Technical Report

TR-2008-015

Sensitivity computation of the l1 minimization problem and its application to dictionary design

by

Lior Horesh, Eldad Haber

Mathematics and Computer Science

EMORY UNIVERSITY

Sensitivity computation of the ℓ_1 minimization problem and its application to dictionary design

L Horesh* and E Haber†

July 29, 2008

Abstract

Recently, there has been a growing interest in application of sparse representation for inverse problems. Most studies have concentrated in devising ways for sparse representation of a solution using a given prototype dictionary. Very few studies have addressed the more challenging problem of constructing an optimal dictionary, and even these were primarily devoted to the simplistic sparse coding application.

In this paper we present a new approach for dictionary design. First, we analyze the sensitivity of the inverse solution with respect to the dictionary. Second, we utilize the derived sensitivity relations for the design of an optimal dictionary.

Our optimality criterion is based on minimizing the empirical risk, given a set of training models. We present a mathematical formulation and an algorithmic framework to achieve this goal. The proposed framework offers incorporation of non-injective operators, where the data and the recovered parameters may reside in different spaces. We test our algorithm and show that it yields improved dictionaries for a diverse set of inverse problems.

keywords sparse representation, sensitivity analysis, empirical risk, constrained optimization, optimal design, bi-level optimization

1 Introduction

We consider a discrete ill-posed inverse problem of the form

$$Am + \epsilon = d,$$

where $m \in \mathbb{R}^n$ is the model, $A : \mathbb{R}^n \to \mathbb{R}^k$ is a forward operator, the acquired data is $d \in \mathbb{R}^k$, and $\epsilon \in \mathbb{R}^k$ is the noise which is assumed to be Gaussian and iid with known variance. We regard the matrix A as ill-posed and under-determined in the usual sense [43].

*Mathematics and Computer Science, Emory University, Atlanta, 30322, GA, USA
†Mathematics and Computer Science, Emory University, Atlanta, 30322, GA, USA

The general aim is recovery of the model m from the noisy data d. However, since the problem is ill-posed, regularization is needed. One possible way to regularize the problem is by using a Tikhonov-like regularization scheme. This can be performed by solving the optimization problem

$$\widehat{m} = \underset{m}{\operatorname{argmin}} \ \frac{1}{2}\|Am - d\|_2^2 + \alpha R(m),$$

where R is a regularization functional, aimed for imposition of a-priori information over the solution and $\alpha > 0$ is a regularization parameter.

A different approach for regularization is based on imposing sparsity, see for example [65, 70, 44, 81, 74, 34, 34, 30, 2, 12, 3, 46, 25, 91, 55] , and references therein. This approach has gained interest and popularity in the past few years. The underlying assumption is that the true solution can be described by a small number of parameters (principle of parsimony); therefore, the model can be represented accurately as a composite of a small set of prototype atoms of an over-complete dictionary D. A common and convenient relation between the dictionary and the sparse code is the linear generative model

$$m = Du,$$

where u is a sparse coefficient vector, i.e., it constitutes mostly zeros besides a small populated subset. Using the linear model it is possible to solve for m through u by solving the optimization problem

$$\min_u \ \frac{1}{2}\|ADu - d\|_2^2 + \alpha\|u\|_p, \tag{1}$$

where α is a regularization parameter and $0 \leq p \leq 1$.

The requirement for $p = 0$ is desirable in many cases. However, computationally, it yields an intractable problem of a non-polynomial runtime complexity. The choice $p = 1$ grants a convex optimization problem that can be solved using numerous algorithms, for instance [57, 14, 32, 24, 50, 83, 31, 48, 42], and references therein. It has been proven recently that for a broad range of problems, the case $p = 1$ provides identical results as of the case where $p = 0$ [23, 22, 82]. This important finding, has motivated many researchers to further consider and direct efforts towards solution of the $p = 1$ problem. For that reason, we focus our attention to this case in the current study.

Sparse representation offers several genuine advantages over the more traditional Tikhonov-like regularization scheme. It offers expressiveness, independence and flexibility by allowing the application and construction of regularization functionals that are not restricted to any specific space. The idea of promoting sparse solutions dates back to the 70's [17]. This notion had gained considerable attention since the work of Olshausen and Field [66, 67] who suggested that the visual cortex in mammalian brain employs a sparse coding strategy.

One can distinguish between two different challenges in the field:

- *Sparse representation* - constructing a sparse vector u given the data d and a *given* overcomplete dictionary D

- *Sensitivity analysis and dictionary design/improvement* - analyzing the sensitivity of the solution with respect to a given dictionary and the construction of an optimal/superior dictionary, D, which in conjunction with the first goal promotes parsimonious representation

Recently, a large volume of studies have primarily addressed the first problem, e.g. [90, 76, 44, 74, 26, 87, 38, 56, 34, 30, 13, 51, 24, 12, 58, 92, 22, 3, 19, 35, 60, 61, 75] (and references therein), while the more intricate problem of dictionary analysis and design was seldom tackled [33, 65, 67, 52, 53, 36, 49, 2, 62, 29].

The goal of this work is to explore a new approach for dictionary design. As a first step we show that it is possible to compute the sensitivity of the solution with respect to the dictionary. We then use the sensitivity relations in order to design an improved dictionary. Similar to [33, 65, 67, 52, 53, 36, 49, 2, 62, 29], we assume the availability of a set of authentic model examples $\{m_1, \ldots, m_s\}$. Our aim is to construct an optimal dictionary for a given set of examples.

There is no work known to the authors in which computation and analysis of the sensitivity of the solution with respect to the dictionary is presented. The computation of such sensitivities offers a reasoned construction of an improved/optimal dictionary in a unique way. As mentioned above, the idea of learning a dictionary from a set of models is not new. Few algorithms for dictionary learning were developed so far. Each one of those was based on optimization of different entity or other heuristics. All work known to us addressed a simplistic version of the inverse problem above, where A was set to be an identity operator. Olshausen and Field [67] developed an Approximate Maximum Likelihood (AML) approach for dictionary update, Lewicki and Sejnowski [52, 53] developed an extension of Independent Component Analysis to overcomplete dictionaries, Girolami developed a variational approach based on Sparse Bayesian Learning [36], a Maximum A - Posteriori (MAP) framework combined with the notion of relative convexity was suggested by Kreutz-Delgado et al. by using the Focal Underdetermined System Solver. This algorithm was further improved by the employment of Column Normalized Dictionary prior (FOCUSS - CNDL) [49] and by incorporation of positivity constraints [62]. The K-SVD algorithm facilitated singular value decomposition for reduction of the residual error [2]. Most recently, an adaptation of the Method of Optimal Directions for dictionary learning was presented by Engan et al [29].

None of the approaches presented above can be easily modified to handle the situation where A is not the identity operator. In fact, the case of underdetermined A is mathematically different than the case of well-posed A. The fundamental difference emerges due to the fact that for a well-posed operator A, recovery of the exact model is possible when the noise ϵ goes to zero, while for ill-posed problems this is obviously not the case. For ill-posed problems, the incorporation of regularization inevitably introduces bias into the solution [7, 45, 37]. Adding the "correct" bias implies retrieving more accurate results, hence, other than merely overcoming noise, the goal of dictionary design in this context is also completion of missing information in the data.

In order to achieve the above goal we base our approach on minimizing the empirical risk. Since the resulting optimization problem is non-smooth, we conduct the learning process by

using a variant of L-BFGS method [11] developed specifically for non-smooth problems.

