FAST APPROXIMATE L0-NORM DECONVOLUTION USING STRUCTURED WAVELET DOMAIN PRIORS

Timothy D. Roberts, Nick Kingsbury

Signal Processing & Communication Lab
Department of Engineering
University of Cambridge, U.K.
Email: {tr331,ngk10}@cam.ac.uk

ABSTRACT

In this paper, we propose a new algorithm to solve linear inverse problems using an approximate $l_0$ penalty on overlapped groups of wavelet coefficients, and apply this to the deconvolution problem specifically. Prior work has shown the improvements gained from using group-sparse penalties over coefficient-sparse penalties for deconvolution and compressed sensing. Instead of minimizing an $l_1$-norm, as done in prior work, we instead design an overlapping group prior which utilizes the Gaussian scale mixture model, and use this to promote $l_0$ sparsity. Using a Bayesian argument, we derive a novel convex penalty function, which is a reweighted $l_2$ approximation to the $l_0$-norm that can be efficiently minimized. We show that the new group-sparse algorithm produces superior deconvolution results compared to the same algorithm utilizing an unstructured coefficient-sparse penalty.

Index Terms—structured models, deconvolution, iterative algorithms, sparsity.

1. INTRODUCTION

Deconvolution of natural images is a well-studied, ill-posed, linear inverse problem. Natural images are sparsely represented by wavelets [1], and superior deconvolution results have been obtained using this knowledge by promoting solutions with sparse wavelet-domain representations [2, 3]. Furthermore, signal models which exploit overlapped parent-child groupings of these sparse wavelet coefficients are potentially advantageous in efficiently solving deconvolution problems, and these topics are considered here.

Owing to the ability of wavelet transforms to localize edges, coefficient magnitudes exhibit strong statistical dependencies across scales [4]. Early work in exploiting this structured sparsity employed the hidden Markov model [5–7]. While this model is effective for capturing some of the non-Gaussian behavior of wavelet magnitudes and promoting the persistence of magnitudes across scales, it requires learning the dependencies using iterative reweighting schemes. This motivates locally-convex quadratic penalties which can be more efficiently solved.

Rao et al [8] have employed a convex overlapped group lasso ($l_1$) penalty and demonstrated significant improvements over coefficient-wise penalties for deconvolution and compressed sensing problems. In [9], Zhang et al have shown improved sparse recovery results over the $l_1$ sparse penalty using the modified subband-adaptive iterative shrinkage/threshold (MSIST) algorithm, which minimizes the reweighted $l_2$ approximation to the $l_0$-norm of wavelet coefficient magnitudes. For this reason, we propose a group sparse solver which incorporates an approximate $l_0$ group sparse penalty, and apply this to the deconvolution problem, building on some of the strengths in both [8] and [9].

In [8], a simple parent-child, overlapping grouping policy is used to facilitate the modeling of groups so that optimization can be performed on a pre-defined structure. Overlapping group structures are useful when the precise group structure or the degree of persistence of wavelet magnitudes across scales is not known a priori [10]. Since the problem of finding the group structure which constitutes the sparse approximation of a signal is NP hard in general [11], we adopt the overlapping group approach in which the interscale group dependencies are gradually learned, starting with a sensible pre-defined group structure.

In [12], Zhang et al have shown promising deconvolution results using the MSIST algorithm. This algorithm is derived from Bayesian principles, utilizing a Gaussian scale mixture (GSM) prior, which is a powerful model for capturing the statistics of wavelet coefficients [13–15].

Combining Bayesian wavelet-based image deconvolution using the GSM prior, quadratic penalties, and group sparsity, we propose a new locally-convex group-sparse solver. Unlike the work in [8], we promote sparse solutions by approximating the $l_0$-norm. Additionally, instead of promoting $l_0$ sparsity at the coefficient-level as in [12], we sparsify at the group level, leading to better estimates of the sparse approximation of the underlying signal. We have chosen the dual-tree complex wavelet transform (DT-CWT) as our sparsifying transformation, because it has the advantages of approximate shift-invariance and directional selectivity for a limited transform domain redundancy. By incorporating the strong parent-child dependency structure of these complex wavelet coefficient magnitudes into our prior model, we show that we can find group-sparse solutions via fast quadratic minimization.

