A(A^TA)^{-1}\phi_j$. Let $r_\Phi = rank(\Phi)$ and assume that any r_Φ columns of Φ are linearly independent.

For any $1 \leq r \leq r_\Phi$ define the r-sparse minimal and maximal eigenvalues of $\Phi^T\Phi$ as

$$\nu_r^2 = \min_{x\in\mathbb{R}^p, \|x\|_0\leq r} \frac{\|\Phi x\|^2}{\|x\|^2}$$

and

$$\kappa_r^2 = \max_{x\in\mathbb{R}^p, \|x\|_0\leq r} \frac{\|\Phi x\|^2}{\|x\|^2}$$

Note that for normalized ϕ_j, $0 < \nu_r^2 \leq 1 \leq \kappa_r^2$.

Fix $1 \leq r \leq r_\Phi/2$ and consider a set of models $\mathcal{M}_r = \{M \subseteq \{1, \ldots, p\} : |M| \leq r\}$ of sizes not larger than r. For a given model $M \in \mathcal{M}_r$ define a diagonal indicator matrix $D_M \in \mathbb{R}^{p\times p}$ with diagonal entries $d_{Mj} = I\{j \in M\}$. The design matrix corresponding to M is then $\Phi_M = \Phi D_M$, while $\Psi_M = \Psi D_M =$

$A(A^TA)^{-1}\Phi_M$. Let $H_M = \Phi_M(\Phi_M^T\Phi_M)^{-1}\Phi_M^T$ be the projection matrix on a span of nonzero columns of Φ_M and

$$f_M = H_M f = \Phi_M(\Phi_M^T\Phi_M)^{-1}\Phi_M^T\ f = \Phi_M(\Phi_M^T\Phi_M)^{-1}\Psi_M^T\ Af$$

be the projection of f on M. Denote

$$z = (A^TA)^{-1}A^Ty = f + \xi,\ \text{where}\ \xi = (A^TA)^{-1}A^T\varepsilon. \tag{2.2}$$

Consider the corresponding projection estimator

$$\hat{f}_M = H_M z = H_M(A^TA)^{-1}A^Ty = \Phi_M(\Phi_M^T\Phi_M)^{-1}\Psi_M^T\ y = \Phi_M\hat{\theta}_M, \tag{2.3}$$

where the vector of projection coefficients $\hat{\theta}_M = (\Phi_M^T\Phi_M)^{-1}\Psi_M^T\ y$. By straightforward calculus, $\hat{f}_M \sim N(f_M, \sigma^2 H_M(A^TA)^{-1}H_M)$ and the quadratic risk

$$E\|\hat{f}_M - f\|^2 = \|f_M - f\|^2 + \sigma^2 \mathrm{Tr}\left((A^TA)^{-1}H_M\right) \tag{2.4}$$

The oracle model is the one that minimizes (2.4) over all models $M \in \mathcal{M}_r$ and the ideal oracle risk is

$$R(\text{oracle}) = \inf_{M\in\mathcal{M}_r} E\|\hat{f}_M - f\|^2 = \inf_M \left\{\|f_M - f\|^2 + \sigma^2\mathrm{Tr}\left((A^TA)^{-1}H_M\right)\right\} \tag{2.5}$$

The oracle risk is unachievable but can be used as a benchmark for a quadratic risk of any available estimator.

3. Model selection by penalized empirical risk

Fix $1 \le r \le r_\Phi/2$ and consider the set of models $\mathcal{M}_r$ of sizes at most r. For a given model $M \in \mathcal{M}_r$ and z defined in (2.2), $\hat{f}_M$ in (2.3) minimizes the corresponding empirical risk $\|z - \tilde{f}_M\|^2$. By Pythagoras' theorem, $\|z - \hat{f}_M\|^2 = \|z\|^2 - \|\hat{f}_M\|^2$. Select a model $\widehat{M}$ by minimizing the penalized empirical risk:

$$\widehat{M} = \operatorname*{argmin}_{M\in\mathcal{M}_r}\left\{\|z - \hat{f}_M\|^2 + \mathrm{Pen}(M)\right\} = \operatorname*{argmin}_{M\in\mathcal{M}_r}\left\{-\|\hat{f}_M\|^2 + \mathrm{Pen}(M)\right\}, \tag{3.1}$$

where $\mathrm{Pen}(M)$ is a penalty function on a model M. The proper choice of $\mathrm{Pen}(M)$ is the core of such an approach.

For *direct* problems ($A = I$), the penalized empirical risk approach, with the complexity type penalties $\mathrm{Pen}(|M|)$ on a model size, has been intensively studied in the literature. In the last decade, in the context of linear regression, the in-depth theories (risk bounds, oracle inequalities, minimaxity) have been developed by a number of authors. See, e.g., [18], [4],[5], [1], [25], [27] among many others.

For inverse problems, [9] considered a truncated orthonormal series estimator, where the cut-off point was chosen by SURE criterion corresponding to the AIC-type penalty $\mathrm{Pen}(M) = 2\sigma^2\mathrm{Tr}((A^TA)^{-1}H_M)$ and established oracle inequalities for the resulting estimator $\hat{f}_{\widehat{M}}$. It was further generalized and improved by risk hull minimization in [8].

To the best of our knowledge, [24] was the first to consider model selections within *overcomplete dictionaries* by empirical risk minimization for statistical inverse problems. She utilized Lasso penalty. However, as usual with Lasso, it required restrictive compatibility conditions on the design matrix Φ (see [24] for more details).

In this paper, we utilize the penalty $\text{Pen}(M)$ that depends on the Frobenius norm of the matrix Ψ_M:

$$\|\Psi_M\|_F^2 = \sum_{j \in M} \|\psi_j\|^2 = \text{Tr}\left((A^T A)^{-1}\Phi_M \Phi_M^T\right)$$

The following theorem provides nonasymptotic upper bounds for the quadratic risk of the resulting estimator $\hat{f}_{\widehat{M}}$ both with high probability and in expectation:

Theorem 1. *Consider the model (2.1) and the penalized empirical risk estimator $\hat{f}_{\widehat{M}}$, where the model $\widehat{M}$ is selected w.r.t. (3.1) with the penalty*

$$Pen(M) \geq \frac{4\sigma^2(\delta+1)}{a\nu_{2r}^2}\ \|\Psi_M\|_F^2\ \ln p \tag{3.2}$$

for some $\delta > 0$ and $0 < a < 1$. Then,

1. *With probability at least $1 - \sqrt{\frac{2}{\pi}}\ p^{-\delta}$*

$$\|\hat{f}_{\widehat{M}} - f\|^2 \leq \frac{1+a}{1-a}\ \min_{M \in \mathcal{M}_r} \left\{\|\hat{f}_M - f\|^2 + \frac{2}{1+a}\ Pen(M)\right\} \tag{3.3}$$

2. *If, in addition, we restrict the set of admissible models to $\mathcal{M}_{r,\gamma} = \{M \in \mathcal{M}_r : \|\Psi_M\|_F^2 \leq \gamma^2 n\}$ for some constant γ,*

$$E\|\hat{f}_{\widehat{M}} - f\|^2 \leq \frac{1+a}{1-a}\ \min_{M \in \mathcal{M}_{r,\gamma}} \left\{E\|\hat{f}_M - f\|^2 + \frac{3}{2(1+a)}\ Pen(M)\right\} + \frac{4\sigma^2\gamma^2}{a(1-a)\nu_{2r}^2}\ n^2\, p^{-\delta/2} \tag{3.4}$$

The additional restriction on the set of models $\mathcal{M}_r$ in the second part of Theorem 1 is required to guarantee that the oracle risk in (2.5) does not grow faster than n.

