OPTIMAL ESTIMATION OF $\ell_1$-REGULARIZATION PRIOR FROM A REGULARIZED EMPIRICAL BAYESIAN RISK STANDPOINT

HUI HUANG
Shenzhen Institute of Advanced Technology
Chinese Academy of Sciences
Shenzhen University Town, 518055, China

ELDAD HABER
Department of Mathematics and Earth and Ocean Science
The University of British Columbia
Vancouver, BC, V6T 1Z2, Canada

LIOR HORESH
Business Analytic and Mathematical Sciences
IBM T J Watson Research Center
26-254, Yorktown Heights, NY, 10598, USA

Abstract. We address the problem of prior matrix estimation for the solution of $\ell_1$-regularized ill-posed inverse problems. From a Bayesian viewpoint, we show that such a matrix can be regarded as an influence matrix in a multivariate $\ell_1$-Laplace density function. Assuming a training set is given, the prior matrix design problem is cast as a maximum likelihood term with an additional sparsity-inducing term. This formulation results in an unconstrained yet nonconvex optimization problem. Memory requirements as well as computation of the nonlinear, nonsmooth sub-gradient equations are prohibitive for large-scale problems. Thus we introduce an iterative algorithm to design efficient priors for such large problems. We further demonstrate that the solutions of ill-posed inverse problems by incorporation of $\ell_1$-regularization using the learned prior matrix perform generally better than commonly used regularization techniques where the prior matrix is chosen a-priori.

1. Introduction. In this study we consider the solution of a discrete inverse problem for a forward problem of the form

$$ y = F(x) + \eta, $$

where $F : \mathbb{R}^m \to \mathbb{R}^n$ is the forward model operator (observation operator) and $m \geq n$, $x \in \mathbb{R}^m$ is the model to be recovered, $y \in \mathbb{R}^n$ is the given noisy data, and $\eta \in \mathbb{R}^n$ represents random measurement noise.

The operators $F$ may vary from an identity operator in simple denoising problems [31, 37, 32] through a linear, constant coefficient form, such as in deblurring and tomography problems [3, 20, 19] to the solution of a diffusive partial differential equation (PDE) system involving Poisson or Maxwell’s differential equations [8, 18]. For most applications the operator $F$ is considered to be ill-posed, that is, it cannot be stably...
inverted in order to recover $x$ given $y$. To obtain stable solutions one typically employs regularization. The objective of regularization is incorporation of “appropriate” a-priori information, and thereby stabilize the inversion process.

Probably the most common type of regularization is the Tikhonov regularization [36], in which the inversion problem is repasted as an optimization problem of the form

$$\min_x T(x) = \frac{1}{2} \|F(x) - y\|^2 + \alpha R(x),$$

where $R : \mathbb{R}^m \to \mathbb{R}^+$ is the regularization operator and $\alpha$ is a regularization parameter. The simplest regularization operator is perhaps of a quadratic form. This reads

$$R(x) = \frac{1}{2} \|Lx\|^2,$$

where $L$ is a regularization matrix, typically, a discretization of the gradient. For problems where the solutions are indeed smooth, such choice of regularization matrix and functional is hard to beat, however, this choice is improper in the presence of model discontinuities [7].

In such circumstances, one typically resort to $\ell_1$-regularization

$$R(x) = \|Dx\|_1.$$ 

If $D$ is chosen to be a derivative operator, such regularization scheme can preserve edges and discontinuous model jumps. This specific configuration corresponds to the well-known total variation regularization [37] (at least in 1D). For more general choices of the prior matrix $D$ other structural properties can be obtained. In particular, it is broadly acknowledged that the $\ell_1$-norm yields sparse solutions where $Dx$ has many zeros and only few non-zeros. This property of the solution has been explored extensively in the last decade [5, 6] in the contexts of sparse representation and compressive sensing. One typically chooses $D$ such that it transforms $x$ into a space that can be expected to be sparse. This had led to many choices of $D$, from wavelets and curvelets to chirplets and other transformations. Evidently the effectiveness of the regularization scheme is highly subjective to that choice. While the aforementioned, model-agnostic transformations may perform reasonably well as generic priors, they are prone to be sub-optimal as rarely the structural assumption they hold upon the model are met fully. In particular, information regarding the model space, the observation operator and the noise characteristics of the data are not accounted for. For that reason, a great diversity of transformations exist in hope that one may incidentally capture some of the problem domain-specific characterizing features.