In sec. 2.1, we introduce the GSM model and the MSIST penalty. We define the group sparse model in sec. 2.2, and extend the MSIST penalty to derive the group sparse MSIST algorithm in sec. 2.3. In sec. 3 we show that, by imposing a simple grouping policy, we can obtain superior results to those obtained without a group structure, using the same basic algorithm and parameters. In sec. 4 we conclude with suggested extensions of this research.

2. METHODS

2.1. Gaussian Scale Mixture (GSM) Model

The GSM model assumes that the statistics of an image can be described locally, so that each wavelet coefficient follows an indepen-
dent Gaussian process with spatially varying statistical parameters. Given an unobserved signal $x \in \mathbb{R}^N$ and an additive white Gaussian noise process $n \sim N(0, \sigma^2)$, we observe

$$y = \Phi x + n$$

(1)

via matrix $\Phi = HW^T$. Usually $H$ is a convolution matrix, and $W^T$ represents an inverse wavelet transform. From a Bayesian point of view, we can reconstruct the image from the observation model, or likelihood $p(y|x)$, and a GSM prior on the signal, $p(x)$:

$$p(y|x) \propto \exp \left( \frac{-\|y - \Phi x\|^2}{2\sigma^2} \right)$$

(2)

$$p(x) \propto \exp \left( -\frac{1}{2}x^T S x \right)$$

(3)

where $S$ is a diagonal matrix formed from the $j$-indexed vector $s$ of inverse variance estimates $\frac{1}{\nu}$ of each element $x_j$ in $x$.

Since we are using the DT-CWT, which transforms real image pixels into complex wavelet coefficients, the redundant tight-frame representation of the transform requires the coefficients to be split into separate real and imaginary parts in $x$, to avoid the need for the awkward “real-part” operator in the inverse transform $W^T$. This can be elegantly handled by regarding the real and imaginary components of each complex coefficient as non-overlapping groups of size 2, and we have incorporated this into our previous work on sparsity-regularized deconvolution [9, 12]. In the present context, we obtain estimates of each $v_j$ by taking the mean of the squares of the real and imaginary parts of each complex coefficient and also including $\epsilon^2$, an estimate of the variance of the noise present in the wavelet coefficients. Hence,

$$v_j = \mathbb{E} \left[ |x_j|^2 \right] \approx \frac{1}{2} \left( x_{j,\text{Re}}^2 + x_{j,\text{Im}}^2 \right) + \epsilon^2$$

(4)

The maximum a posteriori (MAP) estimate $x_{\text{MAP}}$ obtained by maximizing the posterior $p(x|y)$, or by minimizing its negative logarithm, is given by:

$$x_{\text{MAP}} = \arg\min_x \left[ -\ln(p(y|x)p(x)) \right]$$

$$= \arg\min_x \left[ \|\Phi x - y\|^2 + \lambda^2 x^T S x \right]$$

(5)

The expression in eq. (5) can be minimized using fast threshold Landweber iterations [3]. This is the MSIST algorithm, which minimizes the data fidelity term while promoting sparsity in an approximate $l_0$ sense, if the terms in $S$ are equal to the reciprocal of the expected variance of each term in $x$. We now consider how a group-sparse penalty may be incorporated into MSIST, and introduce a separate model for the prior in eq. (3) which incorporates group structure, producing more robust estimates of the $v_j$ in eq. (4).

### 2.2. Overlapping Group Model

In this section, we introduce the notation and terminology for the overlapping group sparse model. In this setup, $x_k$ can belong to two or more separate groups, but a problem exists when we wish to use the statistics of those groups to update the current estimate of $x_k$. As in [8], a strategy we use for dealing with overlapping groups is to introduce a new vector $\hat{x} \in \mathbb{R}^M$ of latent variables comprising multiple copies of each $x_k$. An advantage of this duplication strategy is that it does not require duplicating the columns of $\Phi$ in eq. (1) for each of the latent variables, and thus minimizes computational complexity. Generally $M \geq N$, with equality for the special case when each coefficient comprises a non-overlapping group by itself. The size $M$ of the latent variable vector depends on the grouping policy.

