The Annals of Statistics
2016, Vol. 44, No. 4, 1739–1764
DOI: 10.1214/16-AOS1445

SOLUTION OF LINEAR ILL-POSED PROBLEMS USING OVERCOMPLETE DICTIONARIES

BY MARIANNA PENSKY[1]

University of Central Florida

In memory of Laurent Cavalier

In the present paper, we consider the application of overcomplete dictionaries to the solution of general ill-posed linear inverse problems. In the context of regression problems, there has been an enormous amount of effort to recover an unknown function using an overcomplete dictionary. One of the most popular methods, Lasso and its variants, is based on maximizing the likelihood, and relies on stringent assumptions on the dictionary, the so-called compatibility conditions, for a proof of its convergence rates. While these conditions may be satisfied for the original dictionary functions, they usually do not hold for their images due to contraction properties imposed by the linear operator.

In what follows, we bypass this difficulty by a novel approach, which is based on inverting each of the dictionary functions and matching the resulting expansion to the true function, thus, avoiding unrealistic assumptions on the dictionary and using Lasso in a predictive setting. We examine both the white noise and the observational model formulations, and also discuss how exact inverse images of the dictionary functions can be replaced by their approximate counterparts. Furthermore, we show how the suggested methodology can be extended to the problem of estimation of a mixing density in a continuous mixture. For all the situations listed above, we provide sharp oracle inequalities for the risk in a non-asymptotic setting.

1. Introduction. In this paper, we consider the solution of a general ill-posed linear inverse problem $Qf = q$ where Q is a bounded linear operator that does not have a bounded inverse and the right-hand side q is measured with error. Problems of this kind appear in many areas of application such as astronomy (blurred images), econometrics (instrumental variables), medical imaging (tomography, dynamic contrast enhanced Computerized Tomography and Magnetic Resonance Imaging), finance (model calibration of volatility) and many others.

In particular, we consider the equation

$$y = q + \sigma n^{-1/2}\eta, \qquad q = Qf. \tag{1.1}$$

Received March 2015; revised January 2016.
[1]Supported in part by NSF Grants DMS-11-06564 and DMS-14-07475.
MSC2010 subject classifications. Primary 62G05; secondary 62C10.
Key words and phrases. Linear inverse problems, Lasso, adaptive estimation, oracle inequality.

Here, η is a Gaussian process representing the noise, $\sigma n^{-1/2}$ is the noise level, and $Q : \mathcal{H}_1 \to \mathcal{H}_2$ is a bounded linear operator without a bounded inverse. Hence, problem (1.1) is indeed ill-posed. Here, $\mathcal{H}_1$ and $\mathcal{H}_2$ are Hilbert spaces. We assume that observations are taken as functionals of y

$$\langle y, g\rangle_{\mathcal{H}_2} = \langle Qf, g\rangle_{\mathcal{H}_2} + \sigma n^{-1/2}\eta(g), \qquad g \in \mathcal{H}_2, \tag{1.2}$$

where, for any $g \in \mathcal{H}_2$, $\eta(g)$ is a Gaussian random variable with zero mean and $\mathbb{E}[\eta(g_1)\eta(g_2)] = \langle g_1, g_2\rangle_{\mathcal{H}_2}$, so that $\operatorname{Var}[\eta(g)] = \|g\|^2_{\mathcal{H}_2}$. The formulation above refers to the scenario where one cannot measure function $y(t)$ everywhere: only functionals of y are available. Such functionals may, for example, be coefficients of y in some orthonormal basis (e.g., Fourier, wavelet, eigenbasis of Q). The situation where one observes values of function y at some points is studied in depth in Section 5.

In order to understand the formulation above, consider a common situation where the operator Q is of the form

$$(Qf)(x) = \int_a^b g(x,t)f(t)\,dt, \qquad x \in (c,d), \tag{1.3}$$

and $f(t)$, $g(x,t)$, $q(x)$ and $y(x)$, $t \in (a,b)$, $x \in (c,d)$, are square integrable functions. In this case, $Q : \mathcal{H}_1 \to \mathcal{H}_2$ where $\mathcal{H}_1 = L^2(a,b)$ and $\mathcal{H}_2 = L^2(c,d)$. Formula (1.2) refers to the fact that function $y(x) = (Qf)(x) + \sigma n^{-1/2}\eta(x)$ cannot be measured for all points x. One can only observe linear functionals

$$\langle y, \psi\rangle_{\mathcal{H}_2} = \int_c^d y(x)\psi(x)\,dx = \int_c^d q(x)\psi(x)\,dx + \sigma n^{-1/2}\eta(\psi),$$

where $\eta(\psi) \sim N(0, \|\psi\|^2_{\mathcal{H}_2})$.

Solutions of statistical inverse problem (1.1) usually rely on the reduction of the problem to the sequence model by carrying out the singular value decomposition (SVD) [see, e.g., Cavalier and Golubev (2006), Cavalier et al. (2002), Cavalier and Reiß (2014), Golubev (2010) and Kalifa and Mallat (2003), or its relaxed version, the wavelet-vaguelette decomposition proposed by Donoho (1995) and further studies by Abramovich and Silverman (1998)]. Another general approach is the Galerkin method with subsequent model selection [see, e.g., Cohen, Hoffmann and Reiß (2004), Efromovich and Koltchinskii (2001) and Hoffmann and Reiss (2008)].

The advantage of the methodologies listed above is that they are asymptotically optimal in a minimax sense, and hence, deliver the best possible rates in the "worst case scenario" setting. A function of interest is usually represented via an orthonormal basis which is motivated by the form of the operator Q. However, in spite of being minimax optimal in many contexts, these approaches have two drawbacks. The first one is that, in many situations, these techniques may not be applicable. Indeed, for the majority of linear operators, the SVD decomposition

is unknown, and hence cannot be applied. Wavelet-vaguelette decomposition relies on relatively stringent conditions that are satisfied only for specific operators, mainly, of the convolution type. In addition, wavelet-based methods are advantageous when one recovers a one-dimensional function defined on a finite interval. However, they do not perform as well for a function of several variables or with an infinite domain [see, e.g., Candès (2003) or Le Pennec and Mallat (2005)]. Another shortcoming is that an orthonormal dictionary may not be "rich enough". If the unknown function does not have a relatively compact and accurate representation in the chosen basis, the accuracy of the resulting estimator will be poor. In Section 3, we provide detailed explanations of how application of overcomplete dictionaries can improve the precision of solutions of ill-posed linear inverse problems.

In the last decade, a great deal of effort was spent on the use of overcomplete dictionaries for the recovery of an unknown function f in a regression setting from its noisy observations. In particular, if f has a sparse representation in a dictionary, then it can be recovered with a much better precision than, for example, when it is expanded over an orthonormal basis. The methodology is based on the idea that the error of an estimator of f is approximately proportional to the number of dictionary functions that are used for representing it. Therefore, expanding the function of interest over fewer dictionary elements decreases the estimation error. In order to represent a variety of functions efficiently, one would need to consider a dictionary of much larger size than the number of available observations, the so-called overcomplete dictionary, and also to develop tools for choosing a linear combination of the elements of the dictionary that deliver efficient representation of f.

A variety of techniques have been developed for solution of those problems including likelihood penalization methods and greedy algorithms. The most popular of those methods (due to its computational convenience), Lasso and its variants, has been used for the solution of a number of theoretical and applied statistical problems [see, e.g., Bickel, Ritov and Tsybakov (2009), Bunea, Tsybakov and Wegkamp (2007), Dalalyan, Hebiri and Lederer (2014), Lounici et al. (2011), Yuan and Lin (2006), and also Bühlmann and van de Geer (2011) and references therein]. However, the application of Lasso is based on maximizing the likelihood and, unfortunately, relies on stringent assumptions on the dictionary $\{\varphi_k\}_{j=1}^p$, the so-called compatibility conditions, for a proof of its convergence rates. In the regression set up [with $Q = I$, the identity operator, in (1.1)], as long as one of the compatibility conditions holds, Lasso identifies a linear combination of the dictionary elements which represent the function of interest best of all at the "price" which is proportional to $\sigma n^{-1/2}\sqrt{\log p}$, where p is the dictionary size [see, e.g., Bühlmann and van de Geer (2011)]. Regrettably, while the compatibility conditions may be satisfied for the functions φ_j in the original dictionary, they usually do not hold for their images $Q\varphi_j$ due to contraction imposed by the operator Q. In order to illustrate this issue, one should expand f over the dictionary as $f_{\boldsymbol{\theta}} = \sum_{j=1}^p \theta_j \varphi_j$. Then $q_{\boldsymbol{\theta}} = \sum_{j=1}^p \theta_j u_j$ with $u_j = Q\varphi_j$. In a nutshell, in order

one can prove that Lasso can recover function f with nearly optimal convergence rates, matrix $\mathbf{A}$ with elements $A_{kj} = \langle u_k, u_j \rangle$ should be such that its principal submatrices of a small order have eigenvalues that are uniformly separated from zero and infinity [see, e.g., Bickel, Ritov and Tsybakov (2009)]. The latter usually does not hold for the ill-posed problems where the smallest eigenvalue can decrease polynomially or even exponentially as a function of j.

The objective of this paper is to circumvent this difficulty and apply the Lasso methodology to solution of linear inverse problem (1.1). For this purpose, in Section 4, instead of matching the expansion $q_{\boldsymbol{\theta}}$ to data y, we invert each of the dictionary functions φ_j and match expansion $f_{\boldsymbol{\theta}}$ to the true function f. This approach has several advantages. First, it allows the use of Lasso in a prediction setting where it requires much milder assumptions. In this setting, Lasso converges to the true solution, although at a slow rate, under practically no assumptions on the dictionary. Second, inverting fully known functions φ_j is an easier task than inverting an unknown function measured with noise. In addition, the norms of the inverted images can be viewed as the "price" of including each of the dictionary functions φ_j. In order to ensure that the estimator $f_{\widehat{\boldsymbol{\theta}}}$ attains fast convergence rates, we formulate a compatibility assumption and discuss sufficient conditions that guarantee its validity. Under this compatibility condition, we obtain a sharp oracle inequality for the risk of the Lasso estimator that holds with an arbitrarily large probability in non-asymptotic settings. In particular, we show that our risk bounds are smaller than those derived in Dalalyan and Salmon (2012). In the case of an orthonormal dictionary, they are also simpler than the ones derived in Cohen, Hoffmann and Reiß (2004) and also hold true for any sample size.

The Lasso methodology developed for equations (1.1) and (1.2) allows a variety of generalizations. First, in Section 5, we extend formulations (1.1) and (1.2) to observational models where only the values $y(t_i)$, $i = 1, \ldots, n$, of $y(t)$ are available. Second, in Section 6, we explain how, with very minor modifications, the Lasso technique can be used for estimation of a mixing density in a continuous mixture. Third, in Section 7, we show that, even if the exact inverse images of the dictionary functions do not exist, one can use their approximations and take advantage of the exact knowledge of the dictionary functions, which allows the optimal bias-variance decomposition. This implies that the latter technique, in essence, replaces the Lasso by the elastic net technique.

We would like to emphasize that the Lasso methodology for solution of linear inverse problems can be viewed as an extension of both the Galerkin method and the wavelet-vaguelette decomposition. Really, if instead of an overcomplete dictionary, one were to use an orthonormal basis, then the Lasso methodology just reduces to the Galerkin method with the model selection carried out by a soft thresholding technique. Moreover, if this orthonormal basis is comprised of wavelet functions and conditions for validity of the wavelet-vaguelette decomposition hold, the Lasso penalty just imposes soft thresholding on the wavelet coefficients.

The rest of the paper is organized as follows. After introducing notation (Section 2), Section 3 explains why application of overcomplete dictionaries allows a significant improvement of the estimation accuracy in inverse linear ill-posed problems. Section 4 develops the theoretical foundations of the paper by justifying the application of the Lasso technique to the solution of the general linear inverse problem (1.1). In particular, it introduces a compatibility assumption which guarantees that the Lasso estimator attains fast convergence rates for any function f which has a sparse representation in the dictionary, and thereby also provides a sharp oracle inequality for its risk. The compatibility condition is further discussed in the Appendix. In particular, Section A.1, formulates simpler sufficient conditions under which it holds. Sections 5 and 6 clarify how the theory can be applied to real life observational models and also to the estimation of a mixing density on the basis of observations of a continuous mixture. Section 7 demonstrates how exact inverse images of the dictionary functions can be replaced by their approximate counterparts. Section 8 considers the application of Lasso to the estimation of the unknown density function of the matrix parameter of the Wishart distribution. Section 9 concludes the paper with a discussion of the results. Finally, Appendix contains a detailed discussion of the compatibility condition (Section A.1), and proofs of the statements formulated in earlier sections (Section A.2).

