A GENERAL FRAMEWORK FOR RECONSTRUCTION AND CLASSIFICATION FROM COMPRESSIVE MEASUREMENTS WITH SIDE INFORMATION

Liming Wang†, Francesco Renna⋆, Xin Yuan†, Miguel Rodrigues⋆, Robert Calderbank† and Lawrence Carin†

† Department of Electrical and Computer Engineering, Duke University
⋆ Department of Electronic and Electrical Engineering, University College London

ABSTRACT

We develop a general framework for compressive linear-projection measurements with side information. Side information is an additional signal correlated with the signal of interest. We investigate the impact of side information on classification and signal recovery from low-dimensional measurements. Motivated by real applications, two special cases of the general model are studied. In the first, a joint Gaussian mixture model is manifested on the signal and side information. The second example again employs a Gaussian mixture model for the signal, with side information drawn from a mixture in the exponential family. Theoretical results on recovery and classification accuracy are derived. The presence of side information is shown to yield improved performance, both theoretically and experimentally.

Index Terms— Compressive sensing, side information, dimensionality reduction

1. INTRODUCTION

Dimensionality reduction plays a pivotal role in various machine-learning applications, including compressive feature extraction, supervised dimensionality reduction, and unsupervised dimensionality reduction [1, 2, 3, 4, 5]. These methods often rely on exploiting the inherent structure of the signal, to aid in the dimensionality reduction process, often implying use of a statistical signal model.

Numerous signal models have been proposed, often used when the data are inherently high-dimensional, but possess latent low-dimensional structure. Prominent examples include union-of-subspace models [6], wavelet trees [7], manifolds [8], and closely related Gaussian mixture models with (near) low-rank covariance matrices [9]. In addition to possessing a low-dimensional latent structure (*e.g.*, sparse support of wavelet coefficients), the signal of interest is often accompanied by additional information, with this termed *side information* in information theory [10]. The side information is often manifested in the form of another observed signal, possessing correlation with the principal signal of interest. One is often interested in leveraging the side information to enhance the classification and reconstruction of high-dimensional signals from low-dimensional features (*e.g.*, the signal of interest may be measured compressively [11], and we may also observe another correlated signal).

This problem is reminiscent of well-known challenges in source-coding with side information [12, 10] and distributed source coding [13]. In these problems, the number of extracted signal features may be related to the compression rate, and performance metrics for classification and reconstruction can be regarded as the distortion.

In this work, we propose a general framework for compressive linear-projection measurements with side information, and study the impact of side information on *classification* and *signal recovery* from low-dimensional measurements. The observed compressively measured signal is equivalent to representing the signal in terms of a small set of features, constituted by linear projections. We focus on two special cases of the general model. The first case is referred as the joint Gaussian mixture model (GMM), in which both the signal and side information are drawn from a joint GMM; it generalizes numerous previous models such as [14, 15]. The other example is termed the exponential family (EF) side information model, in which the signal and the side information are drawn from a GMM and a mixture in the exponential family, respectively. These particular examples have important practical applications, such as image reconstruction and text-aided image classification, as demonstrated in the experiments, and further motivated below.

2. SYSTEM MODEL AND PROBLEM STATEMENT

Let $\mathbf{x}_1 \in \mathbb{R}^{n_1}$ and $\mathbf{x}_2 \in \mathbb{R}^{n_2}$ denote two *correlated* signals. In the problem considered here, $\mathbf{x}_1$ is deemed the principal signal of interest and $\mathbf{x}_2$ is side information. Let $C_1 = \{1, \ldots, K_1\}$ and $C_2 = \{1, \ldots, K_2\}$ represent latent underlying indicator variables associated with $\mathbf{x}_1$ and $\mathbf{x}_2$, respectively. Variables C_1 and C_2 may represent class labels, as in a classification problem, and/or they may represent mixture components in a mixture model from which $\mathbf{x}_1$ and $\mathbf{x}_2$ are drawn. We assume that C_1 and C_2 are drawn from an arbitrary discrete joint distribution $p(C_1, C_2)$; $\mathbf{x}_1$ and $\mathbf{x}_2$ are drawn from a mixture model $p(\mathbf{x}_1, \mathbf{x}_2) =$