We may also consider clustering elements of $\hat{x}$ which correspond to each element $x_k$ of $x$. We define the following terms:

- **Group** refers to a subvector $x_{G_k}$ of $x$, or a subvector $\hat{x}_{G_k}$ of $\hat{x}$, indexed by the sets of indices $G_k \subseteq \{1, \ldots, N\}$ and $\hat{G}_k \subseteq \{1, \ldots, M\}$ for $k \in \{1, \ldots, N_x\}$, where $N_x$ is the number of groups. The groups $G_k$ may be overlapping, whereas the groups $\hat{G}_k$ may not. Two groups from the parent+children overlapping group policy are shown in fig. 1 using red and blue undirected edges.

- **Cluster** refers to a subvector $\hat{x}_{G_k}$ of $\hat{x}$, indexed by the set of indices $G_k \subseteq \{1, \ldots, M\}$ for $k \in \{1, \ldots, N_x\}$ such that all elements of $\hat{x}_{G_k}$ correspond to the multiple copies of a single element $x_k$ of $x$. One cluster of size 2 for the overlapped variable $x_2$ is shown in fig. 1 using arrows.

### 2.3. Group Sparse MSIST Algorithm

In this section, we consider sparsifying at the group level while imposing a cluster fidelity constraint in order to keep the master close to the mean of its duplicate elements.

First, adopting a Bayesian interpretation, we estimate $x$ and $\hat{x}$ jointly, maximizing the posterior $p(x, \hat{x}|y)$.

$$[x, \hat{x}]_{\text{MAP}} = \arg\min_{x, \hat{x}} \left[ -\ln(p(y|x, \hat{x})p(x, \hat{x})) \right]$$

(6)

For all $k \in \{1, \ldots, N_x\}$, we assume the conditional distribution of each $x_k$ is normal, with mean $\bar{x}_k$ and variance $v_{G_k}$.

$$p(x_k|G_k) = \frac{1}{\sqrt{2\pi v_{G_k}}} \exp \left( -\frac{(x_k - \bar{x}_k)^2}{2v_{G_k}} \right)$$

(7)
where $\bar{x}_k$, the cluster mean, and $v_{G,k}$, the cluster variance, are estimated from $\hat{\mathbf{x}}_{G|k}$ by

$$
\bar{x}_k = \frac{1}{|G_k|} \sum_{c \in G_k} \hat{x}_c
$$

(8)

$$
v_{G,k} = \frac{1}{|G_k|} \sum_{c \in G_k} |\bar{x}_c - \bar{x}_k|^2
$$

(9)

Under an approximate group-independence assumption, we can obtain the group prior:

$$
p(\hat{\mathbf{x}}) \approx \prod_{i \in \{1 \ldots N_g\}} p(\hat{\mathbf{x}}_{g_i}) \propto \prod_{i \in \{1 \ldots N_g\}} \prod_{k \in G_i} \exp\left(-\frac{x_{g_i,k}^2}{2v_{g_i,k}}\right)
$$

(10)

where $v_{g,i}$ is estimated from $\hat{x}_{g_i}$. We also assume an independent prior $p(v_{g,i}|\nu) \propto \exp(-\nu^2/2v_{g,i})$ for each $v_{g,i}$, as in [9], so that

$$
v_{g,i} = 1/2|g_i| \|\hat{x}_{g_i}\|^2 + \nu^2
$$

(11)