2. Notation. In the paper, we use the following notation.

- For any vector $\mathbf{t} \in \mathbb{R}^p$, denote its ℓ_2, ℓ_1, ℓ_0 and ℓ_∞ norms by, respectively, $\|\mathbf{t}\|_2$, $\|\mathbf{t}\|_1$, $\|\mathbf{t}\|_0$ and $\|\mathbf{t}\|_\infty$. Similarly, for any function f, denote by $\|f\|_2$, $\|f\|_1$ and $\|f\|_\infty$ its L_2, L_1 and L_∞ norms.
- For any matrix $\mathbf{A}$, denote its spectral and Frobenius norms by, respectively, $\|\mathbf{A}\|$ and $\|\mathbf{A}\|_2$. Notation $\mathbf{A} > 0$ or $\mathbf{A} \geq 0$ means, respectively, that $\mathbf{A}$ is positive or nonnegative definite. Denote determinant of $\mathbf{A}$ by $|\mathbf{A}|$ and the largest, in absolute value, element of $\mathbf{A}$ by $\|\mathbf{A}\|_\infty$. Denote the Moore–Penrose inverse of matrix $\mathbf{A}$ by $\mathbf{A}^+$.
- Denote $\mathcal{P} = \{1, \ldots, p\}$. For any subset of indices $J \subseteq \mathcal{P}$, subset J^c is its complement in $\mathcal{P}$ and $|J|$ is its cardinality, so that $|\mathcal{P}| = p$. Let $\mathcal{L}_J = \mathrm{Span}\{\varphi_j, j \in J\}$.
- If $J \subset \mathcal{P}$ and $\mathbf{t} \in \mathbb{R}^p$, then $\mathbf{t}_J \in \mathbb{R}^{|J|}$ denotes reduction of vector $\mathbf{t}$ to subset of indices J.
- Denote by $\lambda_{\min}(m; \mathbf{\Phi})$ and $\lambda_{\max}(m; \mathbf{\Phi})$ the minimum and the maximum restricted eigenvalues of matrix $\mathbf{\Phi}$

$$\lambda_{\min}(m; \mathbf{\Phi}) = \min_{\substack{\mathbf{t} \in \mathbb{R}^p \\ \|\mathbf{t}\|_0 \leq m}} \frac{\mathbf{t}^T \mathbf{\Phi} \mathbf{t}}{\|\mathbf{t}\|_2^2}, \qquad \lambda_{\max}(m; \mathbf{\Phi}) = \max_{\substack{\mathbf{t} \in \mathbb{R}^p \\ \|\mathbf{t}\|_0 \leq m}} \frac{\mathbf{t}^T \mathbf{\Phi} \mathbf{t}}{\|\mathbf{t}\|_2^2}. \tag{2.1}$$

Also, denote by $\varrho(\mathbf{\Phi})$ the maximum of a non-diagonal element of matrix $\mathbf{\Phi}$:

$$\varrho(\mathbf{\Phi}) = \max_{j \neq k} |\Phi_{jk}|. \tag{2.2}$$

Whenever there is no ambiguity, we drop $\mathbf{\Phi}$ in the above notation and write simply $\lambda_{\min}(m)$, $\lambda_{\max}(m)$ and ϱ.

- $a_m \asymp b_m$ means that there exist constants $0 < C_1 < C_2 < \infty$ independent of m such that $C_1 a_m < b_m < C_2 a_m$.

3. Advantages of overcomplete dictionaries. The purpose of this section is to demonstrate how application of a rich overcomplete dictionary can reduce estimation error in inverse linear ill-posed problems. Indeed, if an overcomplete dictionary allows an efficient representation of f, it leads to a smaller estimation error. In order to understand the roots of this phenomenon, consider the situation where operator Q is self-adjoint and has a singular value decomposition $Qe_k = \lambda_k e_k$, $k = 1, 2, \ldots$, and function f can be represented as $f = \sum_k c_k e_k$. Assume, without loss of generality, that for some $\mu > 0$ and $\rho > 0$

$$|c_k| \le C_0 k^{-(\mu+1/2)} \quad \text{and} \quad |\lambda_k| \ge C_\lambda k^{-\rho}. \tag{3.1}$$

In this case, one can construct the SVD estimator $\widehat{f}_{\mathrm{SVD}} = \sum_{k=1}^{m} \lambda_k^{-1} \langle y, e_k \rangle e_k$ of f with the mean squared error (MSE) of the form

$$\begin{aligned} \mathbb{E}\| \widehat{f}_{\mathrm{SVD}} - f \|_2^2 &= \sum_{k=m+1}^{\infty} c_k^2 + \frac{\sigma^2}{n} \sum_{k=1}^{m} \lambda_k^{-2} \asymp m^{-2\mu} + \frac{\sigma^2}{n} m^{1+2\rho} \\ &\asymp n^{-2\mu/(2\mu+2\rho+1)}, \end{aligned} \tag{3.2}$$

where the value of m is chosen to minimize the right-hand side of (3.2). The advantage of the SVD is that its error rates hold in the "worst case" minimax estimation scenario where f is the hardest to estimate in the chosen class of functions.

On the other hand, consider the "best case" scenario where one has an extensive overcomplete dictionary φ_l with $\|\varphi_l\| = 1$, $l = 1, \ldots, p$, and f is proportional to one of the dictionary functions, say, φ_j. Expand dictionary functions φ_l in the eigenbasis e_k and find their inverse images ψ_l obtaining

$$\varphi_l = \sum_{k=1}^{\infty} c_{lk} e_k, \qquad \psi_l = \sum_{k=1}^{\infty} c_{lk} \lambda_k^{-1} e_k.$$

If one had an oracle which identifies the function φ_j that is proportional to f, then $c_{jk} = c_k / \|f\|$ and f would be estimated by $\widehat{f}_{\mathrm{or}} = \langle y, \psi_j \rangle \varphi_j$ with the error

$$\mathbb{E}\| \widehat{f}_{\mathrm{or}} - f \|_2^2 = \frac{\sigma^2}{n} \|\psi_j\|_2^2 = \frac{\sigma^2}{n} \sum_{k=1}^{\infty} \lambda_k^{-2} c_{jk}^2 = \frac{\sigma^2}{n} \|f\|_2^{-2} \sum_{k=1}^{\infty} \lambda_k^{-2} c_k^2. \tag{3.3}$$

Moreover, if $\mu > \rho$ in (3.1), then the series in the right-hand side of (3.3) is convergent and $\widehat{f}_{\mathrm{or}}$ has parametric error rate $\mathbb{E}\| \widehat{f}_{\mathrm{or}} - f \|^2 \asymp n^{-1}$. Otherwise, if $\mu \le \rho$, one can replace ψ_j by

$$\psi_{j,M_j} = \sum_{k=1}^{M_j} c_{jk} e_k \tag{3.4}$$

and estimate f by $\widehat{f}_{\mathrm{or},M} = \langle y, \psi_{j,M_j}\rangle\varphi_j$. It is easy to calculate that

$$\mathbb{E}\|\widehat{f}_{\mathrm{or},M_j} - f\|_2^2 = \left[\frac{\sigma^2}{n}\sum_{k=1}^{M_j}\lambda_k^{-2}c_{jk}^2 + \sum_{k=M_j+1}^{\infty}c_{jk}^2\right] \asymp \frac{{M_j}^{2\rho-2\mu}}{n} + {M_j}^{-2\mu}. \tag{3.5}$$

Choosing M_j that minimizes the right-hand side of (3.5), obtain

$$\mathbb{E}\|\widehat{f}_{\mathrm{or},M_j} - f\|_2^2 \asymp n^{-2\mu/(2\rho)} = o\bigl(n^{-2\mu/(2\mu+2\rho+1)}\bigr) \asymp \mathbb{E}\|\widehat{f}_{\mathrm{SVD}} - f\|_2^2, \qquad n\to\infty,$$

that is, the error of $\widehat{f}_{\mathrm{or},M_j}$ is smaller than the error of the SVD estimator. The advantage comes from the fact that, unlike in (3.2), in the right-hand sides of (3.3) and (3.5), the "large" values λ_k^{-2} are multiplied by "small" values c_{jk}^2 in the expression for the MSE.

One would argue that the assumption that f is proportional to one of the dictionary elements is not very realistic. However, it is very likely that f can be represented by a small subset of the dictionary functions φ_j, $j\in J$, of cardinality $|J| = s$. Then f can be estimated by

$$\widehat{f}_{\mathrm{or},\mathbf{M}} = \sum_{j\in J}\langle y, \psi_{j,M_j}\rangle\varphi_j, \qquad \mathbf{M} = (M_1,\dots,M_s),$$

where ψ_{j,M_j} are defined in (3.4) and the values M_j are found by minimizing the right-hand side of (3.5). If, for example, the dictionary functions are not "much harder" than f, that is, if there exists a constant C_f such that for $j\in J$ one has $c_{jk}^2 \le C_f c_k^2$, then $\mathbb{E}\|\widehat{f}_{\mathrm{or},\mathbf{M}} - f\|^2 \asymp sn^{-1}$ if $\mu > \rho$ and $\mathbb{E}\|\widehat{f}_{\mathrm{or},\mathbf{M}} - f\|^2 \asymp sn^{-2\mu/(2\rho)}$ otherwise. Note also that there is a significant difference between choosing optimal values of M_j in (3.5) and m in (3.2). Indeed, the coefficients of the dictionary functions c_{jk} in (3.5) are known, while coefficients c_k of f in (3.2) are unknown, so the former problem is a straightforward one while the latter one is not.

Since one does not have an oracle which allows to choose the "right" subset of dictionary functions φ_j, $j\in J$, Lasso is instrumental for choosing an appropriate subset such that, even if it does not coincide with the "true" subset J, it provides an estimator of a similar quality.

4. Lasso solution of a general linear inverse problem. Consider equation (1.1) described above with observations defined in (1.2). Denote by Q^* the conjugate operator for Q, so that $\langle Qf, g\rangle_{\mathcal{H}_2} = \langle f, Q^*g\rangle_{\mathcal{H}_1}$ for any $f\in\mathcal{H}_1$ and $g\in\mathcal{H}_2$. Unless there is an ambiguity, in what follows, we denote the scalar product induced norms in both $\mathcal{H}_1$ and $\mathcal{H}_2$ by $\|\cdot\|_2$.

Let $\{\varphi_j, j \in \mathcal{P}\}$ be a dictionary such that $\|\varphi_j\|_{\mathcal{H}_1} = 1$. Denote by f the true solution of the problem (1.1) and by $f_{\boldsymbol{\theta}}$ the projection of this true solution on the linear span of functions $\{\varphi_j, j \in \mathcal{P}\}$ where, for any $\mathbf{t} \in \mathbb{R}^p$, we denote

$$f_{\mathbf{t}} = \sum_{j=1}^{p} t_j \varphi_j. \tag{4.1}$$

If function f were known, we would search for the vector of coefficients $\boldsymbol{\theta}$ of $f_{\boldsymbol{\theta}}$ as a solution of the optimization problem

$$\boldsymbol{\theta} = \arg\min_{\mathbf{t}} \|f - f_{\mathbf{t}}\|_2^2,$$

where $f_{\mathbf{t}}$ is defined in (4.1). Note that, although f is unknown,

$$\|f - f_{\mathbf{t}}\|_2^2 = \|f\|_2^2 + \|f_{\mathbf{t}}\|_2^2 - 2\sum_{j=1}^{p} \langle f, \varphi_j \rangle_{\mathcal{H}_1} t_j \tag{4.2}$$

is the sum of three components where the first one, $\|f\|_2^2$, is independent of $\mathbf{t}$, and the second one, $\|f_{\mathbf{t}}\|_2^2$, is completely known. In order to estimate the last term in (4.2), we assume that the following condition holds:

(A0) There exist $\psi_j \in \mathcal{H}_2$ such that $Q^*\psi_j = \varphi_j$ and $\nu_j = \|\psi_j\|_{\mathcal{H}_2} < \infty$.

