MULTIMODAL SPARSE BAYESIAN DICTIONARY LEARNING APPLIED TO MULTIMODAL DATA CLASSIFICATION

Igor Fedorov *, *Bhaskar D. Rao, Truong Q. Nguyen*

Department of Electrical and Computer Engineering
University of California, San-Diego

ABSTRACT

In this paper, we present a novel multimodal sparse dictionary learning algorithm based on a hierarchical sparse Bayesian framework. The framework allows for enforcing joint sparsity across dictionaries without restricting the actual entries to be equal. We show that the proposed method is able to learn dictionaries of *higher quality* than existing approaches. We validate our claims with extensive experiments on synthetic data as well as real-world data.

Index Terms— Dictionary learning, signal representation, multimodal

1. INTRODUCTION

Due to improvements in sensor technology, computational resources, and communication, acquiring vast amounts of data has become relatively easy [1]. Given the ability to harvest data, the task still remains as to how to extract *relevant information* from it. In most cases, the data is multimodal, which introduces novel challenges in learning from the data. Recently, multimodal dictionary learning (DL) and sparse coding have become popular tools for fusing information from disparate data modalities [2][3][4][5][6].

The multimodal DL problem consists of estimating dictionaries $\{\boldsymbol{D}_j\}_{j=1}^J, \boldsymbol{D}_j \in \mathbb{R}^{N_j \times M}$, and encodings $\{\boldsymbol{x}_j^i\}_{i=1,j=1}^{L,J}$ given data $\{\boldsymbol{y}_j^i\}_{i=1,j=1}^{L,J}$ such that $\boldsymbol{y}_j^i \approx \boldsymbol{D}_j \boldsymbol{x}_j^i$, where L denotes the number of data points and J the number of modalities. In most cases of interest, we are interested in learning *overcomplete* dictionaries because they are much more flexible in the range of signals they can represent [7]. Since $\boldsymbol{y}_j^i$ admits an infinite number of representations under overcomplete (and full-rank) $\boldsymbol{D}_j$, it is often desirable to constrain $\boldsymbol{x}_j^i$ to be *sparse* [8].

Without any further constraints, the multimodal DL problem, as described, can be viewed as a set of J independent unimodal DL problems. To fully capture the multimodal nature of the problem, we must encode our prior knowledge that each set of points $\{\boldsymbol{y}_j^i\}_{j=1}^J$ is generated by some common source which happens to be measured J different ways. For instance, in [9], low and high resolution image patches are modelled as $\boldsymbol{y}_1^i$ and $\boldsymbol{y}_2^i$, respectively, and the association between $\boldsymbol{y}_1^i$ and $\boldsymbol{y}_2^i$ is enforced by the constraint $\boldsymbol{x}_1^i = \boldsymbol{x}_2^i$. The resulting multimodal DL problem is then to solve [10]

$$\Omega^* = \arg\min_{\Omega} \sum_{i=1}^{L} \left\| \tilde{\boldsymbol{y}}^i - \tilde{\boldsymbol{D}} \boldsymbol{x}^i \right\|_2^2 + \lambda \left\| \boldsymbol{x}^i \right\|_1 \qquad (1)$$

where

$$\tilde{\boldsymbol{y}}^i = \begin{bmatrix} \left(\boldsymbol{y}_1^i\right)^T & \cdots & \left(\boldsymbol{y}_J^i\right)^T \end{bmatrix}^T$$
$$\tilde{\boldsymbol{D}} = \begin{bmatrix} \boldsymbol{D}_1^T & \cdots & \boldsymbol{D}_J^T \end{bmatrix}^T$$
$$\Omega = \left\{ \{\boldsymbol{D}_j\}_{j=1}^J, \{\boldsymbol{x}_j^i\}_{i=1,j=1}^{L,J} \right\},$$

the ℓ_1 norm is used as a proxy to the ℓ_0 sparsity measure, and the modality subscript for $\boldsymbol{x}^i$ is omitted because the sparse codes are constrained to be the same. In a classification framework, (1) can be viewed as learning a multimodal feature extractor, where $(\boldsymbol{x}^i)^*$ is the multimodal representation of $\{\boldsymbol{y}_j^i\}_{j=1}^J$ that is fed into a classifier [2][11][3].