Note that for the direct problems, $\Psi_M = \Phi_M$, while $\|\Phi_M\|_F^2 = |M|$ (recall that ϕ_j's are normalized to have unit norms).Thus, the penalty (3.2) is the RIC-type complexity penalty of [18] of the form $\text{Pen}(M) = C|M|\ln p$ and the additional restriction on the set of admissible models required for (3.4) trivially holds with $\gamma = 1$. It is important to note that for inverse problems $Pen(M)$ depends on the model itself rather on its size only. Furthermore, the presence of ν_{2r}^2 in the penalty is essential. As a result, although the risk bounds obtained in Theorem 1 hold for any fixed $1 \le r \le r_\Phi/2$, they increase with r due to the variance component that dominates strongly for large models. However, the use of overcomplete dictionaries allows one to assume that the unknown f has a sparse representation in ϕ_j's and, therefore, makes reasonable to consider $\mathcal{M}_r$ with relatively small r to control the variance.

We can compare the quadratic risk of the proposed estimator with the oracle risk $R(\text{oracle})$ in (2.5). Consider the penalty

$$\text{Pen}(M) = \frac{4\sigma^2(\delta+1)}{a\nu_{2r}^2}\ \|\Psi_M\|_F^2\ \ln p \tag{3.5}$$

for some $0 < a < 1$. Assume that $p \ge n$ (overcomplete dictionary) and choose $\delta \ge 4$. Then, the last term in the RHS of (3.4) turns to be of a smaller order and we obtain

$$E\|\hat{f}_{\widehat{M}} - f\|^2 \le C_1 \min_{M \in \mathcal{M}_{r,\gamma}} \left\{ \|f_M - f\|^2 + C_2 \frac{\gamma\sigma^2}{\nu_{2r}^2} \|\Psi_M\|_F^2 \ln p \right\}$$

for some positive constants C_1, C_2 depending on a and δ only. By standard linear algebra arguments, $\|\Psi_M\|_F^2 = \text{Tr}\left((A^TA)^{-1}\Phi_M\Phi_M^T\right) \le \kappa_{2r}^2\ \text{Tr}\left((A^TA)^{-1}H_M\right)$ and, therefore, the following oracle inequality holds:

Corollary 1. *Assume that $p \ge n$ and consider the penalized empirical risk estimator $\widehat{M}$ from Theorem 1, where $\widehat{M}$ is selected w.r.t. (3.1) over $\mathcal{M}_{r,\gamma}$ with the penalty (3.5) for some $0 < a < 1$ and $\delta \ge 4$. Then,*

$$E\|\hat{f}_{\widehat{M}} - f\|^2 \le C_0 \frac{\kappa_{2r}^2}{\nu_{2r}^2} \ln p\ R(\text{oracle})$$

for some constant $C_0 > 0$ depending on a, δ and γ only.

Thus, the quadratic risk of the proposed estimator $\hat{f}_{\widehat{M}}$ is within $\ln p$-factor of the ideal oracle risk. The $\ln p$-factor is a common closest rate at which an estimator can approach an oracle even in direct (complete) model selection problems (see, e.g., [17], [4], [1], [25] and also [24] for inverse problems). For an *ordered* model selection within a set of nested models, it is possible to construct estimators that achieve the oracle risk within a constant factor (see, e.g., [9] and [8]).

Similar oracle inequalities (even sharp with the coefficient in front of $\|f_M - f\|^2$ equals to one) with high probability were obtained for the Lasso estimator but under the additional compatibility assumption on the matrix Φ (see [24]).

4. Q-aggregation

Note that inequalities (3.3) and (3.4) in Theorem 1 for model selection estimator are not sharp in the sense that the coefficient in front of the minimum is greater than one. In order to derive sharp oracle inequalities both in probability and expectation, one needs to aggregate the entire collection of estimators: $\hat{f} = \sum_{M \in \mathcal{M}_r} \theta_M \hat{f}_M$ rather than to select a single estimator $\hat{f}_{\widehat{M}}$.

[22] considered exponentially weighted aggregation (EWA) with $\theta_M \propto \pi_M \exp\{-\|z - \hat{f}_M\|^2/\beta\}$, where π is some (prior) probability measure on $\mathcal{M}_r$ and $\beta > 0$ is a tuning parameter. They established sharp oracle inequalities in expectation for EWA in direct problems. [14] proved sharp oracle inequalities in expectation for EWA of affine estimators in nonparametric regression model. Their paper offers limited extension of the theory to the case of mildly ill-posed inverse problems where $\mathrm{Var}(\langle y, \psi\rangle_H) = \sigma^2\|\psi_j\|_H^2$ increase at most polynomially with j. However, their results are valid only for the SVD decomposition and require block design which seriously limits the scope of application of their theory. Moreover, [12], [13] and [3] argued that EWA cannot satisfy sharp oracle inequalities with high probability and proposed instead to use Q-aggregation.

Define a general Q-aggregation estimator of f as

$$\hat{f}_{\hat{\theta}} = \sum_{M \in \mathcal{M}_r} \theta_M \hat{f}_M, \tag{4.1}$$

where the vector of weights $\hat{\theta}$ is the solution of the following optimization problem:

$$\hat{\theta} = \operatorname*{argmin}_{\theta \in \Theta_{\mathcal{M}_r}} \left\{ \alpha \sum_{M \in \mathcal{M}_r} \theta_M \|z - \hat{f}_M\|^2 + (1-\alpha) \left\| z - \sum_{M \in \mathcal{M}_r} \theta_M \hat{f}_M \right\|^2 + \mathrm{Pen}(\theta) \right\} \tag{4.2}$$

for some $0 < \alpha < 1$ and a penalty $\mathrm{Pen}(\theta)$, and $\Theta_{\mathcal{M}_r}$ is the simplex

$$\Theta_{\mathcal{M}_r} = \{\theta \in \mathbb{R}^{|\mathcal{M}_r|} : \theta_M \geq 0, \sum_{M \in \mathcal{M}_r} \theta_M = 1\} \tag{4.3}$$

In particular, [12] and [13] considered $\mathrm{Pen}(\theta)$ proportional to the Kullback-Leibler divergence $KL(\theta, \pi)$ for some prior π on $\Theta_{\mathcal{M}_r}$. For direct problems, they derived sharp oracle inequalities both in expectation and with high probability for Q-aggregation with such penalty. In fact, EWA can also be viewed as an extreme case of Q-aggregation for $\alpha = 1$ (see [26]). However, the results for Q-aggregation with Kullback-Leibler-type penalty are not valid for ill-posed problems. A slightly different form of Q-aggregation was considered in [3].