For example, if operator Q is defined by formula (1.3), then ψ_j in Assumption **(A0)** are solutions of the following equations:

$$\big(Q^*\psi_j\big)(t) = \int_c^d g(x,t)\psi_j(x)\,dx = \varphi_j(t), \qquad t \in (a,b). \tag{4.3}$$

Observe that equations resulting from Assumption **(A0)** have completely known right-hand sides. The values of ν_j can be viewed as the "price" of estimating coefficient θ_j of $f_{\boldsymbol{\theta}}$. While, in the regression set up, this "price" is uniform for all coefficients, this is no longer true in the case of ill-posed problems. In general, evaluating ψ_j, $j = 1, \ldots, p$, in Assumption **(A0)** can be computationally expensive. However, in many cases (as it happens, for instance, in Section 8), φ_j, $j = 1, \ldots, p$, can be inverted analytically, so one can evaluate ψ_j, $j = 1, \ldots, p$, directly. If this is impossible, one can evaluate functions ψ_j in advance, in an "off-line" mode, since functions φ_j are data independent. Moreover, since the inversion of each function φ_j is done independently, the procedure can be sped up by the use of parallel computing.

Under Assumption **(A0)**, one can write

$$\beta_j = \langle f, \varphi_j \rangle_{\mathcal{H}_1} = \big\langle f, Q^*\psi_j \big\rangle_{\mathcal{H}_1} = \langle Qf, \psi_j \rangle_{\mathcal{H}_2} = \langle q, \psi_j \rangle_{\mathcal{H}_2}, \tag{4.4}$$

so that

$$\beta_j = \mathbb{E}\langle y, \psi_j \rangle_{\mathcal{H}_2}. \tag{4.5}$$

For this reason, we can replace $\beta_j = \langle f, \varphi_j \rangle_{\mathcal{H}_1}$ in (4.2) by its estimator

$$\widehat{\beta}_j = \langle y, \psi_j \rangle_{\mathcal{H}_2} \tag{4.6}$$

and estimate the vector of coefficients $\boldsymbol{\theta}$ by

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\mathbf{t}} \left\{ \|f_{\mathbf{t}}\|_2^2 - 2\sum_{j=1}^{p} \widehat{\beta}_j t_j + \alpha \sum_{j=1}^{p} \nu_j |t_j| \right\}. \tag{4.7}$$

Note that (4.7) is the weighted Lasso problem with the penalty parameter α. The coefficients ν_j in front of $|t_j|$ are motivated by the fact that $\widehat{\beta}_j$ are centered normal variables with the variances $\nu_j^2 = \|\psi_j\|_2^2$.

In order to reduce optimization problem (4.7) to familiar matrix formulation, we introduce matrix $\boldsymbol{\Phi}$ with elements $\Phi_{jk} = \langle \varphi_j, \varphi_k \rangle$ and vector $\widehat{\boldsymbol{\beta}}$ with elements $\widehat{\beta}_j$. Define matrices $\mathbf{W}$ and $\boldsymbol{\Upsilon}$ by

$$\mathbf{W}^T\mathbf{W} = \boldsymbol{\Phi}, \qquad \boldsymbol{\Upsilon} = \operatorname{diag}(\nu_1, \ldots, \nu_p). \tag{4.8}$$

Then (4.7) can be re-written as

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\mathbf{t}} \{ \mathbf{t}^T \mathbf{W}\mathbf{W}^T \mathbf{t} - 2\mathbf{t}^T \widehat{\boldsymbol{\beta}} + \alpha \|\boldsymbol{\Upsilon}\mathbf{t}\|_1 \}. \tag{4.9}$$

Introducing vector $\boldsymbol{\gamma}$ such that $\mathbf{W}^T\boldsymbol{\gamma} = \widehat{\boldsymbol{\beta}}$ we reduce (4.9) to

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\mathbf{t}} \{ \|\mathbf{W}\mathbf{t} - \boldsymbol{\gamma}\|_2^2 + \alpha \|\boldsymbol{\Upsilon}\mathbf{t}\|_1 \}. \tag{4.10}$$

Here, $\|\boldsymbol{\Upsilon}\mathbf{t}\|_1$ is the weighted Lasso penalty, α is the penalty parameter and $\boldsymbol{\gamma} = (\mathbf{W}^T)^+ \widehat{\boldsymbol{\beta}} = (\mathbf{W}\mathbf{W}^T)^{-1}\mathbf{W}\widehat{\boldsymbol{\beta}}$ is the right-hand side.

Since we are interested in recovering f rather that $\boldsymbol{\theta}$ itself, we are using Lasso for solution of the prediction problem where it requires milder conditions on the dictionary. In particular, estimator $f_{\widehat{\boldsymbol{\theta}}}$ converges to the true function f with no additional assumptions on the dictionary.

THEOREM 1. *Let Assumption* **(A0)** *hold. Let* $\tau > 0$ *and*

$$\alpha_0 = \sigma n^{-1/2} \sqrt{2(\tau + 1) \log p}. \tag{4.11}$$

Then, for any $\alpha \geq \alpha_0$, *with probability at least* $1 - 2p^{-\tau}$, *one has*

$$\|f_{\widehat{\boldsymbol{\theta}}} - f\|_2^2 \leq \inf_{\mathbf{t}} \left[\|f_{\mathbf{t}} - f\|_2^2 + 4\alpha \sum_{j=1}^{p} \nu_j |t_j| \right]. \tag{4.12}$$

If the dictionary is large enough, so that $f_{\boldsymbol{\theta}} = f$ where vector $\boldsymbol{\theta}$ has support J of size $|J| = s$ and components of $\boldsymbol{\theta}$ are uniformly bounded, then with high probability, the error of estimating f by $f_{\widehat{\boldsymbol{\theta}}}$ is $\|f_{\widehat{\boldsymbol{\theta}}} - f\|_2^2 \asymp \sigma n^{-1/2} \sqrt{\log p} \sum_{j \in J} \nu_j$. In the case of the regression problem, $\nu_j = 1$, so that convergence rate appears

as $s\sigma n^{-1/2}\sqrt{\log p}$ and is called the *slow Lasso rate*, in comparison with the *fast Lasso rate* $\sigma^2 s n^{-1}\log p$ that can be obtained only if one of the, so-called, *compatibility* assumptions [see, e.g., Bühlmann and van de Geer (2011)] is satisfied.

In the case of the ill-posed problem (1.1), in order to achieve fast Lasso rate, we also need to formulate a compatibility assumption. For this purpose, we consider a set of p-dimensional vectors

$$\mathcal{J}(\mu, J) = \{\mathbf{d} \in \mathbb{R}^p : \|(\boldsymbol{\Upsilon}\mathbf{d})_{J^c}\|_1 \le \mu \|(\boldsymbol{\Upsilon}\mathbf{d})_J\|_1\}, \qquad \mu > 1, \tag{4.13}$$

where matrix $\boldsymbol{\Upsilon}$ is defined in (4.8). We assume that the following condition holds:

(A) Matrices $\boldsymbol{\Phi}$ and $\boldsymbol{\Upsilon}$ are such that for some $\mu > 1$ and any $J \subset \mathcal{P}$

$$\kappa^2(\mu, J) = \min\left\{\mathbf{d} \in \mathcal{J}(\mu, J), \|\mathbf{d}\|_2 \ne 0 : \frac{\mathbf{d}^T \boldsymbol{\Phi}\mathbf{d} \cdot \operatorname{Tr}(\boldsymbol{\Upsilon}_J^2)}{\|(\boldsymbol{\Upsilon}\mathbf{d})_J\|_1^2}\right\} > 0. \tag{4.14}$$

Assumption (4.14) is not easy to check in practice. For this reason, in Section A.1 in the Appendix, we provide verifiable sufficient conditions that guarantee that condition **(A)** holds with $\kappa^2(\mu, J)$ being uniformly bounded below by a quantity which is separated from zero.

Observe that, in the regression setup, $\boldsymbol{\Upsilon}$ is the identity matrix, and condition **(A)** reduces to the compatibility condition for general sets formulated in Section 6.2.3 of Bühlmann and van de Geer (2011). If one has an orthonormal basis instead of an overcomplete dictionary, then matrix $\boldsymbol{\Phi}$ is an identity matrix and, due to the Cauchy inequality, $\kappa^2(\mu, J) \ge 1$ for any μ and J. On the other hand, for an orthonormal basis, the bias $\|f_{\mathbf{t}} - f\|_2^2$ in (4.12) may be large. Under conditions **(A0)** and **(A)**, one obtains fast convergence rates for the Lasso estimator.

THEOREM 2. *Let Assumptions* **(A0)** *and* **(A)** *hold. Let* $\tau > 0$, $K_0 = 2$ *and* $\alpha = \varpi\alpha_0$ *where* $\varpi \ge (\mu+1)/(\mu-1)$ *and* α_0 *is defined in* (4.11). *Then, with probability at least* $1 - 2p^{-\tau}$, *one has*

$$\begin{aligned}\|f_{\widehat{\boldsymbol{\theta}}} - f\|_2^2 \le \inf_{\mathbf{t}, J\subseteq\mathcal{P}}\Bigg[&\|f_{\mathbf{t}} - f\|_2^2 + 4\alpha\sum_{j\in J^c}\nu_j|t_j| \\ &+ \frac{\sigma^2 K_0(1+\varpi)^2(\tau+1)}{\kappa^2(\mu, J)}\frac{\log p}{n}\sum_{j\in J}\nu_j^2\Bigg].\end{aligned} \tag{4.15}$$

Therefore,

$$\|f_{\widehat{\boldsymbol{\theta}}} - f\|_2^2 \le \inf_{J\subseteq\mathcal{P}}\Bigg[\|f - f_{\mathcal{L}_J}\|_2^2 + \frac{\sigma^2 K_0(1+\varpi)^2(\tau+1)}{\kappa^2(\mu, J)}\frac{\log p}{n}\sum_{j\in J}\nu_j^2\Bigg], \tag{4.16}$$

where $f_{\mathcal{L}_J} = \operatorname{proj}_{\mathcal{L}_J} f$.

Note that inequality (4.16) ensures that, up to a $\log p$ factor, the estimator $f_{\widehat{\boldsymbol{\theta}}}$ attains the minimum possible mean squared error for a particular function of interest f as long as compatibility factor $\kappa^2(\mu, J)$ stays uniformly bounded below. Indeed, if f were known, one would choose $J \subseteq \mathcal{P}$ and estimate f by its projection $\tilde{f}$ on $\mathcal{L}_J$, so that the overall error is bounded below by

$$\mathbb{E}\|\tilde{f} - f\|_2^2 \geq \min_{J \subset \mathcal{P}} \left\{ \|f - f_{\mathcal{L}_J}\|_2^2 + \frac{\sigma^2 \log p}{n\lambda_{\max}(|J|, \boldsymbol{\Phi})} \sum_{j \in J} \nu_j^2 \right\}, \tag{4.17}$$

where $\lambda_{\max}(\cdot)$ is defined in (2.1). If $\kappa^2(\mu, J)$ is bounded below by a constant, then the lower bound in (4.17) differs from the upper bound in (4.16) by a logarithmic factor $\log p$ that serves as a price for choosing a subset of dictionary functions.

Note that oracle inequalities (4.15) and (4.16) are sharp and, if compatibility condition holds, they contain smaller remainder term than the remainder term in Proposition 3 of Dalalyan and Salmon (2012). Indeed, for an arbitrary dictionary, Dalalyan and Salmon (2012) obtained the remainder term which is proportional to $n^{-1} \log p |J| \max_{j \in J} \nu_j^2$ and is larger than $n^{-1} \log p \sum_{j \in J} \nu_j^2$ in Theorem 2.