There are two significant issues with (1):

1. While using the same sparse code for each modality establishes an explicit relationship between the dictionaries for each modality, the same coefficient values may not be able to represent different modalities well.
2. It is often the case that one data modality is less noisy than another and the DL algorithm should be able to leverage the clean modality to learn on the noisy one. Since (1) constrains the regularization parameter λ to be the same for all modalities, it is impossible to incorporate prior knowledge about the noise level of each modality.

One recent approach that builds upon the K-SVD algorithm [8], which we will refer to as Joint ℓ_0 Dictionary Learning (J0DL), seeks to address these issues by framing the problem as [5]

$$\Omega^* = \underset{\Omega:\{\chi_j^i\}_{j=1}^J = \chi^i, |\chi^i| \le s, \forall i}{\arg\min} \sum_{i=1,j=1}^{L,J} \lambda_j \|\boldsymbol{y}_j^i - \boldsymbol{D}_j \boldsymbol{x}_j^i\|_2^2 \qquad (2)$$

*Igor Fedorov was partially supported by the San Diego Chapter of the ARCS Foundation, Inc.

where χ_j^i denotes the support of $\boldsymbol{x}_j^i$ and s the desired sparsity of the solution. This approach seeks to establish a similar relationship between the sparse codes of each modality as (1), while allowing the coefficients themselves to vary. The weights $\{\lambda_j\}_{j=1}^J$ allow for encoding prior information about the noise level of $\boldsymbol{y}_j^i$. The major drawback of J0DL is that, since (2) has an ℓ_0 type constraint, solving it requires a Simultaneous Orthogonal Matching Pursuit (SOMP) [12] algorithm called Distributed Compressive Sensing SOMP (DCS-SOMP)[13]. DCS-SOMP is a greedy algorithm and can produce poor solutions, especially if one modality is much noisier than another.

In this work, we propose a novel multimodal DL approach based on a hierarchical sparse Bayesian model [14][15]. We build upon the seminal Sparse Bayesian Learning (SBL) [16][17] framework along with its extensions to multiple [18][19] and multimodal measurements [5]. Our approach is able to capture the relationship between diverse datasets within the DL process, while addressing both issues above and benefiting from the significant sparse coding performance gains afforded by the Bayesian model [20][5]. Moreover, incorporating DL within the Bayesian framework leads to a joint (dictionary and sparse code) optimization approach with provable convergence guarantees, unlike the block coordinate descent algorithm used to solve (2) [5].

1.1. Contributions

1. We introduce a novel multimodal sparse DL algorithm based on a Gaussian Scale Mixture (GSM) prior on the sparse codes. The Bayesian formulation allows us to force the *support* of the sparse codes for each modality to be the same while allowing the coefficients to be different.

2. We provide a learning procedure for our model and provide a convergence guarantee. We also address significant practical challenges associated with the procedure.

3. We show synthetic data experimental results confirming that the proposed approach is able to leverage information from a clean modality to learn a better dictionary for a noisier one.

4. We apply our approach to a multimodal data classification task and show that the proposed method outperforms the algorithms tested.

2. SPARSE BAYESIAN MULTIMODAL DICTIONARY LEARNING

We begin by specifying the multimodal signal model. Let $\boldsymbol{y}_j$ and $\boldsymbol{x}_j$ be random variables representing $\boldsymbol{y}_j^i$ and $\boldsymbol{x}_j^i$, respectively. The signal model is then given by $\boldsymbol{y}_j = \boldsymbol{D}_j\boldsymbol{x}_j + \boldsymbol{v}_j$, where $\boldsymbol{v}_j \sim \mathsf{N}(\boldsymbol{v}_j; 0, \sigma_j^2\mathbf{I})$. Note that the noise variance is allowed to vary among data modalities. In order to encourage $\boldsymbol{x}_j^i$ to be sparse, we assume a GSM prior on each element of $\boldsymbol{x}_j$ [16][17][21][22], such that

