Multi-Task Compressive Sensing with Dirichlet Process Priors

Yuting Qi YUTING@EE.DUKE.EDU
Dehong Liu LIUDH@EE.DUKE.EDU
Department of Electrical and Computer Engineering, Duke University, Durham, NC, 27708 USA

David Dunson DUNSON@STAT.DUKE.EDU
Department of Statistical Science, Duke University, Durham, NC, 27708 USA

Lawrence Carin LCARIN@EE.DUKE.EDU
Department of Electrical and Computer Engineering, Duke University, Durham, NC, 27708 USA

Abstract

Compressive sensing (CS) is an emerging £eld that, under appropriate conditions, can signi£cantly reduce the number of measurements required for a given signal. In many applications, one is interested in multiple signals that may be measured in multiple CS-type measurements, where here each signal corresponds to a sensing "task". In this paper we propose a novel multi-task compressive sensing framework based on a Bayesian formalism, where a Dirichlet process (DP) prior is employed, yielding a principled means of simultaneously inferring the appropriate sharing mechanisms as well as CS inversion for each task. A variational Bayesian (VB) inference algorithm is employed to estimate the full posterior on the model parameters.

1. Introduction

Over the last two decades researchers have considered sparse signal representations in terms of orthonormal basis functions (*e.g.*, the wavelet transform). For example, consider an m-dimensional real-valued signal $\boldsymbol{u}$ and assume an $m \times m$ orthonormal basis matrix $\boldsymbol{\Psi}$; we may then express $\boldsymbol{u} = \boldsymbol{\Psi}\boldsymbol{\theta}$, where $\boldsymbol{\theta}$ is an m-dimensional column vector of weighting coef£cients. For most natural signals there exists an orthonormal basis $\boldsymbol{\Psi}$ such that $\boldsymbol{\theta}$ is sparse. Consider now an approximation to $\boldsymbol{u}$, $\hat{\boldsymbol{u}} = \boldsymbol{\Psi}\hat{\boldsymbol{\theta}}$, where $\hat{\boldsymbol{\theta}}$ approximates $\boldsymbol{\theta}$ by retaining the largest N coef£cients and setting the remaining $m - N$ coef£cients to zero; due to the aforementioned sparseness properties, $||\boldsymbol{u} - \hat{\boldsymbol{u}}||^2$ is typically very small even for $N \ll m$. Conventional techniques require one to measure the m-dimensional signal $\boldsymbol{u}$ but £nally discard $m - N$ coef£cients (Charilaos, 1999). This sample-then-compress framework is often wasteful since the signal acquisition is potentially expensive, and only a small amount of data N is eventually required for the accurate approximation $\hat{\boldsymbol{u}}$. One may therefore consider the following fundamental question: Is it possible to directly measure the informative part of the signal? Recent research in the £eld of compressive sensing shows that this is indeed possible (Candes, 2006)(Donoho, 2006).

Exploiting the same sparseness properties of $\boldsymbol{u}$ employed in transform coding ($\boldsymbol{u} = \boldsymbol{\Psi}\boldsymbol{\theta}$ with $\boldsymbol{\theta}$ sparse), in compressive sensing one measures $\boldsymbol{v} = \boldsymbol{\Phi}\boldsymbol{\theta}$, where $\boldsymbol{v}$ is an n-dimensional vector with $n < m$, and $\boldsymbol{\Phi}$ is the $n \times m$ sensing matrix. There are several ways in which $\boldsymbol{\Phi}$ may be constituted, with the reader referred to (Donoho, 2006) for details. In most cases $\boldsymbol{\Phi}$ is represented as $\boldsymbol{\Phi} = \mathbf{T}\boldsymbol{\Psi}$, where $\mathbf{T}$ is an $n \times m$ matrix with components constituted randomly (Tsaig & Donoho, 2006); hence, the CS measurements correspond to projections of $\boldsymbol{u}$ with the rows of $\mathbf{T}$: $\boldsymbol{v} = \mathbf{T}\boldsymbol{u} = \mathbf{T}\boldsymbol{\Psi}\boldsymbol{\theta} = \boldsymbol{\Phi}\boldsymbol{\theta}$, which is an underdetermined problem. Assuming the signal $\boldsymbol{u}$ is N-sparse in $\boldsymbol{\Psi}$, implying that the coef£cients $\boldsymbol{\theta}$ only have N nonzero values (Candes, 2006) (Donoho, 2006), Candès, Romberg and Tao in (Candes et al., 2006) show that, with overwhelming probability, $\boldsymbol{\theta}$ (and hence $\boldsymbol{u}$) is recovered via

$$\min ||\boldsymbol{\theta}||_{l_1}, \quad \text{s.t.}, \quad \boldsymbol{v} = \boldsymbol{\Phi}\boldsymbol{\theta}, \tag{1}$$

if the number of CS measurements $n > C \cdot N \cdot \log m$ (C is a small constant); if N is small (*i.e.*, if u is highly compressible in the basis $\boldsymbol{\Psi}$) then $n \ll m$. In practice the signal $\boldsymbol{u}$ is not exactly sparse, but a large number of coef£cients in the basis $\boldsymbol{\Psi}$ may be discarded with minimal error in reconstructing $\boldsymbol{u}$; in this practical case the CS framework has also been shown to operate effectively.

Appearing in *Proceedings of the 25^{th} International Conference on Machine Learning*, Helsinki, Finland, 2008.