If one imposes a somewhat stronger condition,

$$\max_{j'} \sum_{j \neq j'} |\Phi_{jj'}| \leq 1 - \kappa_0 < 1 \tag{4.18}$$

for some $\kappa_0 \geq 0$, then $\lambda_{\min}(\boldsymbol{\Phi}) \geq \kappa_0 \geq \kappa^2(\mu, J)$ and Assumption **(A)** holds. This is a "low-dimensional" application of Lasso technique which, however, may be of use in some practical situations. In particular, if one uses an orthonormal dictionary, then $\kappa_0 = 1$ for any μ and J. Applying Theorem 2 with $\mu = 3$ and $\varpi = 2$, we obtain that

$$\|f_{\widehat{\boldsymbol{\theta}}} - f\|_2^2 \leq \inf_{\mathbf{t}, J \subseteq \mathcal{P}} \Bigg[\|f_{\mathbf{t}} - f\|_2^2 + 8\sigma n^{-1/2} \sqrt{2(\tau+1)\log p} \sum_{j \in J^c} \nu_j |t_j| + 18(\tau+1)\sigma^2 n^{-1} \log p \sum_{j \in J} \nu_j^2 \Bigg]$$

and

$$\|f_{\widehat{\boldsymbol{\theta}}} - f\|_2^2 \leq \inf_{J \subseteq \mathcal{P}} \left\{ \|f - f_{\mathcal{L}_J}\|_2^2 + 18(\tau+1)\sigma^2 n^{-1} \log p \sum_{j \in J} \nu_j^2 \right].$$

The bounds above are simpler than the ones derived in Cohen, Hoffmann and Reiß (2004) and, in addition, they hold for any p and n.

5. Observational model. Consider a real-life observational model corresponding to equation (1.1)

$$y_i = q(x_i) + \xi_i, \qquad i = 1, \ldots, n, \tag{5.1}$$

where ξ_i are i.i.d. centered sub-Gaussian random variables such that for some σ and any t

$$\mathbb{P}(|\xi_i| > t) \le \exp(-t^2/2\sigma^2). \tag{5.2}$$

Assume that $x_i \in \mathcal{X}$, $i = 1, \dots, n$, are fixed nonrandom points where $y(x)$ in equation (1.1) is measured. To be more specific, we consider the case when $\mathcal{X} = [a, b]$ is an interval, $a = x_0 < x_1 < \cdots < x_n = b$ and $\mathcal{H}_2 = L^2[a, b]$, so that

$$\beta_j = \langle q, \psi_j \rangle_{\mathcal{H}_2} = \int_a^b q(x)\psi_j(x)\,dx. \tag{5.3}$$

Denote $T = b - a$, $\Delta x_i = x_i - x_{i-1}$. Observe that, similarly to the white noise model, estimator $\widehat{\boldsymbol{\theta}}$ can be recovered as a solution of optimization problem (4.7) with the only difference that β_j in (5.3) can no longer be estimated by (4.6). Instead, we approximate β_j in (5.3) using rectangle formula and then replace $q(x_i)$ by y_i, $i = 1, \dots, n$, obtaining new values of ν_j and $\widehat{\beta}_j$

$$\widehat{\beta}_j = \frac{1}{n}\sum_{i=1}^{n} y_i \psi_j(x_i)\Delta x_i, \qquad \nu_j^2 = \frac{T^2}{n}\sum_{i=1}^{n} \psi_j^2(x_i). \tag{5.4}$$

We search for $\widehat{\boldsymbol{\theta}}$ as a solution of optimization problem (4.7) [or (4.9)] with $\widehat{\beta}_j$ and ν_j given by (5.4). If $|\Delta x_i|$ are small and n is large enough, then rectangle rule approximations of β_j, $j = 1, \dots, p$, have good precisions, so one expects to estimate f on the basis of discrete data in (5.1) as well as on the basis of the white noise model (1.1). Theorem 3 below shows that this indeed is true.

THEOREM 3. *Let Assumptions* **(A0)** *and* **(A)** *hold and* $\tau > 0$ *be an arbitrary constant. Denote*

$$\aleph = \max_{1 \le j \le p}\left[\frac{1}{\nu_j}\max_{x \in \mathcal{X}}\left|\frac{d[q(x)\psi_j(x)]}{dx}\right|\right]. \tag{5.5}$$

If for some nonnegative constant ϑ, *one has*

$$\max_i |\Delta x_i| \le \vartheta\frac{T}{n} \quad \text{and} \quad n \ge \mathcal{N} = \frac{T^4\aleph^2}{4K_0\sigma^2(\tau+1)\log p}, \tag{5.6}$$

then, for $\alpha_0 = 2\vartheta n^{-1/2}\sigma\sqrt{2(\tau+1)\log p}$, $\alpha = \varpi\alpha_0$ *with* $\varpi \ge (\mu+1)/(\mu-1)$ *and* $K_0 = 8\vartheta^2$, *inequalities* (4.15) *and* (4.16) *hold with probability at least* $1 - ep^{-\tau}$.

Note that the estimator $f_{\widehat{\boldsymbol{\theta}}}$ is fully adaptive since α_0 is known. The lower bound $n \ge \mathcal{N}$ for n is motivated by the fact that the rectangular rule approximations of the integrals in (5.3) should be close in value to those integrals. In addition, if functions ψ_j and q are smooth, so that functions $q\psi_j$ have uniformly bounded second derivatives, one can replace the rectangular rule for calculating β_j by the trapezoid rule. In this case, oracle inequalities in Theorem 3 can be obtained with a smaller value of $\mathcal{N}$.

6. Lasso recovery of a mixing density from a continuous mixture. In this section, we show that, with a small modification, the method used in the previous sections, can be applied to estimation of the mixing density in a continuous mixture. Consider the situation when one observes a random sample $Y_1, Y_2, \dots, Y_n$ of a random variable Y with an unknown probability density function $q(y)$, $y \in \mathcal{Y}$, of the form

$$q(y) = \int_{\mathcal{X}} g(y \mid x) f(x)\, dx, \qquad y \in \mathcal{Y}, \tag{6.1}$$

where $g(y \mid x)$ is a known conditional density of Y given $X = x$, $x \in \mathcal{X}$, and $f(x)$, $x \in \mathcal{X}$, is an unknown mixing density of interest. If $g(y \mid x) = g(y - x)$, then problem (6.1) reduces to the extensively studied density deconvolution problem [see, e.g., Meister (2009) and references therein]. In a general set up, problem (6.1) is usually solved by expanding f over some orthonormal dictionary and then recovering coefficients of the expansion [see, e.g., Comte and Genon-Catalot (2015), Hengartner (1997) and Walter (1981)], by the kernel method [see, e.g., Goutis (1997)] or by maximizing the likelihood [see, e.g., Liu, Levine and Zhu (2009)]. It is easy to see that when the conditional density $g(y \mid x)$ is known, the problem of recovering f in (6.1) on the basis of observations from q can be viewed as a particular case of the linear inverse problem (5.1) with the main difference that one can sample from the p.d.f. q instead of having noisy observations of the values of q. Therefore, one can easily estimate any linear functional of q, so that, indeed, observations are taken in the form (1.2). For this reason, in this set up, one again can benefit from using a large overcomplete dictionary which allows a compact representation of f.

Let, as before, $\{\varphi_k\}_{k=1}^p$ be a dictionary and function f be expanded over this dictionary yielding its approximation (4.1). The goal is to recover the vector of coefficients $\boldsymbol{\theta}$. By introducing Hilbert spaces $\mathcal{H}_1 = L^2(\mathcal{X})$ and $\mathcal{H}_2 = L^2(\mathcal{Y})$ and a linear operator $Q : \mathcal{H}_1 \to \mathcal{H}_2$ given by

$$(Qf)(y) = \int_{\mathcal{X}} g(y \mid x) f(x)\, dx \qquad \text{with } \big(Q^* u\big)(x) = \int_{\mathcal{Y}} g(y \mid x) u(y)\, dy, \tag{6.2}$$

one can essentially reduce the problem (6.1) to (1.1). Indeed, observe that in (4.4) one has

$$\beta_j = \langle f, \varphi_j \rangle_{\mathcal{H}_1} = \langle q, \psi_j \rangle_{\mathcal{H}_2} = \mathbb{E}\big[\psi_j(Y_1)\big].$$

Hence, once again we search for $\widehat{\boldsymbol{\theta}}$ as a solution of optimization problem (4.7) [or (4.9)] with $\widehat{\beta}_j$ and ν_j given by

$$\widehat{\beta}_j = \frac{1}{n} \sum_{i=1}^n \psi_j(Y_i), \qquad \nu_j = \|\psi_j\|_\infty. \tag{6.3}$$

Since $\widehat{\beta}_j$ are unbiased estimators of β_j, $j = 1, \dots, p$, with variances bounded by ν_j^2, then, for n large enough, the values of $\widehat{\beta}_j$ are uniformly close to β_j, $j = 1, \dots, p$, so that following statement is true.

THEOREM 4. *Let Assumptions* **(A0)** *and* **(A)** *hold. Let* $\widehat{\beta}_j$ *and* ν_j *be defined in* (6.3), τ *be any positive constant and* $\alpha_0 = 2n^{-1/2}\sqrt{(\tau+1)\log p}$. *Let* $\alpha = \varpi\alpha_0$ *with* $\varpi \geq (\mu+1)/(\mu-1)$. *If* $n \geq \mathcal{N}_0 = 16/9(\tau+1)\log p$, *then with probability at least* $1 - 2p^{-\tau}$, *inequalities* (4.15) *and* (4.16) *hold with* $K_0 = 4$.

Note that, despite the fact that the idea of this section seems to be similar to Bunea et al. (2010), we consider a different problem and apply a completely novel approach. Indeed, in Bunea et al. (2010), the authors assumed that observations from the density of interest are available which makes their problem similar to the regression problem. On the contrary, in our case, observations from the density of interest are unavailable which leads to the difficulties that are experienced in the context of the ill-posed linear inverse problems. Really, though expansion (4.1) leads to $q = \sum_j \theta_j u_j$ with $u_j = Q\varphi_j$, due to contraction imposed by operator Q, the system of functions $\{u_j, j \in \mathcal{P}\}$ does not meet compatibility condition even if $\{\varphi_j, j \in \mathcal{P}\}$ does. On the other hand, if one starts with an incoherent dictionary $\{u_j, j \in \mathcal{P}\}$, the system of functions $v_j = Q^{-1}u_j$ may be totally inappropriate for estimating f.

REMARK 1 (Smaller penalties). Note that $\nu_j^2 = \|\psi_j\|_\infty^2$ in (6.3) can be replaced by a smaller value $\nu_j^2 = \mathrm{Var}[\psi_j(Y_1)]$ which leads to a smaller overall error, provided the number of observations n is large enough, in particular,

$$n \geq \mathcal{N}_1 = \max_{1\leq j\leq p}\left[\frac{16(\tau+1)\log p\|\psi_j\|_\infty^2}{9\,\mathrm{Var}[\psi_j]}\right].$$

The latter guarantees that the values of $\sqrt{n}|\widehat{\beta}_j - \beta_j|/\nu_j$ are uniformly bounded above with high probability.

Note that, though $\mathrm{Var}[\psi_j(Y_1)]$ is unavailable (since f is unknown), one can easily construct an upper bound for ν_j^2

$$\nu_j^2 \leq \max_{x\in\mathcal{X}}\left[\int_{\mathcal{Y}} g(y\mid x)\psi_j^2(y)\,dy\right] \tag{6.4}$$

or estimate $\mathrm{Var}[\psi_j(Y_1)]$ from observations.

REMARK 2 (Estimation by a density function). Estimator $f_{\widehat{\boldsymbol{\theta}}}$ obtained as a solution of optimization problem (4.7) with $\widehat{\beta}_j$ and ν_j given by (6.3) is not necessarily a probability density function since we do not require the dictionary functions to be nonnegative and the weights to be such that $f_{\widehat{\boldsymbol{\theta}}}$ integrates to one. This, however, can be easily accomplished in the context of Lasso estimator if one uses dictionary functions that are p.d.f.s themselves and adds the constraints that the coefficients are nonnegative and sum to one. Note that since we are using the weighted Lasso penalty, those constraints do not allow to get rid of the penalty term altogether though the nonnegativity condition should make compatibility assumption **(A)** weaker. However, pursuing this extension of the Lasso solution is a matter of a future investigation.

7. Approximate inverse images of the dictionary functions. Condition **(A0)** requires that each dictionary function φ_j has an exact inverse image ψ_j such that $Q^*\psi_j = \varphi_j$ and $\nu_j = \|\psi_j\|_2 < \infty$. Sometimes this is not true since functions ψ_j may not be easy to construct or they may have infinite norms. In this situation, arguments of Section 3 suggest that exact inverse images ψ_j can be replaced by approximate ones $\psi_{j,\delta}$.