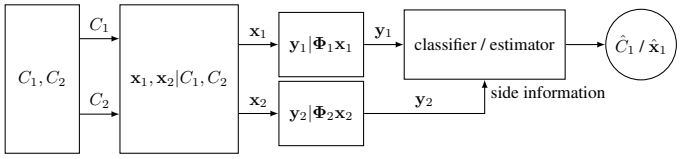

Fig. 1: General compressive feature extraction with side information model. The decoder attempts to formulate $\hat{C}_1$ of the index of the component from which the input signal $\mathbf{x}_1$ was drawn (classification) or it aims to generate an estimate $\hat{\mathbf{x}}_1$ of the input signal itself (reconstruction) based on the observation of both feature vectors $\mathbf{y}_1$ and $\mathbf{y}_2$.

$\sum_{i=1,k=1}^{K_1,K_2} p(C_1 = i, C_2 = k) p(\mathbf{x}_1, \mathbf{x}_2 | C_1 = i, C_2 = k)$. At this point $p(\mathbf{x}_1, \mathbf{x}_2 | C_1 = i, C_2 = k)$ is treated as general, and in Sections 3 and 4 we consider special cases with practical relevance (subsequently demonstrated in the experiments in Section 5). We assume that $p(\mathbf{x}_1, \mathbf{x}_2 | C_1, C_2)$ and $p(C_1, C_2)$ are known *a priori*, *e.g.*, can be estimated from training data. Rather than directly observing $\mathbf{x}_1$, we observe a compressed version $\mathbf{y}_1 \in \mathbb{R}^{m_1}$, where typically $m_1 \ll n_1$. Further, rather than directly observing side information $\mathbf{x}_2$, we may observe a compressed version $\mathbf{y}_2 \in \mathbb{R}^{m_2}$, with $m_2 \leq n_2$. Conditioned on $\mathbf{x}_1$ and $\mathbf{x}_2$, $\mathbf{y}_1$ and $\mathbf{y}_2$ are drawn independently from $p(\mathbf{y}_1; \mathbf{\Phi}_1 \mathbf{x}_1)$ and $p(\mathbf{y}_2; \mathbf{\Phi}_2 \mathbf{x}_2)$ *i.e.,* $\mathbf{y}_1, \mathbf{y}_2 | \mathbf{x}_1, \mathbf{x}_2 \sim p(\mathbf{y}_1; \mathbf{\Phi}_1 \mathbf{x}_1) p(\mathbf{y}_2; \mathbf{\Phi}_2 \mathbf{x}_2)$ where $p(\mathbf{y}_1; \mathbf{\Phi}_1 \mathbf{x}_1)$ and $p(\mathbf{y}_2; \mathbf{\Phi}_2 \mathbf{x}_2)$ denote respective distributions for $\mathbf{y}_1$ and $\mathbf{y}_2$ with parameters $\mathbf{\Phi}_1 \mathbf{x}_1$ and $\mathbf{\Phi}_2 \mathbf{x}_2$; $\mathbf{\Phi}_1 \in \mathbb{R}^{m_1 \times n_1}$ and $\mathbf{\Phi}_2 \in \mathbb{R}^{m_2 \times n_2}$ are *projection* matrices associated with $\mathbf{x}_1$ and $\mathbf{x}_2$, respectively. In Figure 1, we summarize the proposed model.