The problem in (1) may be solved by linear programming (S. Chen & Saunders, 1999) and greedy algorithms (Tropp & Gilbert, 2005) (Donoho et al., 2006). A Bayesian compressive sensing (BCS) methodology is proposed in (Ji et al., 2007b), by posing the CS inversion problem as a linear-regression problem with a sparseness prior on the regression weights $\boldsymbol{\theta}$. One advantage of BCS is that this framework may be extended to multi-task compressive sensing (Ji et al., 2007a), in which each CS measurement $\boldsymbol{v}_i = \boldsymbol{\Phi}_i \boldsymbol{\theta}_i$ represents a sensing "task" and the objective is to jointly invert for all $\{\boldsymbol{\theta}_i\}_{i=1,M}$, through an appropriate sharing of information between the M data collections. In multi-task CS, one may potentially reduce the number of measurements required for each task by exploiting the statistical relationships among the tasks, for example, "Distributed Compressed Sensing" (DCS) (Baron et al., 2005), an empirical Bayesian strategy "Simultaneous Sparse Approximation" in (Wipf & Rao, 2007), and a hierarchical Bayesian model for multi-task CS (Ji et al., 2007a). However, these multi-task algorithms assume all tasks are appropriate for sharing, which may not be true in many practical applications. In this paper we introduce a Dirichlet process (DP) prior (West et al., 1994) to the hierarchical BCS model, which can simultaneously perform the inversion of the underlying signals *and* infer the appropriate sharing/clustering structure across the M tasks.

As detailed below, an important property of DP for the work presented here is that it provides a tool for semi-parametric clustering (*i.e.*, the number of clusters need not be set in advance). The DP-based hierarchical model is employed to realize the desired property of simultaneously clustering and CS inversion of the M measurements $\{\boldsymbol{v}_i\}_{i=1,M}$. A variational Bayes (Blei & Jordan, 2004) inference algorithm is considered, yielding a full posterior over the model parameters $\boldsymbol{\theta}_i$.

2. Multi-Task CS Modeling with DP Priors

2.1. Multi-Task CS Formulation for Global Sharing

Let $\boldsymbol{v}_i$ represent the CS measurements associated with task i, and assume a total of M tasks. The i-th CS measurement may be represented as

$$\boldsymbol{v}_i = \boldsymbol{\Phi}_i \boldsymbol{\theta}_i + \boldsymbol{\epsilon}_i, \tag{2}$$

where the CS measurements $\boldsymbol{v}_i$ are characterized by an n_i-dimensional real vector, the sensing matrix $\boldsymbol{\Phi}_i$ corresponding to task i is of size $n_i \times m$, and $\boldsymbol{\theta}_i$ is the set of (sparse) transform coef£cients associated with task i. The j^{th} coef£cient of $\boldsymbol{\theta}_i$ is denoted $\theta_{i,j}$. The residual error vector $\boldsymbol{\epsilon}_i \in \mathbb{R}^{n_i}$ is modeled as n_i *i.i.d.* draws from a zero-mean Gaussian distribution with an unknown precision α_0 (variance $1/\alpha_0$); the residual corresponds to the error imposed by setting the small transform coef£cients exactly to zero when performing the CS inversion.

We impose a hierarchical sparseness prior on the parameters $\boldsymbol{\theta}_i$, the lower level of which is

$$p(\boldsymbol{\theta}_i|\boldsymbol{\alpha}_i) = \prod_{j=1}^{m} \mathcal{N}(\theta_{i,j}|0, \alpha_{i,j}^{-1}), \tag{3}$$

where $\alpha_{i,j}$ is the j^{th} component of the vector $\boldsymbol{\alpha}_i$. To impose sparseness, on a layer above a Gamma hyperprior is employed independently on the precisions $\alpha_{i,j}$. The likelihood function for the parameters $\boldsymbol{\theta}_i$ and α_0, given the CS measurements $\boldsymbol{v}_i$, may be expressed as

$$p(\boldsymbol{v}_i|\boldsymbol{\theta}_i, \alpha_0) = (\frac{2\pi}{\alpha_0})^{-\frac{n_i}{2}} \exp(-\frac{\alpha_0}{2}\|\boldsymbol{v}_i - \boldsymbol{\Phi}_i\boldsymbol{\theta}_i\|_2^2). \tag{4}$$

Concerning the aforementioned hyperprior, for the multi-task CS model proposed in (Ji et al., 2007a), the parameters $\boldsymbol{\alpha}_i = \boldsymbol{\alpha}$, for $i = 1, \cdots, M$, and $\boldsymbol{\alpha} \sim \prod_{j=1}^{m} Ga(c, d)$. In this framework the CS data from all M tasks are used to jointly infer the hyper-parameters $\boldsymbol{\alpha}$ (global processing). However, the assumption in such a setting is that it is appropriate to employ all of the M tasks jointly to infer the hyper-parameters. One may envision problems for which the M tasks may be clustered into several sets of tasks (with the union of these sets constituting the M tasks), and data sharing may only be appropriate within each cluster. Through use of the Dirichlet process (DP) (Escobar & West, 1995) employed as the prior over $\boldsymbol{\alpha}_i$, we simultaneously cluster the multi-task CS data, and within each cluster the CS inversion is performed jointly. Consequently, we no longer need assume that all CS data from the M tasks are appropriate for sharing.

2.2. Dirichlet Process for Clustered Sharing

The Dirichlet process, denoted as $DP(\lambda, G_0)$, is a measure on measures, and is parameterized by a positive scaling parameter λ and the base distribution G_0. Assume we have $\{\boldsymbol{\alpha}_i\}_{i=1,M}$ and each $\boldsymbol{\alpha}_i$ is drawn identically from G, and G itself is a random measure drawn from a Dirichlet process,

$$\begin{aligned} \boldsymbol{\alpha}_i|G &\overset{iid}{\sim} G, \quad i = 1, \cdots, M, \\ G &\sim DP(\lambda, G_0), \end{aligned} \tag{5}$$

where G_0 is a non-atomic base measure.

Sethuraman (Sethuraman, 1994) provides an explicit characterization of G in terms of a stick-breaking construction. Consider two in£nite collections of independent random variables π_k and $\boldsymbol{\alpha}_k^*$, $k = 1, 2, \cdots, \infty$, where the π_k are drawn *i.i.d.* from a Beta distribution, denoted $Beta(1, \lambda)$, and the $\boldsymbol{\alpha}_k^*$ are drawn *i.i.d.* from the base distribution G_0.