First, let us consider the setting of Section 4 where observations are taken in the form (1.2) and $\mathcal{H}_1$ and $\mathcal{H}_2$ are spaces of square integrable functions. Let functions $\varphi_{j,\delta}$ be such that $\|\varphi_{j,\delta} - \varphi_j\|_2 \le \delta_j$ and $\psi_{j,\delta}$ be the solutions of the equations $Q^*\psi_{j,\delta} = \varphi_{j,\delta}$ with $\nu_{j,\delta} = \|\psi_{j,\delta}\|_2 < \infty$. Then β_j can be estimated by $\widehat{\beta}_{j,\delta} = \langle y, \psi_{j,\delta}\rangle$, so that

$$\widehat{\beta}_{j,\delta} = \beta_j + \sigma n^{-1/2}\nu_{j,\delta}\eta_j + h_{j,\delta}, \qquad h_{j,\delta} = \langle q, \psi_{j,\delta} - \psi_j\rangle, \tag{7.1}$$

where η_j are standard normal variables. Hence, application of Lemma 2 with $K = \sqrt{2}$, $\tau > 0$,

$$\begin{aligned} C_{h\delta} &= \max_{1\le j\le p}\left[\frac{|h_{j,\delta}|\sqrt{n}}{\sigma\nu_{j,\delta}\sqrt{2(\tau+1)\log p}}\right], \\ C_{\alpha\delta} &= \sqrt{2(\tau+1)}(1 + C_{h\delta}), \end{aligned} \tag{7.2}$$

$\alpha_0 = C_{\alpha\delta}\sigma n^{-1/2}\sqrt{\log p}$ and $\alpha = \varpi\alpha_0$ where $\varpi \ge (\mu+1)/(\mu-1)$, yields with probability at least $1 - 2p^{-\tau}$, that (4.12), (4.15) and (4.16) hold [with $K_0 = 2(1 + C_{h\delta})^2$].

In order to construct functions $\varphi_{j,\delta}$ and $\psi_{j,\delta}$ for a given dictionary function φ_j and a given operator Q, consider operator $QQ^* : \mathcal{H}_1 \to \mathcal{H}_1$ and a parameter $\delta > 0$. Construct function $\psi_{j,\delta} = (QQ^* + \delta I)^{-1}Q\varphi_j$ where $I : \mathcal{H}_1 \to \mathcal{H}_1$ is the identity operator. Since the relation (7.1) holds for any $\delta > 0$, the value of δ can be chosen so to minimize the mean squared error of estimating β_j by $\widehat{\beta}_{j,\delta}$

$$\mathbb{E}(\widehat{\beta}_{j,\delta} - \beta_j)^2 = \sigma^2 n^{-1}\nu_{j,\delta}^2 + h_{j,\delta}^2 = \sigma^2 n^{-1}\nu_{j,\delta}^2 + \big[\langle q, \psi_{j,\delta} - \psi_j\rangle\big]^2. \tag{7.3}$$

Note that the values of $\nu_{j,\delta}$ in (7.3) are completely known. The values of $h_{j,\delta}$ are unknown but can be estimated from observations. Indeed, if observations are available in the form (1.2) or (5.1), one can construct a kernel or a projection estimator $\hat{q}$ of q and then replace q by $\hat{q}$ in (7.3). In the case of recovery of a mixing density in a continuous mixture considered in Section 6, the values of $h_{j,\delta}$ in (7.3) are of the forms

$$h_{j,\delta} = \mathbb{E}(\widehat{\beta}_{j,\delta}) - \beta_j = \mathbb{E}\big[\psi_{j,\delta}(Y) - \psi_j(Y)\big], \qquad j = 1, \dots, p,$$

and can be estimated by their sample averages.

8. Estimation of the density of the matrix parameter of the Wishart distribution. In this section, we consider an application of the theory above. In order to show capabilities of the Lasso technique, in Section 8, we study estimation of the unknown density function of the matrix parameter of the Wishart distribution. This type of problems is very hard to handle by traditional methods due to the curse of dimensionality.

Let $\mathbf{Y} \mid \mathbf{X} \sim \text{Wishart}(m, \mathbf{X})$, where $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{r\times r}$ are symmetric positive definite r-dimensional matrices:

$$g(\mathbf{Y} \mid \mathbf{X}) = \frac{|\mathbf{Y}|^{(m-r-1)/2}2^{-mr/2}}{|\mathbf{X}|^{m/2}\Gamma_r(m/2)} \exp\biggl\{-\frac{1}{2}\operatorname{Tr}(\mathbf{X}^{-1}\mathbf{Y})\biggr\}, \tag{8.1}$$
$$\mathbf{X}, \mathbf{Y} > 0, m > 3r,$$

where $\Gamma_r(a)$ is the multivariate gamma function [see, e.g., Gupta and Nagar (2000), Section 1.4]

$$\Gamma_r(a) = \pi^{r(r-1)/4} \prod_{l=1}^{r} \Gamma\biggl(\frac{2a-l+1}{2}\biggr). \tag{8.2}$$

Consider the situation when, given $\mathbf{X}_i = \mathbf{X}$, matrix $\mathbf{Y}_i$ has the Wishart p.d.f. of the form (8.1), $i = 1, \dots, n$, and matrices $\mathbf{X}_1, \dots, \mathbf{X}_n$ are independent with the common unknown p.d.f. $f(\mathbf{X})$. Here, matrices $\mathbf{Y}_1, \dots, \mathbf{Y}_n$ are available for observation but $\mathbf{X}_1, \dots, \mathbf{X}_n$ are not. The objective is to estimate the p.d.f. $f(\mathbf{X})$ of the unknown matrix parameter $\mathbf{X}$ on the basis of observations $\mathbf{Y}_1, \dots, \mathbf{Y}_n$ of $\mathbf{Y}$. This problem appears, for example, when one has several equal size samples from the multivariate normal distributions with the different unknown covariance matrices $\mathbf{X}_1, \dots, \mathbf{X}_n$ that are related by a common p.d.f. $f(\mathbf{X})$. An estimator $\widehat{f}(\mathbf{X})$ of $f(\mathbf{X})$ can be used, for example, as a prior distribution in the subsequent Bayesian inference.

It is a well-known fact that, even for moderate values of r, an estimator will suffer from the curse of dimensionality. In order to circumvent this difficulty, we estimate $f(\mathbf{X})$ using an overcomplete dictionary. In this example, $\mathcal{X} = \mathcal{Y}$ are the spaces of symmetric nonnegative definite matrices in $\mathbb{R}^{r\times r}$ and $\mathcal{H}_1 = \mathcal{H}_2$ are the Hilbert spaces of square integrable functions on $\mathcal{X} = \mathcal{Y}$. We choose a dictionary that consists of a collection of mixtures of inverse Wishart densities since this is a wide class, so that, the true density $f(\mathbf{X})$ either belongs to this class or is well approximated by it. In particular, we choose the dictionary functions of the form

$$\varphi_j(\mathbf{X}) = C_{\mathbf{A}_j,\gamma_j} u(\mathbf{X} \mid \mathbf{A}_j, \gamma_j), \qquad j = 1, \dots, p,$$

with $2r < \gamma_j < m - r$, where $u(\mathbf{X} \mid \mathbf{A}, \gamma)$ is the inverse Wishart density and $C_{\mathbf{A},\gamma}$ is the normalizing constant, such that $u(\mathbf{X} \mid \mathbf{A}, \gamma)$ has the unit L^2-norm:

$$\varphi(\mathbf{X}) = \varphi(\mathbf{X} \mid \mathbf{A}, \gamma) = C_{\mathbf{A},\gamma} u(\mathbf{X} \mid \mathbf{A}, \gamma) \qquad \text{with } \bigl\|\varphi(\mathbf{X})\bigr\|_2 = 1,$$
$$u(\mathbf{X} \mid \mathbf{A}, \gamma) = \frac{2^{-(\gamma-r-1)r/2}|\mathbf{A}|^{(\gamma-r-1)/2}}{\Gamma_r((\gamma-r-1)/2)|\mathbf{X}|^{\gamma/2}} \exp\biggl\{-\frac{1}{2}\operatorname{Tr}(\mathbf{X}^{-1}\mathbf{A})\biggr\}, \qquad \mathbf{X}, \mathbf{A} > 0. \tag{8.3}$$

By direct calculations (see Section A.3), it is easy to check that

$$C_{\mathbf{A},\gamma} = \Gamma_r\left(\frac{\gamma - r - 1}{2}\right)\left[\Gamma_r\left(\frac{2\gamma - r - 1}{2}\right)\right]^{-1/2} 2^{r(\gamma-r-1)/2}|\mathbf{A}|^{(r+1)/4}, \tag{8.4}$$

and that matrix $\mathbf{\Phi}$ in (4.8) has components of the forms

$$\begin{aligned}\Phi_{i,j} &= \langle \varphi_i, \varphi_j \rangle \\ &= \frac{\Gamma_r((\gamma_i + \gamma_j - r - 1)/2)2^{r(\gamma_i+\gamma_j-r-1)/2}}{[\Gamma_r((2\gamma_i - r - 1)/2)\Gamma_r(2\gamma_j - r - 1/2)]^{1/2}} \\ &\quad \times \frac{|\mathbf{A}_i|^{(2\gamma_i-r-1)/4}|\mathbf{A}_j|^{(2\gamma_j-r-1)/4}}{|\mathbf{A}_i + \mathbf{A}_j|^{(\gamma_i+\gamma_j-r-1)/2}}.\end{aligned} \tag{8.5}$$

Functions $\psi_j(\mathbf{Y})$ in (6.3) are solutions of equations $Q^*\psi_j = \varphi_j$ where operator Q^* is defined in (6.2). It is easy to verify (see Section A.3) that functions $\psi_j(\mathbf{Y})$ are of the forms

$$\psi_j(\mathbf{Y}) = \psi(\mathbf{Y} \mid \mathbf{A}_j, \gamma_j) = C_{\mathbf{A}_j,\gamma_j} v(\mathbf{Y} \mid \mathbf{A}_j, \gamma_j), \qquad j = 1, \ldots, p, \tag{8.6}$$

where $C_{\mathbf{A},\gamma}$ is defined in (8.4) and $v(\mathbf{Y} \mid \mathbf{A}, \gamma)$ is the solution of the equation

$$\int_{\mathcal{Y}} g(\mathbf{Y} \mid \mathbf{X}) v(\mathbf{Y} \mid \mathbf{A}, \gamma)\, d\mathbf{Y} = u(\mathbf{X} \mid \mathbf{A}, \gamma). \tag{8.7}$$

Here, $g(\mathbf{Y} \mid \mathbf{X})$ and $u(\mathbf{X} \mid \mathbf{A}, \gamma)$ are defined by, respectively, formulae (8.1) and (8.3), and the integral is calculated over the space $\mathcal{Y}$ of all $(r \times r)$ symmetric non-negative definite matrices. By straightforward calculus (see Section A.3), derive that

$$\begin{aligned}\psi(\mathbf{Y} \mid \mathbf{A}, \gamma) &= \frac{\Gamma_r(m/2)2^{\gamma r/2}|\mathbf{A}|^{(2\gamma-r-1)/4}}{\Gamma_r(m - \gamma/2)\sqrt{\Gamma_r(2\gamma - r - 1/2)}} \\ &\quad \times \frac{|\mathbf{Y} - \mathbf{A}|^{(m-\gamma-r-1)/2}}{|\mathbf{Y}|^{(m-r-1)/2}} \mathbb{I}(\mathbf{Y} - \mathbf{A} > 0).\end{aligned} \tag{8.8}$$

Then Theorem 4 yields the following corollary.

