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 Analytics and Mathematical Sciences
IBM T J Watson Research Center
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 non-convex optimization problem. Memory requirements as well as computation of the nonlinear, non-smooth subgradient 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 $F : \mathbb{R}^m \rightarrow \mathbb{R}^n$ 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 [45, 55, 46] through a linear, constant coefficient form, such as in deblurring and tomography problems [3, 27, 26] to a non-linear (function of the model parameters $x$) operator, corresponding to the discretization of a set of partial differential equations (PDEs) involving for instance, Poisson or Maxwell’s differential equations [11, 25]. For a broad range of applications the operator $F$ is considered to be ill-posed, that is, it cannot be stably inverted in order to recover the model parameters $x$ given the observables.

2000 Mathematics Subject Classification. Primary: 49N45; Secondary: 62F15.

Key words and phrases. $\ell_1$-Regularization Prior Matrix, Empirical Bayesian Estimation, Optimal Design, Nonlinear and Nonsmooth Optimization.
y. To obtain stable solutions one often resort to the incorporation of regularization. The objective of regularization is incorporation of “appropriate” a-priori information, that restricts the solution space and thereby stabilizes the inversion process.

Probably the most common type of regularization is the Tikhonov regularization [52], in which the inversion problem is recasted as an optimization problem of the form

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

where $R : \mathbb{R}^m \rightarrow \mathbb{R}^+$ 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 [9].

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 [55] (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 [6, 7] 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 through shearlets, contourlets and bandelets to chirplets and other transformations [7, 8, 15, 33, 34, 39].

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 exists 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 (bijective) then we can rewrite the problem (1) in its synthesis form

$$\min_{z} \; T(x) \equiv \frac{1}{2} \|F(D^{-1}z) - y\|^2 + \alpha \|z\|_1,$$

whereas the formulation (1) is known as the analysis form [17, 43].

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 [35, 23, 32, 40, 18, 17] can be utilized. For problems where $F$ is an ill-posed, linear operator and the matrix $D$ is over-complete, the algorithmic formulation proposed in [28] can be employed. Nonetheless, this formulation entails a non-smooth bi-level optimization problem that is very difficult to solve.

In this research work the analysis 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. In the first subsection we shall layout the derivation of a maximum-a posteriori estimate of the model parameters $x$. Then in the subsection that follows, we shall derive an exemplar-based (regularized) maximum likelihood estimate for the matrix $D$.

2.1. Maximum A Posteriori Model Estimation. Assuming $x$ and $y$ are jointly distributed continuous random variables, the posteriori 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

$$\pi(y|x) = \frac{1}{(2\pi\sigma_\eta^2)^{n/2}} \exp \left( -\frac{\|F(x) - y\|^2}{2\sigma_\eta^2} \right).$$

The Laplace prior coincides with the analysis form of $\ell_1$-regularization. There are numerous publications in which some version of this form is utilized [51, 22]. The basic claim is that Laplace distribution does not penalize the tail of the distribution as much as say, the commonly used Gaussian distribution, thus, it serves as a natural choice for incorporation of sparseness in the problem. For this reason, we assume here that the prior is defined by the $\ell_1$-Laplace density function with zero mean [29]

$$\pi(x) = \frac{1}{\sqrt{2} \det(D^{-1})} \exp(-\sqrt{2} |Dx|_1),$$

(3)
where $D$ is invertible and the variance is $V = (D^T D)^{-1}$. The $\ell_1$-Laplace distribution defined in this way differs from the standard generalized multivariate Laplace distributions whose negative-log-likelihood are usually proportional to the $\ell_2$-norm [31]. Here, if we set $D$ to be the identity matrix, the negative-log-likelihood of (3) is indeed proportional to the $\ell_1$-norm [17]. The posterior distribution of $x$ given $y$ is

$$\pi(x | y) \propto |\det(D)| \exp \left( -\frac{\|F(x) - y\|^2}{2\sigma^2} \right) \exp(-\sqrt{2} D|x|_1).$$

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

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

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).

### 2.2. Maximum Likelihood Prior Matrix Estimation

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 $x \in X$ where $X$ 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.