The paper is organized as follows: In Section 2, we describe a computational framework for the solution of the sparse representation problem (the inverse problem) (1), i.e., finding u for a given D. In Section 3 we discuss the computation of the sensitivity of the solution with respect to the dictionary. In Section 4, the mathematical and statistical frameworks for a novel dictionary design approach are introduced. This formulation is based on the sensitivity computation elucidated in the previous section. In Section 5 several computational and numerical aspects of the proposed optimization framework are discussed. In Section 6 we bring some numerical results for problems of different scales, for a non-injective limited angle tomography transformation A as well as for an injective Gaussian kernel. Finally, in Section 7 the study is summarized.

2 Solving the ℓ_1 minimization problem

In this section we briefly review and discuss the solution of the inverse problem (1). This problem can be solved by numerous possible ways (see, for example, [90, 76, 31, 83] and reference within). Here we focus on a simple strategy that was recently investigated in [31]. This approach was found to be particularly useful for sensitivity computation as presented in the next section.

The non-smooth ℓ_1-norm is replaced by a smooth optimization problem with inequality constraints. By setting $u = p - q$ with both $p, q \geq 0$, it is easy to show that the inverse problem in (1) is equivalent to the following optimization problem

$$\begin{aligned} \min_{p,q} \quad & \frac{1}{2}\|AD(p-q) - b\|^2 + \alpha e^\top (p+q) \\ \text{s.t} \quad & p, q \geq 0, \end{aligned} \tag{2}$$

where $e = [1, \ldots, 1]^\top$. In [31] a variant of the projected-gradient method was proposed for the solution of this optimization problem. We have experimented with this approach[1] over various problems and settings. Our findings indicated that for some problems, and in particular, those which are characterized by high level of sparsity, convergence is typically very prompt.

3 Sensitivity with respect to the dictionary

In order to obtain the sensitivities of the inverse solution with respect to the dictionary we use the decomposition of u into $p - q$. One can readily verify that the optimality conditions

[1] The code of the method described in [31] can be downloaded from `www.lx.it.pt/~mtf/GPSR/`

for a minimum are

$$D^\top A^\top (AD(p-q)-b) + \alpha e - \lambda_p = 0 \tag{3a}$$
$$-D^\top A^\top (AD(p-q)-b) + \alpha e - \lambda_q = 0 \tag{3b}$$
$$\lambda_p \odot p = 0 \tag{3c}$$
$$\lambda_q \odot q = 0 \tag{3d}$$
$$p, q, \lambda_p, \lambda_q \geq 0. \tag{3e}$$

The following lemma can now be proved

Lemma 1 *Let $p^*, q^*, \lambda_p^*, \lambda_q^*$ be a solution of the system (3). If $\alpha > 0$ then $p \odot q = 0$.*

Proof: Summation of the pointwise multiplication of (3a) and (3b) with $p \odot q$ yields

$$2\alpha p \odot q - \lambda_p \odot p \odot q - \lambda_q \odot p \odot q = 0.$$

By using relations (3c) and (3d) one obtains

$$2\alpha p \odot q = 0,$$

since $\alpha > 0$, we get $p \odot q = 0$ ■

Using Lemma 1 we can now rewrite the conditions for a minimum, while eliminating λ_p and λ_q from the equations. Let $\mathcal{I}$ and $\mathcal{J}$ be the inactive sets obtained at the minimum for the vectors p and q. Also, let the matrices $P_\mathcal{I}$ and $P_\mathcal{J}$ be the selection matrices which mark the inactive indices, that is $P_\mathcal{I} p = p_\mathcal{I}$ and $P_\mathcal{J} q = q_\mathcal{J}$. Then, the system (3) can be written as $F(p, q; D) = 0$, where

$$F(p,q;D) = \begin{cases} P_\mathcal{I} D^\top A^\top A D P_\mathcal{I}^\top p_\mathcal{I} - P_\mathcal{I}(D^\top A^\top b - \alpha e) \\ P_\mathcal{J} D^\top A^\top A D P_\mathcal{J}^\top q_\mathcal{J} - P_\mathcal{J}(D^\top A^\top b - \alpha e), \end{cases} \tag{4}$$

and

$$p_i = 0 \quad i \notin \mathcal{I} \tag{5}$$
$$q_j = 0 \quad j \notin \mathcal{J}. \tag{6}$$

The sensitivities of p and q with respect to the jth column of D, D_j, can be computed using implicit differentiation (see for example [41, 73]). Differentiation of F with respect to p, q and D_j provides

$$F_{p_\mathcal{I}} \delta p_\mathcal{I} + F_{D_j} \delta D_j = 0 \quad \text{and} \quad F_{q_\mathcal{J}} \delta q_\mathcal{J} + F_{D_j} \delta D_j = 0,$$

which implies that the sensitivity of p and q with respect to D_j is

$$\frac{\partial p_\mathcal{I}}{\partial D_j} = -F_{p_\mathcal{I}}^{-1} F_{D_j}, \qquad \frac{\partial q_\mathcal{J}}{\partial D_j} = -F_{q_\mathcal{J}}^{-1} F_{D_j},$$
$$\frac{\partial p_k}{\partial D_j} = 0 \quad k \notin \mathcal{I}, \qquad \frac{\partial q_k}{\partial D_j} = 0 \quad k \notin \mathcal{J}.$$

Computation of the derivatives with respect to $p_{\mathcal{I}}$ and $q_{\mathcal{J}}$ provides

$$F_{p_{\mathcal{I}}} = P_{\mathcal{I}} D^\top A^\top A D P_{\mathcal{I}}^\top \quad \text{and} \quad F_{q_{\mathcal{J}}} = P_{\mathcal{J}} D^\top A^\top A D P_{\mathcal{J}}^\top.$$

We would now describe how the derivatives of F with respect to the j^{th} column in D, D_j, can be obtained. First note that for a given vector w, differentiation of the product Dw with respect to the j^{th} column reads

$$\frac{\partial [Dw]}{\partial D_j} = w_j I.$$

where w_j is the j^{th} entry in w. Similarly, for a vector z we have

$$\frac{\partial [D^\top z]}{\partial D_j} = Z_j,$$

where

$$Z_j = \begin{pmatrix} 0 \\ \vdots \\ z^\top \\ 0 \\ \vdots \end{pmatrix}.$$

By using the above derivatives and the product rule, it is possible to verify that

$$\frac{\partial [P D^\top A^\top A D P^\top w]}{\partial D_j} = [P^\top w]_j P D^\top A^\top A + P G_{w_j},$$

$$\frac{\partial [P D^\top A^\top b]}{\partial D_j} = P B_j,$$

where G_{w_j} and B_j comprise the vectors $A^\top A D P^\top w$ and $A^\top b$ in their j^{th} row and zero elsewhere, respectively.

We summarize the observations above in the following theorem

Theorem 1 *Let u^* be the solution of the optimization problem (1) and let $p_{\mathcal{I}} = u^*_+$ be the positive entries in u^* and $q_{\mathcal{J}} = -u^*_-$ be the negative entries in u^*. Then assuming that $ADP_{\mathcal{I},\mathcal{J}}$ is full rank and that the length of $p_{\mathcal{I}}$ or $q_{\mathcal{J}}$ is smaller than the rank of $ADP_{\mathcal{I}}$ and $ADP_{\mathcal{J}}$. The sensitivities of $p_{\mathcal{I}}$ and $q_{\mathcal{J}}$ with respect to the j^{th} column in D are given by*

$$\frac{\partial p_{\mathcal{I}}}{\partial D_j} = -(P_{\mathcal{I}} D^\top A^\top A D P_{\mathcal{I}}^\top)^{-1} \left([P_{\mathcal{I}}^\top p_{\mathcal{I}}]_j P_{\mathcal{I}} D^\top A^\top A + P_{\mathcal{I}} G_{p_{\mathcal{I}},j} - P_{\mathcal{I}} B_j \right) \tag{8a}$$

$$\frac{\partial q_{\mathcal{J}}}{\partial D_j} = -(P_{\mathcal{J}} D^\top A^\top A D P_{\mathcal{J}}^\top)^{-1} \left([P_{\mathcal{J}}^\top q_{\mathcal{J}}]_j P_{\mathcal{J}} D^\top A^\top A + P_{\mathcal{J}} G_{q_{\mathcal{J}},j} - P_{\mathcal{J}} B_j \right) \tag{8b}$$

At this stage we would like to highlight several interesting observations.