COROLLARY 1. *Let Assumption* **(A)** *hold with the matrix* $\mathbf{\Phi}$ *defined in* (8.5) *and*

$$\begin{aligned}\nu_j &= \|\psi_{\mathbf{A}_j,\gamma_j}\|_\infty \\ &= \frac{\Gamma_r(m/2)(m - \gamma_j - r - 1)^{r(m-\gamma_j-r-1)/2}(2\gamma_j)^{r\gamma_j/2}}{\Gamma_r((m - \gamma_j)/2)\sqrt{\Gamma_r((2\gamma_j - r - 1)/2)}(m - r - 1)^{r(m-r-1)/2}} \\ &\quad \times |\mathbf{A}_j|^{-(r+1)/4}.\end{aligned} \tag{8.9}$$

Let $\widehat{\beta}_j$ *be given by* (6.3), τ *be any positive constant,* $\alpha_0 = 2n^{-1/2}\sqrt{(\tau+1)\log p}$ *and* $\alpha = \varpi\alpha_0$ *with* $\varpi \geq (\mu+1)/(\mu-1)$. *If* $n \geq \mathcal{N}_0 = 16/9(\tau+1)\log p$, *then with probability at least* $1 - 2p^{-\tau}$, *inequalities* (4.15) *and* (4.16) *hold with* $K_0 = 4$.

9. Discussion. In the present paper, we consider application of the Lasso technique to a general linear inverse problem. The approach is based on expanding the unknown function over an overcomplete dictionary and inverting of each of the dictionary functions in order to match the resulting expansion to the true function f. We investigate the white noise formulation of the problem and further extend the theory to the case of discrete observations with Gaussian or sub-Gaussian noise. In addition, we explain how this methodology can be used when the inverse images of the dictionary functions are replaced by their approximate versions. We also show how the technique suggested in the paper can be extended to the problem of estimation of a mixing density in a continuous mixture.

Using an example of the Laplace convolution equation, we study performance of the Lasso-based estimators via simulations and compare their precisions with the SVD estimators, the wavelet-vaguelette estimators and the estimators based on the expansion of the unknown function via the Laguerre functions basis. We show that as long as the function of interest f has an efficient representation in the overcomplete dictionary, the Lasso estimator yields satisfactory reconstruction.

Although in the paper we assume that the linear operator Q is completely known, the theory can be extended to the case when operator Q is measured with error or is estimated from the data. The advantage of the approach of the paper is that it naturally partitions the problem of solution of a linear inverse problem with a noisy operator and a right-hand side measured with error into two easier problems: solution of an inverse linear problem with the noisy operator and completely known right-hand side, and estimation of the linear functional of the right-hand side on the basis of its noisy version. However, solution of a general linear ill-posed problems with a noisy operator lies outside the scope of the present paper and will be treated in future.

APPENDIX

A.1. Discussion of the compatibility condition. Note that condition (4.14) is guaranteed by combination of two kinds of assumptions.

The first condition needs to ensure that the dictionary $\{\varphi_j, j \in \mathcal{P}\}$ is incoherent. The latter can be warranted by one of the following alternative assumptions introduced in Bickel, Ritov and Tsybakov (2009). In what follows, $\lambda_{\min}$, $\lambda_{\max}$ and ϱ refer to matrix $\mathbf{\Phi}$.

(A1(a)) For some s, $1 \le s \le p/2$, some $m \ge s$ such that $s + m \le p$ and some constant C_0 one has

$$m\lambda_{\min}(s+m) > C_0^2 s \lambda_{\max}(m), \tag{A.1}$$

where $\lambda_{\min}(s+m)$ and $\lambda_{\max}(m)$ are restricted eigenvalues defined in (2.1).

(A1(b)) For some s, $1 \le s \le p/2$, and some constant C_0 one has

$$\varrho < \big[s(2C_0+1)\big]^{-1}, \tag{A.2}$$

where ϱ is defined in (2.2).

Under Assumption **(A1(a))** or **(A1(b))**, small modifications of Lemma 4.1 of Bickel, Ritov and Tsybakov (2009) leads to the following result:

LEMMA 1 [Lemma 4.1 of Bickel, Ritov and Tsybakov (2009)]. *Let Assumption* **(A1(a))** *or* **(A1(b))** *be valid with* $C_0 = \mu C_\nu$. *Then, for any set* $J \in \mathcal{G}(C_\nu)$ *of cardinality* $|J| \le s$, *Assumption* **(A)** *holds with* $\kappa^2(\mu, J) \ge \vartheta(s, m, \mu, C_\nu)$ *where*

$$\vartheta(s, m, \mu, C_\nu) = \begin{cases} \lambda_{\min}(s+m)\left(1 - \dfrac{\mu C_\nu \sqrt{s\lambda_{\max}(m)}}{\sqrt{m\lambda_{\min}(s+m)}}\right)^2, & \text{if } \textbf{(A1(a))} \text{ holds}, \\ 1 - \left[s(2\mu C_\nu + 1)\right]^{-1}, & \text{if } \textbf{(A1(b))} \text{ holds}. \end{cases} \tag{A.3}$$

The second assumption guarantees that the function of interest is not too hard to estimate. As we have already mentioned, since the "price" of estimating coefficients varies from one dictionary function to the other, one needs to make sure that Lasso selects coefficients with relatively low variances and sets to zero the ones with high variances. This would be useful if the true function f does not have those components. For this purpose, we consider the set of subsets $J \subset \mathcal{P}$ such that

$$\mathcal{G}(C_\nu) = \left\{ J \subseteq \mathcal{P} : \max_{j \in J, j' \in J^c} \frac{\nu_j}{\nu_{j'}} \le C_\nu \right\}. \tag{A.4}$$

We assume that the true function f is such that its best approximation can be achieved using $J \in \mathcal{G}(C_\nu)$.

(A2) For some $\mu > 0$, $C_\nu > 0$ and some $H_0 > 0$, one has

$$\widehat{J} = \arg\min\left\{ J \subset \mathcal{P} : \|f - f_{\mathcal{L}_J}\|_2^2 + \frac{H_0\sigma^2}{\vartheta(s, m, \mu, C_\nu)} \frac{\log p}{n} \sum_{j \in J} \nu_j^2 \right\} \in \mathcal{G}(C_\nu). \tag{A.5}$$

Note that Assumption **(A2)** is natural and is similar to the usual assumptions that f is smooth and does not have fast oscillating components. In the context of the ill-posed problems, Assumption **(A2)** means that f is not "too hard" to estimate.

If Assumption **(A2)** is valid, then one can replace $J \subset \mathcal{P}$ by $J \in \mathcal{G}(C_\nu)$ in the inequality (4.16). For $J \in \mathcal{G}(C_\nu)$, Assumption **(A1(a))** [or **(A1(b))**] yields convenient lower bound (A.3) on the compatibility factor $\kappa^2(\mu, J)$ in Assumption **(A)**. Combination of (4.16) and (A.3) ensures that if f allows sparse representation in the dictionary $\{\varphi_j, j \in \mathcal{P}\}$, so that set $\widehat{J}$ in Assumption **(A2)** has at most s components, then Lasso provides an optimal (up to a logarithmic factor) representation of the function f. In particular, the following Corollary of Theorem 2 is valid.

COROLLARY 2. *Let $\tau > 0$ and $\alpha = \varpi\alpha_0$ where α_0 is defined in (4.11). Let Assumptions* **(A0)**, **(A1(a))** *or* **(A1(b))** *and* **(A2)** *hold with some $\mu > 0$ and $C_\nu > 0$, $|\widehat{J}| \le s$, $C_0 = \mu C_\nu$ and $H_0 \ge 8(1+\varpi)^2(\tau+1)$. If $\varpi \ge (\mu+1)/(\mu-1)$, then, with probability at least $1 - 2p^{-\tau}$, one has*

$$\|f_{\widehat{\boldsymbol{\theta}}} - f\|_2^2 \le \inf_{J\subseteq\mathcal{P}}\left\{\|f - f_{\mathcal{L}_J}\|^2 + \frac{H_0\sigma^2}{\vartheta(s,m,\mu,C_\nu)}\frac{\log p}{n}\sum_{j\in J}\nu_j^2\right\}. \tag{A.6}$$

Finally, we comment about the choice of m in Assumption **(A1(a))**. Similarly to regression set up, this choice depends on how fast the minimal eigenvalues of the order m sub-matrices of $\boldsymbol{\Phi}$ are decreasing as functions of m [see, e.g., Bickel, Ritov and Tsybakov (2009)].

A.2. Proofs. Validity of Theorems 1–4 rely on the following Lemma, the proof of which follows the lines of reasoning in Dalalyan, Hebiri and Lederer (2014). However, since we are interested in weighted Lasso and allow for non-centered errors, for completeness, we provide the proof of the lemma below.

LEMMA 2. *Let f be the true function and $f_{\boldsymbol{\theta}}$ be its projection onto the linear span of the dictionary $\mathcal{L}_\mathcal{P}$. Consider solution of the weighted Lasso problem (4.9) with $\boldsymbol{\Phi} = \mathbf{W}^T\mathbf{W}$, $\boldsymbol{\beta} = \boldsymbol{\Phi}\boldsymbol{\theta}$ and $\widehat{\boldsymbol{\beta}} = \mathbf{W}^T\boldsymbol{\gamma}$. Let*

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \sqrt{\varepsilon}\boldsymbol{\Upsilon}\boldsymbol{\eta} + \mathbf{h}, \qquad \boldsymbol{\eta}, \mathbf{h} \in \mathbb{R}^p, \tag{A.7}$$

where $\mathbf{h}$ is a nonrandom vector, $\mathbb{E}\boldsymbol{\eta} = 0$ and components η_j of $\boldsymbol{\eta}$ are sub-Gaussian random variables satisfying, for some $K > 0$ and any t,

$$\mathbb{P}\big(|\eta_i| > t\big) \le 2\exp\big(-t^2/K^2\big). \tag{A.8}$$

Choose $\tau > 0$ and denote

$$C_h = \max_{1\le j\le p}\left[\frac{|h_j|}{\nu_j\sqrt{\varepsilon\log p}}\right], \qquad C_\alpha = K\sqrt{\tau+1} + C_h. \tag{A.9}$$

If $\alpha_0 = C_\alpha\sqrt{\varepsilon\log p}$, then for any $\tau > 0$ and any $\alpha \ge \alpha_0$, with probability at least $1 - 2p^{-\tau}$, one has

$$\|f_{\widehat{\boldsymbol{\theta}}} - f\|_2^2 \le \inf_{\mathbf{t}}\big[\|f_{\mathbf{t}} - f\|_2^2 + 4\alpha\|\boldsymbol{\Upsilon}\mathbf{t}\|_1\big]. \tag{A.10}$$

Moreover, if Assumption **(A)** *holds and $\alpha = \varpi\alpha_0$ where $\varpi \ge (\mu+1)/(\mu-1)$, then for any $\tau > 0$ with probability at least $1 - 2p^{-\tau}$, one has*

$$\begin{aligned}\|f_{\widehat{\boldsymbol{\theta}}} - f\|_2^2 \le \inf_{\mathbf{t}, J\subseteq\mathcal{P}}\Big[&\|f_{\mathbf{t}} - f\|_2^2 + 4\alpha\big\|(\boldsymbol{\Upsilon}\mathbf{t})_{J^c}\big\|_1 \\ &+ \frac{(1+\varpi)^2C_\alpha^2}{\kappa^2(\mu,J)}\varepsilon\log p\sum_{j\in J}\nu_j^2\Big].\end{aligned} \tag{A.11}$$

PROOF. Following Dalalyan, Hebiri and Lederer (2014), by K–K–T condition, we derive for any $\mathbf{t} \in \mathbb{R}^p$

$$\widehat{\boldsymbol{\theta}}^T(\widehat{\boldsymbol{\beta}} - \boldsymbol{\Phi}\widehat{\boldsymbol{\theta}}) = \alpha\|\boldsymbol{\Upsilon}\widehat{\boldsymbol{\theta}}\|_1, \qquad \mathbf{t}^T(\widehat{\boldsymbol{\beta}} - \boldsymbol{\Phi}\widehat{\boldsymbol{\theta}}) \le \alpha\|\boldsymbol{\Upsilon}\mathbf{t}\|_1,$$

so that, subtracting the first line from the second, we obtain

$$(\widehat{\boldsymbol{\theta}} - \mathbf{t})^T(\boldsymbol{\Phi}\widehat{\boldsymbol{\theta}} - \widehat{\boldsymbol{\beta}}) \le \alpha\big(\|\boldsymbol{\Upsilon}\mathbf{t}\|_1 - \|\boldsymbol{\Upsilon}\widehat{\boldsymbol{\theta}}\|_1\big). \tag{A.12}$$