In this section we propose a different type of penalty for Q-aggregation in (4.2) that is specifically designed for the solution of inverse problems. In particular, this penalty allows one to obtain sharp oracle inequalities both in expectation and with high probability in both mild and severe ill-posed linear inverse problems (see Introduction). We consider the penalty $Pen(\theta)$ of the form

$$\mathrm{Pen}(\theta) = \sum_{M \in \mathcal{M}_r} \theta_M Pen(M), \tag{4.4}$$

where $Pen(M) \geq \frac{4\sigma^2(\delta+1)}{a\nu_{2r}^2} ||\Psi_M||_F^2 \ln p$ for some $\delta > 0$ and $0 < a < 1$ as in (3.2).

For such type of a penalty $Pen(\theta)$ and $\alpha = 1/2$, the resulting $\hat{\theta}$ is

$$\hat{\theta} = \underset{\theta \in \Theta_{\mathcal{M}_r}}{\operatorname{argmin}} \left\{ \sum_{M \in \mathcal{M}_r} \theta_M \left(\|z - \hat{f}_M\|^2 + Pen(M) \right) + \left\| z - \sum_{M \in \mathcal{M}_r} \theta_M \hat{f}_M \right\|^2 \right\} \tag{4.5}$$

Note that the first term in the minimization criteria (4.5) is the same as in model selection (3.1). The presence of the second term is inherent for Q-aggregation. In fact, the model selection estimator $\hat{f}_{\widehat{M}}$ from Section 3 is a particular case of a Q-aggregate estimator $\hat{f}_{\hat{\theta}}$ with the weights obtained by solution of problem (4.2) with $\alpha = 1$.

The non-asymptotic upper bounds for the quadratic risk of $\hat{f}_{\hat{\theta}}$, both with high probability and in expectation, are given by the following theorem :

Theorem 2. *Consider the model (2.1) and the Q-aggregate estimator $\hat{f}_{\theta}$ given by (4.1), where the weights θ are selected as a solution of the optimization problem (4.5) with the penalty (4.4).*

1. *Then, with probability at least $1 - \sqrt{\frac{2}{\pi}}\, p^{-\delta}$*

$$\|\hat{f}_{\hat{\theta}} - f\|^2 \le \min_{M \in \mathcal{M}_r} \left\{ \|\hat{f}_M - f\|^2 + 2\ Pen(M) \right\} \tag{4.6}$$

2. *If, in addition, we restrict the set of admissible models to $\mathcal{M}_{r,\gamma}$ defined in Theorem 1 for some γ, then*

$$E\|\hat{f}_{\hat{\theta}} - f\|^2 \le \min_{M \in \mathcal{M}_{r,\gamma}} \left\{ E\|\hat{f}_M - f\|^2 + \frac{3}{2}\ Pen(M) \right\} + \frac{4\sigma^2\gamma^2}{\nu_{2r}^2}\ n^2 p^{-\delta/2} \tag{4.7}$$

Unlike Theorem 1 for model selection estimator, inequalities in both (4.6) and (4.7) for Q-aggregation are sharp. For $A = I$, the results of Theorem 2 are similar to those obtained in [13] and [3] for Q-aggregation in direct problems.

5. Simulation study

In this section we present results of a simulation study that illustrates the performance of the model selection estimator $\widehat{M}$ from (3.1) with the penalty (3.5) and the Q-aggregation estimator $\hat{f}_{\hat{\theta}}$ given by (4.1) with the weights defined in (4.5).

The data were generated w.r.t. a (discrete) ill-posed statistical linear problem (2.1) corresponding to the convolution-type operator $Af(t) = \int_0^t e^{-(t-x)} f(x)dx$, $0 \le t \le 1$, on a regular grid $t_i = i/n$:

$$A_{ij} = e^{-\frac{i-j}{n}}\, I(j \le i), \quad i, j = 1, \ldots, n,$$

where $I(\cdot)$ is the indicator function and $n = 128$.

We considered the dictionary obtained by combining two wavelet bases of different type: the Daubechies 8 wavelet basis $\{\phi^D_{3,0}, ..., \psi^D_{3,0}, ..., \psi^D_{6,63}\}$ and the Haar basis $\{\phi^H_{3,0}, ..., \psi^H_{3,0}, ..., \psi^H_{5,53}\}$, with the overall dictionary size $p = 128 + 64 = 192$. In our notations, ϕ^D and ϕ^H are the scaling functions, while ψ^D and ψ^H are the wavelets functions of the Daubechies and Haar bases respectively with the initial resolution level $J_0 = 3$.

In order to investigate the behavior of the estimators, we considered test functions of various sparsity and several noise levels. In particular, we used four test functions presented in Figure 2, that correspond to different sparsity scenarios:

1. $f_1 = \phi^D_{3,4} + \phi^H_{3,0}$ (high sparsity)
2. $f_2 = \phi^D_{3,0} + \phi^D_{3,6} + \psi^D_{3,7} + \phi^H_{3,6}$ (moderate sparsity)
3. $f_3 = \phi^D_{3,1} + \phi^D_{3,5} + \phi^D_{3,7} + \psi^D_{3,0} + \psi^D_{3,3} + \psi^D_{3,5} + \phi^H_{3,0} + \phi^H_{3,3}$ (low sparsity)
4. f_4 is the well-known HeaviSine function from [17] (uncontrolled sparsity)

For each test function, we used three different values of σ that were chosen to ensure a signal-to-noise ratios $SNR = 10, 7, 5$, where $SNR(f) = \|f\|/\sigma$.

The accuracy of each estimator was measured by its relative integrated error:

$$R(\hat{f}) = \|f - \hat{f}\|^2/\|f\|^2$$

Since the model selection estimator $\hat{f}_{\widehat{M}}$ involves minimizing a cost function of the form $-\|\hat{f}_M\|^2 + 4\sigma^2\lambda \ln p \|\Psi_M\|^2_F$ over the entire model space $\mathcal{M}_r$ of a very large size, we used a Simulated Annealing (SA) stochastic optimization algorithm for an approximate solution. The SA algorithm is a kind of a Metropolis sampler where the acceptance probability is "cooled down" by a synthetic temperature parameter (see [6], Chapter 7, Section 8). More precisely, if $M^{(r)}$ is a solution at step $r > 0$ of the algorithm, at step $r + 1$ a tentative solution M^* is selected according to a given symmetric proposal distribution and it is accepted with probability

$$a(M^*, M^{(r)}) = \min\left\{1, \exp\left(-\frac{\pi(M^*) - \pi(M^{(r)})}{T^{(r)}}\right)\right\}. \tag{5.1}$$

where $T^{(r)}$ is a temperature parameter at step r. The expression (5.1) is motivated by the fact that while M^* is always accepted if $\pi(M^{(r)}) \geq \pi(M^*)$, it can still be accepted even if $\pi(M^{(r)}) < \pi(M^*)$ in spite of being worse than the current one. The chance of acceptance of M^* for the same value of $\pi(M^{(r)}) - \pi(M^*) < 0$ diminishes at every step as the temperature $T^{(r)}$ decreases with r. The law that reduces the temperature is called the cooling schedule, in particular, here we choose $T^{(r)} = 1/(1 + \log(r))$.