$$p(\boldsymbol{x}_j|\boldsymbol{\gamma}) = \prod_{m=1}^{M} \mathsf{N}(\boldsymbol{x}_j[m]; 0, \boldsymbol{\gamma}[m])$$

where $\boldsymbol{x}_j[m]$ denotes the m'th element of the random vector $\boldsymbol{x}_j$ and the choice of $p(\boldsymbol{\gamma}[m])$ determines the marginal density of $\boldsymbol{x}_j[m]$. The key is that $\boldsymbol{\gamma}$ is shared by the sparse codes for each modality, encoding our prior knowledge that the sparse codes share the same support while retaining the freedom to have different coefficient values. This model is equivalent to the one used in [23][18][5], although $\{\boldsymbol{D}_j\}_{j=1}^J$ is assumed known there.

We adopt an empirical Bayesian strategy and seek to jointly estimate $\theta = \{\{\boldsymbol{D}_j\}_{j=1}^J, \{\boldsymbol{\gamma}^i\}_{i=1}^L\}$, which is done by maximizing the data log-likelihood given θ:

$$\theta^* = \arg\max_{\theta} \log p(\{\boldsymbol{y}_j^i\}_{i=1,j=1}^{L,J}|\theta). \tag{3}$$

In order to maximize (3), we employ the Expectation-Maximization (EM) algorithm [17], where we consider $\{\{\boldsymbol{y}_j^i\}_{i=1,j=1}^{L,J}, \{\boldsymbol{x}_j^i\}_{i=1,j=1}^{L,J}, \theta\}$ as the complete data and $\{\boldsymbol{x}_j^i\}_{i=1,j=1}^{L,J}$ as the latent data. In the E-step, we compute $Q(\theta, \theta^t)$, given by the expected value of the complete data under the posterior of the latent data given the observations and θ^t, the estimate of θ at iteration t. We assume non-informative priors on $\boldsymbol{\gamma}^i[m], \forall i, m$ [16], and $\boldsymbol{D}_j, \forall j$ [15]. The posterior needed to compute $Q(\theta, \theta^t)$ is given by $p(\boldsymbol{x}_j|\boldsymbol{y}_j^i, \theta) = \mathsf{N}(\boldsymbol{x}_j^i; \boldsymbol{\mu}_j^i, \boldsymbol{\Sigma}_j^i)$, where

$$\boldsymbol{\Sigma}_j^i = \boldsymbol{\Gamma}^i - \boldsymbol{\Gamma}^i\boldsymbol{D}_j^T\left(\sigma_j^2\mathsf{I} + \boldsymbol{D}_j\boldsymbol{\Gamma}^i\boldsymbol{D}_j^T\right)^{-1}\boldsymbol{D}_j\boldsymbol{\Gamma}^i \tag{4}$$

$$\boldsymbol{\mu}_j^i = \sigma_j^{-2}\boldsymbol{\Sigma}_j^i\boldsymbol{D}_j^T\boldsymbol{y}_j^i \tag{5}$$

and $\boldsymbol{\Gamma}^i = \mathrm{diag}(\boldsymbol{\gamma}^i)$. In the M-step, we maximize $Q(\theta, \theta^t)$ with respect to θ, leading to the update rules

$$\left(\boldsymbol{\gamma}^i[m]\right)^{t+1} = \frac{1}{J}\left(\sum_{j=1}^{J}\boldsymbol{\Sigma}_j^i[m,m] + \boldsymbol{\mu}_j^i[m]^2\right) \tag{6}$$

$$\boldsymbol{D}_j^{t+1} = \boldsymbol{Y}_j\boldsymbol{U}_j^T\left(\boldsymbol{U}_j\boldsymbol{U}_j^T + \sum_{i=1}^{L}\boldsymbol{\Sigma}_j^i\right)^{-1} \tag{7}$$

where $\boldsymbol{Y}_j = \begin{bmatrix}\boldsymbol{y}_j^1 & \cdots & \boldsymbol{y}_j^L\end{bmatrix}$ and $\boldsymbol{U}_j = \begin{bmatrix}\boldsymbol{\mu}_j^1 & \cdots \boldsymbol{\mu}_j^L\end{bmatrix}$. Note that the update rules given in (6) and (7) are equivalent to those shown in [15] for $J = 1$. Unlike[5][14][10], our approach is not a block-coordinate descent algorithm.