- One may notice that the sensitivity of the solution with respect to a column in D is non-intuitive. The sensitivity expression involves the observation (forward) operator A, as well as the dictionary D. This dependency clearly reapproves that dictionary design of the inverse problem outlined in (1) with $A \neq I$ differs intrinsically from that of all previous work, for which the particular case where A is an identity matrix was considered. For example, it may well be that a particular atom in D is desirable for the case $A = I$ but for the case where $A \neq I$ and underdetermined, this atom may reside in the null space of A and thus would not be considered as useful in the solution process.

- The computation of the inverse of the Hessian of the inactive set times another dense matrix is required for derivation of the sensitivities. However, these calculations need not require an explicit construction of the Hessian and can be conducted by using matrix-vector products merely. Furthermore, all the ingredients of the sensitivity relations are generated as byproducts of solving the design optimization problem, and therefore, can be obtained with no additional cost.

- Another important observation that requires special attention is the fact that the sensitivity relations are not continuously differentiable. Indeed, if a small change in D is considered, one may expect a consequent change in the active sets for p and q, and thus, their associated derivatives are accordingly non-continuous. This attribute is well known and broadly studied [9] in the context of sensitivity analysis of optimization problems. Therefore, appropriate precautions are required while dealing with the sensitivities in the process of solving an optimization problem.

4 Mathematical Framework for Dictionary Design and Improvement

A direct application of sensitivity analysis is dictionary design and its improvement. While we focus our attention on design, it is important to note that in many cases a reasonable dictionary already exists. Improvement upon a given dictionary may suffice in practice.

The general idea here is to compute an optimal, or an improved dictionary given some a-priori information about the type of models we seek. Frequently, dictionaries are based on wavelet-like bases [44, 27, 69, 32, 31]. This choice offers rapid computation of the products Du and $D^\top m$. Nevertheless, predefining a particular basis or dictionary may not serve well for all problems. For an instance, it is rather unlikely that the same basis (or dictionary) would be optimal for deblurring MRI images as well as for deblurring astronomical star cluster images. The two model problems are characterized by different features, some of these features may be particularly popular in one model problem, but, potentially absent from another and vice versa.

In this study, we focused our attention on design of over-complete dictionaries that are learned from examples. The main technical issue in dealing with such dictionaries is efficiency

in computation of D and $D^\top$ times a vector. It is possible to set the dimensions of each atom in D to match the dimensions of the entire model. However, such choice may inherently leads to construction of a dense matrix D and therefore can be only used effectively whenever the model dimensions are modest. In order to circumvent this limitation, we adopted an approach of defining the atom dimensions to correspond to sub-domains of uniform size. This approach, as suggested by [78, 28, 47, 86] yields sparse dictionaries D. For large-scale models, it can be useful to take advantage of repetition of local features. This allows for use of simple dictionaries which depend on a smaller number of parameters. Recent work using this approach yielded excellent results for the denoising problem [28, 10].

In order to construct a sparse D the model m is divided into overlapping domains of uniform dimensions. Let Q_j extract the j^{th} domain of the model which is a vector of size r_1. Then, as in [78, 47, 86] we assume that the j^{th} domain in m can be written as

$$Q^j m = \mathbf{\Phi} u^j,$$

where $\mathbf{\Phi} \in \mathbb{R}^{r_1 \times r_2}$ is a local dictionary which is assumed to be invariant for the entire model. The size of the local dictionary is governed by the dimensions of the local domains, r_1 and the number of atoms in $\mathbf{\Phi}$, r_2. These are user dependent parameters (see [71, 8]).

Assume that there are n_p domains. By grouping the domains together the model can be expressed by

$$Qm = (I \otimes \mathbf{\Phi})u,$$

where $Q = \text{diag}(Q^j)$, I is an identity matrix of size n_p and $u = [u^1, \ldots, u^{n_p}]$. In this sequel, by assuming consistency we can rewrite m as

$$m = (Q^\top Q)^{-1} Q^\top (I \otimes \mathbf{\Phi})u = D(\mathbf{\Phi})u,$$

where we have defined

$$D(\mathbf{\Phi}) := (Q^\top Q)^{-1} Q^\top (I \otimes \mathbf{\Phi}). \tag{9}$$

Given this particular form (parametrization) of dictionary, the central question we ask is what would be the optimal dictionary for a particular inverse problem we have in mind? In order to answer this question, we first need to better define the model space. Let us consider a family of models $\mathcal{M}$. The family $\mathcal{M}$ can be defined, for example, by some probability density function (PDF) or by a convex set. Either ways, sampling is mandatory. Under more realistic settings $\mathcal{M}$ would normally not be explicitly specified. Yet, instead, a set of examples $M_s = \{m_1, \ldots, m_s\}$ belongs to the space $\mathcal{M}$ and assumed to be iid, can be provided by a specialist. Such a set is often referred to as a training set. This situation is common in many geophysical and medical imaging applications; for example, one may have access to numerous MRI images, albeit obtaining their associated probability density function may be nontrivial. Yet, this set of model examples may represent reliably the statistics of a given anomaly. Regardless of the origin of M_s, either obtained by sampling a probability density function or given by a training set introduced by an expert, the salient goal is to obtain an optimal dictionary for the set M_s. The statistical interpretation of the optimum

may differ depending on the origin of M_s, nonetheless, the computational gear for obtaining this optimal dictionary is identical. Typical aspects of unsupervised learning have to be addressed. These aspects include cross-validation and performance assessment. We discuss these aspects in Section 6.

The dictionary $D(\mathbf{\Phi})$ described by equation (9) depends only on our choice of local dictionary $\mathbf{\Phi}$. The number of parameters comprising the local dictionary is typically substantially smaller than the number of elements in a single model. Therefore, it is unlikely that any $\mathbf{\Phi}$ would recover any single model exactly. The goal here is to develop a mathematical framework for the estimation of $\mathbf{\Phi}$ given the training set M_s. The construction of an optimal $\mathbf{\Phi}$ requires an optimality criterion. One such obvious criterion is the effectiveness of a dictionary in solving the desired inverse problem given a training set M_s.

To do so, we define the loss function

$$L(m, \mathbf{\Phi}) = \frac{1}{2}\|\widehat{m}(D(\mathbf{\Phi}), m) - m\|_2^2, \tag{10}$$

where $\widehat{m}$ is obtained by the solution u of the optimization problem

$$\min_u \ \frac{1}{2}\|AD(\mathbf{\Phi})u - d(m)\|_2^2 + \alpha\|u\|_1, \tag{11}$$

and $d(m) = Am + \boldsymbol{\epsilon}$ is the observation model.

Various alternative loss functions, other than the one prescribed above can be considered, e.g., a semi-norm that focuses on a specific region of interest, or a distance measure for edges. Obviously, such choice needs to be customized individually according to the specific requirements of the application.