Given compressed measurement $\mathbf{y}_1$ and compressed side information $\mathbf{y}_2$, the decoder aims to formulate an estimate $\hat{\mathbf{x}}_1$ for $\mathbf{x}_1$; this is termed the "signal recovery problem". In particular, we consider a minimum mean square error (MMSE) estimator $\hat{\mathbf{x}}_1$ obtained by the conditional mean estimator that minimizes the reconstruction error:

$$\hat{\mathbf{x}}_1(\mathbf{y}_1, \mathbf{y}_2) = \mathbb{E}[\mathbf{x}_1 | \mathbf{y}_1, \mathbf{y}_2] = \int \mathbf{x}_1 p(\mathbf{x}_1 | \mathbf{y}_1, \mathbf{y}_2) d\mathbf{x}_1,$$

where $p(\mathbf{x}_1 | \mathbf{y}_1, \mathbf{y}_2)$ is the posterior density of $\mathbf{x}_1$ given the observations $\mathbf{y}_1$ and $\mathbf{y}_2$.

Alternatively, we may seek an estimate $\hat{C}_1$ for the latent indicator C_1, to identify the component from which $\mathbf{x}_1$ was drawn; this is termed the "classification problem". The minimum average error probability in classifying C_1 from $\mathbf{y}_1$ and $\mathbf{y}_2$ is achieved by the maximum *a posterior* (MAP) classifier [16], given by

$$\begin{aligned}\hat{C}_1 &= \arg \max_{i \in \{1,\ldots,K_1\}} p(C_1 = i | \mathbf{y}_1, \mathbf{y}_2) \\ &= \arg \max_i \sum_{k=1}^{K_2} p(C_1 = i, C_2 = k) p(\mathbf{y}_1, \mathbf{y}_2 | C_1 = i, C_2 = k),\end{aligned}$$

where $p(C_1 = i | \mathbf{y}_1, \mathbf{y}_2)$ is the posterior probability of class $C_1 = i$ conditioned on $\mathbf{y}_1$ and $\mathbf{y}_2$.

The above framework is general, and we now concentrate on two special cases, for which important applications apply. Specifically, in Section 3 we assume that $p(\mathbf{x}_1, \mathbf{x}_2 | C_1, C_2)$ is a multivariate Gaussian, with Gaussian compressive measurement models $p(\mathbf{y}_1 | \mathbf{\Phi}_1 \mathbf{x}_1)$ and $p(\mathbf{y}_2 | \mathbf{\Phi}_2 \mathbf{x}_2)$. As we demonstrate when presenting experiments, this has important applications in image processing, and associated compressive sensing of images, with image-based side information. In the second application, considered in Section 4, it is assumed that $\mathbf{x}_1$ is drawn from a GMM, and $\mathbf{x}_2$ is drawn from a mixture in the exponential family.

3. JOINT GMM SIGNAL MODEL

Conditioned on $(C_1, C_2) = (i, k)$, the source distribution is

$$p(\mathbf{x}_1, \mathbf{x}_2 | C_1 = i, C_2 = k) = \mathcal{N}(\boldsymbol{\mu}_{\mathbf{x}}^{(ik)}, \mathbf{\Sigma}_{\mathbf{x}}^{(ik)}), \quad (1)$$

where

$$\boldsymbol{\mu}_{\mathbf{x}}^{(ik)} = \begin{bmatrix} \boldsymbol{\mu}_{\mathbf{x}_1}^{(ik)} \\ \boldsymbol{\mu}_{\mathbf{x}_2}^{(ik)} \end{bmatrix}, \mathbf{\Sigma}_{\mathbf{x}}^{(ik)} = \begin{bmatrix} \mathbf{\Sigma}_{\mathbf{x}_1}^{(ik)} & \mathbf{\Sigma}_{\mathbf{x}_{12}}^{(ik)} \\ \mathbf{\Sigma}_{\mathbf{x}_{21}}^{(ik)} & \mathbf{\Sigma}_{\mathbf{x}_2}^{(ik)} \end{bmatrix}, \quad (2)$$