Therefore we propose a novel and effective approach from a different angle. Rather than choosing $D$ out of broad set of possible transformations, we wish to design the prior matrix $D$ and tailor it to our specific problem domain. For clarity, we shall distinguish between the problem associated with estimation of the model $x$, hereafter the inverse problem, and the problem of estimating the prior matrix $D$, hereafter, the design problem. Although it will become apparent later, in some cases the latter can be regarded as an inverse problem by itself.

Let us consider now the case where a particular family of models $x \in \mathcal{X}$ (the space $\mathcal{X}$ is specified in the next section) is available, and assume that an $\ell_1$-regularization scheme is utilized in order to recover $x$ given some data $y$. The question is, how should we optimally design the regularization matrix $D$?
Interestingly, this question is also related to a somewhat different formulation. If $D$ is square and invertible then we can rewrite the problem (1) in its analysis form
\[
\min_z T(z) = \frac{1}{2} \|F(D^{-1}z) - y\|^2 + \alpha \|z\|_1,
\]
whereas the formulation (1) is known as the synthesis form [11].

Given some information about the space $X$, one can opt for obtaining an optimal $D^{-1}$ for this formulation. In the simplest settings where $F$ is merely the identity operator, techniques such as those described in [25, 16, 24, 28, 12, 11] can be utilized. For problems where $F$ is linear and the matrix $D$ is overcomplete, the algorithmic formulation proposed in [21] can be employed. Nonetheless, this formulation entails a nonsmooth bi-level optimization problem that is very difficult to solve.

In this research work the synthesis formulation (1) with $R(x) = \|Dx\|_1$ is considered. For these settings we develop a methodology for designing an optimal prior matrix $D$. The remainder of the paper is organized as follows. In Section 2 we review the principles of the Bayesian estimation framework for solution of inverse problems. We then show that one can obtain an optimization problem from this point of view to recover $D$. In Section 3 we discuss a numerical procedure that enable us to evaluate $D$. In Section 4 we test the validity of the proposed concepts along with their algorithmic implementation over several simplistic as well as over some realistic problems. Finally, in Section 5 we summarize the study and discuss the results.

2. Bayesian Framework for Estimation. In order to estimate the matrix $D$, we address the problem from a Bayesian viewpoint. Assuming $x$ and $y$ are jointly distributed continuous random variables, the posterior probability density function is given by
\[
\pi(x|y) = \frac{\pi(y|x)\pi(x)}{\pi(y)},
\]
where $\pi(x)$ and $\pi(y|x)$ are known as the prior and likelihood respectively. If $y = F(x) + \eta$ and $\eta \sim \text{Normal}(0, \sigma_\eta^2 I)$, then, the likelihood is
\[
\pi(y|x) = \frac{1}{(2\pi\sigma_\eta^2)^n/2} \exp \left( -\frac{\|F(x) - y\|^2}{2\sigma_\eta^2} \right).
\]

Consider that Laplace prior is identical to the analysis format of $\ell_1$-regularization. There are numerous papers using some version of this format [35, 15]. The basic claim is that Laplace does not penalize the tail of the distribution as much as say, the Gaussian distribution and thus is a very natural choice for incorporating sparseness in the problem. So here we assume that the prior is defined by the $\ell_1$-Laplace density function with zero mean [22]
\[
\pi(x) = \frac{1}{\sqrt{2}\det(D^{-1})} \exp(-\sqrt{2}|Dx|_1),
\]
where $D$ is invertible and the variance is $V = (D^TD)^{-1}$. The $\ell_1$-Laplace distribution defined in this way differs from standard generalized multivariate Laplace distributions whose negative-log-likelihood are usually proportional to the $\ell_2$-norm [23]. Here, if we set $D$ to be the identity matrix, the negative-log-likelihood of (3) is indeed proportional to the $\ell_1$-norm [11]. The posteriori distribution of $x$ given $y$ is
\[
\pi(x|y) \propto |\det(D)| \exp \left( -\frac{\|F(x) - y\|^2}{2\sigma_\eta^2} \right) \exp(-\sqrt{2}|Dx|_1).
\]
One of the several possible estimates of $x$ from $\pi(x|y)$ is the so-called maximum a-posteriori (MAP) estimate, which seeks for $x$ that maximizes

$$\hat{x} = \arg \max_x \{ \pi(x|y) \} = \arg \min_x \{ -\log(\pi(x|y)) \} = \arg \min_x \{ -\log(\pi(y|x)) - \log(\pi(x)) \},$$

where we can identify the first log term on the right hand side as a misfit term and the second as the $\ell_1$-regularization term in (1).