The stick-breaking representation of G is then defined as

$$G = \sum_{k=1}^{\infty} w_k \delta_{\boldsymbol{\alpha}_k^*}, \quad \text{with} \tag{6}$$

$$w_k = \pi_k \prod_{i=1}^{k-1}(1-\pi_i), \tag{7}$$

where $\pi_k|\lambda \overset{iid}{\sim} Beta(1,\lambda)$ and $\boldsymbol{\alpha}_k^*|G_0 \overset{iid}{\sim} G_0$. This representation makes explicit that the random measure G is discrete with probability one and the support of G consists of an infinite set of atoms located at $\boldsymbol{\alpha}_k^*$, drawn independently from G_0. The mixing weights w_k for atom $\boldsymbol{\alpha}_k^*$ are given by successively breaking a unit length "stick" into an infinite number of pieces, with $0 \leq w_k \leq 1$ and $\sum_{k=1}^{\infty} w_k = 1$.

2.3. Multi-Task CS with DP Priors

We employ a DP prior with stick-breaking representation for $\boldsymbol{\alpha}_i$ in the model in (3), which assumes that $\boldsymbol{\alpha}_i|G \sim G$ and $G = \sum_{k=1}^{\infty} w_k \delta_{\boldsymbol{\alpha}_k^*}$. The base distribution G_0 corresponds to the sparseness promoting representation discussed in Sec 2.1. To facilitate posterior computation we introduce an indicator variable z_i with $z_i = k$ indicating $\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_k^*$. Therefore the DP multi-task CS model is expressed as

$$\begin{aligned}
\boldsymbol{v}_i|\boldsymbol{\theta}_i, \alpha_0 &\sim \mathcal{N}(\boldsymbol{\Phi}_i\boldsymbol{\theta}_i, \alpha_0^{-1} I), \\
\theta_{i,j}|z_i, \{\boldsymbol{\alpha}_k^*\}_{k=1,K} &\sim \mathcal{N}(0, {\alpha_{z_i,j}^*}^{-1}), \\
z_i|\{w_k\}_{k=1,K} &\overset{iid}{\sim} Multinomial(\{w_k\}_{k=1,K}), \\
w_k &= \pi_k \prod_{l=1}^{k-1}(1-\pi_l), \\
\pi_k &\overset{iid}{\sim} Beta(1,\lambda), \\
\lambda|e,f &\sim Ga(e,f), \\
\boldsymbol{\alpha}_k^*|c,d &\overset{iid}{\sim} \prod_{j=1}^{m} Ga(c,d), \\
\alpha_0 &\sim Ga(a,b),
\end{aligned} \tag{8}$$

where $i = 1, \cdots, M$, $j = 1, \cdots, m$, $k = 1, \cdots, K$, $1 \leq K \leq \infty$, and $\alpha_{i,j}$ is the j-th element of $\boldsymbol{\alpha}_i$. For convenience, we denote the model in (8) as DP-MT CS. In practice K is chosen as a relatively large integer (*e.g.*, $K = M$ if M is relatively large) which yields a negligible difference compared to the true DP (Ishwaran & James, 2001), while making the computation practical.

The choice of G_0 here is consistent with the sparseness-promoting hierarchical prior discussed in Section II-A. Consider task i and assume $\boldsymbol{\alpha}_i$ takes value $\boldsymbol{\alpha}_k^*$; the prior distribution over $\boldsymbol{\theta}_i$ is then

$$p(\boldsymbol{\theta}_i|c,d) = \prod_{j=1}^{m} \int \mathcal{N}(\theta_{i,j}|0, {\alpha_{k,j}^*}^{-1}) Ga(\alpha_{k,j}^*|c,d) d\alpha_{k,j}^*. \tag{9}$$

Equation (9) is a type of automatic relevance determination (ARD) prior which enforces the sparsity over $\boldsymbol{\theta}_i$ (Tipping, 2001). We usually set c and d very close to zero (*e.g.*, 10^{-4}) to make a broad prior over $\boldsymbol{\alpha}_k^*$, which allows the posteriors on many of the elements of $\boldsymbol{\alpha}_k^*$ to concentrate at very large values, consequently the posteriors on the associated elements of $\boldsymbol{\theta}_i$ concentrate at zero, and therefore the sparseness of $\boldsymbol{\theta}_i$ is achieved (MacKay, 1994) (Neal, 1996). Since these posteriors have "heavy tails" compared to a Gaussian distribution, they allow for more robust shrinkage and borrowing of information. Similarly, hyper-parameters a, b, e, and f are all set to a small value to have a non-informative prior over α_0 and λ respectively.

3. Variational Bayesian Inference

One may perform inference via MCMC (Gilks et al., 1996), however this requires vast computational resources and MCMC convergence is often difficult to diagnose (Gilks et al., 1996). Variational Bayes inference is therefore introduced as a relatively efficient method for approximating the posterior. From Bayes' rule, we have

$$p(\mathbf{H}|\mathbf{V}, \Upsilon) = \frac{p(\mathbf{V}|\mathbf{H})p(\mathbf{H}|\Upsilon)}{\int p(\mathbf{V}|\mathbf{H})p(\mathbf{H}|\Upsilon)d\mathbf{H}}, \tag{10}$$

where $\mathbf{V} = \{\boldsymbol{v}_i\}_{i=1,M}$ are CS measurements from M CS tasks, $\mathbf{H} = \{\alpha_0, \lambda, \boldsymbol{\pi}, \{z_i\}_{i=1,M}, \{\boldsymbol{\theta}_i\}_{i=1,M}, \{\boldsymbol{\alpha}_k^*\}_{k=1,K}\}$ are hidden variables (with $\boldsymbol{\pi} = \{\pi_k\}_{k=1,K}$) and $\Upsilon = \{a,b,c,d,e,f\}$ are known hyper-parameters. The integration in the denominator of (10), called the *marginal likelihood*, or *"evidence"* (Beal, 2003), is generally intractable to compute analytically. Instead of directly estimating $p(\mathbf{H}|\mathbf{V}, \Upsilon)$, variational methods seek a distribution $q(\mathbf{H})$ to approximate the true posterior distribution $p(\mathbf{H}|\mathbf{V}, \Upsilon)$. Consider the log marginal likelihood