so that $p(\mathbf{x}_1 | C_1 = i, C_2 = k) = \mathcal{N}(\boldsymbol{\mu}_{\mathbf{x}_1}^{(ik)}, \mathbf{\Sigma}_{\mathbf{x}_1}^{(ik)})$ and $p(\mathbf{x}_2 | C_1 = i, C_2 = k) = \mathcal{N}(\boldsymbol{\mu}_{\mathbf{x}_2}^{(ik)}, \mathbf{\Sigma}_{\mathbf{x}_2}^{(ik)})$; $\boldsymbol{\mu}_{\mathbf{x}_1}^{(ik)}$ and $\mathbf{\Sigma}_{\mathbf{x}_1}^{(ik)}$ are the mean and covariance matrix of $\mathbf{x}_1$ conditioned on the pair of classes (i, k), respectively, and $\boldsymbol{\mu}_{\mathbf{x}_2}^{(ik)}$ and $\mathbf{\Sigma}_{\mathbf{x}_2}^{(ik)}$ play the same role for $\mathbf{x}_2$. The term $\mathbf{\Sigma}_{\mathbf{x}_{12}}^{(ik)}$ is the cross-covariance matrix between $\mathbf{x}_1$ and $\mathbf{x}_2$, conditioned on (i, k). We further assume that $p(\mathbf{y}_1 | \mathbf{\Phi}_1 \mathbf{x}_1) = \mathcal{N}(\mathbf{\Phi}_1 \mathbf{x}_1, \mathbf{I} \cdot \sigma^2)$ and $p(\mathbf{y}_2 | \mathbf{\Phi}_2 \mathbf{x}_2) = \mathcal{N}(\mathbf{\Phi}_2 \mathbf{x}_2, \mathbf{I} \cdot \sigma^2)$, where $\mathbf{I}$ denotes the identity matrix (the noise variance of these compressive measurements may readily be imposed as distinct for $\mathbf{y}_1$ and $\mathbf{y}_2$, and are made equal here for simplicity). Equivalently, we may compactly express the model of the observed data as

$$\mathbf{y} = \mathbf{\Phi}\, \mathbf{x} + \mathbf{w}, \qquad \mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix}, \mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix}, \mathbf{w} = \begin{bmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \end{bmatrix} \quad (3)$$

where

$$\mathbf{\Phi} = \begin{bmatrix} \mathbf{\Phi}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{\Phi}_2 \end{bmatrix}, \ \mathbf{w} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \cdot \mathbf{I}). \quad (4)$$

It is straightforward to see that the joint pdf of $\mathbf{x}_1$ and $\mathbf{x}_2$ follows a GMM. Hence, the MMSE estimator $\hat{\mathbf{x}}_1(\mathbf{y})$ and the MAP classifier $\hat{C}_1$ possess analytical expressions, akin to [9].

Let $\text{MMSE}(\sigma^2) = \mathbb{E}[\|\mathbf{x} - \hat{\mathbf{x}}_1(\mathbf{y})\|^2]$ denote the mean square error of the MMSE estimator $\hat{\mathbf{x}}_1$, which is a function of σ^2. We present a theoretical result fully characterizing the behavior of $\text{MMSE}(\sigma^2)$ in the low-noise regime, *i.e.,* when $\sigma^2 \to 0$.

Theorem 1. *Assume the joint GMM as summarized in* (1) *and* (3)*. If*

$$m_1 > r_{\mathbf{x}_1}^{(ik)} \text{ or } \begin{cases} m_1 > r_{\mathbf{x}}^{(ik)} - r_{\mathbf{x}_2}^{(ik)} \\ m_1 + m_2 > r_{\mathbf{x}}^{(ik)} \end{cases}, \forall (i, k) \in \mathcal{S}, \quad (5)$$

then, with probability 1*, we have* $\lim_{\sigma^2\to 0} MMSE(\sigma^2) = 0$*, where* $\mathcal{S} = \{(i,k) \in \{1,\dots,K_1\} \times \{1,\dots,K_2\} : p(C_1 = i, C_2 = k) > 0\}$*,* $r_{\mathbf{x}_1}^{(ik)} = rank(\mathbf{\Sigma}_{\mathbf{x}_1}^{(ik)})$*,* $r_{\mathbf{x}_2}^{(ik)} = rank(\mathbf{\Sigma}_{\mathbf{x}_2}^{(ik)})$ *and* $r_{\mathbf{x}}^{(ik)} = rank(\mathbf{\Sigma}_{\mathbf{x}}^{(ik)})$.