The use of training data in the context of matrix design is not a new idea. It has been broadly established and matured over the years in the machine learning community [54]. Previous art involves exemplar based covariance matrices (or their inverses) design [5, 51], optimal basis selection (screening) from a given set of candidates [50, 57, 58] and data-driven design of over-complete dictionaries for sparse coding [35, 23, 32, 40, 18, 17, 38, 30, 49]. As opposed to the growing body of studies addressing data-driven over-complete dictionary design, very few studies tackled the more intricate situation where a non-trivial, potentially ill-posed observation operator is considered [28, 43]. Under these general settings, a far broader set of inverse problems can be addressed consistently. In this study, such goal-oriented approach is considered, however, since a bijective prior matrix is considered, a single level optimization can be devised and be used for either the analysis or the synthesis forms. Also, as elucidated below, implicit structure (in the form of sparsity) is also imposed upon the designed matrix, as advocated in different contexts [21, 44, 49].
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

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

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

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 [12, 14]. 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

$$\min_D -\log(L(D)) + \rho \|D\|_1 = \min_D -k \log(|\det(D)|) + \sqrt{2}\|DS\|_1 + \rho \|D\|_1. \quad (5)$$

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) [24], Akaike Information Criterion (AIC) [1] or Bayesian Information Criterion (BIC) [47]. 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 [21, 37]. 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 [42]. Since

$$\log(|\det(D)|) = \log(\det(D^TD)^{\frac{1}{2}}) = \frac{1}{2} \log(\det(D^TD)),$$

we have

$$\partial(\log(|\det(D)|)) = \frac{1}{2} \det(D^TD)^{-1} \partial(\det(D^TD)).$$

Now consider

$$\partial(\det(D^TD)) = \det(D^TD)(D^TD)^{-1} \partial(\det(D^TD)) = \det(D^TD)(D^TD)^{-1} \partial D + (D^TD)^{-1}D^TD \partial D = \det(D^TD)(D^TD)^{-1} \partial D + D^{-\top} \partial D = \det(D^TD)2D^{-\top} \partial D.$$

It yields that

$$\partial(\log(|\det(D)|)) = D^{-\top}.$$

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

$$-kD^{-\top} + \sqrt{2} \text{sign}(DS)S^\top + \rho \text{sign}(D) = 0. \quad (6)$$

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: $D \leftarrow \text{EL1RP}(S, D_0, \rho)$

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

for $i = 1$ to $N$

$D \leftarrow D_0$ and $\rho \leftarrow \rho(i)$

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)

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

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. [56] 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

$$BIC(\rho) = -\log(|\det(\hat{D})|) + \frac{\sqrt{2}}{k} \|\hat{D}S\|_1 + p \log(k) \frac{\log(k)}{k^2}, \quad (7)$$
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, the training set \( S \) is generated from sample signals with 256 variables and \( \ell_1 \)-Laplace distribution determined by the variance matrix \( B \); an optimal parameter \( \rho^* \) is then sharply obtained at 2.8. The original variance matrix \( B \) and the computed \( D(2.8) \) are spied in Fig. 2. Clearly, the estimated prior \( D \) is sparser than but has very similar structure with \( B \). Although there are some shrinkages, the scales of \( D \) and \( B \) are in the same range.

We have found that the overall BIC algorithm is rather insensitive to the choice of the threshold value \( \epsilon \) as long as it is sufficiently small. Here we used \( \epsilon = 10^{-2} \) which seemed to provide satisfactory results.

Then we consider how to efficiently and accurately solve the sub-gradient equation (6) for \( D(\rho) \), which is nonlinear and nonsmooth. In [36], 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 [48] 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 employ 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. The initial guess, \( D_0 \), is set as an identity matrix and a discrete sequence of \( \varrho = [0.1 : 0.3 : 10] \).
Figure 2. Comparing the sparsity structure of estimated prior matrix $D$ with the original matrix $B$, where the BIC selected prior parameter $\rho = 2.8$ (see Fig. 1).

<table>
<thead>
<tr>
<th>nVars</th>
<th>SD</th>
<th>Iter</th>
<th>Time</th>
<th>LSD</th>
<th>Iter</th>
<th>Time</th>
<th>PCG</th>
<th>Iter</th>
<th>Time</th>
<th>L-BFGS</th>
<th>Iter</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>64</td>
<td>3000</td>
<td>26</td>
<td>2370</td>
<td>19</td>
<td>204</td>
<td>2.8</td>
<td>145</td>
<td>1.5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>256</td>
<td>3000</td>
<td>184</td>
<td>2113</td>
<td>129</td>
<td>295</td>
<td>41</td>
<td>197</td>
<td>23</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1024</td>
<td>3000</td>
<td>4098</td>
<td>1791</td>
<td>2458</td>
<td>577</td>
<td>1868</td>
<td>295</td>
<td>621</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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.

Table 2. Comparing the average reconstruction errors with different variable sizes (nVars) and numerical solvers.