2.1. Simulated Annealing of σ_j^2

Although the parameters $\{\sigma_j\}_{j=1}^J$ can be readily estimated within the EM framework above, there are serious identifiability issues even in the unimodal case when the dictionary

is known [19]. To motivate our approach, we consider the maximum a-posteriori (MAP) estimate of $\boldsymbol{x}_j^i$ given $\boldsymbol{y}_j^i$:

$$\left(\boldsymbol{x}_j^i\right)^{MAP} = \arg\min_{\boldsymbol{x}_j^i} \|\boldsymbol{y}_j^i - \boldsymbol{D}_j \boldsymbol{x}_j^i\|^2 - 2\sigma_j^2 \log p(\boldsymbol{x}_j^i).$$

This shows that σ_j^2 can be thought of as a regularization parameter which controls the trade-off between sparsity and signal reconstruction error. We propose to adopt an annealing strategy for σ_j^2, summarized by

$$\sigma_j^{t+1} = \alpha \sigma_j^t, \alpha < 1. \tag{8}$$

The motivation for annealing σ_j^2 is that the quality of $\{\boldsymbol{D}_j\}_{j=1}^J$ increases with t, so seeking sparse $\boldsymbol{x}_j^i$ for small t does not make sense and can force the EM algorithm to converge to a poor local minimum.

2.2. Complete Algorithm Specification

The complete algorithm is summarized in Alg. 1. In practice, at each iteration, the dictionaries $\{\boldsymbol{D}_j\}_{j=1}^J$ are normalized to unit ℓ_2 column norm in order to prevent scaling instabilities.

Algorithm 1 Multimodal DL algorithm

Require: $\{\boldsymbol{y}_j^i\}_{i=1,j=1}^{L,J}, \{\sigma_j^0\}_{j=1}^J, \alpha, \{\sigma_j^{min}\}_{j=1}^J$
1: **while** not converged **do**
2: Update $\{\boldsymbol{\Sigma}_j^i\}_{i=1,j=1}^{L,J}$ using (4)
3: Update $\{\boldsymbol{\mu}_j^i\}_{i=1,j=1}^{L,J}$ using(5)
4: Update $\{\boldsymbol{\gamma}^i\}_{i=1}^L$ using (6)
5: Update $\{\boldsymbol{D}_j\}_{j=1}^J$ using (7)
6: Update σ_j using (8) unless $\sigma_j \leq \sigma_j^{min}$
7: **end while**
8: **return** $\{\boldsymbol{D}_j\}_{j=1}^J$

2.3. Analysis

We now provide a valuable theoretical result pertaining to the proposed approach.

Theorem 1. *Algorithm 1 is guaranteed to converge to a stationary point of* (3)*.*

Proof. For fixed σ_j, Algorithm 1 is guaranteed to converge to a stationary point of (3) because $Q(\theta, \theta^t)$ satisfies the conditions of (Theorem 2 [24]). To extend this result to the case of varying σ_j, we note that σ_j is annealed for a pre-specified, finite number of iterations, after which Algorithm 1 is executed until convergence without modifying σ_j. ■