Conversely, if we $\lim_{\sigma^2\to 0} MMSE(\sigma^2) = 0$*, then, with probability* 1*, we have*

$$m_1 \geq r_{\mathbf{x}_1}^{(ik)} \text{ or } \begin{cases} m_1 \geq r_{\mathbf{x}}^{(ik)} - r_{\mathbf{x}_2}^{(ik)} \\ m_1 + m_2 \geq r_{\mathbf{x}}^{(ik)} \end{cases}, \forall (i,k) \in \mathcal{S}. \quad (6)$$

We note that the above theorem is derived from another theorem on the accuracy of $\hat{C}_1$, which is omitted due to the space limit. It is interesting to note that the necessary conditions for the phase transition of the MMSE are one feature away from the corresponding sufficient conditions. Theorem 1 provides a sharp characterization of the region associated to the phase transition of the MMSE of the joint GMM.

4. EXPONENTIAL FAMILY SIDE INFORMATION

We now assume $p(\mathbf{x}_1, \mathbf{x}_2|C_1, C_2) = p(\mathbf{x}_1|C_1, C_2)p(\mathbf{x}_2|C_1, C_2)$, where $p(\mathbf{x}_2|C_1 = i, C_2 = j)$ is a point mass, *i.e.,* $p(\mathbf{x}_2|C_1 = i, C_2 = j) = \delta(\mathbf{x}_2 = \mathbf{x}_2^{(ij)})$ and thus $p(\mathbf{x}_2) = \sum_{i=1}^{K_1}\sum_{j=1}^{K_2} p(C_1 = i, C_2 = j)\delta(\mathbf{x}_2 = \mathbf{x}_2^{(ij)})$. We further assume $p(\mathbf{x}_1|C_1 = i, C_2 = j) = \sum_{s=1}^{N_{ij}} \pi_s^{(ij)} \mathcal{N}(\boldsymbol{\mu}_s^{(ij)}, \mathbf{\Sigma}_s^{(ij)})$, *i.e.,* $\mathbf{x}_1|(C_1 = i, C_2 = j)$ is a GMM in which $\pi_s^{(ij)}$, $s = 1,\dots,N_{ij}$ are the GMM coefficients within class (i,j) and $\boldsymbol{\mu}_s^{(ij)}$ and $\mathbf{\Sigma}_s^{(ij)}$, $s = 1,\dots,N_{ij}$ are the means and variances of respective N_{ij} Gaussian components. Hence, $\mathbf{x}_1$ is drawn from a GMM, and in the experiments $\mathbf{x}_1$ will be composed of a concatenation of vectors, each of which is drawn i.i.d. from the associated GMM (patch model of an image [17]). The compressed measurement $\mathbf{y}_1$ is obtained via $\mathbf{y}_1 = \mathbf{\Phi}_1\mathbf{x}_1 + \mathbf{w}_1$ with $\mathbf{w}_1 \sim \mathcal{N}(\mathbf{0}; \sigma^2\mathbf{I})$.

The side information is modeled as $\mathbf{y}_2|\mathbf{x}_2 \sim \prod_{i=1}^{m_2} p_{ef}(\cdot\,; (\mathbf{\Phi}_2\mathbf{x}_2)_i)$, where $(\cdot)_i$ denotes the i-th entry of the argument vector and $p_{ef}(\cdot\,; x) = \exp\left(t(\cdot)\cdot\eta(x) - F(\eta) + k(\cdot)\right)$ denotes an arbitrary scalar exponential family (EF) distribution associated with the natural parameter $\eta(x)$, the sufficient statistics $t(\cdot)$, the log-normalizer $F(\eta)$ and the carrier measure $k(\cdot)$ [18].