Besides L-BFGS, we have experimented with the standard steepest descent (SD) and its variant lagged steepest descent (LSD) [4] and the smooth nonlinear conjugate gradient (NLCG) [10] with diagonal preconditioner (PCG) has also been tested. The Lagged Steepest Descent (LSD) approach has been successfully implemented for general purpose problems involving a quadratic term and simple (box) bounds constraints, characterized by non-continuously differentiable objectives [13], such as compressed sensing [53] [20]. The method essentially employs a step size computed using the past rather than the present information. The smoothed NLCG, is an extension of a smoothing projected gradient approach (SPG) [59] in which a conjugate gradient of
Optimal Estimation of $\ell_1$-Regularization Prior

A smoothing function is used as a search direction. Doing so at each iteration guarantees that any accumulation point of a sequence generated by this method is a Clarke stationary point of the non-smooth and non-convex optimization problem.

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 apply the same initial guess and 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) for each testing set. 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.

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. Note that here the variable size is tunable, and all sample signals have $\ell_1$-Laplace distribution derived from the pre-defined variance matrix $B$. Then we solve the sub-gradient equation (6) using L-BFGS, resulting in an estimate of the matrix $D$, see a comparison in Fig. 2. 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): $R(x) = \|Wx\|_1$, where $W$ is discrete wavelet transform matrix formulated from Daubechies wavelets [19].

To generate the input data, we blur a true signal $x^*$ with a kernel function, see the left of Fig. 3. The corresponding measurement matrix $F(x) = A$ is depicted on the right of Fig. 3. We then solve the Tikhonov optimization problem (1) and reconstruct an estimate $\hat{x}$ of the true signal $x^*$. Note that there is another regularization parameter $\alpha$ that controls the balance between the data fidelity term and regularization term to
be determined. In general, the optimization problem (1) must be solved many times by the processes described above for different regularization parameter values $\alpha$. This contributes in a major way to the total cost of the solution process. A solution is accepted when some stopping criterion is satisfied. This is not our focus in the present article, and thus we apply a simple criterion, namely, the discrepancy principle [55], to choose the best $\alpha$ for all reconstruction examples we presented in the paper. Specifically here, we found from numerical experiments that with the same kernel function in Fig. 3, L2 and TV have the best performance with $\alpha = 1$, and DWT and L1D produce the best results with $\alpha = 0.1$ in general.

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 very 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 $X$ 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 $X$ 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 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 and the computational time is similar as the record in Table 1 with respect to $n\text{Vars} = 1024$. The corresponding prior parameter $\rho$ is automatically selected by (7) and L-BFGS stops when the change rate of the object function is less than $10^{-4}$.

For the first 2D example we use the MRI data set from MATLAB. This set contains 27 image slices of a human head, see the top of Fig. 4. We first estimate $\hat{D}$ with 1024$^2$ variables using only 256 patches (32 × 32) randomly taken from the whole set, and then use it to restore the corrupted MRI data in the middle of Fig. 4, where the applied standard deviation of Gaussian blur kernel is 2 and the added white noise level is about 5%. Our reconstruction with L1D regularization is presented at the bottom of Fig. 4. To compare, in addition to algorithms with $L_2$, TV and DWT regularizations, we have also run the deblurring with a recently proposed adaptive regularization (AR) technique [16] on the same data. Note that this approach is also training-based and derived from a pre-collected dataset of example image patches. 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. In comparison, the training-based AR method is less robust and the average reconstruction error on the set MRI-Diff is much higher than on the set MRI-Diff.
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 [41]. 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. 5(a); secondly, blur it with Gaussian kernel (the standard deviation is 3) and add 10% white noise to create an input data set presented in Fig. 5(c) 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, see Fig. 5(b). Similarly as with the experiments on MRI, L2 and TV regularizations fail to
Figure 5. True, training and corrupted input data sets for the following deconvolution.
Figure 6. Comparing reconstructions on car images with different regularization schemes.
provide reasonably good image recovery, comparing the average reconstruction errors collected in Table 4. The reconstruction with both DWT regularization and adaptive regularization in Figs. 6(a) and 6(b) are not visually gratifying either. The noise is actually amplified during the DWT deblurring process. With a smaller $\alpha$, such noise amplification could be attenuated, but the reconstructed images would remain blurry. The deblurred images with adaptive regularization are also not sharp enough and there are noticeable defects appearing on the boundary of images. On the other hand, the deconvolution with our L1D regularization in Fig. 6(c) 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 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 minFunc, which has been modified to solve our unconstrained optimization problem. We also thank the authors of [16] for making their codes available online. The car database 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.

REFERENCES