$$\log p(\mathbf{V}|\Upsilon) = \mathcal{F}(q(\mathbf{H})) + \mathcal{D}_{KL}(q(\mathbf{H})||p(\mathbf{H}|\mathbf{V}, \Upsilon)), \tag{11}$$

where

$$\mathcal{F}(q(\mathbf{H})) = \int q(\mathbf{H}) \log \frac{p(\mathbf{V}|\mathbf{H}|, \Upsilon)p(\mathbf{H}, \Upsilon)}{q(\mathbf{H})} d\mathbf{H}, \tag{12}$$

and $\mathcal{D}_{KL}(q(\mathbf{H})||p(\mathbf{H}|\mathbf{V}, \Upsilon))$ is the KL divergence between $q(\mathbf{H})$ and $p(\mathbf{H}|\mathbf{V}, \Upsilon)$. The approximation of $p(\mathbf{H}|\mathbf{V}, \Upsilon)$ using $q(\mathbf{H})$ can be achieved by maximizing $\mathcal{F}(q(\mathbf{H}))$, which forms a strict lower bound on $\log p(\mathbf{V}|\Upsilon)$. In this way estimation of $q(\mathbf{H})$ may be made computationally tractable. In particular, for computational convenience,

$q(\mathbf{H})$ is expressed in a factorized form, with the same functional form as the priors $p(\mathbf{H}|\Upsilon)$. For the model in (8), we assume

$$q(\mathbf{H}) = q(\alpha_0)q(\lambda)q(\boldsymbol{\pi})\prod_{i=1}^{M} q(z_i)\prod_{i=1}^{M} q(\boldsymbol{\theta}_i)\prod_{k=1}^{K} q(\boldsymbol{\alpha}_k^*), \tag{13}$$

where $q(\alpha_0) \sim Ga(\tilde{a}, \tilde{b})$, $q(\lambda) \sim Ga(\tilde{e}, \tilde{f})$, $q(\boldsymbol{\pi}) \sim \prod_{k=1}^{K-1} Beta(\tau_{1k}, \tau_{2k})$, $q(z_i) \sim Multinomial(\boldsymbol{w})$, $q(\boldsymbol{\theta}_i) \sim \mathcal{N}(\mu_i, \boldsymbol{\Gamma}_i)$, $q(\boldsymbol{\alpha}_k^*) \sim \prod_{j=1}^{m} Ga(\alpha_{k,j}^*|\tilde{c}_{k,j}, \tilde{d}_{k,j})$, with $\boldsymbol{w} = \{w_k\}_{k=1,K}$.

By substituting (13) and (8) into (12), the lower bound $\mathcal{F}(q)$ is readily obtained. The optimization of the lower bound $\mathcal{F}(q)$ is realized by taking functional derivatives with respect to each of the $q(\cdot)$ distributions while £xing the other q distributions, and setting $\partial\mathcal{F}(q)/\partial q(\cdot) = 0$ to £nd the distribution $q(\cdot)$ that increases $\mathcal{F}$ (Beal, 2003). The update equations for the variational posteriors are summarized in the Appendix. The convergence of the algorithm is monitored by the increase of the lower bound $\mathcal{F}$. One practical issue of the variational Bayesian inference is that the VB algorithm converges to a local maximum of the lower bound of the marginal log-likelihood since the true posterior usually is multi-modal. Therefore the average of multiple runs of the algorithm from different starting points may avoid this issue and yield better performance.

4. Experimental Results

4.1. Synthetic data

In the £rst set of examples we consider synthesized data to examine the sharing mechanisms associated with the DP-MT CS inversion. In the £rst example we generate data with 10 underlying clusters. Figure 1 shows ten "templates", each corresponding to a 256-length signal, with 30 non-zero components (the values of those non-zero components are randomly drawn from $\mathcal{N}(0, 1)$). The non-zero locations are chosen randomly for each template such that the correlation between these sparse templates is zero. For each template, £ve sparse signals (each with 256 samples) are generated by randomly selecting three non-zero elements from the associated template and setting the coef£cients to zero, and three zero-amplitude points in the template are randomly now set to be non-zero (each of these three non-zero values again drawn from $\mathcal{N}(0, 1)$). In this manner the sparseness properties of the £ve signals generated from a given template are highly related, and the ten clusters of sparse signals have distinct sparseness properties. For each sparse signal a set of CS random projections are performed, with the components of each projection vector drawn randomly from $\mathcal{N}(0, 1)$(Donoho, 2006). The reconstruction error is de£ned as $||\hat{\boldsymbol{u}} - \boldsymbol{u}||_2/||\boldsymbol{u}||_2$, where $\hat{\boldsymbol{u}}$ is the recovered signal and $\boldsymbol{u}$ is the original one.

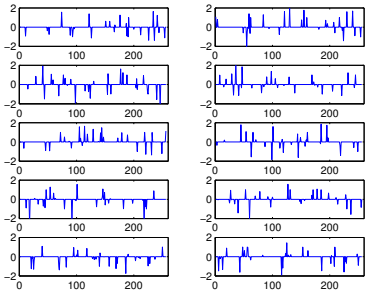

Figure 1. Ten template signals for 10-cluster case.

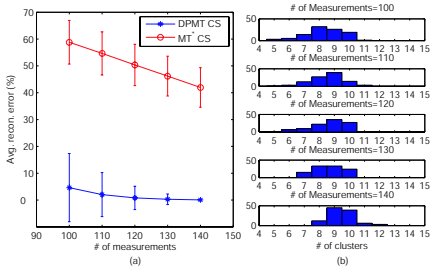

Figure 2. Multi-task CS inversion error (%) for DP-MT and MT* CS for the ten-cluster case. (a) Reconstruction errors, (b) histograms of the number of clusters yielded by DP-MT.