Given a model m and a noise realization ϵ, an optimal dictionary should provide superior model reconstruction over any other dictionary that fails to comply with the optimality criterion. There are two problems in using the loss function as a criterion for optimality. First, the function depends on the random variable ϵ and second, the problem depends on the particular (unknown) model. It is fairly plausible that a dictionary would be particularly effective in recovering a certain model while performing poorly for others. The dependency of the objective function with respect to the noise can be eliminated by considering the expected value of the loss, as defined by the risk

$$\text{risk}(m, D(\mathbf{\Phi})) = \frac{1}{2}\ \mathbf{E}_{\boldsymbol{\epsilon}}\, \|\widehat{m}(D(\mathbf{\Phi}), m) - m\|_2^2, \tag{12}$$

where $\mathbf{E}_{\boldsymbol{\epsilon}}$ is the expectation with respect to the noise. Since analytical calculation of this expectation is difficult, we use the common approximation

$$\text{risk}(m, D(\mathbf{\Phi})) \approx \frac{1}{n_\epsilon}\sum_j^{n_\epsilon} \|\widehat{m}(D(\mathbf{\Phi}), m, \epsilon_j) - m\|_2^2, \tag{13}$$

where ϵ_j is a noise realization and n_ϵ is the number of noise realizations.

While considering the expectation may eliminate the uncertainly associated with the noise, yet, we are still left with a measure which depends on the unknown model. In order to eliminate the unknown model dependency from the risk we average over the model space $\mathcal{M}$ by using the training set M_s. Note that for situations where a PDF for $\mathcal{M}$ is available, this procedure is equivalent to a Monte-Carlo integration over the space $\mathcal{M}$. Thus, we define the optimal $\mathbf{\Phi}$ as the one that yields a dictionary that minimizes

$$\min_{\mathbf{\Phi}} \boldsymbol{J} = \frac{1}{2sn_\epsilon} \sum_{i=1}^{s} \sum_{j=1}^{n_\epsilon} \|D(\mathbf{\Phi})\widehat{u}_{ij}(D(\mathbf{\Phi}), m_i) - m_i\|_2^2 \tag{14a}$$

$$\text{s.t} \qquad \widehat{u}_{ij} = \underset{u_{ij}}{\operatorname{argmin}} \ \frac{1}{2}\|AD(\mathbf{\Phi})u_{ij} - Am_i - \epsilon_j\|_2^2 + \alpha\|u_{ij}\|_1. \tag{14b}$$

A few comments are in order

- So far, the choice of the regularization parameter α was not discussed. However, it can be easily observed that the optimal $\mathbf{\Phi}$ scales the regularization term such that it fits best the data. This become evident, if one notes that for $\beta \neq 0$

  $$v = D(\mathbf{\Phi})u = D(\beta\mathbf{\Phi})(\beta^{-1}u).$$

  Thus, dictionary scaling automatically results in subsequent rescaling of u, which in term is equivalent to modification of the regularization parameter α. This scaling obviously works on average for the training models. This choice is reasonable since the design process is performed *previous* to collection of any data.

- Even though the problem was formulated as a minimization problem, in many cases one may have already considered a particular choice of $\mathbf{\Phi}$. In such cases, improvement of the given $\mathbf{\Phi}$ may suffice in practice. This implies that a low accuracy solution to the optimization problem may be satisfactory in effect.

- The underlying assumption in this process is that a training set M_s of plausible models is obtainable, and that these models are essentially samples taken from some common space $\mathcal{M}$. Although it is difficult to verify this assumption in practice, it has been used successfully in the past [72]. In fact, a similar assumption forms the basis for empirical risk minimization and Support Vector Machines (SVM) [84, 77]. A discussion regarding suitable methods for extraction of such a set is beyond the scope of this study.

- Following the above formulation, it is evident that an optimal dictionary for a particular forward problem, would differ from an optimal dictionary of another, even if an identical training set M_s is utilized. Thus, the forward problem, rather than merely the model space by itself, plays a primary role in the design of an optimal dictionary. Such prominent property is absent when the forward operator is taken to be a unit matrix (image denoising), as has been exercised by previous authors [67, 53, 49, 2, 62].

5 Solving the Dictionary Design Problem

In the followings we describe a methodology for the solution of the design problem. The design problem is formulated as a bi-level optimization problem. The solution of such problems can be difficult. An important feature of the suggested formulation is that the inner optimization problem is convex. Thus, numerical methods for the solution of the outer optimization problem can be derived, avoiding local minima convergence jeopardy within the solution of the inner optimization problem [6, 85, 4, 59, 20, 40, 5, 21, 18].

We utilize the sensitivity calculation described in Section 3 in order to compute the reduced gradient of (14). Let p, q represent the decomposition of the solution u at the minimum. We compute the gradient of the loss function

$$L(D) = \frac{1}{2}\|D(p(D) - q(D)) - m\|_2^2$$

with respect to a perturbation in the j^{th} column of D, δD_j. By means of Taylor's expansion we can express $L(D + \delta D_j e_j^\top)$ as

$$\begin{aligned}
&\frac{1}{2}\|(D + \delta D_j e_j^\top)(p(D + \delta D_j e_j^\top) - q(D + \delta D_j e_j^\top)) - m\|_2^2 \approx \\
&\frac{1}{2}\|(D + \delta D_j e_j^\top)(p(D) + p_{D_j}\delta D_j - q(D) - q_{D_j}\delta D_j) - m\|_2^2 \approx \\
&\frac{1}{2}\|(D(p(D) - q(D)) + [p(D) - q(D)]_j \delta D_j + D(p_{D_j} - q_{D_j})\delta D_j - m\|_2^2.
\end{aligned}$$

By defining the sensitivity of $m = D(p - q)$ with respect to D_j as

$$J_{D_j} = [p(D) - q(D)]_j I + D(p_{D_j} - q_{D_j}), \tag{15}$$

we can now rewrite the perturbed loss, L, as

$$\frac{1}{2}\|(D + \delta D_j e_j^\top)(p(D + \delta D_j e_j^\top) - q(D + \delta D_j e_j^\top)) - m\|_2^2 = \frac{1}{2}\|D(p - q) + J_{D_j}\delta D_j - m\|_2^2,$$

which implies that the gradient of the loss L with respect to D_j is nothing but

$$\boldsymbol{\nabla}_{D_j} L = J_{D_j}^\top (D(p - q) - m).$$

Finally, given the parametrization $D = D(\boldsymbol{\Phi})$, we compute $\frac{\partial D_j}{\partial \boldsymbol{\Phi}}$ and sum over all the columns of D, the training models and the noise realizations to obtain the gradient of the objective function $\boldsymbol{J}$ in (14) with respect to $\boldsymbol{\Phi}$

$$\boldsymbol{\nabla}_{\boldsymbol{\Phi}} \mathbf{J} = \sum_{ijk} \left(\frac{\partial D_j}{\partial \boldsymbol{\Phi}}\right)^\top J_{D_j}^\top (D(p_{ik} - q_{ik}) - m_i). \tag{16}$$

Although computation of a gradient is feasible, the employment of conventional optimization techniques for the solution of the design problem is still problematic. One needs to

recall that the solution is not continuously differentiable with respect to the dictionary and therefore, only methods which were designed to deal with non-smooth optimization problems are permissible. One recent method was investigated in [11]. Its main appeal is that only simple modification of a gradient descent or the L-BFGS [64] algorithm is needed for obtaining an approximate solution.

6 Numerical Studies

In order to test our algorithm we have considered two different inverse problems.