The Bayesian interpretation of regularization implies that $D$ is given by the prior; if the prior is specified by the distribution (3) and one can identify that prior then, $D$ is uniquely defined. Nonetheless, in realistic applications the prior is rarely known analytically and therefore, a numerical evaluation is required. In the following we describe how such numerical evaluation, also known as empirically-based framework, is conducted in practice.

To evaluate $D$, let us assume that $x \in \mathcal{X}$ where $\mathcal{X}$ is the probability space given by the distribution in (3) with an unknown invertible square matrix $D$. Further, assume also that we are able to draw $k$ samples from that space, obtaining a set of “training models” $\{x_1, \ldots, x_k\}$. Given these training models we would like to approximate $D$.

We believe that this scenario is rather realistic and even common in practice. For example, in image deblurring problems one often has access to a clean set of images that can be considered as a training set. Similarly, in medical imaging one possesses a-priori information about the human body. For such problems it is possible to obtain training sets with relative ease.

Assuming that $x_1, \ldots, x_k$ are independent and identically distributed with the $\ell_1$-Laplace distribution (3), the likelihood function for $D$ reads

$$L(D) = 2^{-\frac{mk}{2}} \prod_{i=1}^{k} |\det(D)| \exp(-\sqrt{2} |D x_i|_1)$$

$$= 2^{-\frac{mk}{2}} |\det(D)|^k \exp(-\sqrt{2} \sum_{i=1}^{k} |D x_i|_1) = 2^{-\frac{mk}{2}} |\det(D)|^k \exp(-\sqrt{2} \|DS\|_1),$$

where the training set $S = [x_1, \ldots, x_k] \in \mathbb{R}^{m \times k}$ and the $\ell_1$ matrix norm $\| \cdot \|_1$ denotes the sum of the absolute values of the matrix elements:

$$\|Y\|_1 = \sum_{i,j} |Y_{ij}|.$$ 

It is possible to obtain a maximum likelihood estimator for $D$ by minimizing the negative-log-likelihood (4). However, this approach is in general ill-advised. The main difficulty is that the number of samples is typically rather small compared to the dimension of $D$. Thus, although minimization of the negative-log-likelihood in such settings is unbiased, it prone to comprise a very high variance. This undesirable behavior is notoriously known for a related problem, namely, the evaluation of inverse covariance matrices for Gaussian priors [9, 10]. We therefore propose to deal with the problem in a similar way and minimize a penalized negative-log-likelihood.

Many different regularization choices for this design problem can be sought. However, one which is of particular appeal, would be to promote sparseness in $D$ itself by imposition of an $\ell_1$-norm penalty again. Yet, as oppose to the traditional use of $\ell_1$-norm penalty for promoting sparseness in the model solution $x$, this time the penalty
is applied upon the prior matrix \( D \) directly. This leads to the following optimization problem

\[
\min_D -\log(L(D)) + \rho \| D \|_1 = \min_D -k \log(\det(D)) + \sqrt{2} \| DS \|_1 + \rho \| D \|_1.
\]  

(5)

Here \( \rho \) is a regularization parameter. For a larger \( \rho \) value, \( D \) is increasingly sparse whereas for smaller \( \rho \) values it becomes denser. Selection of the (design) regularization parameter \( \rho \) can be obtained by any of the usual techniques for parameter estimation, such as generalized cross validation (GCV) [17], Akaike Information Criterion (AIC) [1] or Bayesian Information Criterion (BIC) [33]. We further elaborate on this matter in the next section.

At this point it is worthwhile to draw a comparison to the common problem of inverse covariance estimation [14, 27]. In a sense the technique discussed here is a simple extension of covariance estimation to an \( \ell_1 \)-prior. Similar ideas have recently addressed other priors [2]. Although the motivation for all these problems is rather similar, the problem here is much more challenging. In fact, it is clear that the design problem is nonconvex and that there is no unique minimizer for the objective function. Nonetheless, it is important to acknowledge that in many cases, sufficient improvement over an existing prior matrix \( D \) can be paramount as it can yield better reconstructions compared to \( D \)s that are chosen naively.