Figure 2 shows the reconstruction errors of the CS inversion by DP-MT CS as well as the global-sharing MT CS discussed in Sec 2.1 (denoted as MT* CS for simplicity), as a function of the number of CS measurements. Both CS algorithms are based on the VB DP-MT algorithm described in Sec 3, however for MT*, we set $\kappa_{i,1} = 1$, and $\kappa_{i,k} = 0$ for $k > 1$ for all tasks and £x the values of $\kappa_{i,k}$ in each iteration without update. The experiment was run 100 times (with 100 different random generations of random projection as well as initial membership), and the error bars in Figure 2 represent the standard deviation about the mean. From Figure 2 the advantage of the DP-based formulation is evident. In Figure 2 we also present histograms for the number of different clusters inferred by the DP-MT CS. It is clear from Figure 2 that the algorithm tends to infer about 10 clusters, but there is some variation, with the variation in the number of clusters increasing with decreasing number of CS measurements.

To further examine the impact of the number of underlying clusters, we now consider examples for which the data are generated for 5, 3, 2 and 1 underlying. For each of templates, £ve sparse signals are generated randomly, in the manner discussed above for the ten-cluster case. In Figures 3-6 are shown results in the form considered in Figure 2, for the case of 5, 3, 2 and 1 underlying clusters for data generation. One notes the following phenomenon: As the number of underlying clusters diminishes, the difference between DP-MT and MT* CS algorithms diminishes, with almost identical performance witnessed for the case

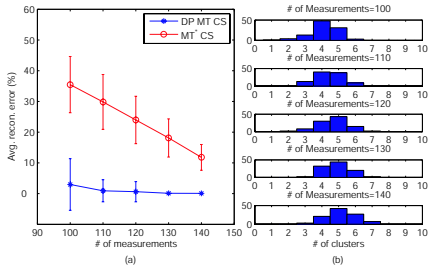

Figure 3. Multi-task CS inversion error (%) for DP-MT and MT* CS for the £ve-cluster case. (a) Reconstruction errors, (b) histograms of the number of clusters yielded by DP-MT.

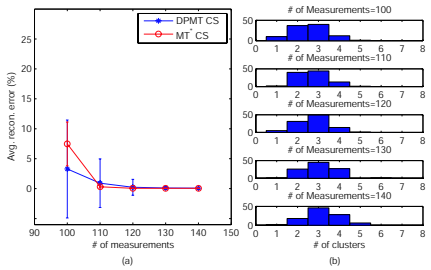

Figure 4. Multi-task CS inversion error (%) for DP-MT and MT* CS for the three-cluster case. (a) Reconstruction errors, (b) histograms of the number of clusters yielded by DP-MT.

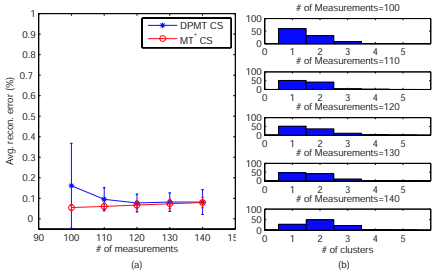

Figure 5. Multi-task CS inversion error (%) for DP-MT and MT* CS for the two-cluster case. (a) Reconstruction errors, (b) histograms of the number of clusters yielded by DP-MT.

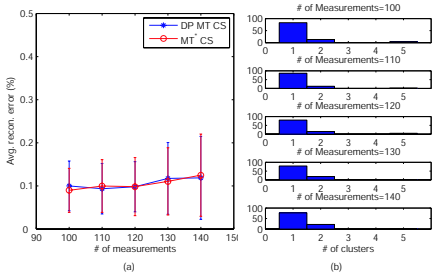

Figure 6. Multi-task CS inversion error (%) for DP-MT and MT* CS for the one-cluster case. (a) Reconstruction errors, (b) histograms of the number of clusters yielded by DP-MT.

of three and two clusters; this phenomenon is particularly evident as the number of CS measurements increases. As an aside, we also note that the DP-based inference of the number of underlying clusters adapts well to the underlying data generation.

We now provide an explanation for the relationships between the DP-MT and MT* CS algorithms. For two sparse signals like those in Figure 1, they have distinct non-zero coef£cients and therefore one would typically infer that they have dissimilar sparseness properties. However, they share many zero-amplitude coef£cients. If we consider M sparse signals, and if *all* of the M signals share the same large set of zero-amplitude coef£cients, then they are appropriate for sharing even if the associated (small number of) non-zero coef£cients are entirely distinct. For the 10-cluster case, because of the large number of clusters, the templates do not cumulatively share the same set of zero-amplitude coef£cients; in this case global sharing for CS inversion is inappropriate, and the same is true for the 5-cluster case. However, for the 3 and 2 cluster cases, the templates share a signi£cant number of zero-amplitude coef£cients, and therefore global sharing is appropriate. This underscores that global sharing across M tasks is appropriate when there is substantial sharing of zero-amplitude coef£cients, even when all of the non-zero-amplitude coef£cients are distinct. However, one typically does not know *a priori* if global sharing is appropriate (as it was *not* in Figures 2 and 3), and therefore the DP-based formulation offers generally high-quality results when global sharing is appropriate *and* when it is not.

We consider the sharing mechanisms manifested for two examples from the three-cluster case considered in Figure 4. The truncation level K can be set either to a large number or be estimated in principle by increase the number of sticks included until the log-marginal likelihood (the lower bound) in the VB algorithm starts to decrease. In this example we choose the number of sticks in the DP formulation to $K = 8$ which corresponds to the upper bound of the log-marginal likelihood, and we show the stick (cluster) with which each of the 15 tasks were grouped at the end of the inference process. These examples were selected because they both yielded roughly the same average CS inversion accuracy across the 15 CS inversions (0.40% and 0.38% error), but these two runs yield distinct clusterings. This example emphasizes that because the underlying signals are very sparse and they have signi£cant overlap in the set of zero-amplitude coef£cients, the particular clustering manifested by the DP formulation is not particularly important for the £nal CS-inversion quality.