In this paper we adopted the classical symmetric uniform proposal distribution and selected a starting solution $M^{(0)}$ according to the following initial probability

$$p(j) = C \exp\{\langle \psi_j, y\rangle^2 - c\|\psi_j\|^2\}, \quad \text{for } j = 1, \dots, p \tag{5.2}$$

where $c = \sum_j \langle \psi_j, y\rangle^2 / \sum_j \|\psi_j\|^2$ and $C = \left[\sum_{j=1}^{p} \exp\{\langle \psi_j, y\rangle^2 - c\|\psi_j\|^2\}\right]^{-1}$ is the normalizing constant. Observe that the argument in the exponent in (5.2) is the difference of $\langle \psi_j, y\rangle^2$ and $\|\psi_j\|^2$ where the first term $\langle \psi_j, y\rangle^2$ is the squared j-th empirical coefficient while the second term $\|\psi_j\|^2$ is the increase in the variance due to the addition of the j-th dictionary function. Hence, the prior $p(j)$ is more likely to choose dictionary functions with small variances that are highly correlated to the true function f.

Thus, the adopted SA procedure can be summarized as follows:

- generate a random number $m \leq n/\log(p)$. Set $T^{(1)} = 1$
- generate a starting solution $M^{(0)}$ with $\text{card}(M^{(0)}) = m$ by sampling indices $j \in \{1, \dots, p\}$ according to the probability given by equation (5.2)
- repeat for $r = 1, 2, \dots r_{max}$
  1. generate a variable $j^* \sim$ Uniform $(1, \dots, p)$
  2. *if* $j^* \notin M^{(r)}$ propose $M^* = M^{(r)} \cup \{j^*\}$
     else
     propose $M^* = M^{(r)} - \{j^*\}$
  3. with probability $a(M^*, M^{(r)})$ given in equation (5.1) assign $M^{(r+1)} = M^*$, otherwise $M^{(r+1)} = M^{(r)}$
  4. update the temperature parameter $T^{(r+1)} = 1/(1 + \log(r+1))$

While various stopping criteria could be used in the SA procedure, we found $r_{max} = 100,000$ to be sufficient for obtaining a good approximation of the global minimum in (3.1). Once the algorithm is terminated, we evaluated $\hat{f}_{\widehat{M}}$, where $\widehat{M} = \arg\min_{0 \leq r \leq r_{max}} \pi(M^{(r)})$ was the "best" model in the chain of models generated by SA algorithm.

Similarly, the Q-aggregation estimator $\hat{f}_{\hat{\theta}}$ involves computationally expensive aggregation of estimators over the entire model space $\mathcal{M}_r$. We, therefore, approximated it by aggregating over the subset $\mathcal{M}'_r$ of the last 50 "visited" models in the SA chain, i.e. $\hat{f}_{\hat{\theta}} = \sum_{M \in \mathcal{M}'_r} \hat{\theta}_M \hat{f}_M$ with $\hat{\theta}$ being a solution of (4.5).

For f_1, f_2 and f_3 we also considered the oracle projection estimator $\hat{f}_{oracle}$ based on the true model. In addition, we compared the proposed estimators with the Lasso-based estimator $\hat{f}_{Lasso} = \sum_{j=1}^{p} \hat{\theta}_j \phi_j$ of [24], where the vector of coefficients $\hat{\theta}$ is a solution of the following optimization problem

$$\hat{\theta} = \arg\min_{\theta} \left\{ \left\| \sum_{j=1}^{p} \theta_j \phi_j \right\|^2 - 2\sum_{j=1}^{p} \theta_j \langle y, \psi_j\rangle + \lambda \sum_{j=1}^{p} |\theta_j|\, \|\psi_j\|^2 \right\},$$

and λ is a tuning parameter.

The tuning parameters λ for $\hat{f}_{\widehat{M}}$ and $\hat{f}_{Lasso}$, were chosen by minimizing the error on a grid of possible values. To reduce heavy computational costs we used the same λ of $\hat{f}_{\widehat{M}}$ for all 50 aggregated models used for calculating $\hat{f}_{\hat{\theta}}$.

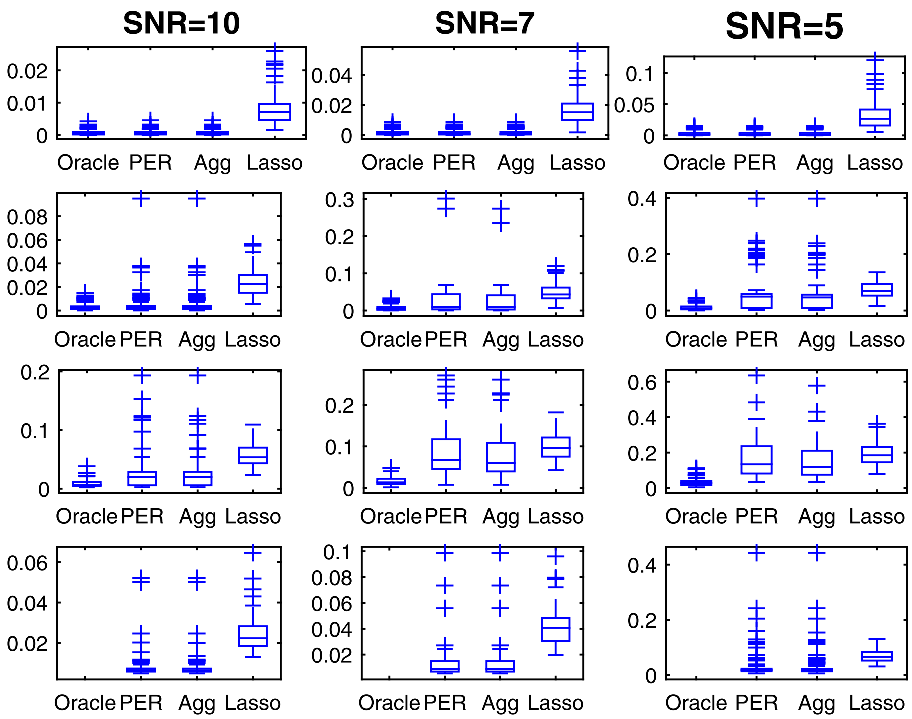

FIG 1. *The boxplots of the relative integrated errors of $\hat{f}_{oracle}$, $\hat{f}_{\widehat{M}}$, $\hat{f}_{\hat{\theta}}$ and $\hat{f}_{Lasso}$ over 100 independent runs. Top row: f_1, second row: f_2, third row f_3, bottom row f_4.*

Figure 1 presents the boxplots of $R(\hat{f})$ over 100 independent runs for $\hat{f}_{oracle}$ (for f_1, f_2 and f_3), $\hat{f}_{\widehat{M}}$, $\hat{f}_{\hat{\theta}}$ and $\hat{f}_{Lasso}$. Performances of all estimators deteriorate as SNR decreases especially for the less sparse test functions. The estimators $\hat{f}_{\widehat{M}}$ and $\hat{f}_{\hat{\theta}}$ always outperform $\hat{f}_{Lasso}$ and, as it is expected from our theoretical statements, $\hat{f}_{\hat{\theta}}$ yields better results than $\hat{f}_{\widehat{M}}$. We expect that the differences in precisions of $\hat{f}_{\widehat{M}}$ and $\hat{f}_{\hat{\theta}}$ would be more significant if we carried out aggregation over a larger portion of the model space than the last 50 visited models. Figure 2 illustrates these conclusion by displaying examples of the estimators for $SNR = 5$. We should also mention that estimator $\hat{f}_{\widehat{M}}$ was usually more sparse than $\hat{f}_{Lasso}$.