In order to carry out gradient-based optimization, computation of gradients of the objective function (5) is needed. The derivation of a sub-gradient of the matrix norm \( \| \cdot \|_1 \) with respect to \( D \) in (5) is straightforward, however, derivation of the gradient of \( \log(\det(D)) \) is more complicated, see [30]. Since

\[
\log(\det(D)) = \log(\det(D^\top D)^{\frac{1}{2}}) = \frac{1}{2} \log(\det(D^\top D)),
\]

we have

\[
\partial(\log(\det(D))) = \frac{1}{2} \det(D^\top D)^{-1} \partial(\det(D^\top D)).
\]

Now consider

\[
\partial(\det(D^\top D)) = \det(D^\top D)(D^\top D)^{-1} \partial(D^\top D) = \det(D^\top D)(D^\top D)^{-1}(I \otimes D^\top \partial D + D^\top \otimes I \partial D^\top) = \det(D^\top D)(D^\top D)^{-1} \partial D + D^\top D)^{-1} \partial D + D^\top \partial D) = \det(D^\top D)2D^{-\top} \partial D.
\]

It yields that

\[
\partial(\log(\det(D))) = D^{-\top}.
\]

Thus, we require that the sub-gradient of the negative-log-likelihood (5) satisfies

\[
-kD^{-\top} + \sqrt{2} \text{sign}(DS)S^\top + \rho \text{sign}(D) = 0.
\]

(6)

Here, we use the notation \([\text{sign}(A)]_{ij} = \text{sign}(A_{ij})\). In order to find a critical point of the function we need to find the zeros of the discontinuous function (6). This is discussed next.
Algorithm 1 Estimating the $\ell_1$-regularization prior: $D \leftarrow \text{EL1RP}(S, D_0, \varrho)$

Input: a training set $S$, initial guess $D_0$ and a sequence $\varrho$ with $N$ elements

for $i = 1$ to $N$ do
    $D \leftarrow D_0$ and $\rho \leftarrow \varrho(i)$
    Solve the equation (6) for $D(\rho)$
    Threshold $D(\rho)$, set $D(\rho)_{ij} = 0$ if $|D(\rho)_{ij}| < \epsilon$
    $i \leftarrow i + 1$
end for

Select $\rho^*$ by BIC (7)

$D \leftarrow D(\rho^*)$

Figure 1. Illustration of the Bayesian Information Criterion (7) for selection of the regularization parameter $\rho$ as applied to the design problem.

3. Solving the Optimization Problem. We now discuss the solution of the non-smooth optimization problem (5). Solution of this problem can be highly challenging due to the fact that the problem is not continuously differentiable as well as it large-dimensionality. Furthermore, since the regularization parameter $\rho$ is unknown a-priori, some criteria is needed for judicial selection. Recently, Wang et al. [38] demonstrated that the tuning parameters selected by a BIC type criterion can identify the true model consistently. Their theoretical results further extends not only the scope of applicability of the traditional BIC type criteria but also that of shrinkage estimation methods. The simulations clearly indicated the overall superiority of the proposed BIC in comparison with cross-validation and AIC. Therefore here, in this context we rationally select the optimal regularization parameter $\rho$ using a BIC-type criterion by minimizing

$$BIC(\rho) = -\log(|\det(\hat{D})|) + \frac{\sqrt{2}}{k} \|\hat{D}S\|_1 + p \frac{\log(k)}{k^2},$$

(7)

where $p$ is the number of nonzero entries in the estimated prior matrix $\hat{D}$. Fig. 1 displays the behavior of the proposed BIC selector for the 1D signal reconstruction example presented in the next section. In this case, an optimal parameter $\rho^*$ is sharply obtained at 2.8. We have found that the overall BIC algorithm is rather insensitive
OPTIMAL ESTIMATION OF $\ell_1$-REGULARIZATION PRIOR

SD

LSD

PCG

L-BFGS

| nVars | SD | LSD | PCG | L-BFGS |
|-------|----|-----|-----|--------|
| 64    | 9.09e-01 | 2.23e-01 | 8.68e-02 | 6.45e-02 |
| 256   | 8.41e-01 | 1.92e-01 | 8.64e-02 | 5.47e-02 |
| 1024  | 6.33e-01 | 1.31e-01 | 6.33e-02 | 3.94e-02 |

Table 1. Comparing the number of iterations and computing CPU time (in seconds) with different variable sizes (nVars) and numerical solvers. SD hits the upper bound of 3000 iterations before convergence.