4.2. Real images

In the following examples, applied to imagery, we perform comparisons between DP-MT, MT*, and also a single-task Bayesian CS (ST), in which the CS inversion is performed independently on each of the tasks. ST CS is realized with

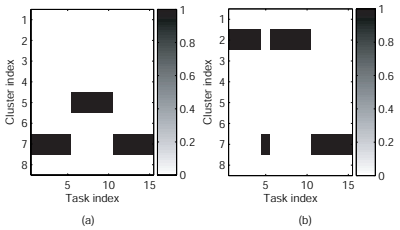

Figure 7. 2 example runs of the DP-MT CS clustering for the 3-cluster case (100 CS measurements). The grey scale denotes the probability that a given task is associated with a particular cluster. (a) Reconstruction error was 0.40%, (b) reconstruction error of 0.38%.

the same algorithm as DP-MT and MT*, but set $\kappa_{i,1} = 1$, and $\kappa_{i,k} = 0$ for $k > 1$, and consider only one CS task at a time ($M = 1$)..

We conduct two examples on CS reconstruction of typical imagery from "natural" scenes. All the images in these examples are of size 256×256 and are highly compressible in a wavelet basis. We choose the "Daubechies 8" wavelet as our orthonormal basis, and the sensing matrix $\mathbf{\Phi}$ is constructed in the same manner as in Sec 4.1. In this experiment we adopt a hybrid CS scheme, in which using CS we measure only £ne-scale wavelet coef£cients, while retaining all coarse-scale coef£cients (no compression in the coarse scale) (Tsaig & Donoho, 2006). We also assume all the wavelet coef£cients at the £nest scale are zero and only consider (estimate) the other 4096 coef£cients. In both examples, the coarsest scale is $j_0 = 3$, and the £nest scale is $j_1 = 6$. We use the mean of the posterior over $\boldsymbol{\theta}$ to perform the image reconstruction. The reconstruction error is de£ned as $||\hat{\boldsymbol{u}} - \boldsymbol{u}||_2/||\boldsymbol{u}||_2$, where $\hat{\boldsymbol{u}}$ is the reconstructed image and $\boldsymbol{u}$ is the original one.

In the £rst example, we choose 12 images from three different scenes. To reconstruct the image, we perform an inverse wavelet transform on the CS-estimated coef£cients. In Figure 8 (a) we show the reconstructed images with all 4096 measurements using linear reconstruction ($\boldsymbol{\theta} = \mathbf{\Phi}^T \boldsymbol{v}$), which is the best possible performance. Figure 8 (b)-(d) represent the reconstructed images by the DP-MT, MT*, and the ST algorithms, respectively, with the number of CS measurements $n = 1764$ (1700 measurements in the £ne scales and 64 in the coarse scale) for each task. The reconstruction errors for these four methods are compared in Table 1. We notice that the DP-MT algorithm reduces the reconstruction error compared to the ST method, which indicates that the multi-task CS inversion shares information among tasks and therefore requires less measurements than the single task learning does to achieve the same performance. In addition to the CS inversion, the DP-MT also yield task clustering, with this inferred simultaneously with the CS inversion; while this clustering is not the £nal product of interest, it is informative, with results shown in Figures 9. Note that the algorithms infer three clusters, each corresponding to a particular class of imagery. By contrast the MT* algorithm imposes complete sharing among all tasks, and the results in Table I indicate that this undermines performance.

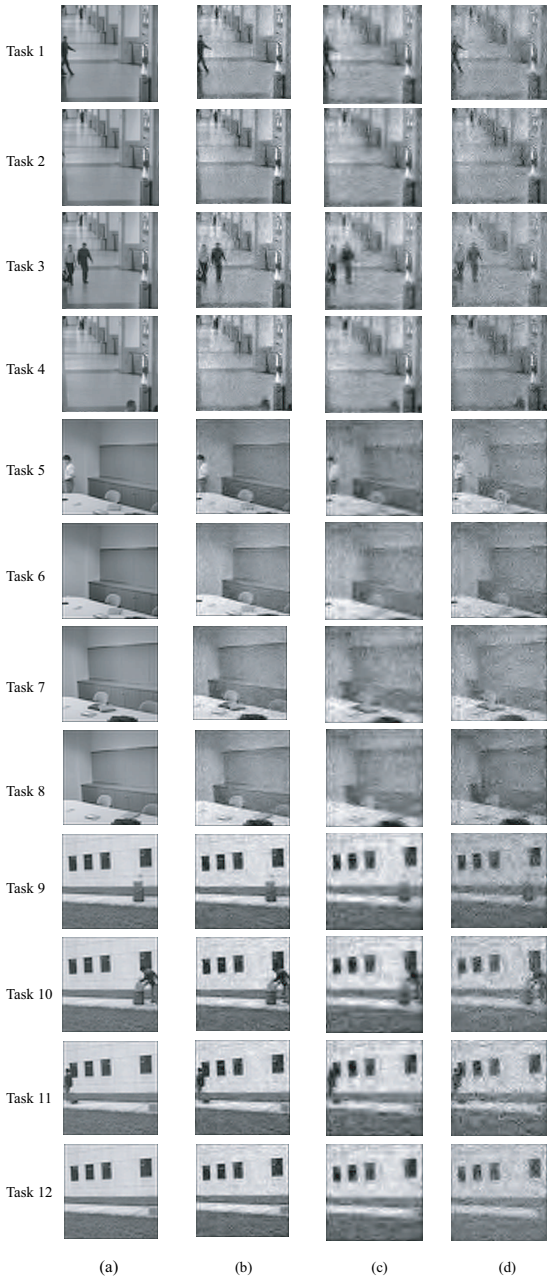

Figure 8. CS recon., (a) Linear, (b) DP-MT, (c) MT*, (d) ST

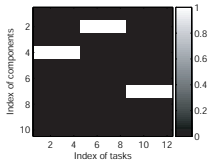

Figure 9. Sharing mechanism for 12 tasks in Figure 8 yielded by DP-MT CS.

Table 1. Reconstruction error (%) for the example in 8.