- 2D image deblurring problem. This problem (see [63, 68]) is of relevance for a broad range of real-life applications, e.g. optics, aerospace, machine vision and microscopy.
- The 2D limited-angle ray tomography problem. This problem is a common geophysical test problem. It has been used by various authors; see for example [80, 1, 39, 79, 15, 89, 93, 54, 88].

The algorithm testing procedure consisted of two separate stages: dictionary learning and performance assessment. In the first stage, the learning phase, a dictionary was trained using an initial prototype dictionary D_0 and a training data set $\{d_1, .., d_s\}$, which corresponded to a particular training model set $\{m_1, .., m_s\}$ subject to the application of the observation operator A.

In the second stage, the performance of the acquired trained dictionary D_t was compared with that of the original prototype dictionary D_0 by solving the inverse problem for a given separate set of data $\{d_{s+1}, \ldots, d_{s+k}\}$, which corresponded to an unseen set of models $M_k = \{m_{s+1}, \ldots, m_{s+k}\}$. Hereafter, the set M_k and its corresponding data is referred to as a validation set. The validation set was not used for dictionary training purposes, and was solely used for performance assessment purposes. In order to quantify the performance of the optimal dictionary we define the relative risk measure

$$\text{relative risk} = \frac{R(D_t)}{R(D_0)}.$$

The risk was recorded over the training and validation sets. By definition, for the training set, the relative risk is smaller than one. If a similar number is attained for the validation set, then the obtained dictionary is reasonably generalized for the addressed problem. Conversely, if the relative risk of the validation set is far from the relative risk of the training set, then we may concur that either the training or the validation set fails to represent the problem. The size of the dictionary is assumed to be considerably smaller than the size of the models for all problems, therefore, over-fitting can be excluded.

6.1 Results for image deblurring

The first training set was generated by splitting a 256×256 head MRI scan of a mid-lateral sagittal projection into 64 sub-images of 32×32, out of which subsets of 15 sub-images

were randomly chosen. Only sub-images with variance exceeding 20% of the overall mean variance in the training model were considered. This way, sub-images of smooth background which are characterized by poor feature content were excluded from the training set. These sub-images conveyed a portion of 20% of the entire training image. As test data, 17 head MRI image slices from lateral sagittal projections of 256×256 were used. Two different operators A were applied over this data set: a 4×4 averaging operator and a Gaussian point spread function operator ([16]). Dictionary training was performed using 32×128 overcomplete DCT, feature and random prototype dictionaries. Convergence was reached after 16-22 iterations, in which the relative empirical risk reduced by 115%-150%. The norm of the relative dictionary change $\frac{\|D_t - D_0\|}{\|D_0\|}$ ranged between 0.69 - 0.76. Similar improvement factor was obtained for the validation set.

A recovered MRI image, using blurred data and a trained dictionary is presented in figure 1. The improvement in the quality of the recovered images was visually apparent, although, less dramatic than the quantitative improvement. These findings are sensible due to two reasons, first, the obtained recovered models were relatively quite successful, even by using a prototype dictionary, and thus, a radical improvement was not feasible. Second, our improvement measure is related to the risk, which in this study was elected to be an ℓ_2-norm, which differs from the eye-norm. This issue is discussed further in the next section.

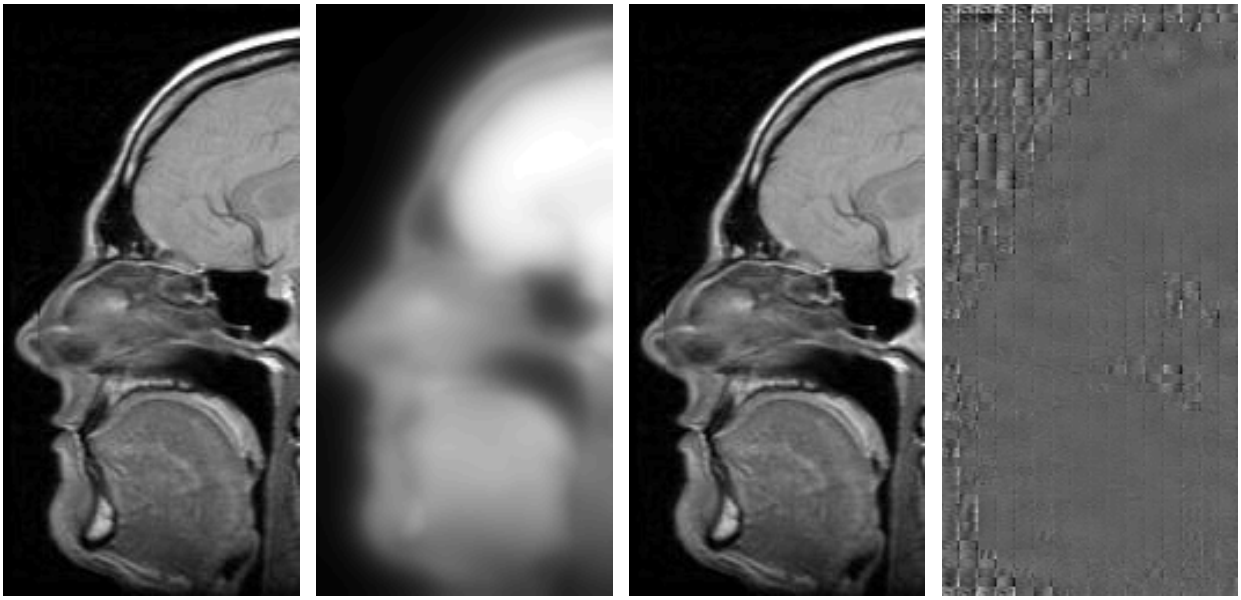

Figure 1: Dictionary learning for MRI image with Gaussian PSF operator. Left to right: true models m, data (model after application of A), recovered model using trained dictionary $D_t u$, error $\|m - D_t u\|^2$ (gray scale of the error were rescaled for display purposes)

6.2 Results for the 2D geophysical tomography problem

The second training set was generated from a 122×384 Marmousi hard model. Four 60×60 sub-models were arbitrarily elected for that purpose. These sub-models conveyed a portion of 42% of the entire model. As test data, four other 60×60 sub-models were elected. A limited-angle ray tomography operator A was applied over the model set. Dictionary training was conducted using a 36×128 overcomplete DCT prototype dictionary. After 24 design iterations convergence was achieved (see Figure 6.2). Within that process, the empirical risk reduced by a factor of 420% and the norm of the relative dictionary change was 0.56. Similar improvement factor was expected for the corresponding validation set. Models recovered using the prototype dictionary and model recovered using the trained dictionary can be found in figure 3. For this problem, greater improvement in model recovery was achieved by dictionary training. We attribute this improvement to the fact that this problem is more severely ill-posed. Hence, there was more breadth for improvement from the relatively poorly recovered models that were obtained using the original prototype dictionary.

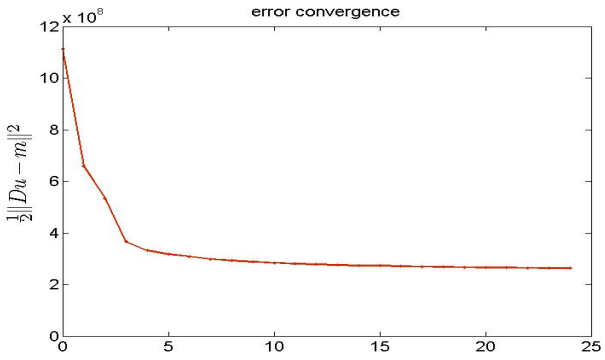

Figure 2: Risk convergence during the dictionary learning process for four sub-models of Marmousi hard model with limited-angle ray tomography data

7 Conclusions and Future Challenges

A comprehensive sensitivity analysis and an innovative method for dictionary design for solving generic inverse problems by means of sparse representation were presented here.