Then we consider how to efficiently and accurately solve the sub-gradient equation (6) for $D(\rho)$, which is nonlinear and nonsmooth. In [26], following careful investigation Lewis and Overton claimed that the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm with an inexact line search can effectively solve nonsmooth and nonconvex functions. The method consistently converges to local minimizers, or at least to points that are apparently Clarke stationary, on various difficult classes of examples. The convergence rate varies consistently with the problem parameters and has been observed to be linear with respect to the number of function evaluations. Thus, to solve our optimization problem, the limited memory BFGS with Shanno-Phua scaling [34] is employed to compute the search direction. Once a search direction has been obtained bracketing is performed with consequent line search seeking for a point satisfying the strong Wolfe conditions. For enhanced rootedness, within the line search, we insert a safeguarded cubic interpolation to generate trial values, and the scheme switches to an Armijo back-tracking line search on iterations where the objective function enters a region where the parameters do not produce real valued output.

Besides L-BFGS, we have experimented with the standard steepest descent (SD) and its variant lagged steepest descent (LSD) [4], where the current step size is determined using the past rather than present information. The nonlinear conjugate gradient with diagonal preconditioner (PCG) has also been tested. In order to compare fairly, we monitor the objective function value at each L-BFGS iteration and terminate the iterative process when its minimization change rate is less than $10^{-4}$. For SD, LSD and PCG on the same problem, we stop the iterations when the corresponding objective function value reaches the value where L-BFGS is terminated.

Algorithm 1 is implemented in MATLAB and all computations are performed on a Mac with an Intel Core 2 Duo 2.66 GHz processor. Tables 1 and 2 display numerical comparisons with various number of variables and iterative solvers on three related aspects: iterations, computing times and average reconstruction errors. The
corresponding 1D example is described below, and the regularization parameter $\rho$ is selected by BIC (7) above. Clearly, the collected computational results demonstrate the L-BFGS generally outperforms the SD, LSD and PCG, and is capable of solving problems within a reasonable amount of time.

4. Numerical Examples. In this section we test the performance of the designed (learned) $D$ on various problems. We use some “toy” problems in 1D to experiment with a controlled environment and then use more realistic examples in 2D to test our approach in practice.

```
nVars = 256; nSamps = nVars/4;
B = speye(nVars)+sprandn(nVars,nVars,0.001); B = B'*B;
S = zeros(nVars,nSamps);
for i=1:nSamps
    u = rand(nVars,1)-0.5;
    b = -sign(u).*log(1-2*abs(u));
    S(:,i) = sqrt(2)*B\b;
end
```

4.1. 1D examples. For 1D examples, we use $\ell_1$-regularization based on the prior matrix $D$, referred to as L1D, to restore signals from the given degraded, corrupted data. To estimate prior $D$, we first generate a training set $S$ as shown in the box above, where all sample signals have $\ell_1$-Laplace distribution determined by the variance matrix $B$. Then solve the sub-gradient equation (6) using L-BFGS, resulting in an estimate of the matrix $D$. To test the quality of our prior estimation, we compare the $L1D$ regularization $R(x) = \|Dx\|_1$ with the following commonly used regularization techniques:

- $\ell_2$-regularization (L2): $R(x) = \frac{1}{2}\|\nabla x\|^2$.
- Total variation regularization (TV): $R(x) = \|\nabla x\|_1$.
- Discrete wavelet regularization (DWT): $R(x) = \|Wx\|_1$.

Here $W$ is discrete wavelet transform matrix formulated from Daubechies wavelets [13]. To generate the input data, we blur a true signal $x^*$ with a kernel function, see the left of Fig. 2. The corresponding measurement matrix $F(x) = A$ is depicted on the right of Fig. 2. We then solve the Tikhonov optimization problem (1) and reconstruct an estimate $\hat{x}$ of the true signal $x^*$.