Next, we form diagonal matrices $\hat{\mathbf{S}}(N \times N)$ and $\hat{\mathbf{S}}(M \times M)$ from vectors $\hat{\mathbf{s}}$ and $\hat{\mathbf{s}}$. The entries $\hat{s}_n$ of $N$-length vectors $\hat{\mathbf{s}}$ and $\hat{\mathbf{s}}$ are the cluster inverse variance estimates $v_{G,k}$, and the entries $\hat{s}_n$ of $M$-length vector $\hat{\mathbf{s}}$ are the group inverse variance estimates $v_{g,i}$. When $N_g < M$, $\hat{\mathbf{S}}$ is formed from an expanded version of $\hat{\mathbf{s}}$ so that $\hat{s}_{ij} = \hat{s}_i$ for each $j \in g$. Thus we have the following penalty function derived from eq. (6):

$$
J(\mathbf{x}, \hat{\mathbf{x}}, \hat{\mathbf{S}}, \bar{\mathbf{x}}, \hat{\mathbf{S}}) = \|\mathbf{x} - \bar{\mathbf{x}}\|^2 + \nu^2\left[(\mathbf{x} - \bar{\mathbf{x}})^T \hat{\mathbf{S}}(\mathbf{x} - \bar{\mathbf{x}}) - \text{ln}(\|\hat{\mathbf{S}}\|)\right]
$$

$$
+ \nu^2\left[\mathbf{x}^T \hat{\mathbf{S}} \mathbf{x} - \text{ln}(\|\hat{\mathbf{S}}\|) + \nu^2 \text{tr}(\hat{\mathbf{S}})\right]
$$

$$
+ (\mathbf{x} - \bar{\mathbf{x}})^T \hat{\mathbf{A}}_n (\mathbf{x} - \bar{\mathbf{x}}) - \|\hat{\mathbf{S}}(\mathbf{x} - \bar{\mathbf{x}})\|^2
$$

(12)

The penalty function in eq. (6) includes the cluster fidelity constraint (second line) and group sparsity penalty (third line). In practice we estimate $\bar{\mathbf{x}}$ and $\hat{\mathbf{S}}$ from $\hat{\mathbf{x}}$, but for generality eq. (12) is left as a function of all three quantities. The Majorization-Minimization (MM) terms (fourth line) are not part of the Bayesian derivation, but serve to remove the difficult cross-coupling term $\mathbf{x}^T \hat{\mathbf{S}} \mathbf{F} \mathbf{x}$ which results from expanding the first line. The MM term is the first line, and causes $J(\cdot)$ to upper-bound the MAP penalty (the third line), as long as the diagonal matrix $\hat{\mathbf{A}}_n$ is chosen such that $(\mathbf{A}_n - \mathbf{F}^T \hat{\mathbf{F}})$ is positive definite [3,12].

In order to minimize eq. (12) we consider estimating the quantities $\bar{\mathbf{x}}$ and $\hat{\mathbf{S}}$ in terms of $\hat{\mathbf{x}}$. The vector of cluster means $\bar{\mathbf{x}} \in \mathbb{R}^{N_g}$ can be written using an explicit matrix form:

$$
\bar{\mathbf{x}} = \Lambda_{G|k} \mathbf{D} \hat{\mathbf{x}} = \mathbf{A} \hat{\mathbf{x}}
$$

(13)

where $\mathbf{D}$ is an $N \times M$ matrix which performs the sum in eq. (8), whose $(k,m)^{th}$ element is given by

$$
D_{k,m} = \begin{cases} 1, & \text{if } \bar{x}_m \in C_{G_k} \\ 0, & \text{otherwise} \end{cases}
$$