Appendix

The proofs of the main results are based on the following auxiliary lemmas.

Lemma 1. *For any $x > 0$,*

$$P\left(\sup_{M \in \mathcal{M}_r} \left\{\|\Psi_M^T \varepsilon\|^2 - 2\sigma^2 \|\Psi_M\|_F^2 (\ln p + x)\right\} \leq 0\right) \geq 1 - \sqrt{\frac{2}{\pi}}\, e^{-x}$$

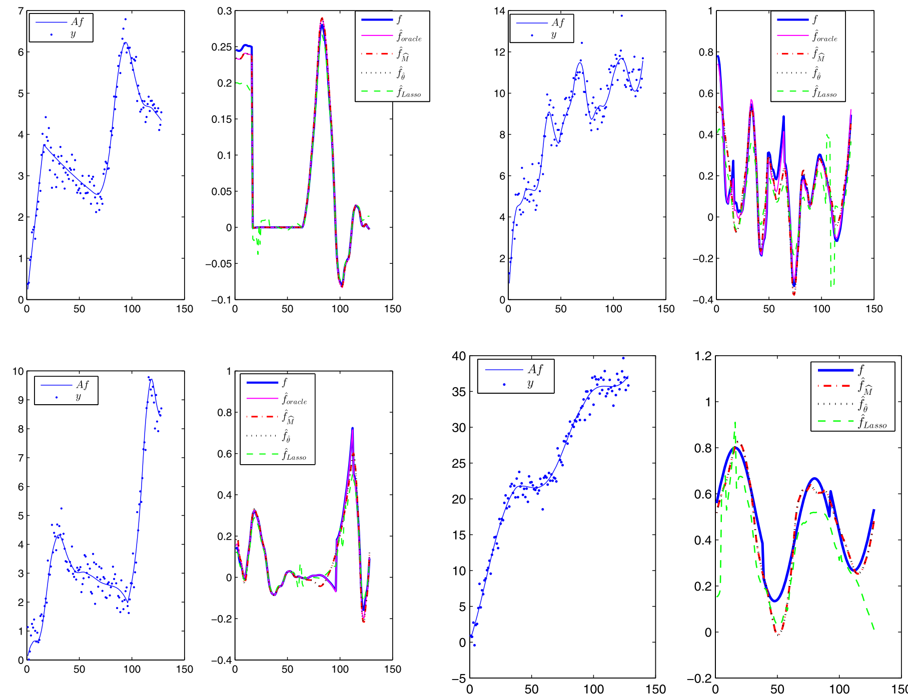

Fig 2. *The test functions, observed data and estimators for $SNR = 5$. In each block of figures corresponding to four test functions, the left panel shows the data y and the true Af; the right panel shows the true f and the four estimators. Top left: f_1; top right: f_2; bottom left: f_3; bottom right: f_4.*

Proof. For any model $M \in \mathcal{M}_r$, $\|\Psi_M^T \varepsilon\|^2 = \sum_{j\in M}(\psi_j^T\varepsilon)^2$, where $\psi_j^T\varepsilon \sim N(0,\sigma^2\|\psi_j\|^2)$. By Mill's ratio

$$P\left((\psi_j^T\varepsilon)^2 > 2\sigma^2\|\psi_j\|^2(\ln p + x)\right) \le \sqrt{\frac{2}{\pi}}\; p^{-1}e^{-x}$$

for any $x > 0$. Then,

$$\begin{aligned} &P\left(\bigcap_{j\in M}\left\{(\psi_j^T\varepsilon)^2 - 2\sigma^2\|\psi_j\|^2(\ln p + x) \le 0\right\}\right) \\ &\quad\ge 1 - \sum_{j\in M} P\left((\psi_j^T\varepsilon)^2 - 2\sigma^2\;\|\psi_j\|^2(\ln p + x) > 0\right) \ge 1 - \sqrt{\frac{2}{\pi}}\; e^{-x} \end{aligned}$$

and, therefore,

$$P\left(\sup_{M\in\mathcal{M}_r}\left\{\|\Psi_M^T\varepsilon\|^2-2\sigma^2\|\Psi_M\|_F^2(\ln p+x)\le 0\right\}\right)$$
$$\ge P\left(\bigcap_{M\in\mathcal{M}_r}\bigcap_{j\in M}\left\{(\psi_j^T\varepsilon)^2-2\sigma^2\ \|\psi_j\|^2(\ln p+x)\le 0\right\}\right)$$
$$= P\left(\bigcap_{j=1}^{p}\left\{(\psi_j^T\varepsilon)^2-2\sigma^2\ \|\psi_j\|^2(\ln p+x)\le 0\right\}\right)\ge 1-\sqrt{\frac{2}{\pi}}\ e^{-x} \qquad \square$$

Lemma 2. *For any $M_1, M_2 \in \mathcal{M}_r$, one has*

$$||\Psi_{M_1\cup M_2}||_F^2 \quad\le\quad ||\Psi_{M_1}||_F^2+||\Psi_{M_2}||_F^2 \tag{5.3}$$

Proof. The proof of (5.3) is straightforward by noting that

$$\mathrm{Tr}(\Psi_{M_1\cup M_2}^T\Psi_{M_1\cup M_2})=\sum_{j\in M_1\cup M_2}\|\psi_j\|^2\le \mathrm{Tr}(\Psi_{M_1}^T\Psi_{M_1})+\mathrm{Tr}(\Psi_{M_2}^T\Psi_{M_2}) \quad \square$$

Lemma 3. *If $\theta\in\Theta_{\mathcal{M}_r}$ where $\Theta_{\mathcal{M}_r}$ is defined in (4.3), then, for any function $\tilde f$*

$$\sum_{M\in\mathcal{M}_r}\hat\theta_M\|\tilde f-\hat f_M\|^2=\|\tilde f-\hat f_{\hat\theta}\|^2+\sum_{M\in\mathcal{M}_r}\hat\theta_M\|\hat f_{\hat\theta}-\hat f_M\|^2. \tag{5.4}$$

Proof. Note that for any $\tilde f$ one has

$$\sum_{M\in\mathcal{M}_r}\hat\theta_M\|\tilde f-\hat f_M\|^2 \quad=\quad \sum_{M\in\mathcal{M}_r}\hat\theta_M\|\tilde f-\hat f_{\hat\theta}\|^2+\sum_{M\in\mathcal{M}_r}\hat\theta_M\|\hat f_M-\hat f_{\hat\theta}\|^2$$
$$+\quad 2\langle\tilde f-\hat f_{\hat\theta}\ ,\sum_{M\in\mathcal{M}_r}\hat\theta_M(\hat f_{\hat\theta}-\hat f_M\rangle,$$

where for $\theta\in\Theta_{\mathcal{M}_r}$, the scalar product term in the last identity is equal to zero. $\square$