Table 3 displays the average reconstruction errors $\|\hat{x} - x^*\|/\|x^*\|$ out of 50 tests for all four regularizations. The proposed estimated prior $D$ performs well. Surprisingly, when the sample size is much smaller than the variable size, e.g., 64 vs. 256, the prior matrix $D$ estimated with less computation effort performs generally better instead, comparing the first and second rows with the third and fourth rows in Table 3.

In the above experiment we have chosen $\mathcal{X}$ to be drawn from a Laplace distribution. However, in realistic application this may not be appropriate. Our goal in the second test is to experiment our methodology when $\mathcal{X}$ has a different distribution and thus our Laplace distribution assumption does not hold. We therefore run the same scheme on another 50 test problems, where the true signal $x^*$ is of random uniform distribution instead of Laplace distribution. Theoretically, the estimated prior should fail to be optimal, as we are dealing with a wrong distribution. However, the L1D regularization scheme actually performs reasonably well, and still retains the lowest reconstruction errors, compared to other regularization schemes considered, as listed in the second and
Figure 2. Illustration of the Gaussian point spread function and the corresponding measurement matrix $F(x) = A$ used in 1D example with 256 variables.

Table 3. Comparing the average reconstruction errors with different sample sizes (nSamps), signal distributions (Laplace, Uniform) and regularization techniques (L2, TV, DWT, L1D).

| $\|x - x^*\|/\|x^*\|$ | nSamps | nVars | L2   | TV   | DWT  | L1D  |
|-----------------|--------|-------|------|------|------|------|
| Laplace         | 64     | 256   | 9.86e-01 | 9.45e-01 | 9.28e-01 | 3.68e-02 |
| Uniform         | 64     | 256   | 8.82e-01 | 8.52e-01 | 8.42e-01 | 4.93e-01 |
| Laplace         | 256    | 256   | 9.85e-01 | 9.48e-01 | 9.27e-01 | 6.13e-02 |
| Uniform         | 256    | 256   | 8.24e-01 | 8.58e-01 | 7.85e-01 | 5.74e-01 |

fourth row of the Table 3. This demonstrates that our prior behaves rather robustly for 1D signal restoration. Note that objectively, for reconstruction problems, the average performance rather than the results for certain specific examples matters. Different regularizations can be applied for the better recovery of different models or for different applications, but the general performance, behaviors, or say the average errors, indicate the overall quality and robustness of the selected regularization.

4.2. **2D examples.** In the next set of examples we use 2D images with unknown distributions. We consider the problem of image deconvolution. The image size is 128-by-128. So the optimization problem is very large and contains $128^4$ unknowns. As the sub-gradient equation involves $D^{-1}$, we cannot solve the problem efficiently without taking further assumptions. One alternative method is to estimate $\hat{D}$ on image patches, e.g., 32-by-32, rather than over the whole images. Then get the complete $D$ by a Kronecker product

$$D = \text{kron}(I_{32}, \hat{D}) + \text{kron}(\hat{D}, I_{32}).$$

Such a patch-wise approach is justifiable, since the influence on one image pixel comes mostly from its local neighborhood, and randomly picking many large patches to form a training set avoids the possibility of missing long distance interaction of image pixels, such as eyes. Moreover, this method can provide sufficient number of samples to guarantee estimation accuracy when the available image training set is too small. Note that even on 32-by-32 image patches, $\hat{D}$ has more than one million variables, still a fairly
large problem. Our proposed algorithm works well on this case. The regularization parameter $\rho$ is selected by (7) and L-BFGS stops when the change rate of the object function is less than $10^{-4}$.

As a training set we use the MRI data set from MATLAB. This set contains 27 image slices of a human head, see the top of Fig. 3. We first estimate $\hat{D}$ with $1024^2$ variables using only 256 patches ($32 \times 32$) randomly taken from the whole set, and then use it to restore the given corrupted, noisy and blurry MRI data in the middle of Fig. 3. Our reconstruction with L1D regularization is presented at the bottom of Fig. 3. To compare, we have also run deblurring algorithms with $L_2$, TV and DWT regularizations, respectively on the same data. The average reconstruction errors are recorded in the first row of Table 4. Furthermore, to check if our proposed prior estimation scheme is robust for 2D examples, we divide the MRI data set into two subsets, one containing odd-numbered images and the other including even-numbered images. The odd set serves as a training set, on which we implement the prior estimation. Then, the even set is used to generate input data and be compared with resulting reconstruction. Our L1D regularization leads to the smallest reconstruction error on this case as well, see the second row of Table 4.