Our numerical experiments demonstrated that regardless the choice of initial dictionary, the least square error of models recovered using trained dictionaries were consistently smaller than that of models recovered using the original dictionaries. Moreover, a comparison of the error (risk) reduction over the testing data versus the validation data revealed that the acquired trained dictionaries were sufficiently general to provide equivalent results over unseen data.

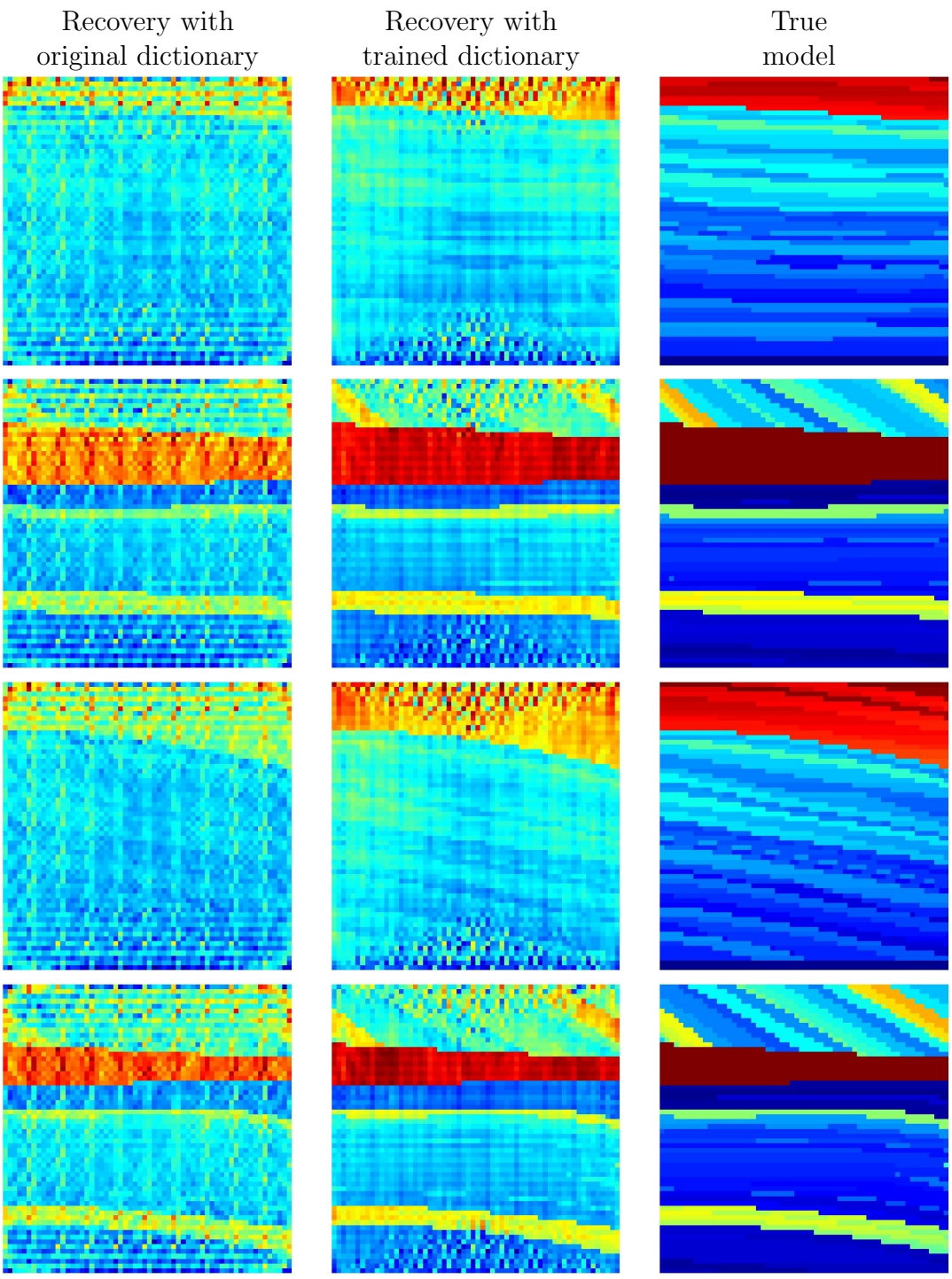

Figure 3: Dictionary learning for four sub-models of Marmousi hard model recovered from limited angle ray tomography data. Left to right: true models $m_{1..4}$, data (model after application of A), recovered models using original dictionary $D_0 u_{1..4}(D_0)$, recovered models using trained dictionary $D_t u_{1..4}(D_t)$

An important observation is that despite the relatively large percentage improvement in the least square ℓ_2-norm error (up to 500% in some cases) in using a trained dictionary over a prototype dictionary, such improvement was less apparent when assessed by the appraisal of the eye (sometimes referred to as the "eyeball norm"). This discrepancy can be attributed mainly to the fact that the considered loss measure, i.e. least square ℓ_2-norm, differs substantially from the error measure employed by our vision. Nevertheless, the methodology proposed here allows incorporation of any other derivable loss expression.

Several future questions need to be addressed. The two principal issues are to explore which algorithm performs best for the design problem, and to prescribe the minimum bound for the number of training models that are required for obtaining robust results. We intend to pursue these challenging questions in our future work.

8 Acknowledgements

The authors wish to express their great gratitude to Michael Friedlander and Luis Tenorio for their detailed remarks and thorough reviews. In addition, we wish to thank Michele Benzi, Miki Elad, Raya Horesh, Jim Nagy and Steve Wright for their valuable advices.

This research was supported by NSF grants DMS-0724759, CCF-0427094, CCF-0728877 and by DOE grant DE-FG02-05ER25696.

References

[1] R. Abgrall and J.D. Benamou. Big ray-tracing and eikonal solver on unstructured grids: Application to the computation of a multivalued traveltime field in the marmousi model. *Geophysics*, 64(1):230–239, 1999.

[2] M. Aharon, M. Elad, and A. Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on signal processing*, 54(11):4311–4322, 2006.

[3] M. Akcakaya and V. Tarokh. Performance study of various sparse representation methods using redundant frames. In *41st Annual Conference on Information Sciences and Systems, 2007. CISS '07*, pages 726–729, 14-16 March 2007.

[4] N. Alexandrov and J. E. Dennis. Algorithms for bilevel optimization. *AIAA/USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization, 5th, Panama City Beach, FL*, pages 810–816, 1994.

[5] J.F. Bard. *Practical bilevel optimization: Algorithms and applications (Nonconvex optimization and its applications)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[6] J.F. Bard and J.T. Moore. A branch and bound algorithm for the bilevel programming problem. *SIAM Journal on Scientific and Statistical Computing*, 11(2):281–292, 1990.

[7] J. Bee Bednar, L. R. Lines, R. H. Stolt, and A. B. Weglein. *Geophysical inversion*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1992.

[8] R. Bierman and R. Singh. Influence of dictionary size on the lossless compression of microarray images. In *Twentieth IEEE International Symposium on Computer-Based Medical Systems, 2007. CBMS '07*, pages 237–242, 20-22 June 2007.

[9] J.F. Bonnans and A. Shapiro. *Perturbation analysis of optimization problems*. Springer, New York, 2000.

[10] A.M. Bruckstein, D.L. Donoho, and M. Elad. From sparse solutions of systems of equations to sparse modeling of signals and images. *to appear in SIAM Review*, 2008.