3. RESULTS

3.1. Synthetic Data

In order to validate how well the proposed algorithm is able to learn unimodal and multimodal dictionaries, we conducted a series of experiments on synthetic data. We adopt the experimental setup given in [14] and begin by generating the elements of the ground-truth dictionaries $\boldsymbol{D}_j^* \in \mathbb{R}^{20\times 50}$ by sampling from a $\mathsf{N}(0,1)$ distribution and scaling the resulting matrices to have unit ℓ_2 column norm. We then generate $\boldsymbol{x}_j^i$ by randomly selecting $s = 5$ indices and generating the non-zero entries by drawing samples from a $\mathsf{N}(0,1)$ distribution. The supports of $\{\boldsymbol{x}_j^i\}_{j=1}^J$ are constrained to be the same, while the coefficients themselves are not. Finally, $\boldsymbol{v}_j^i$ is generated by drawing samples from a $\mathsf{N}(0,1)$ distribution and scaling the resulting vector in order to achieve a specified Signal-to-Noise Ratio (SNR). We use $L = 1000$ and $J = 2$ and simulate a dataset consisting of a noisy modality with 10 dB SNR and a clean modality with 30 dB SNR. In order to measure the performance of the recovered dictionaries $\{\hat{\boldsymbol{D}}_j\}_{j=1}^J$, we compute the probability of successfully recovering $\{\boldsymbol{D}_j^*\}_{j=1}^J$, given by

$$L_j = \frac{1}{M}\sum_{m=1}^{M} \mathbb{1}\left[\max_{1\leq k\leq M} \frac{|\langle \boldsymbol{d}_j^m, \hat{\boldsymbol{d}}_j^k\rangle|}{\|\boldsymbol{d}_j^m\|_2^2 \|\hat{\boldsymbol{d}}_j^k\|_2^2} > 0.99\right]$$

where $\langle\cdot,\cdot\rangle$ denotes an inner product and $\mathbb{1}[\cdot]$ the indicator function. The experiment is performed 50 times and averaged results are reported. We compare the proposed method with commonly used sparse DL algorithms, including the ℓ_1-norm based method for solving (1) [10], K-SVD [8], and J0DL. The regularization parameters λ in (1) and $\{\lambda_j\}_{j=1}^J$ in (2) were selected by a grid search and both K-SVD and J0DL were given the true sparsity parameter s. For multimodal DL, K-SVD was given data in the format $\{\tilde{\boldsymbol{y}}^i\}_{i=1}^L$. J0DL was implemented in a block-coordinate descent fashion where the sparse coding step was computed using DC-SOMP and the dictionary update step was done using the same procedure as in K-SVD [5]. All algorithms were run for 1000 iterations.

The synthetic data results are shown in Fig. 1. For unimodal data, all of the algorithms recover the true dictionary almost perfectly when the SNR $= 30$ dB, with the exception of J0DL and K-SVD. On the other hand, for SNR $= 10$ dB, all of the tested algorithms perform poorly, with the proposed method outperforming all of the tested methods.

In the multimodal scenario, the proposed method clearly distinguishes itself from the other methods tested. Not only does the proposed method (almost perfectly) recover the clean data dictionary, but it achieves an accuracy of 93.5% on the noisy data dictionary, which is 28.8% better (in absolute terms) than the next best method. The performance of the proposed method is even more impressive considering that the ℓ_1 and K-SVD algorithms were not able to attain more than 0.2% accuracy in recovering either the clean or noisy dictionaries. J0DL was able to capture some of the multimodal information in learning the noisy dictionary, but the noisy dictionary accuracy only reaches 65.1%.

3.2. Wikipedia Dataset Multimodal Classification

The Wikipedia dataset [25] consists of 2,866 Wikipedia articles grouped into 10 categories (art, history, biology, etc.),

		Method			
DL scheme	Feature Type	Proposed	J0DL	ℓ_1 [2]	K-SVD
Unimodal	Images	24.24	**25.53**	24.96	—
	Text	70.27	69.26	**70.42**	—
Multimodal	Images	**29.58**	23.23	25.11	25.69
	Text	**69.99**	66.09	68.83	66.52
	Joint sparse coding	**71.00**	62.63	66.52	54.55

Table 1: Wikipedia dataset classification accuracy results (%).

with each article represented as an image with corresponding text. In order to facilitate document classification, we use the 10-dimensional Latent Dirichlet Allocation (LDA) [26] text features and 128-dimensional SIFT [27] image features provided by the authors of [25]. This dataset provides a good testbed for multimodal learning algorithms because the text modality is much less noisy than the image modality and can be leveraged to obtain better representations of images.