We have also conducted tests on other images derived from a more mundane source. A set of car images was generated from video sequences taken in Boston and Cambridge over the course of a couple of days in 1999 [29]. The pose of the cars in the database includes both frontal and rear views. Firstly, we randomly pick 45 car images out of 516 as a true set presented in Fig. 4(a); secondly, blur it with Gaussian kernel and add white noise to create an input data set presented in Fig. 4(b) for the following deconvolution tests. To estimate priors, we train $D$ on both true set and another 45 car images randomly selected from the rest of the database. Similarly as with the experiments on MRI, L2 and TV regularizations fail to provide reasonably good image recovery, comparing the average reconstruction errors collected in Table 4. The reconstruction with DWT regularization in Fig. 5(a) is not visually gratifying either. The noise is actually amplified during the deblurring process. On the other hand, the deconvolution with our L1D regularization in Fig. 5(b) provides rather visually satisfactory results for the entire 45 image set. The corresponding lowest average reconstruction errors in Table 4 quantitatively show the achieved improvement.

5. Conclusions. In this work we have addressed the problem of designing an $\ell_1$-norm regularization matrix. We have considered an empirical Bayesian framework which allowed us to interpret the regularization matrix as a-priori component. This formulation leads to a difficult, nonconvex and nonsmooth optimization problem. We

| $\|\hat{x} - x^*\|/\|x^*\|$ | L2 | TV | DWT | L1D |
|-----------------|-----|-----|-----|-----|
| MRI–Same        | 2.60e-01 | 2.30e-1 | 1.60e-01 | 1.03e-01 |
| MRI–Diff        | 2.62e-01 | 2.91e-01 | 2.52e-01 | 1.26e-01 |
| CAR–Same        | 1.51e-01 | 1.85e-01 | 1.12e-01 | 9.21e-02 |
| CAR–Diff        | 1.51e-01 | 1.85e-01 | 1.12e-01 | 9.30e-02 |

Table 4. Average reconstruction errors for MRI and car image sets. “Same” indicates that the training set is equal to the validation set; “Diff” refers to disjoint training and validation set.
have used the L-BFGS algorithm to solve the problem, coupled with BIC for the selection of regularization parameter.

To test our methodology we have evaluated the performance of the designed \( D \) matrices on both controlled numerical experiments as well as realistic data sets. Results demonstrate that learning the prior matrix \( D \) from available training models can substantially improve recovery of solutions to inverse problems.

Acknowledgments. Thanks go to Mark Schmidt for his Matlab code \textit{minFunc}, which has been modified to solve our unconstrained optimization problem. The car database

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure3.png}
\caption{True MRI images are on the top; corrupted data are in the middle; our reconstructions with L1D regularization are at the bottom.}
\end{figure}
Figure 4. True data set and input data set for the following deconvolution.

is downloaded from the Center for Biological and Computational Learning (CBCL) at MIT. This study was partially supported by a Design in Inversion Open Collaborative Research (OCR) program between IBM TJ Watson Research Center, the University of British Colombia and MITACS.
Figure 5. Comparing reconstructions on car images with different regularization schemes.