Proof of Theorem 1

Let z be defined in (2.2). Since $\widehat M$ is the minimizer in (3.1), for any given model $M\in\mathcal{M}_r$

$$\|z-\hat f_{\widehat M}\|^2+\mathrm{Pen}(\widehat M)\le\|z-\hat f_M\|^2+\mathrm{Pen}(M)$$

and, by a straightforward calculus, one can easily verify that

$$\|\hat f_{\widehat M}-f\|^2\le\|\hat f_M-f\|^2+2\langle\xi,\hat f_{\widehat M}-\hat f_M\rangle+\mathrm{Pen}(M)-\mathrm{Pen}(\widehat M) \tag{5.5}$$

Denote $\widetilde M=\widehat M\cup M$ and recall that, by the definition of $\mathcal{M}_r$, $|\widetilde M|\le 2r$. By the Cauchy-Schwarz inequality

$$2\langle\xi,\hat f_{\widehat M}-\hat f_M\rangle\ =\ 2\langle H_{\widetilde M}\xi,\hat f_{\widehat M}-\hat f_M\rangle\ \le\ 2||H_{\widetilde M}\xi||\ ||\hat f_{\widehat M}-\hat f_M||, \tag{5.6}$$

where by the definition of ν_{2r}^2

$$||H_{\widetilde{M}}\xi||^2 = ||H_{\widetilde{M}}(A^TA)^{-1}A^T\varepsilon||^2 = \varepsilon^T\Psi_{\widetilde{M}}(\Phi_{\widetilde{M}}^T\Phi_{\widetilde{M}})^{-1}\Psi_{\widetilde{M}}^T\varepsilon \ \le \ \frac{1}{\nu_{2r}^2}||\Psi_{\widetilde{M}}^T\varepsilon||^2 \tag{5.7}$$

Using inequalities $2\sqrt{uv} \le au + \frac{1}{a}v$ for any positive $a, u, v > 0$ and $\|u - v\|^2 \le 2(\|u\|^2 + \|v\|^2)$, from (5.6) and (5.7) obtain

$$2\langle \xi, \hat{f}_{\widehat{M}} - \hat{f}_M \rangle \ \le \ a\|\hat{f}_{\widehat{M}} - f\|^2 + a\|\hat{f}_M - f\|^2 + \frac{2}{a\nu_{2r}^2}\|\Psi_{\widetilde{M}}^T\varepsilon\|^2$$

Thus, from (5.5) it follows that for any $0 < a < 1$

$$(1-a)\|\hat{f}_{\widehat{M}} - f\|^2 \le (1+a)\|\hat{f}_M - f\|^2 + \mathrm{Pen}(M) + \frac{2}{a\nu_{2r}^2}\|\Psi_{\widetilde{M}}^T\varepsilon\|^2 - \mathrm{Pen}(\widehat{M}) \tag{5.8}$$

Applying Lemma 1 with $x = \delta \ln p$ for any $\delta > 0$, one has

$$\|\Psi_{\widetilde{M}}^T\varepsilon\|^2 \le 2\sigma^2\|\Psi_{\widetilde{M}}\|_F^2(\delta+1)\ln p \le 2\sigma^2\left(\|\Psi_{\widehat{M}}\|_F^2 + \|\Psi_M\|_F^2\right)(\delta+1)\ln p \tag{5.9}$$

w.p. at least $1 - \sqrt{\frac{2}{\pi}}\ p^{-\delta}$. Hence, for the penalty $Pen(M)$ in (3.2), after a straightforward calculus, (5.8) and (5.9) imply that w.p. at least $1 - \sqrt{\frac{2}{\pi}}\ p^{-\delta}$,

$$(1-a)\|\hat{f}_{\widehat{M}} - f\|^2 \le (1+a)\|\hat{f}_M - f\|^2 + 2\ Pen(M) \tag{5.10}$$

simultaneously for all models $M \in \mathcal{M}_r$ that proves (3.3).

To prove the oracle inequality in expectation (3.4), consider again (5.8) with $Pen(M)$ in (3.2) and note that $\|\Psi_{\widetilde{M}}^T\varepsilon\|^2 \le \|\Psi_{\widehat{M}}^T\varepsilon\|^2 + \|\Psi_M^T\varepsilon\|^2$. Therefore, taking expectation in (5.8), for the penalty $Pen(M)$ in (3.2) derive

$$\begin{aligned}(1-a)\,E\|\hat{f}_{\widehat{M}} - f\|^2 \le& (1+a)E\|\hat{f}_M - f\|^2 + \mathrm{Pen}(M) + \frac{2\sigma^2}{a\nu_{2r}^2}\|\Psi_M\|_F^2 + E\Delta \\ \le& (1+a)E\|\hat{f}_M - f\|^2 + \frac{3}{2}\mathrm{Pen}(M) + E\Delta,\end{aligned} \tag{5.11}$$

where $\Delta \le \frac{2}{a\nu_{2r}^2}\|\Psi_{\widehat{M}}^T\varepsilon\|^2 - Pen(\widehat{M})$. By Lemma 1, $P(\Delta > 0) \le \sqrt{\frac{2}{\pi}}p^{-\delta}$, so that for $\widehat{M} \in \mathcal{M}_{r,\gamma}$, by straightforward calculus,

$$\begin{aligned}E\Delta \le E\left(\Delta \cdot I\{\Delta > 0\}\right) &\le \frac{2\gamma^2 n}{a\nu_{2r}^2}\sqrt{E\|\varepsilon\|^4}\ \sqrt{P(\Delta > 0)} \\ &\le \frac{2\gamma^2\sigma^2}{a\nu_{2r}^2}\, n\sqrt{n(n+2)}\left(\frac{2}{\pi}\right)^{1/4} p^{-\delta/2} \le \frac{4\gamma^2\sigma^2}{a\nu_{2r}^2}\, n^2 p^{-\delta/2}\end{aligned}$$

Combining the last inequality with (5.11) obtain

$$(1-a)\,E\|\hat{f}_{\widehat{M}} - f\|^2 \le (1+a)\|\hat{f}_M - f\|^2 + \frac{3}{2}\ \mathrm{Pen}(M) + + \frac{4\gamma^2\sigma^2}{a\nu_{2r}^2}\, n^2 p^{-\delta/2}$$

which yields (3.4). □

Proof of Theorem 2

The beginning of the proof goes along the lines of the proof of Theorem 1 of [13] for Q-aggregation in direct problems.

Denote

$$\widehat{S}(\theta) = \frac{1}{2} \sum_{M \in \mathcal{M}_r} \theta_M \|z - \hat{f}_M\|^2 + \frac{1}{2}\|z - \hat{f}_\theta\|^2$$

and

$$S(\theta) = \frac{1}{2} \sum_{M \in \mathcal{M}_r} \theta_M \|f - \hat{f}_M\|^2 + \frac{1}{2}\|f - \hat{f}_\theta\|^2$$