Since $\boldsymbol{\Phi}\boldsymbol{\theta} = \boldsymbol{\beta}$, (A.12) yields

$$(\widehat{\boldsymbol{\theta}} - \mathbf{t})^T\boldsymbol{\Phi}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \le \sqrt{\varepsilon}(\widehat{\boldsymbol{\theta}} - \mathbf{t})^T\boldsymbol{\Upsilon}\boldsymbol{\eta} + (\widehat{\boldsymbol{\theta}} - \mathbf{t})^T\mathbf{h} + \alpha\big(\|\boldsymbol{\Upsilon}\mathbf{t}\|_1 - \|\boldsymbol{\Upsilon}\widehat{\boldsymbol{\theta}}\|_1\big).$$

Since for any $\mathbf{u}, \mathbf{v} \in \mathbb{R}^p$ one has $\mathbf{v}^T\boldsymbol{\Phi}\mathbf{u} = \frac{1}{2}[\mathbf{v}^T\boldsymbol{\Phi}\mathbf{v} + \mathbf{u}^T\boldsymbol{\Phi}\mathbf{u} - (\mathbf{v} - \mathbf{u})^T\boldsymbol{\Phi}(\mathbf{v} - \mathbf{u})]$, choosing $\mathbf{v} = \widehat{\boldsymbol{\theta}} - \mathbf{t}$ and $\mathbf{u} = \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}$ and observing that for any $\mathbf{t}$ (and, in particular, for $\mathbf{t} = \widehat{\boldsymbol{\theta}}$), $\|f_{\mathbf{t}} - f\|_2^2 = (\mathbf{t} - \boldsymbol{\theta})^T\boldsymbol{\Phi}(\mathbf{t} - \boldsymbol{\theta}) + \|f_{\boldsymbol{\theta}} - f\|_2^2$, for any $\mathbf{t} \in \mathbb{R}^p$, one obtains

$$\begin{aligned}&\|f_{\widehat{\boldsymbol{\theta}}} - f\|_2^2 + (\widehat{\boldsymbol{\theta}} - \mathbf{t})^T\boldsymbol{\Phi}(\widehat{\boldsymbol{\theta}} - \mathbf{t})\\ &\quad\le \|f_{\mathbf{t}} - f\|_2^2 + \sqrt{\varepsilon}(\widehat{\boldsymbol{\theta}} - \mathbf{t})^T\boldsymbol{\Upsilon}\boldsymbol{\eta} + (\widehat{\boldsymbol{\theta}} - \mathbf{t})^T\mathbf{h} + 2\alpha\big(\|\boldsymbol{\Upsilon}\mathbf{t}\|_1 - \|\boldsymbol{\Upsilon}\widehat{\boldsymbol{\theta}}\|_1\big).\end{aligned} \tag{A.13}$$

By setting $t = K\sqrt{(\tau+1)\log p}$ in (A.8) and using (A.9), observe that, on the set

$$\Omega = \Big\{\omega : \max_{1\le j\le p}|\eta_j| \le K\sqrt{(\tau+1)\log p}\Big\} \qquad \text{with } \mathbb{P}(\Omega) \ge 1 - 2p^{-\tau} \tag{A.14}$$

one has $|\sqrt{\varepsilon}(\widehat{\boldsymbol{\theta}} - \mathbf{t})^T\boldsymbol{\Upsilon}\boldsymbol{\eta} + (\widehat{\boldsymbol{\theta}} - \mathbf{t})^T\mathbf{h}| \le \sqrt{\varepsilon\log p}(K\sqrt{\tau+1} + C_h)\|\boldsymbol{\Upsilon}(\widehat{\boldsymbol{\theta}} - \mathbf{t})\|_1 = \alpha_0\|\boldsymbol{\Upsilon}(\widehat{\boldsymbol{\theta}} - \mathbf{t})\|_1$. Combining the last inequality with (A.13), obtain that, for any $\alpha > 0$, on the set Ω,

$$\begin{aligned}&\|f_{\widehat{\boldsymbol{\theta}}} - f\|^2 + (\widehat{\boldsymbol{\theta}} - \mathbf{t})^T\boldsymbol{\Phi}(\widehat{\boldsymbol{\theta}} - \mathbf{t})\\ &\quad\le \|f_{\mathbf{t}} - f\|^2 + 2\alpha\big(\|\boldsymbol{\Upsilon}\mathbf{t}\|_1 - \|\boldsymbol{\Upsilon}\widehat{\boldsymbol{\theta}}\|_1\big) + 2\alpha_0\big\|\boldsymbol{\Upsilon}(\widehat{\boldsymbol{\theta}} - \mathbf{t})\big\|_1.\end{aligned} \tag{A.15}$$

Application of inequality $\|\boldsymbol{\Upsilon}(\widehat{\boldsymbol{\theta}} - \mathbf{t})\|_{t1} \le |\boldsymbol{\Upsilon}\mathbf{t}\|_1 + \|\boldsymbol{\Upsilon}\widehat{\boldsymbol{\theta}}\|_1$ combined with $\alpha \ge \alpha_0$ completes the proof of inequality (A.10).

In order to prove inequality (A.11), denote $\mathbf{d} = \widehat{\boldsymbol{\theta}} - \mathbf{t}$ and observe that, due to $|t_j| - |\widehat{\theta}_j| \le |\widehat{\theta}_j - t_j|$ and $|\widehat{\theta}_j| \ge |\widehat{\theta}_j - t_j| - |t_j|$, inequality (A.15) implies that, for any set $J \subseteq \mathcal{P}$, one obtains

$$\begin{aligned}&\|f_{\widehat{\boldsymbol{\theta}}} - f\|_2^2 + \mathbf{d}^T\boldsymbol{\Phi}\mathbf{d}\\ &\quad\le \|f_{\mathbf{t}} - f\|_2^2 + 4\alpha\big\|(\boldsymbol{\Upsilon}\mathbf{t})_{J^c}\big\|_1 + 2(\alpha + \alpha_0)\big\|(\boldsymbol{\Upsilon}\mathbf{d})_J\big\|_1\\ &\qquad - 2(\alpha - \alpha_0)\big\|(\boldsymbol{\Upsilon}\mathbf{d})_{J^c}\big\|_1.\end{aligned} \tag{A.16}$$

Now, we consider two possibilities. If $(\alpha + \alpha_0)\|(\boldsymbol{\Upsilon}\mathbf{d})_J\|_1 \le (\alpha - \alpha_0)\|(\boldsymbol{\Upsilon}\mathbf{d})_{J^c}\|_1$, then $\|f_{\widehat{\boldsymbol{\theta}}} - f\|_2^2 + \mathbf{d}^T\boldsymbol{\Phi}\mathbf{d} \le \|f_{\mathbf{t}} - f\|_2^2 + 4\alpha\|(\boldsymbol{\Upsilon}\mathbf{t})_{J^c}\|_1$ and (A.11) is valid. Otherwise, since $\alpha = \varpi\alpha_0$ with $\varpi \ge (\mu+1)/(\mu-1)$ implies that $\mu \ge (\alpha + \alpha_0)/(\alpha -$

$\alpha_0)$, one has $\mathbf{d} \in \mathcal{J}(\mu, J)$. Therefore, due to compatibility condition (4.14) and inequality $2ab \le a^2 + b^2$, one derives

$$2(\alpha+\alpha_0)\|(\boldsymbol{\Upsilon}\mathbf{d})_J\|_1 \le 2(\alpha+\alpha_0)\sqrt{\operatorname{Tr}(\boldsymbol{\Upsilon}_J^2)\mathbf{d}^T\boldsymbol{\Phi}\mathbf{d}}/\kappa(\mu, J)$$
$$\le \mathbf{d}^T\boldsymbol{\Phi}\mathbf{d} + (\alpha+\alpha_0)^2\operatorname{Tr}(\boldsymbol{\Upsilon}_J^2)/\kappa^2(\mu, J).$$

Plugging the latter into (A.16) and using $\alpha = \varpi\alpha_0$, obtain that (A.11) holds for any $\mathbf{t}$. □

PROOF OF THEOREM 1. Let $\boldsymbol{\beta}$ and $\widehat{\boldsymbol{\beta}}$ be the vectors with components $\boldsymbol{\beta}_j = \langle f, \varphi_j\rangle_{\mathcal{H}_1}$ and $\widehat{\boldsymbol{\beta}}_j = \langle y, \psi_j\rangle_{\mathcal{H}_2}$, $j = 1, \dots, p$. Then, due to (4.5), one has $\widehat{\boldsymbol{\beta}}_j = \boldsymbol{\beta}_j + \sqrt{\varepsilon}\nu_j\boldsymbol{\eta}_j$ where $\boldsymbol{\eta}_j$ are standard normal variables and $\varepsilon = \sigma/\sqrt{n}$. Moreover, if f is the true function and $f_{\boldsymbol{\theta}}$ is its projection onto the span of the dictionary $\mathcal{L}_{\mathcal{P}}$, then, $\langle f - f_{\boldsymbol{\theta}}; \varphi_j\rangle = 0$ for $j = 1, \dots, p$, and $\boldsymbol{\beta} = \boldsymbol{\Phi}\boldsymbol{\theta}$. Therefore, validity of Theorem 1 follows from Lemma 2 with $K = \sqrt{2}$, $\mathbf{h} = 0$ and $C_h = 0$ in (A.9). □

PROOF OF THEOREM 2. Validity of (4.15) follows from Lemma 2 with $\varepsilon = \sigma/\sqrt{n}$, $K = \sqrt{2}$, $\mathbf{h} = 0$ and $C_h = 0$ in (A.9), so that $C_\alpha = \sqrt{2(\tau+1)}$ and $K_0 = 2$ in (4.15). In order to prove (4.16), choose $f_{\mathbf{t}} = \operatorname{proj}_{\mathcal{L}_J} f$, then $t_j = 0$ for $j \in J^c$. □

PROOF OF THEOREM 3. Note that vector $\widehat{\boldsymbol{\beta}}$ has components $\widehat{\beta}_j = \beta_j + \delta_{j1} + \delta_{j2}$, $j = 1, \dots, p$, where

$$\delta_{j1} = \sum_{i=1}^n \xi_i\psi_j(x_i)\Delta x_i,$$
$$\delta_{j2} = \sum_{i=1}^n q(x_i)\psi_j(x_i)\Delta x_i - \int_{\mathcal{X}} q(x)\psi_j(x)\,dx,$$

are, respectively, the random error component and the bias of $\widehat{\beta}_j$. In order to bound above the random term, apply Proposition 5.10 of Vershynin (2012) which implies that, for any vector $\mathbf{a}$ and any $z > 0$, one has

$$\mathbb{P}\left(\left|\sum_{i=1}^n a_i\xi_i\right| > z\right) \le e\exp\left(-\frac{z^2}{2\sigma^2\|\mathbf{a}\|_2^2}\right).$$

Choosing $a_j = \psi_j(x_i)\Delta x_i$ and $z = \sigma\nu_j t/\sqrt{n}$ and noting that, by assumption (5.6), one has $\|\mathbf{a}\|_2^2 \le n^{-1}\nu_j^2\vartheta^2$, obtain

$$\mathbb{P}\left(|\delta_{j1}| > \frac{\sigma\nu_j t}{\sqrt{n}}\right) \le e\exp\left\{-\frac{t^2}{2\vartheta^2}\right\}.$$

Also, it is known that the error of the rectangular approximation of an integral obeys $|\delta_{j2}| \le \nu_j \aleph T^2/(2n)$. Apply Lemma 2 with $\varepsilon = \sigma^2/n$, $K = \vartheta\sqrt{2}$, $h_j = \delta_{j2}$ and $C_h = T^2\vartheta\aleph/(2\sigma\sqrt{n\log p})$ and observe that for $n \ge \mathcal{N}$, one has $K\sqrt{\tau+1} \ge C_h$. Then, for $C_\alpha = 2\vartheta\sqrt{2(\tau+1)}$ and $K_0 = 8\vartheta^2$, obtain that inequalities (4.15) and (4.16) hold with probability at least $1 - ep^{-\tau}$. □