REFERENCES

[1] H. Akaike. A new look at the statistical model identification. *IEEE Trans. Automat. Control*, 19(6):716–723, 1974.
[2] A. Aravkin. A robust kalman smoother. *IEEE Transactions on Automatic Control*, 2011. In press.
[3] U. Ascher, E. Haber, and H. Huang. On effective methods for implicit piecewise smooth surface recovery. *SIAM J. Sci. Comput.*, 28(1):339–358, 2006.
[4] U. Ascher, K. van den Doel, H. Huang, and B. Svaiter. Gradient descent and fast artificial time integration. *M2AN*, 43:689–708.
[5] A. M. Bruckstein, D. L. Donoho, and M. Elad. From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM Review*, 51(1):34–81, 2009.
[6] E. J. Candès and M. B. Wakin. An introduction to compressive sampling. *IEEE In Signal Processing Magazine*, 25(2):21–30, 2008.
[7] T. Chan and J. Shen. *Image Processing and analysis: variational, PDE, wavelet, and stochastic methods*. SIAM, 2005.
[8] M. Cheney, D. Isaacson, and J. C. Newell. Electrical impedance tomography. *SIAM Review*, 41:85–101, 1999.
[9] J. Dahl, L. Vandenberghe, and V. Roychowdhury. Covariance selection for non-chordal graphs via chordal embedding. *Optimization Methods and Software*, 23(4):501–520, 2008.
[10] A. d’Aspremont, O. Banerjee, and L. E. Ghaoui. First-order methods for sparse covariance selection. *SIAM J. Matrix Anal. Appl.*, 30:56–66, 2008.
[11] M. Elad, P. Milanfar, and R. Rubinstein. Analysis versus synthesis in signal priors. *Inverse Problems*, 23:947–968.
[12] K. Engan, K. Skretting, and J. H. Husoy. Family of iterative ls-based dictionary learning algorithms, ils-dla, for sparse signal representation. *Digit. Signal Process.*, 17(1):32–49, 2007.
[13] 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(4):586–597, 2007.
[14] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9:432–441, 2008.
[15] A. Genkin, D. Lewis, and D. Madigan. Large-scale bayesian logistic regression for text categorization. *Technometrics*, 49(3):291–304, 2007.
[16] M. Girolami. A variational method for learning sparse and overcomplete representations. *Neural Computation*, 13(11):2517–2532, 2001.
[17] G. H. Golub, M. T. Heath, and G. Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21:215–223, 1979.
[18] E. Haber, U. Ascher, and D. Oldenburg. Inversion of 3D electromagnetic data in frequency and time domain using an inexact all-at-once approach. *Geophysics*, 69:1216–1228, 2004.
[19] E. Haber, D. Oldenburg, and A. Celler. Recovery of dynamic parameters in SPECT. *IEEE on NUC SCI*, 67:129–135, 1997.
[20] P. C. Hansen. *Rank Deficient and Ill-Posed Problems*. SIAM, Philadelphia, 1998.
[21] L. Horesh and E. Haber. Sensitivity computation of the $\ell_1$ minimization problem and its application to dictionary design of ill-posed problems. *Inverse Problems*, 25(9):095009.
[22] N. Johnson and S. Kotz. *Continuous Multivariate Distributions*. Wiley, New York, 1972.
[23] S. Kotz, T. J. Kozubowski, and K. Podgórski. *Laplace distribution and generalizations: a revisit with applications to communications, economics, engineering, and finance*. Birkhäuser, Boston, 2001.
[24] 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, 2003.
[25] M. S. Lewicki and T. J. Sejnowski. Learning overcomplete representations. *Neural Computation*, 12(2):337–365, 2000.
[26] A. S. Lewis and M. L. Overton. Nonsmooth optimization via BFGS. *SIAM J. Optimization*, 2011. In press.
[27] Z. Lu. Smooth optimization approach for sparse covariance selection. *SIAM J. Optim.*, 19:1807–1827, 2009.
[28] J. F. Murray and K. Kreutz-Delgado. Learning sparse overcomplete codes for images. *Journal of VLSI Signal Processing*, 45(1):97–110, 2006.
[29] C. Papageorgiou and T. Poggio. A trainable object detection system: Car detection in static images. Technical Report 1673, October 1999. (CBCL Memo 180).
[30] A. Pázman. *Foundations of Optimum Experimental Design*. Springer-Verlag, New York, 1986.
[31] G. Sapiro. *Geometric Partial Differential Equations and Image Analysis*. Cambridge, 2001.
[32] O. Scherzer. Scale-space methods and regularization for denoising and inverse problems. *Advances in Image and Electron Physics*, 128:445–530, 2003.
[33] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.

[34] D. F. Shanno and K. Phua. Matrix conditioning and nonlinear optimization. *Mathematical Programming*, 14:149–160.

[35] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B.*, 58(1):267–288, 1996.

[36] A. N. Tychonoff. Solution of incorrectly formulated problems and the regularization method. *Soviet Math Dokl*, 4:1035–1038, 1963.

[37] C. R. Vogel. *Computational Methods for Inverse Problem*. SIAM, Philadelphia, 2002.

[38] H. Wang, B. Li, and C. Leng. Shrinkage tuning parameter selection with a diverging number of parameters. *J. R. Statist. Soc. B*, 71(3):671–683, 2009.

E-mail address: hui.huang@siat.ac.cn
E-mail address: haber@math.ubc.ca
E-mail address: lhoresh@us.ibm.com