Hence,

$$S(\hat{\theta}) - S(\theta) = \frac{1}{2} \sum_{M \in \mathcal{M}_r} (\hat{\theta}_M - \theta_M)\|f - \hat{f}_M\|^2 + \frac{1}{2}\left(\|f - \hat{f}_{\hat{\theta}}\|^2 - \|f - \hat{f}_\theta\|^2\right) \tag{5.12}$$

and also

$$\widehat{S}(\theta) = S(\theta) + \|z\|^2 - \|f\|^2 - 2\langle \xi, \hat{f}_\theta \rangle, \tag{5.13}$$

where ξ is defined in (2.2). By definition of $\hat{\theta}$, one has

$$\widehat{S}(\hat{\theta}) + \sum_{M \in \mathcal{M}_r} \hat{\theta}_M Pen(M) \le \widehat{S}(\theta) + \sum_{M \in \mathcal{M}_r} \theta_M Pen(M) \tag{5.14}$$

where $Pen(M) = 4\nu_r^{-2}\, \sigma^2(\delta + 1)\ \|\Psi_M\|_F^2\ \ln p$. Then, (5.13) and (5.14) yield

$$S(\hat{\theta}) - S(\theta) \le \sum_{M \in \mathcal{M}_r} (\theta_M - \hat{\theta}_M) Pen(M) + 2\langle \xi, \hat{f}_{\hat{\theta}} - \hat{f}_\theta \rangle \tag{5.15}$$

Fix $\beta \in (0,1)$ and a model $M_0 \in \mathcal{M}_r$. Let $e_{M_0} \in \mathbb{R}^{|\mathcal{M}_r|}$ be a vector from a canonical basis in $\mathbb{R}^{|\mathcal{M}_r|}$ corresponding to M_0 and consider a vector of weights $\tilde{\theta} = (1-\beta)\hat{\theta} + \beta e_{M_0}$. Thus,

$$\hat{f}_{\tilde{\theta}} = (1-\beta)\hat{f}_{\hat{\theta}} + \beta \hat{f}_{M_0} \tag{5.16}$$

Since $\|f - \hat{f}_{M_0}\|^2 - \|f - \hat{f}_{\hat{\theta}}\|^2 - \|\hat{f}_{\hat{\theta}} - \hat{f}_{M_0}\|^2 = 2\langle f - \hat{f}_{\hat{\theta}}, \hat{f}_{\hat{\theta}} - \hat{f}_{M_0}\rangle$, one can write

$$\begin{aligned}
&\|(1-\beta)(f - \hat{f}_{\hat{\theta}}) + \beta(f - \hat{f}_{M_0})\|^2 \\
&\quad = (1-\beta)\|f - \hat{f}_{\hat{\theta}}\|^2 - \beta(1-\beta)\|\hat{f}_{\hat{\theta}} - \hat{f}_{M_0}\|^2 + \beta\|f - \hat{f}_{M_0}\|^2
\end{aligned}$$

and combining the last equality with (5.16) obtains

$$\begin{aligned}
&\|f - \hat{f}_{\hat{\theta}}\|^2 - \|f - \hat{f}_{\tilde{\theta}}\|^2 \\
&\quad = \beta\left(\|f - \hat{f}_{\hat{\theta}}\|^2 + \|\hat{f}_{\hat{\theta}} - \hat{f}_{M_0}\|^2 - \|f - \hat{f}_{M_0}\|^2\right) - \beta^2\|\hat{f}_{\hat{\theta}} - \hat{f}_{M_0}\|^2.
\end{aligned}$$

Plugging the last equality into (5.12) derive

$$\frac{1}{\beta}\left(S(\hat{\theta})-S(\tilde{\theta})\right)=\frac{1}{2}\|f-\hat{f}_{\hat{\theta}}\|^2-\frac{1}{2}\|f-\hat{f}_{M_0}\|^2+\frac{1-\beta}{2}\|\hat{f}_{\hat{\theta}}-\hat{f}_{M_0}\|^2+\frac{1}{2}\Delta_1, \tag{5.17}$$

where by Lemma 3 with $\tilde{f}=f$,

$$\begin{aligned}\Delta_1&=\frac{1}{\beta}\sum_{M\in\mathcal{M}_r}(\hat{\theta}_M-\tilde{\theta}_M)\|f-\hat{f}_M\|^2=\sum_{M\in\mathcal{M}_r}\hat{\theta}_M\|f-\hat{f}_M\|^2-\|f-\hat{f}_{M_0}\|^2\\&=\sum_{M\in\mathcal{M}_r}\hat{\theta}_M\|\hat{f}_{\hat{\theta}}-\hat{f}_M\|^2\end{aligned}$$

Combining the last formulae with (5.17) yields

$$\begin{aligned}\frac{1}{\beta}\left(S(\hat{\theta})-S(\tilde{\theta})\right)&=\|f-\hat{f}_{\hat{\theta}}\|^2-\|f-\hat{f}_{\hat{\theta}}\|^2\\&\quad+\frac{1}{2}\left\{\sum_{M\in\mathcal{M}_r}\hat{\theta}_M\|\hat{f}_{\hat{\theta}}-\hat{f}_M\|^2+(1-\beta)\|\hat{f}_{\hat{\theta}}-\hat{f}_{M_0}\|^2\right\}\end{aligned} \tag{5.18}$$

From (5.16) one has

$$2\langle\xi,\hat{f}_{\hat{\theta}}-\hat{f}_{\tilde{\theta}}\rangle=2\beta\langle\xi,\hat{f}_{\hat{\theta}}-\hat{f}_{M_0}\rangle \tag{5.19}$$

Similarly,

$$\sum_{M\in\mathcal{M}_r}(\hat{\theta}_M-\tilde{\theta}_M)Pen(M)=\beta\left(\sum_{M\in\mathcal{M}_r}\hat{\theta}_M Pen(M)-Pen(M_0)\right). \tag{5.20}$$

Plugging (5.18)–(5.20) into (5.15) and setting $\beta\to 0$ imply

$$\|f-\hat{f}_{\hat{\theta}}\|^2\le\|f-\hat{f}_{M_0}\|^2+Pen(M_0)-\sum_{M\in\mathcal{M}_r}\hat{\theta}_M Pen(M)+2\langle\xi,\hat{f}_{\hat{\theta}}-\hat{f}_{M_0}\rangle-\frac{1}{2}\Delta_2 \tag{5.21}$$

where, applying Lemma 3 with $\tilde{f}=\hat{f}_{M_0}$,

$$\Delta_2=\sum_{M\in\mathcal{M}_r}\hat{\theta}_M\|\hat{f}_{\hat{\theta}}-\hat{f}_M\|^2+\|\hat{f}_{\hat{\theta}}-\hat{f}_{M_0}\|^2=\sum_{M\in\mathcal{M}_r}\hat{\theta}_M\|\hat{f}_M-\hat{f}_{M_0}\|^2 \tag{5.22}$$

Consider now the inner product term $2\langle\xi,\hat{f}_{\hat{\theta}}-\hat{f}_{M_0}\rangle=2\sum_{M\in\mathcal{M}_r}\hat{\theta}_M\langle\xi,\hat{f}_M-\hat{f}_{M_0}\rangle$ in (5.21). Repeating the arguments in the proof of Theorem 1 we have

$$2\langle\xi,\hat{f}_M-\hat{f}_{M_0}\rangle\le\frac{2}{\nu_{2r}^2}||\Psi_{\widetilde{M}}^T\varepsilon||^2+\frac{1}{2}||\hat{f}_M-\hat{f}_{M_0}||^2, \tag{5.23}$$

where $\widetilde{M}=M\bigcup M_0$, and by (5.9),

$$||\Psi_{\widetilde{M}}^T\varepsilon||^2\le 2\sigma^2(||\Psi_M||_F^2+||\Psi_{M_0}||_F^2)(\delta+1)\ln p$$

w.p. at least $1-\sqrt{\frac{2}{\pi}}p^{-\delta}$ simultaneously for all M and M_0. Hence, with this probability, for the penalties $Pen(M)$ in Theorem 2,

$$2\langle \xi, \hat{f}_{\hat{\theta}} - \hat{f}_{M_0}\rangle \leq Pen(M_0) + \sum_{M\in\mathcal{M}_r} \hat{\theta}_M Pen(M) + \frac{1}{2}\sum_{M\in\mathcal{M}_r} ||\hat{f}_M - \hat{f}_{M_0}||^2$$

that together with (5.21) and (5.22) imply $\|f-\hat{f}_{\hat{\theta}}\|^2 \leq ||f-\hat{f}_{M_0}||^2 + 2\ Pen(M_0)$ for all $M_0 \in \mathcal{M}_r$ and, therefore, (4.6) holds.