PROOF OF THEOREM 4. To prove the theorem, apply Lemma 2 with $\mathbf{h} = 0$ and $\eta_j = \sqrt{n}(\widehat{\beta}_j - \beta_j)/\nu_j$ in (A.7). The main difference between the proof of this theorem and Theorem 2 is that we need to establish existence of set Ω in (A.14) directly instead of relying on assumption (A.8). For this purpose, we observe that

$$\eta_j = n^{-1}\sum_{i=1}^{n} z_{ij} \qquad \text{with } z_{ij} = \frac{\sqrt{n}}{\nu_j}\psi_j(Y_i) - \frac{\sqrt{n}}{\nu_j}\mathbb{E}\psi_j(Y_i)$$

with $\mathbb{E}z_{ij} = 0$, $\mathbb{E}z_{ij}^2 = \sigma_z^2 = n\nu_j^{-2}\operatorname{Var}[\psi_j(Y_1)] \le 1$ and $\|z_{ij}\|_\infty = \max|z_{ij}| < 2\sqrt{n}\|\psi_j\|_\infty/\nu_j$. Applying Bernstein inequality, we obtain

$$\mathbb{P}(|\eta_j| > z) \le 2\exp\left\{-\frac{z^2}{2}\left(1 + \frac{2z\|\psi_j\|_\infty}{3\sqrt{n}\nu_j}\right)^{-1}\right\}. \tag{A.17}$$

Choosing $z = 2\sqrt{(\tau+1)\log p}$ in (A.17) and noting that $2z\|\psi_j\|_\infty/(3\sqrt{n}\nu_j) \le 1$ for $n \ge \mathcal{N}_0$, we obtain (A.14) with $K = 2$. Application of Lemma 2 completes the proof. □

PROOF OF COROLLARY 1. In order to prove validity of the corollary, we just need to verify the expression for ν_j in (8.9). For simplicity, we drop the index j. Observe that since $\mathbf{A}$ is symmetric and positive definite, there exists a symmetric square root $\sqrt{\mathbf{A}} = \mathbf{A}^{1/2}$ and that expression (8.8) can be re-written as

$$\begin{aligned}\psi(\mathbf{Y} \mid \mathbf{A}, \gamma) = {} & \frac{\Gamma_r(m/2)2^{\gamma r/2}|\mathbf{A}|^{-(r+1)/4}}{\Gamma_r((m-\gamma)/2)\sqrt{\Gamma_r(2\gamma - r - 1/2)}} \\ & \times \frac{|\mathbf{A}^{-1/2}\mathbf{Y}\mathbf{A}^{-1/2} - \mathbf{I}|^{(m-\gamma-r-1)/2}}{|\mathbf{A}^{-1/2}\mathbf{Y}\mathbf{A}^{-1/2}|^{(m-r-1)/2}}\mathbb{I}(\aleph),\end{aligned}$$

where $\aleph = \{\omega : \mathbf{A}^{-1/2}\mathbf{Y}\mathbf{A}^{-1/2} - \mathbf{I} > 0\}$. Furthermore, note that matrix $\mathbf{A}^{-1/2}\mathbf{Y}\mathbf{A}^{-1/2}$ is symmetric, so that there exists a diagonal matrix $\mathbf{D} \in \mathbb{R}^{r\times r}$ with components $D_k > 1$, $k = 1, \dots, r$, due to $\mathbf{A}^{-1/2}\mathbf{Y}\mathbf{A}^{-1/2} > \mathbf{I}$, and an orthogonal matrix $\mathbf{U}$ such that $\mathbf{A}^{-1/2}\mathbf{Y}\mathbf{A}^{-1/2} = \mathbf{U}\mathbf{D}\mathbf{U}^T$. Using the fact that $|\mathbf{U}| = 1$, obtain that

$$\frac{|\mathbf{A}^{-1/2}\mathbf{Y}\mathbf{A}^{-1/2} - \mathbf{I}|^{(m-\gamma-r-1)/2}}{|\mathbf{A}^{-1/2}\mathbf{Y}\mathbf{A}^{-1/2}|^{(m-r-1)/2}} = \prod_{k=1}^{r}\left[(D_k - 1)^{(m-\gamma-r-1)/2}(D_k)^{-(m-r-1)/2}\right].$$

Maximizing the last expression with respect to $D_k > 1$ and noting that

$$\max_{x>1}\bigl[(x-1)^{(m-\gamma-r-1)/2}x^{-(m-r-1)/2}\bigr] = (m-\gamma-r-1)^{m-\gamma-r-1}\gamma^{\gamma}(m-r-1)^{m-r-1},$$

obtain (8.9). □

A.3. Calculations for Section 8. First, let us show that $C_{\mathbf{A},\gamma}$ is given by formula (8.4). For this purpose, note that

$$\int_{\mathcal{Y}} u(\mathbf{X}\mid\mathbf{A}_i,\gamma_i)u(\mathbf{X}\mid\mathbf{A}_j,\gamma_j)\,d\mathbf{X} = \frac{2^{(d-1)d/2}\Gamma_r(\gamma_i+\gamma_j-r-1/2)}{\Gamma_r((\gamma_i-r-1)/2)\Gamma_r(\gamma_j-r-1/2)}F(\mathbf{A}_i,\mathbf{A}_j),$$

where $F(\mathbf{A}_i,\mathbf{A}_j) = |\mathbf{A}_i|^{(\gamma_i-r-1)/2}\mathbf{A}_j|^{(\gamma_j-r-1)/2}|\mathbf{A}_i+\mathbf{A}_j|^{-(\gamma_i+\gamma_j-r-1)/2}$. By setting $\mathbf{A}_i = \mathbf{A}_j = \mathbf{A}$, $\gamma_1 = \gamma_2 = \gamma$ and noting that $\|\varphi(\mathbf{X}\mid\mathbf{A},\gamma)\|_2 = 1$, we derive that $C_{\mathbf{A},\gamma}$ is indeed of the form (8.4). Now, in order to verify expression (8.5), it is sufficient to note that

$$\Phi_{i,j} = C_{\mathbf{A}_i,\gamma_i}C_{\mathbf{A}_j,\gamma_j}\int_{\mathcal{Y}} u(\mathbf{X}\mid\mathbf{A}_i,\gamma_i)u(\mathbf{X}\mid\mathbf{A}_j,\gamma_j)\,d\mathbf{X}.$$

Next, let us verify expression (8.8). For this purpose, we re-arrange equation (8.7), arriving at

$$\int_{\mathcal{Y}} \frac{|\mathbf{Y}|^{(m-r-1)/2}2^{-mr/2}}{|\mathbf{X}|^{(m-\gamma)/2}\Gamma_r(m/2)}v(\mathbf{Y}\mid\mathbf{A},\gamma)\exp\bigl\{-\operatorname{Tr}(\mathbf{X}^{-1}(\mathbf{Y}-\mathbf{A})/2)\bigr\}\,d\mathbf{Y} = |\mathbf{A}|^{(\gamma-r-1)/2}2^{-(\gamma-r-1)r/2}\biggl[\Gamma_r\biggl(\frac{\gamma-r-1}{2}\biggr)\biggr]^{-1}.$$

By comparing the integrand with the p.d.f. of the Wishart distribution, we gather that

$$v(\mathbf{Y}\mid\mathbf{A},\gamma) = \frac{\Gamma_r(m/2)2^{(r+1)r/2}|\mathbf{A}|^{(\gamma-r-1)/2}}{\Gamma_r((m-\gamma)/2)\Gamma_r(\gamma-r-1/2)} \times \frac{|\mathbf{Y}-\mathbf{A}|^{(m-\gamma-r-1)/2}}{|\mathbf{Y}|^{(m-r-1)/2}}\mathbb{I}(\mathbf{Y}-\mathbf{A}>0). \tag{A.18}$$

Combination of (8.4) and (A.18) yields (8.8).

Acknowledgements. The author would like to thank SAMSI for providing support which allowed her participation in the 2013–14 LDHD program that was instrumental for writing this paper. The author also wants to thank the Associate Editor and the anonymous reviewers whose valuable remarks and suggestions helped to improve this paper.

REFERENCES

ABRAMOVICH, F. and SILVERMAN, B. W. (1998). Wavelet decomposition approaches to statistical inverse problems. *Biometrika* **85** 115–129. MR1627226

BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469

BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data*: *Methods, Theory and Applications*. Springer, Heidelberg. MR2807761

BUNEA, F., TSYBAKOV, A. and WEGKAMP, M. (2007). Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.* **1** 169–194. MR2312149

BUNEA, F., TSYBAKOV, A. B., WEGKAMP, M. H. and BARBU, A. (2010). Spades and mixture models. *Ann. Statist.* **38** 2525–2558. MR2676897

CANDÈS, E. J. (2003). Ridgelets: Estimating with ridge functions. *Ann. Statist.* **31** 1561–1599. MR2012826

CAVALIER, L. and GOLUBEV, YU. (2006). Risk hull method and regularization by projections of ill-posed inverse problems. *Ann. Statist.* **34** 1653–1677. MR2283712

CAVALIER, L. and REISS, M. (2014). Sparse model selection under heterogeneous noise: Exact penalisation and data-driven thresholding. *Electron. J. Stat.* **8** 432–455. MR3195122

CAVALIER, L., GOLUBEV, G. K., PICARD, D. and TSYBAKOV, A. B. (2002). Oracle inequalities for inverse problems. *Ann. Statist.* **30** 843–874. Dedicated to the memory of Lucien Le Cam. MR1922543

COHEN, A., HOFFMANN, M. and REISS, M. (2004). Adaptive wavelet Galerkin methods for linear inverse problems. *SIAM J. Numer. Anal.* **42** 1479–1501 (electronic). MR2114287

COMTE, F. and GENON-CATALOT, V. (2015). Adaptive Laguerre density estimation for mixed Poisson models. *Electron. J. Stat.* **9** 1113–1149. MR3352069

DALALYAN, A. S., HEBIRI, M. and LEDERER, J. (2014). On the prediction performance of the Lasso. Preprint. Available at arXiv:1402.1700.

DALALYAN, A. S. and SALMON, J. (2012). Sharp oracle inequalities for aggregation of affine estimators. *Ann. Statist.* **40** 2327–2355. MR3059085

DONOHO, D. L. (1995). Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition. *Appl. Comput. Harmon. Anal.* **2** 101–126. MR1325535

EFROMOVICH, S. and KOLTCHINSKII, V. (2001). On inverse problems with unknown operators. *IEEE Trans. Inform. Theory* **47** 2876–2894. MR1872847

GOLUBEV, Y. (2010). On universal oracle inequalities related to high-dimensional linear models. *Ann. Statist.* **38** 2751–2780. MR2722455

GOUTIS, C. (1997). Nonparametric estimation of a mixing density via the kernel method. *J. Amer. Statist. Assoc.* **92** 1445–1450. MR1615254

GUPTA, A. K. and NAGAR, D. K. (2000). *Matrix Variate Distributions*. Chapman & Hall/CRC, Boca Raton, FL. MR1738933

HENGARTNER, N. W. (1997). Adaptive demixing in Poisson mixture models. *Ann. Statist.* **25** 917–928. MR1447733

HOFFMANN, M. and REISS, M. (2008). Nonlinear estimation for linear inverse problems with error in the operator. *Ann. Statist.* **36** 310–336. MR2387973

KALIFA, J. and MALLAT, S. (2003). Thresholding estimators for linear inverse problems and deconvolutions. *Ann. Statist.* **31** 58–109. MR1962500

LE PENNEC, E. and MALLAT, S. (2005). Sparse geometric image representations with bandelets. *IEEE Trans. Image Process.* **14** 423–438. MR2128287

LIU, L., LEVINE, M. and ZHU, Y. (2009). A functional EM algorithm for mixing density estimation via nonparametric penalized likelihood maximization. *J. Comput. Graph. Statist.* **18** 481–504. MR2548074

LOUNICI, K., PONTIL, M., VAN DE GEER, S. and TSYBAKOV, A. B. (2011). Oracle inequalities and optimal inference under group sparsity. *Ann. Statist.* **39** 2164–2204. MR2893865

MEISTER, A. (2009). *Deconvolution Problems in Nonparametric Statistics*. *Lecture Notes in Statistics* **193**. Springer, Berlin. MR2768576

VERSHYNIN, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing* 210–268. Cambridge Univ. Press, Cambridge. MR2963170

WALTER, G. G. (1981). Orthogonal series estimators of the prior distribution. *Sankhyā Ser. A* **43** 228–245. MR0666383

YUAN, M. and LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68** 49–67. MR2212574

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF CENTRAL FLORIDA
ORLANDO, FLORIDA 32816-1354
USA
E-MAIL: Marianna.Pensky@ucf.edu