We adopt the experimental setup of [2] and split the dataset into a training set, consisting of about 75% of the documents, and testing set, consisting of the rest of the documents. The training data is used to learn dictionaries for the text and image modalities. To classify a given test document, the learned dictionaries are used to extract sparse codes for the text and image modalities. The extracted sparse codes are then used as inputs to a one-vs-all support vector machine (SVM) [28] classifier (trained on the training data sparse codes) with radial basis function kernel. The tuning parameters of the classifier are estimated using cross-validation.

The classification results are reported in Table 1. In the unimodal setting, the dictionaries for text and images are learned independently. During testing, a given document is classified using either the image or text sparse code. The parameters λ in (1) and $\{\lambda_j\}_{j=1}^{J}$ in (2) were selected by grid search. We use $s = 10$ for the ℓ_0 based methods. Note that K-SVD is equivalent to J0DL in this scenario, so K-SVD results are omitted. As expected, all of the tested algorithms perform nearly identically. In the multimodal setting, the text and image dictionaries are learned jointly. At test time, the learned dictionaries are used to extract the image and text sparse codes, or to extract both jointly (i.e. by using the same procedure as the one used to learn the dictionaries, but with the dictionaries fixed). The proposed method outperforms the other tested methods in the multimodal scenario. Moreover, the proposed method is able to leverage the text modality to learn a *higher quality* image dictionary and achieve an improvement of 5.34% over the unimodal case in classifying images, with the other methods showing only slight improvement or performance loss. As in the synthetic data case, the proposed method maintains its performance on the clean (text) modality, whereas the other algorithms show a drop in performance. The best overall performance is achieved by the proposed method in the joint sparse coding setting.

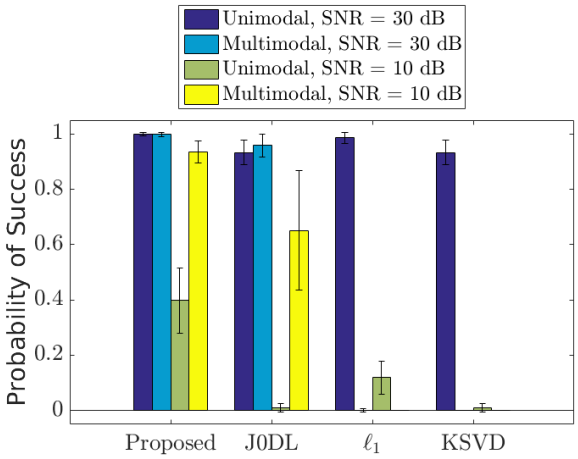

Fig. 1: Synthetic data results with one standard deviation error bars.

4. CONCLUSION

We have detailed a novel sparse multimodal DL algorithm. Our approach incorporates the main features of existing methods, which establish a correspondence between the elements of the dictionaries for each modality, while addressing the major drawbacks of previous algorithms. Our method enjoys the theoretical guarantees and superior sparse recovery rates associated with the sparse Bayesian learning framework.

5. REFERENCES

[1] James M. Tien, "Big data: Unleashing information," *Journal of Systems Science and Systems Engineering*, vol. 22, no. 2, pp. 127–151, 2013.

[2] Miriam Cha, Youngjune Gwon, and HT Kung, "Multimodal sparse representation learning and applications," *arXiv preprint arXiv:1511.06238*, 2015.

[3] S. Shekhar, V. M. Patel, N. M. Nasrabadi, and R. Chellappa, "Joint sparse representation for robust multimodal biometrics recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 1, pp. 113–126, Jan 2014.

[4] Juan C Caicedo and Fabio A González, "Multimodal fusion for image retrieval using matrix factorization," in

Proceedings of the 2nd ACM international conference on multimedia retrieval. ACM, 2012, p. 56.

[5] Yacong Ding and Bhaskar D Rao, "Joint dictionary learning and recovery algorithms in a jointly sparse framework," in *2015 49th Asilomar Conference on Signals, Systems and Computers*. IEEE, 2015, pp. 1482–1486.