	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Task 7	Task 8	Task 9	Task 10	Task 11	Task 12
DP-MT	8.79	7.89	9.69	8.04	14.33	13.22	15.18	14.54	15.51	16.71	16.11	15.19
MT*	10.19	9.14	11.49	9.18	16.94	15.59	17.46	16.50	18.62	19.82	19.34	18.03
ST	10.28	10.37	12.81	10.28	18.37	16.18	18.65	17.67	20.77	22.24	21.19	19.59
Linear	6.66	6.20	7.08	6.14	12.41	11.70	12.43	11.99	13.83	14.41	14.10	13.53

In the second example we consider 11 images from three scenes. The reconstructed images are shown in Figure 10 by the linear reconstruction, DP-MT, MT* and ST algorithms; the reconstruction errors are listed in Table 2 for all four methods. As expected, the multi-task CS inversion algorithm yields smaller reconstruction error than the single task algorithm. The clustering result is shown in Figure 11, in which images 1-4 and 9-11 are clustered together by DP-MT. However, recall the simple example considered in Figure 7. The DP-based algorithm seeks to share the underlying sparseness of the images, even though the images themselves may appear distinct. In fact, the results in Figure 11 motivated the simple example considered in Figure 7.

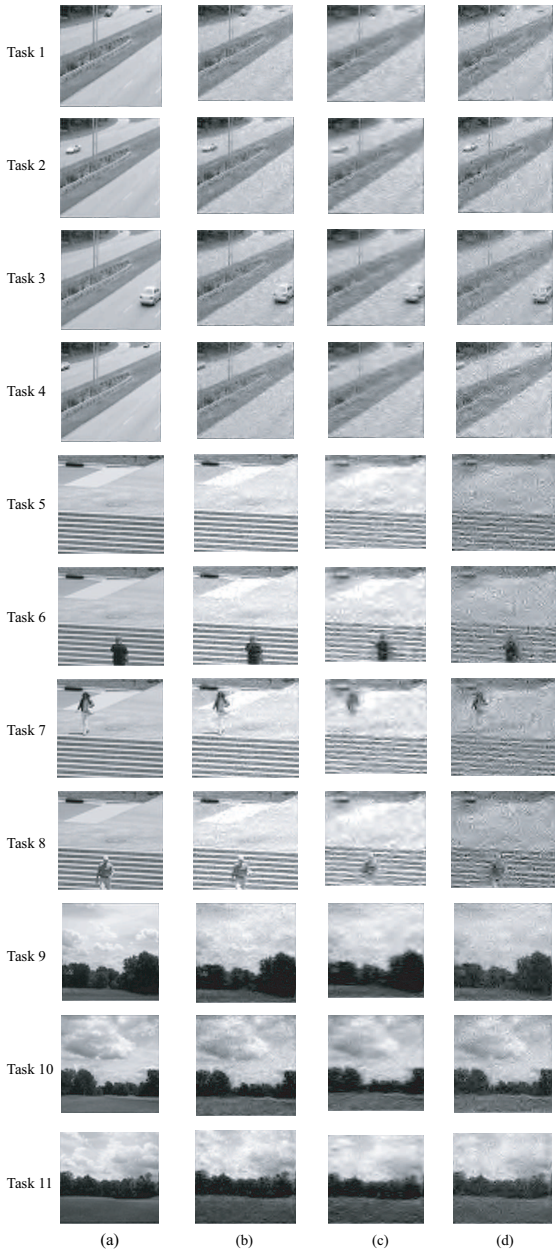

Figure 10. CS recon. (a) Linear, (b) DP-MT, (c) MT*, (d) ST

5. Conclusions

Hierarchical Dirichlet process (DP) priors are considered for the imposition of sparseness on the transform coef£cients in the context of inverting multiple CS measurements. An independent zero-mean Gaussian prior is placed on each transform coef£cient of each CS task and the task-dependent precision parameters are assumed drawn from a distribution G, where G is drawn from a Dirichlet process (DP); the base distribution of the DP is a product of Gamma distributions. The DP framework imposes the belief that many of the tasks may share underlying sparseness properties, and the objective is to cluster the CS measurements, where each cluster constitutes a particular form of sparseness. The DP formulation is non-parametric, in the sense that the number of clusters is not set *a priori* and is inferred from the data. A computationally ef£cient variational Bayesian inference has been considered on all model parameters. For all examples considered, the DP-MT CS inversion performed at least as well as ST CS inversion and CS inversion based on global sharing. Especially when global sharing was inappropriate, the DP-based inversion is signi£cantly better.

In future research, we may consider correlation between spatially and spectrally adjacent transformation coef£cients and remove the assumption of exchangeability employed within the DP, which in practice may not be true.

Appendix: Update Equations in VB DP MT

The updated hyperparameters for all $q(\cdot)$ in Sec 3 are

- $\tilde{a} = a + \frac{1}{2}\sum_{i=1}^{M} n_i$ and $\tilde{b} = b + \frac{1}{2}\sum_{i=1}^{M}\left[tr(\boldsymbol{\Phi}_i\boldsymbol{\Gamma}_i^{-1}\boldsymbol{\Phi}_i^T)+(\boldsymbol{\Phi}_i\boldsymbol{\mu}_i-\boldsymbol{v}_i)^T(\boldsymbol{\Phi}_i\boldsymbol{\mu}_i-\boldsymbol{v}_i)\right]$.

- $\tilde{e} = e+K-1$ and $\tilde{f} = f-\sum_{k=1}^{K-1}\left[\psi(\tau_{2k})-\psi(\tau_{1k}+\tau_{2k})\right]$, where $\psi(x) = \frac{\partial}{\partial x}\log\boldsymbol{\Gamma}(x)$.