It is easy to see that the MAP estimator $\hat{C}_1$ possesses an analytic expression. However, the MMSE estimator $\hat{\mathbf{x}}_1$ for this case does not possess an analytic expression in general, and a Monte-Carlo integration method can be utilized to calculate the conditional expectation $\mathbb{E}[\mathbf{x}_1|\mathbf{y}_1, \mathbf{y}_2]$ numerically [2]. We may alternatively resort to a sub-optimal scheme in which the MAP estimate $\hat{C}_1$ is manifested first to determine the GMM from which $\mathbf{x}_1$ was drawn; once C_1 is so estimated, $\mathbf{x}_1$ may be estimated analytically. Such a two-step approach has been considered in [19] for the model considered in Section 3, but not that of this section. A recovery performance bound in the low-noise regime can be consequently established with high probability via the results in [19], provided that the classification accuracy of the MAP classifier $\hat{C}_1$ is high.

Rather than seeking the exact misclassification probability P_{err}, which is intractable, we consider an upper bound $\bar{P}_{err}$ for P_{err} via the Bhattacharyya misclassification bound [16, 20]. An expression for $\bar{P}_{err}$ is presented via the following theorem, which also justifies the contribution of side information by showing that the upper bound of misclassification probability $\bar{P}_{err}$ always decreases in the presence of side information.

Theorem 2. *Assume the EF side information model as described above.* $\bar{P}_{err}$ *can be expressed as*

$$\bar{P}_{err} = \sum_{i=1}^{K_1}\sum_{\substack{j=1\\ j\neq i}}^{K_1} p_{C_1}(i) \sum_{k.\ell=1}^{K_2} \sqrt{p_{C_2|C_1}(k|i)p_{C_2|C_1}(\ell|j)} \times e^{-B_1-B_2}, \quad (7)$$

where B_1 *is a function of all the model parameters associated with* $\mathbf{y}_1$, i.e., $\pi^{(ij)}$, $\boldsymbol{\mu}^{(ij)}$, $\mathbf{\Sigma}^{(ij)}$, N_{ij}, $\mathbf{\Phi}_1$ *and* σ^2*.* B_2 *can be expressed as*

$$B_2 = \sum_{s=1}^{m_2} J_F(\eta((\mathbf{\Phi}_2\mathbf{x}_2^{(ik)})_s), \eta((\mathbf{\Phi}_2\mathbf{x}_2^{(jl)})_s)),$$

where $J_F(x,y) = \frac{1}{2}F(x) + \frac{1}{2}F(y) - F(\frac{1}{2}x + \frac{1}{2}y)$*.* F *and* η *are the log-normalizer and the natural parameter associated with the underlying exponential family, respectively.*

In particular, let $\bar{P}'_{err}$ *denote the Bhattacharyya misclassification upper bound when the side information is ignored,* i.e., *the classification is performed via the MAP classifier* $\hat{C}'_1 = \arg\max_{C_1} p(C_1|\mathbf{y}_1)$*. Then we have* $\bar{P}_{err} \leq \bar{P}'_{err}$*.*

5. EXPERIMENTS

We first apply proposed framework to the image recovery task. The principal signal is represented by the RGB "castle" image with size $480 \times 320 \times 3$, and side information is the *gray-scale* "castle" image with size 120×80, a low-resolution version of the same subject. Both images are partitioned into non-overlapping patches, so that the input data $\mathbf{x}_1$ represents $8 \times 8 \times 3$ patches extracted from the RGB image and $\mathbf{x}_2$ represents 2×2 patches extracted from the small gray-scale image.