[6] Nicolas Seichepine, Slim Essid, Cédric Févotte, and Olivier Cappé, "Soft nonnegative matrix co-factorization," *IEEE Transactions on Signal Processing*, vol. 62, no. 22, pp. 5940–5949, 2014.

[7] D. L. Donoho, M. Elad, and V. N. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Transactions on Information Theory*, vol. 52, no. 1, pp. 6–18, Jan 2006.

[8] M. Aharon, M. Elad, and A. Bruckstein, "k -svd: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, Nov 2006.

[9] Jianchao Yang, John Wright, Thomas S Huang, and Yi Ma, "Image super-resolution via sparse representation," *Image Processing, IEEE Transactions on*, vol. 19, no. 11, pp. 2861–2873, 2010.

[10] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro, "Online dictionary learning for sparse coding," in *Proceedings of the 26th annual international conference on machine learning*. ACM, 2009, pp. 689–696.

[11] Youngjune Gwon, William Campbell, Kevin Brady, Douglas Sturim, Miriam Cha, and H.T. Kung, "Multimodal sparse coding for event detection," *NIPS MMML*, 2015.

[12] Joel A Tropp, Anna C Gilbert, and Martin J Strauss, "Simultaneous sparse approximation via greedy pursuit," in *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.* IEEE, 2005, vol. 5, pp. v–721.

[13] Dror Baron, Marco F Duarte, Michael B Wakin, Shriram Sarvotham, and Richard G Baraniuk, "Distributed compressive sensing," *arXiv preprint arXiv:0901.3403*, 2009.

[14] Linxiao Yang, Jun Fang, and Hongbin Li, "Sparse bayesian dictionary learning with a gaussian hierarchical model," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 2564–2568.

[15] Mark Girolami, "A variational method for learning sparse and overcomplete representations," *Neural computation*, vol. 13, no. 11, pp. 2517–2532, 2001.

[16] Michael E Tipping, "Sparse bayesian learning and the relevance vector machine," *The journal of machine learning research*, vol. 1, pp. 211–244, 2001.

[17] David P Wipf and Bhaskar D Rao, "Sparse bayesian learning for basis selection," *Signal Processing, IEEE Transactions on*, vol. 52, no. 8, pp. 2153–2164, 2004.

[18] Shihao Ji, David Dunson, and Lawrence Carin, "Multitask compressive sensing," *IEEE Transactions on Signal Processing*, vol. 57, no. 1, pp. 92–106, 2009.

[19] Z. Zhang and B. D. Rao, "Sparse signal recovery with temporally correlated source vectors using sparse bayesian learning," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 5, pp. 912–926, Sept 2011.

[20] R. Giri and B. Rao, "Type I and Type II Bayesian Methods for Sparse Signal Recovery Using Scale Mixtures," *IEEE Transactions on Signal Processing*, vol. 64, no. 13, pp. 3418–3428, July 2016.

[21] David P Wipf and Bhaskar D Rao, "An empirical bayesian strategy for solving the simultaneous sparse approximation problem," *Signal Processing, IEEE Transactions on*, vol. 55, no. 7, pp. 3704–3716, 2007.

[22] David P Wipf, Bhaskar D Rao, and Srikantan Nagarajan, "Latent variable bayesian models for promoting sparsity," *Information Theory, IEEE Transactions on*, vol. 57, no. 9, pp. 6236–6255, 2011.

[23] Yu Wang, David Wipf, Jeong-Min Yun, Wei Chen, and IJ Wassell, "Clustered sparse bayesian learning," in *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2015.

[24] CF Jeff Wu, "On the convergence properties of the em algorithm," *The Annals of statistics*, pp. 95–103, 1983.

[25] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G.R.G. Lanckriet, R. Levy, and N. Vasconcelos, "A New Approach to Cross-Modal Multimedia Retrieval," in *ACM International Conference on Multimedia*, 2010, pp. 251–260.

[26] David M Blei, Andrew Y Ng, and Michael I Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

[27] David G Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[28] Chih-Chung Chang and Chih-Jen Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.