[11] J.V. Burke, A.S. Lewis, and M.L. Overton. A robust gradient sampling algorithm for nonsmooth, nonconvex optimization. *SIAM J. Optimization*, 15:751–779, 2005.

[12] E. Candes, N. Braun, and M. Wakin. Sparse signal and image recovery from compressive samples. In *4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, 2007. ISBI 2007*, pages 976–979, 12-15 April 2007.

[13] J. Chen and X. Huo. Sparse representations for multiple measurement vectors (mmv) in an over-complete dictionary. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05)*, volume 4, pages iv/257–iv/260, 18-23 March 2005.

[14] S.S. Chen, D.L. Donoho, and M.A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1998.

[15] Y. Choi, C. Shin, D.J. Min, and T. Ha. Efficient calculation of the steepest descent direction for source-independent seismic waveform inversion: an amplitude approach. *J. Comput. Phys.*, 208(2):455–468, 2005.

[16] J. Chung, E. Haber, and J. Nagy. Numerical methods for coupled super-resolution. *Inverse Problems*, 22:1261–1272, 2006.

[17] J. Claerbout and F. Muir. Robust modeling with erratic data. *Geophysics*, 38:826–844, 1973.

[18] B. Colson, P. Marcotte, and G. Savard. An overview of bilevel optimization. *Annals of Operations Research*, 153,1:235–256, 2007.

[19] L. Danhua, S. Guangming, and G. Dahua. A new method for signal sparse decomposition. In *International Symposium on Intelligent Signal Processing and Communication Systems, 2007. ISPACS 2007*, pages 750–753, Nov. 28 2007-Dec. 1 2007.

[20] S. Dempe. A bundle algorithm applied to bilevel programming problems with non-unique lower level solutions. *Comput. Optim. Appl.*, 15(2):145–166, 2000.

[21] S. Dempe and N. Gadhi. Necessary optimality conditions for bilevel set optimization problems. *J. of Global Optimization*, 39(4):529–542, 2007.

[22] D.L. Donoho. For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics*, 59(6):797–829, 2006.

[23] D.L. Donoho and M. Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via ell 1 minimization. *PNAS*, 100(5):2197–2202, 2003.

[24] D.L. Donoho and J. Tanner. Thresholds for the recovery of sparse solutions via l1 minimization. In *40th Annual Conference on Information Sciences and Systems, 2006*, pages 202–206, 22-24 March 2006.

[25] F.X. Dupe, M.J. Fadili, and J.L. Starch. Image deconvolution under poisson noise using sparse representations and proximal thresholding iteration. In *IEEE International Conference on Acoustics, Speech and Signal Processing, 2008. ICASSP 2008*, pages 761–764, March 31 2008-April 4 2008.

[26] M. Elad and A.M. Bruckstein. On sparse signal representations. In *International Conference on Image Processing, 2001*, volume 1, pages 3–6vol.1, 7-10 Oct. 2001.

[27] L.O. Endelt and A. la Cour-Harbo. Wavelets for sparse representation of music. In *Proceedings of the Fourth International Conference on Web Delivering of Music, 2004. WEDELMUSIC 2004*, pages 10–14, 2004.

[28] K. Engan and K. Skretting. A novel image denoising technique using overlapping frames. In *Visualization, Imaging, and Image Processing*, 2002.

[29] K. Engan, K. Skretting, and J. Håkon Husoy. Family of iterative ls-based dictionary learning algorithms, ils-dla, for sparse signal representation. *Digit. Signal Process.*, 17(1):32–49, 2007.

[30] M.J. Fadili and J.L. Starck. Em algorithm for sparse representation-based image inpainting. In *IEEE International Conference on Image Processing, 2005. ICIP 2005*, volume 2, pages II–61–4, 11-14 Sept. 2005.

[31] M.A.T. Figueiredo, R.D. Nowak, and S.J. Wright. Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. *IEEE Journal of Selected Topics in Signal Processing*, 1:586–597, 2007.

[32] S. Fischer, G. Cristöbal, and R. Redondo. Sparse overcomplete gabor wavelet representation based on local competitions. *IEEE Trans. Image Process.*, 15(2):265–272, Feb. 2006.

[33] P. Földiák. Forming sparse representations by local anti-hebbian learning. *Biological. Cybernetics*, 64:165–170, 1990.

[34] J.J. Fuchs. Recovery of exact sparse representations in the presence of noise. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04)*, volume 2, pages ii–533–6vol.2, 17-21 May 2004.

[35] J.J. Fuchs and C. Guillemot. Fast implementation of a penalized sparse representations algorithm: Applications in image denoising and coding. In *Conference Record of the Forty-First Asilomar Conference on Signals, Systems and Computers, 2007. ACSSC 2007*, pages 508–512, 4-7 Nov. 2007.

[36] M. Girolami. A variational method for learning sparse and overcomplete representations. *Neural Computation*, 13(11):2517–2532, 2001.

[37] W.P. Gouveia and J.A. Scales. Resolution of seismic waveform inversion: Bayes versus occam. *Inverse Problems*, 13(2):323–349, 1997.

[38] R. Gribonval and M. Nielsen. Sparse decompositions in "incoherent" dictionaries. In *International Conference on Image Processing, 2003. ICIP 2003. Proceedings. 2003*, volume 1, pages I–33–6vol.1, 14-17 Sept. 2003.

[39] Z. Guanquan and Z. Wensheng. Parallel implementation of 2-d prestack depth migration. In *The Fourth International Conference/Exhibition on High Performance Computing in the Asia-Pacific Region, 2000. Proceedings*, volume 2, pages 970–975, 14-17 May 2000.

[40] Z.H. Gümüş and C.A. Floudas. Global optimization of nonlinear bilevel programming problems. *J. of Global Optimization*, 20(1):1–31, 2001.

[41] E. Haber, U.M. Ascher, and D. Oldenburg. On optimization techniques for solving nonlinear inverse problems. *Inverse Problems*, 16:1263–1280, 2000.

[42] Wotao Y. Hale, E.T. and Z. Yin. Fixed-point continuation for l1-minimization: methodology and convergence. Technical report, Rice, 2008.

[43] P.C. Hansen. *Rank-deficient and discrete ill-posed problems: Numerical aspects of linear inversion*. Philadelphia, 1998.

[44] A. Hyvarinen, P. Hoyer, and E. Oja. Sparse code shrinkage: denoising of nongaussian data by maximum likelihood estimation. *Neural Computation*, 11(7):1739–1768, Oct 1999.

[45] T.A. Johansen. On tikhonov regularization, bias and variance in nonlinear system identification. *Automatica*, 33(3):441–446, 1997.

[46] C.Y. Jong, Y.L. Su, and Y. Bresler. Exact reconstruction formula for diffuse optical tomography using simultaneous sparse representation. In *5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, 2008. ISBI 2008*, pages 1621–1624, 14-17 May 2008.

[47] C. Kervrann and J. Boulanger. Optimal spatial adaptation for patch-based image denoising. *IEEE Trans Image Process*, 15(10):2866–78, 2006.

[48] K. Koh, S.J. Kim, and S. Boyd. An interior-point method for large-scale l1-regularized logistic regression. *J. Mach. Learn. Res.*, 8:1519–1555, 2007.

[49] K. Kreutz-Delgado, J.F. Murray, B.D. Rao, K. Engan, T.W. Lee, and T.J. Sejnowski. Dictionary learning algorithms for sparse representation. *Neural Computation*, 15(2):349–396, February 2003.