Table 2. Reconstruction Error (%) for the example in Figure 10

	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Task 7	Task 8	Task 9	Task 10	Task 11
DP-MT	6.50	6.41	6.89	6.86	15.81	15.09	15.91	14.74	7.74	8.05	8.50
MT*	7.79	7.76	8.12	8.32	18.17	17.70	18.66	17.13	8.87	9.16	9.97
ST	8.31	8.23	8.81	9.23	19.79	19.74	20.36	18.88	8.77	9.58	9.62
Linear	4.78	4.77	5.01	5.15	15.39	14.49	15.18	14.06	6.10	5.72	6.72

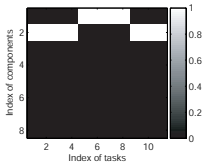

Figure 11. Sharing mechanism for 11 tasks in Figure 10 yielded by DP-MT

- $\tau_{1k} = 1 + \sum_{i=1}^{M} \kappa_{i,k}$ and $\tau_{2k} = \frac{\tilde{e}}{\tilde{f}} + \sum_{i=1}^{M} \sum_{l=k+1}^{K} \kappa_{i,k}$, where $\kappa_{i,k} = q(z_i = k)$.
- $\tilde{c}_{k,j} = c + \frac{1}{2} \sum_{i=1}^{M} \kappa_{i,k}$ and $\tilde{d}_{k,j} = d + \frac{1}{2} \sum_{i=1}^{M} \kappa_{i,k}(\sigma_{i,j} + \mu_{i,k,j}^2)$, where $[\sigma_{i,1}, \cdots, \sigma_{i,m}]$ is the diagonal elements of $\mathbf{\Gamma}_i^{-1}$ and $\mu_{i,k,j}$ is the j^{th} element of vector $\boldsymbol{\mu}_i$.
- $\mathbf{\Gamma}_i = \sum_{k=1}^{K} \kappa_{i,k} \mathbf{\Lambda}_k + \frac{\tilde{a}}{\tilde{b}} \mathbf{\Phi}_i^T \mathbf{\Phi}_i$ and $\boldsymbol{\mu}_i = \frac{\tilde{a}}{\tilde{b}} \mathbf{\Gamma}_i^{-1} \mathbf{\Phi}_i^T \boldsymbol{v}_i$, where $\mathbf{\Lambda}_k = diag(\tilde{c}_{k,1}/\tilde{d}_{k,1}, \cdots, \tilde{c}_{k,m}/\tilde{d}_{k,m})$ is a diagonal matrix of $m \times m$.
- $\kappa_{i,k} = \frac{e^{\lambda_{i,k}}}{\sum_{l=1}^{K} e^{\lambda_{i,l}}}$, where $\lambda_{i,k} = \sum_{l=1}^{k-1} \left[\psi(\tau_{2l}) - \psi(\tau_{1l} + \tau_{2l})\right] + \left[\psi(\tau_{1k}) - \psi(\tau_{1k} + \tau_{2k})\right] - \frac{1}{2}\Big\{ \sum_{j=1}^{m} \left[\ln 2\pi - \psi(\tilde{c}_{k,j}) + \ln(\tilde{d}_{k,j})\right] + tr(\mathbf{\Gamma}_i^{-1} \mathbf{\Lambda}_k) + \boldsymbol{\mu}_i^T \mathbf{\Lambda}_k \boldsymbol{\mu}_i \Big\}$.

References

Baron, D., Wakin, M. B., Duarte, M. F., Sarvotham, S., & Baraniuk, R. G. (2005). Distributed compressed sensing.

Beal, M. J. (2003). *Variational algorithms for approximate bayesian inference*. Doctoral dissertation, Gatsby Computational Neuroscience Unit, Univ. College London.

Blei, D., & Jordan, M. (2004). Variational methods for the dirichlet process. *ICML*.

Candes, E. (2006). Compressive sensing. *Proc. of ICM*, *3*, 1433–1452.

Candes, E., Romberg, J., & Tao, T. (2006). Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. on Inform. Theory*, *52*, 489–502.

Charilaos, C. (1999). Jpeg2000 tutorial. *ICIP*.

Donoho, D. L. (2006). Compressed sensing. *IEEE Trans. Information Theory*, *52*, 1289–1306.

Donoho, D. L., Tsaig, Y., Drori, I., & Starck, J.-C. (2006). Sparse solution of underdetermined linear equations by stagewise orthogonal matching pursuit.

Escobar, M. D., & West, M. (1995). Bayesian density estimation and inference using mixtures. *JASA*, *90*, 577–588.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). Introducing markov chain monte carlo. In *Markov chain monte carlo in practice*. London, U.K.: Chapman Hall.

Ishwaran, H., & James, L. F. (2001). Gibbs sampling methods for stick-breaking priros. *JASA*, *96*, 161–173.

Ji, S., Dunson, D., & Carin, L. (2007a). Multi-task compressive sensing. *Submit to IEEE Trans. on SP*.

Ji, S., Xue, Y., & Carin, L. (2007b). Bayesian compressive sensing. *Accepted to IEEE Trans. on SP*.

MacKay, D. J. C. (1994). Bayesian methods for backpropagation networks. In E. Domany, van J. L. Hemmen and K. Schulten (Eds.), *Models of neural networks III*. New York: Springer-Verlag.

Neal, R. M. (1996). *Bayesian learning for neural networks*. Springer.

S. Chen, D. L. D., & Saunders, M. A. (1999). Atomic decomposition by basis pursuit. *SIAM Journal on Scienti£c Computing*, *20*, 33–61.

Sethuraman, J. (1994). A constructive de£nition of the dirichlet prior. *Statistica Sinica*, *2*, 639–650.

Tipping, M. E. (2001). Sparse bayesian learning and the relevance vector machine. *JMLR*, *1*, 211–244.

Tropp, J. A., & Gilbert, A. C. (2005). Signal recovery from partial information via orthogonal matching pursuit.

Tsaig, Y., & Donoho, D. L. (2006). Extensions of compressed sensing. *Signal Processing*, *86*, 549–571.

West, M., Muller, P., & Escobar, M. (1994). Hierarchical priors and mixture models with applications in regression and density estimation. In P. R. Freeman and A. F. Smith (Eds.), *Aspects of uncertainty*, 363–386. John Wiley.

Wipf, D. P., & Rao, B. D. (2007). An empirical bayesian strategy for solving the simultaneous sparse approximation problem. *IEEE Trans. on SP*, *55*, 3704–3716.