(14)

and $\Lambda_{G|k}$ is an $N \times N$ diagonal matrix of $v_{G,k}^{-1}$. In practice we can achieve a similar effect of reducing the second line in eq. (12) by minimizing the $l_2$ norm using a user-defined Lagrangian parameter $\tau^2$. In this way, we can control the relative strength of the closeness of $x_k$ to the mean of $\bar{x}_{G_k}$ using $\tau^2$ while encouraging sparseness at the group level using $\nu^2$. In practice, we find that using $\tau^2$ instead of $\nu^2$ produces better deconvolution results. Additionally, this simplification avoids computing eq. (9) on each iteration. Substituting $\tau^2 \mathbf{I}$ for $\nu^2 \hat{\mathbf{S}}$ in eq. (12), we have

$$
J(\mathbf{x}, \mathbf{z}, \hat{\mathbf{x}}, \hat{\mathbf{S}}, \hat{\mathbf{S}}) = \|\mathbf{x} - \bar{\mathbf{x}}\|^2 + \tau^2 \|\mathbf{x} - \bar{\mathbf{x}}\|^2 + \text{const}
$$

$$
+ \nu^2 \left[(\hat{\mathbf{x}}^T \hat{\mathbf{S}} \hat{\mathbf{x}} - \text{ln}(\|\hat{\mathbf{S}}\|) + \nu^2 \text{tr}(\hat{\mathbf{S}})\right]
$$

$$
+ (\mathbf{x} - \bar{\mathbf{x}})^T \mathbf{A}_n (\mathbf{x} - \bar{\mathbf{x}}) - \|\hat{\mathbf{S}}(\mathbf{x} - \bar{\mathbf{x}})\|^2
$$

(15)

Using eq. (13) where appropriate, we derive update rules to minimize eq. (15) at each iteration $n$ by differentiating $J(\cdot)$ with respect to each variable in turn and setting the derivatives to 0:

$$
\frac{\partial J(\cdot)}{\partial \mathbf{x}} = 2\left[-\mathbf{F}^T \mathbf{F} \mathbf{z} + \mathbf{F}^T \mathbf{A} \mathbf{x} + \nu^2 \hat{\mathbf{x}}\right]
$$

$$
x_{n+1} = (\mathbf{A} + \nu^2 \mathbf{I})^{-1} \left[\mathbf{F}^T \mathbf{F} \mathbf{z} + (\mathbf{A} - \nu^2 \mathbf{I})\mathbf{z}_n + \tau^2 \mathbf{x}_n\right]
$$

$$
\frac{\partial J(\cdot)}{\partial \mathbf{z}} = 2(\mathbf{A} - \nu^2 \mathbf{I})\mathbf{z}_n
$$

$$
\frac{\partial J(\cdot)}{\partial \hat{\mathbf{x}}_{g,i}} = \frac{1}{v_{g,i}} - \frac{1}{\nu^2}
$$

$$
\mathbf{S}_{k,n+1} = \frac{1}{v_{g,i}} \nu^2 \hat{\mathbf{S}}_{k,n}, \ \forall i \in \{1 \ldots N_g\}
$$

(19)

Let $\hat{\mathbf{S}}$ be the $M \times M$ block-diagonal matrix $\tau^2 \mathbf{A}^T \mathbf{A} + \nu^2 \hat{\mathbf{S}}_n$ in eq. (18). The $(p, q)^{th}$ entry of each of $N$ $|G_k| \times |G_k|$ blocks, $\hat{\mathbf{S}}_k$, of $\hat{\mathbf{S}}$ is given by

$$
\hat{\mathbf{S}}_{k,(p,q)} = \begin{cases} \tau^2_{k} \hat{\mathbf{S}}_{k}^{-1} \mathbf{A}^T \mathbf{x}_n, & \text{if } p = q \\
\frac{\tau^2_{k} \hat{\mathbf{S}}_{k}^{-1} \mathbf{A}^T \mathbf{x}_n}{\sqrt{\nu^2}}, & \text{otherwise} \end{cases}
$$

where $\hat{s}_{k,(p,q)}$ is the group inverse variance estimate corresponding to the $(p,q)^{th}$ element of $\hat{\mathbf{S}}_{G_k}$. Thus, the update of $\hat{\mathbf{x}}$ in eq. (18) can be accomplished via

$$
\hat{\mathbf{x}}_{k,n+1} = \tau^2 \hat{\mathbf{S}}_{k(n,q)}^{-1} \mathbf{A}^T \mathbf{x}_n
$$

(21)

which requires small matrix inverse operations and $M$ vector-multiply operations.