[50] Chinh La and M.N. Do. Tree-based orthogonal matching pursuit algorithm for signal reconstruction. In *IEEE International Conference on Image Processing, 2006*, pages 1277–1280, 8-11 Oct. 2006.

[51] H. Lee, A. Battle, R. Raina, and Y.N. Andrew. Efficient sparse coding algorithms. pages 801–808, 2006.

[52] M.S. Lewicki and B.A. Olshausen. Inferring sparse, overcomplete image codes using an efficient coding framework. pages 815–821, 1998.

[53] M.S. Lewicki and T.J. Sejnowski. Learning overcomplete representations. *Neural Computation*, 12(2):337–365, 2000.

[54] Y. Ma and G.F. Margrave. Seismic depth imaging with the gabor transform. *SEG Technical Program Expanded Abstracts*, 25(1):2504–2508, 2006.

[55] J. Mairal, M. Elad, and G. Sapiro. Sparse representation for color image restoration. *IEEE Trans. Image Process.*, 17(1):53–69, Jan 2008.

[56] D.M. Malioutov, M. Cetin, and A.S. Willsky. Optimal sparse representations in general overcomplete bases. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04)*, volume 2, pages ii–793–6vol.2, 17-21 May 2004.

[57] S. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.

[58] L. Mancera and J. Portilla. L0-norm-based sparse representation through alternate projections. In *IEEE International Conference on Image Processing, 2006*, pages 2089–2092, 8-11 Oct. 2006.

[59] A. Migdalas, P.M. Pardalos, and P. Värbrand. *Multilevel optimization: Algorithms and applications*, volume 20 of *Nonconvex Optimization and Its Applications*. Springer, 1998.

[60] M. Mishali and Y.C. Eldar. The continuous joint sparsity prior for sparse representations: Theory and applications. In *2nd IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing, 2007. CAMPSAP 2007*, pages 125–128, 12-14 Dec. 2007.

[61] G.H. Mohimani, M. Babaie-Zadeh, and C. Jutten. Complex-valued sparse representation based on smoothed &sup¿0norm. In *IEEE International Conference on Acoustics, Speech and Signal Processing, 2008. ICASSP 2008*, pages 3881–3884, March 31 2008-April 4 2008.

[62] J.F. Murray and K. Kreutz-Delgado. Learning sparse overcomplete codes for images. *Journal of VLSI Signal Processing*, 45, 1:97–110, 2006.

[63] J.G. Nagy and D.P. O'Leary. Image deblurring: I can see clearly now. *Computing in Science and Engg.*, 5(3):82–84, 2003.

[64] J. Nocedal and S. Wright. *Numerical Optimization*. New York: Springer, 1999.

[65] B. A. Olshausen and D. J. Field. Natural image statistics and efficient coding. *Network*, 7(2):333–339, May 1996.

[66] B.A. Olshausen and D.J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, June 1996.

[67] B.A. Olshausen and D.J. Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision Research*, 37:3311–3325, 1997.

[68] D.P. O'Leary P.C. Hansen, J.G. Nagy. *Deblurring images: Matrices, spectra and filtering*, volume 75. Society for Industrial and Applied Mathematics, 08 2006.

[69] L. Peotta, L. Granai, and P. Vandergheynst. Image compression using an edge adapted redundant dictionary and wavelets. *Signal Process.*, 86(3):444–456, 2006.

[70] T. Poggio and F. Girosi. A sparse representation for function approximation. *Neural Computation*, 10(6):1445–1454, 1998.

[71] L. Qiangsheng, W. Qiao, and W. Lenan. Size of the dictionary in matching pursuit algorithm. *IEEE Transactions on Signal Processing*, 52, 12:3403–3408, 2004.

[72] A. Rakhlin. *Applications of empirical processes in learning theory: algorithmic stability and generalization bounds*. PhD thesis, Dept. of Brain and Cognitive Sciences, Massachusetts Institute of Technology, 2006.

[73] S.M. Robinson. Perturbed kuhn-tucker points and rates of convergence for a class of nonlinear-programming algorithms. *Math. Programming*, 7:1–16, 1974.

[74] T. Ryen, S.O. Aase, and J.H. Husoy. Finding sparse representation of quantized frame coefficients using 0-1 integer programming. In *Proceedings of the 2nd International Symposium on Image and Signal Processing and Analysis, 2001. ISPA 2001*, pages 541–544, 19-21 June 2001.

[75] R. Saab, R. Chartrand, and O. Yilmaz. Stable sparse approximations via nonconvex optimization. In *IEEE International Conference on Acoustics, Speech and Signal Processing, 2008. ICASSP 2008*, pages 3885–3888, March 31 2008-April 4 2008.

[76] M.D. Sacchi and T.J Ulrych. Improving resolution of radon operators using a model re-weighted least squares procedure. *Journal of Seismic Exploration*, 4:315–328, 1995.

[77] J. Shawe-Taylor and N. Cristianini. *Support vector machines and other kernel-based learning methods*. Cambridge University Press, 2000.

[78] K. Skretting, K. Engan, Husy J. H., and S.O. Aas. Sparse representation of images using overlapping frames. In *SCIA*, Bergen, Norway, June 2001.

[79] P. Sukjoon, S. Changsoo, M. Dong-Joo, and H. Taeyoung. Refraction traveltime tomography using damped monochromatic wavefield. *Geophysics*, 70(2):U1–U7, 2005.

[80] P. Thierry, S. Operto, and G. Lambaré. Fast 2-d ray + born migration/inversion in complex media. *Geophysics*, 64(1):162–181, 1999.

[81] M.E. Tipping. Sparse bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.*, 1:211–244, 2001.

[82] Y. Tsaig and D.L. Donoho. Breakdown of equivalence between the minimal l1-norm solution and the sparsest solution. *Signal Processing*, 86(3):533–548, Mar 2006.

[83] E. van den Berg and M. P. Friedlander. In pursuit of a root. Technical report, Department of Computer Science, University of British Columbia, June 2007.

[84] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

[85] L.N. Vicente and P.H. Calamai. Bilevel and multilevel programming: A bibliography review. *Journal of Global Optimization*, 5(3):291–306, October 1994.

[86] J. E. Vila-Forcn, O. Voloshynovskiy, S.and Koval, and T. Pun. Facial image compression based on structured codebooks in overcomplete domain. *EURASIP Journal on Applied Signal Processing*, ID 69042:11, 2006.

[87] C. Vogel. *Computational methods for inverse problem*. SIAM, Philadelphia, 2001.

[88] J. Wang and M.D. Sacchi. High-resolution wave-equation amplitude-variation-with-ray-parameter (avp) imaging with sparseness constraints. *Geophysics*, 72(1):S11–S18, 2007.

[89] Z. Wang. Factor analysis for ocean remote sensing. In *OCEANS 2006*, pages 1–6, Sept. 2006.

[90] K.P. Whittall and D.W. Oldenburg. *Inversion of magnetotelluric data for a one dimensional conductivity*, volume 5. SEG monograph, 1992.

[91] T. Yardibi, J.Li, P. Stoica, and L.N. Cattafesta. Sparsity constrained deconvolution approaches for acoustic source mapping. *J. Acoust. Soc. Am.*, 123(5):2631–2642, May 2008.

[92] H. Zhaoshui, X. Shengli, and F. Yuli. Sparse representation of complex valued signals. In *International Conference on Computational Intelligence and Security, 2006*, volume 2, pages 1008–1011, 3-6 Nov. 2006.

[93] W. Zhenhai and C.H. Chen. Factor analysis for geophysical signal processing with seismic profiles. In *International Joint Conference on Neural Networks, 2006. IJCNN '06*, pages 2555–2560, 16-21 July 2006.