We now prove (4.7) for $M \in \mathcal{M}_{r,\gamma}$. From (5.21)-(5.23) it follows that

$$\begin{aligned} ||f-\hat{f}_{\hat{\theta}}||^2 \leq & ||f-\hat{f}_{M_0}||^2 + Pen(M_0) + \sum_{M\in\mathcal{M}_{r,\gamma}} \hat{\theta}_M \left(\frac{2}{\nu_{2r}^2}||\Psi_{\widetilde{M}}^T\varepsilon||^2 - Pen(M)\right) \\ \leq & ||f-\hat{f}_{M_0}||^2 + Pen(M_0) + \sup_{M\in\mathcal{M}_{r,\gamma}} \left(\frac{2}{\nu_{2r}^2}||\Psi_{\widetilde{M}}^T\varepsilon||^2 - Pen(M)\right) \\ \leq & ||f-\hat{f}_{M_0}||^2 + Pen(M_0) + \frac{2}{\nu_{2r}^2}||\Psi_{M_0}^T\varepsilon||^2 + \Delta, \end{aligned} \tag{5.24}$$

where $\Delta = \sup_{M\in\mathcal{M}_{r,\gamma}}\left(\frac{2}{\nu_{2r}^2}||\Psi_{\widetilde{M}}^T\varepsilon||^2 - Pen(M)\right)$. By Lemma 1, $P(\Delta > 0) \leq \sqrt{\frac{2}{\pi}}p^{-\delta}$. Taking the expectations in both sides of (5.24) and repeating the arguments in the proof of (3.4) in Theorem 1 for this Δ obtain (4.7). ☐

Acknowledgments

The authors would like to thank the anonymous referee for valuable remarks.

References

[1] Abramovich, F. and Grinshtein, V. (2010). MAP model selection in Gaussian regression. *Electr. J. Statist.* **4**, 932–949. MR2721039

[2] Abramovich F. and Silverman B.W. (1998). Wavelet decomposition approaches to statistical inverse problems. *Biometrika* **85**, 115–129. MR1627226

[3] Bellec, P.C. (2018). Optimal bounds for aggregation of affine estimators. *Ann. Statist.* **46**, 30–59. MR3766945

[4] Birgé, L. and Massart, P. (2001). Gaussian model selection. *J. Eur. Math. Soc.* **3**, 203–268. MR1848946

[5] Birgé, L. and Massart, P. (2007). Minimal penalties for Gaussian model selection. *Probab. Theory Relat. Fields* **138** 33–73. MR2288064

[6] Brémaud, P. (1999). *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues*, Springer, New York.

[7] Cavalier, L. (2011). Inverse problems in staistics. *Inverse Problems and High Dimensional Estimation* (Eds. Alquier, P., Gautier, E. and Gilles, G.), Lecture Notes in Statistics **203**, Springer, 3–96. MR2867623

[8] Cavalier, L. and Golubev, Yu. (2006) Risk hull method and regularization by projections of ill-posed inverse problems. *Ann. Statist.* **34**, 1653–1677. MR2283712

[9] Cavalier, L., Golubev, G.K., Picard, D. and Tsybakov, A.B. (2002). Oracle inequalities for inverse problems. *Ann. Statist.* **30**, 843–874. MR1922543

[10] Cavalier, L. and Tsybakov, A. (2002). Sharp adaptation for inverse problems with random noise, *Probab. Theory Related Fields* **123**, 323—354. MR1918537

[11] Chen, S., Donoho, D.L. and Saunders, M. A. (1999). Atomic decomposition by basic pursuit. *SIAM J. Scient. Comput.* **20**, 33–61.

[12] Dai, D., Rigolette, P. and Zhang, T. (2012) Deviation optimal leaning using greedy Q-aggregation. *Ann. Statist.* **40**, 1878–1905.

[13] Dai, D., Rigolette, P., Xia, L. and Zhang, T. (2014) Aggregation of affine estimators. *Electr. J. Statist.* **8**, 302–327. MR3766945

[14] Dalalyan, A. and Salmon, J. (2012) Sharp oracle inequalities for aggregation of affine estimators. *Ann. Statist.* **40**, 2327–2355. MR3059085

[15] Donoho, D.L. (1995). Nonlinear solutions of linear inverse problems by wavelet-vaguelette decomposition. *Appl. Comput. Harmon. Anal.* **2**, 101–126. MR1325535

[16] Donoho, D.L. and Elad, M. (2003). Optimally sparse representation in general (nonorthogonal) dictionaries via l_1 minimization. *PNAS* **100**, 2197–2202.

[17] Donoho, D.L. and Johnstone, I.M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425–455. MR1311089

[18] Foster, D.P. and George, E.I. (1994) The risk inflation criterion for multiple regression. *Ann. Statist.* **22**, 1947–1975.

[19] Johnstone, I.M., Kerkyacharian, G., Picard, D. and Raimondo, M. (2004) Wavelet decomposition in a periodic setting. *J. R. Statist. Soc. B* **66**, 547–573 (with discussion). MR1329177

[20] Johnstone, I.M. and Silverman, B.W. (1990) Speed of estimation in positron emission tomography and related inverse problems. *Ann. Statist.* **18**, 251–280. MR1041393

[21] Kalifa, J. and Mallat, S. (2003) Thresholding estimators for linear inverse problems and deconvolution. *Ann. Statist.* **31**, 58–109. MR1962500

[22] Leung, G. and Barron, A.R. (2006) Information theory and mixing least-squares regressions. *IEEE Trans. Inf. Theory* **52**, 3396–3410. MR2242356

[23] Mallat, S. and Zhang, Z. (1993). Matching pursuit with time-frequency dictionaries. *IEEE Trans. in Signal Proc.* **41**, 3397–3415.

[24] Pensky, M. (2016) Solution of linear ill-posed problems using overcomplete dictionaries. *Ann. Statist.* **44**, 1739–1754. MR3519939

[25] Rigollet, P. and Tsybakov, A. (2011) Exponential screening and optimal rates of sparse estimation. *Ann. Statist.* **39**, 731–771. MR2816337

[26] Rigollet, P. and Tsybakov, A. (2012) Sparse estimation by exponential weighting. *Statist. Science* **27**, 558–575. MR3025134

[27] Verzelen, N. (2012). Minimax risks for sparse regressions: Ultra-high dimensionals phenomenon. *Electr. J. Statist.* **6**, 38–90. MR2879672