### 3. RESULTS

In this section, we test our claim that the algorithm which incorporates the group sparse prior in eq. (10) and minimizes the penalty in eq. (15), referred to as Group MSIST (MG), outperforms the coefficient-sparse MSIST algorithm, referred to as Coefficient MSIST (MC), which solves eq. (5). We have used the parent+4children grouping, as in fig. 1, for these experiments.

Using the standard Cameraman test image, we have simulated various BSNR levels and blurring operations and compared with
previously reported MSIST deconvolution results [12]. For these tests, we have used the 9×9 uniform and the 15×15 Gaussian (σ = 1) blur kernels for H and the DT-CWT as our redundant sparsifying transform W to produce Φ, and added Gaussian noise n to produce the observation y, as in eq. (1). Additionally, we’ve used the same values for Λ as reported in [12] for the 9 × 9 uniform blur test case. By varying the nominal noise variance σ, we’ve produced several blurred signal-to-noise ratio (BSNR) test scenarios. BSNR is computed using

\[
\text{BSNR} = 10 \log_{10} \frac{||H_{u_r} - H_{u_c}||^2}{N_p \sigma^2}
\] (22)

where \(u_r = W^T x \in \mathbb{R}^{N_p}\) is the original ground truth signal before applying the over-complete transform W and \(H_{u_r}\) is the mean of \(H_{u_c}\).

We’ve seeded both algorithms with \(x_0\) obtained by using the regularized Wiener filter:

\[
x_0 = W \left[ (H^T H + 10^{-3} \sigma^2 I)^{-1} H^T y \right]
\] (23)

and on each iteration, we’ve computed the improvement in signal-to-noise ratio (ISNR) via

\[
\text{ISNR}_n = 10 \log_{10} \frac{||y - u_r||^2}{||W^T x_n - u_r||^2}
\] (24)

and use this to compare the performance of our algorithm to the MC benchmark.

For almost all test cases, incorporating the group sparse penalty leads to improved deconvolution results in terms of visual quality and final ISNR. Fig. 3 shows the final visual result obtained from applying the MC and MG algorithms for 200 iterations. In Fig. 2 and tables 1 and 2, we show average ISNR values at iteration n obtained from repeating our experiments over 30 noise realizations. We’ve ensured that the parameters ν and ϵ for the coefficient MSIST and group MSIST experiments were the same for each test scenario. In particular, we’ve employed geometric homotopy continuation rules for these parameters with a shrinkage factor of 0.8 per iteration, so that the reweighted \(l_0\) penalty \(\gamma^T \mathbf{S}\mathbf{x}\) more closely approximates the \(l_0\)-norm as ϵ is decreased, which is explained in detail in [12]. For all of these experiments, we have used τ = 0.1.

While this algorithm generally requires \(N\) block inverses, with this chosen grouping policy the maximum block size is \(2 \times 2\) (\(G_x \leq 2\)), so that the matrix inverse has an explicit form which permits efficient vector multiply operations. Thus, the per-iteration computational overhead is not significantly greater for MG than MC.

4. CONCLUSION

In this paper, we’ve shown improved deconvolution results by incorporating group sparsity using a Bayesian argument. These results come at the added per-iteration expense of doing \(N\) small inverses as part of the variable duplication approach. However, when small cluster sizes are employed, this expense can be minimized. Future work may include faster approximation of these inverses, or employing a strategy which simply optimizes the coefficient-wise variances in the standard GSM model using group information, potentially avoiding the need for latent variables. Using other conjugate priors, such as the Gamma prior, and variational Bayesian methods for estimating the wavelet coefficient densities, may lead to additional gains. The ISNR results for this and other sparsity-based methods do not compare favorably with the state-of-the-art results of BM3D [16]. However, relatively efficient algorithms which do not rely on block matching and also leverage the strengths of statistical modelling and wavelet domain thresholding, such as SURE-LET [17], show promise for general linear inverse problems and tractable extensions to three-dimensional deconvolution.
5. REFERENCES