The signals $\{\mathbf{x}_1, \mathbf{x}_2\}$ are assumed to follow the joint GMM as described in Section 3, with $K_1 = K_2 = 10$. The latent indicators C_1 and C_2 are assumed to be perfectly correlated, so that $p(C_1 = i, C_2 = k) = 0$, if $i \neq k$. The parameters of the joint GMM are learned from *other images* in Caltech 101 [21], via the expectation-maximization (EM) algorithm.

In order to fit the trained model to an exact low-rank GMM, to showcase further how Theorem 1 aligns with practice, we modify the covariance matrices $\mathbf{\Sigma}_{\mathbf{x}}$ by retaining

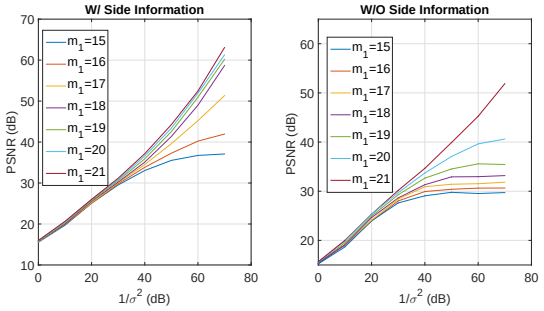

Fig. 2: Results of numerical experiments to demonstrate phase transition, depicting PSNR vs. $1/\sigma^2$ for reconstruction of the RGB image "castle" without and with side information ($\mathbf{\Phi}_2 = \mathbf{I}$).

only the first 20 principal components and setting to zero all remaining eigenvalues.

The compression is implemented by the projection $\mathbf{\Phi}_1 \in \mathbb{R}^{m_1 \times n_1}$, which has i.i.d. Gaussian entries with zero-mean and unit variance (as in [22]), and $\mathbf{\Phi}_2 = \mathbf{I}$. Figure 2 plots the PSNR vs. $1/\sigma^2$ curves obtained for different numbers of linear features extracted from the test image patches.

For comparison, we also show the corresponding curves for the case of reconstruction without side information [19]. We clearly observe how the phase transition for the reconstruction obtained with natural images matches the predictions obtained with the mathematical analysis developed in Section 3. Specifically, when side information is not available, the phase transition occurs at $m_1 > \max_{(i,k)} r_{\mathbf{x}_1}^{(ik)} = 20$ [19], which matches the right plot in Figure 2. On the other hand, the presence of side information allows for reliable reconstruction in the low-noise regime with a reduced number of features extracted from each patch. Namely, in the case under consideration, the conditions (6) are equivalent to $m_1 > 16$, which is shown to match well the PSNR behavior in the left plot of Figure 2. Thus, this natural image example has verified our Theorem 1.

The above example demonstrates that side information can help image reconstruction of the joint GMM (considered theoretically in Section 3). We now concentrate on another example: the classification problem of the EF side information model investigated theoretically Section 4. Specifically, we assume that the data $\mathbf{x}_1$ follows a GMM (for a patch model of an image) and side information $\mathbf{x}_2$ follows a *Poisson* distribution (corresponding to a count vector, associated with a text document).

Specifically, we consider the image classification problem with the text annotations of each image as side information. In particular, we consider the LabelMe data [23]. Five outdoor image classes: "coast", "highway", "insidecity", "mountain", "street" are used in our testing. Following the settings of images and annotations in [23], each patch (represented as a segment of the vector $\mathbf{x}_1$) is drawn from a GMM, and thus an image $\mathbf{x}_1$ is drawn from the product of these independent GMM components. We aim to perform the classification on images in the presence of these annotations side information.

We use the bag-of-words model [24] for the side information, and the collected word counts follow $\mathbf{y}_2 \sim \text{Pois}(\mathbf{\Phi}_2 \mathbf{x}_2)$. For each image class $C_1 = i$, $i = 1, \ldots, K_1$, we associate it with one Poisson rate vector $\{\mathbf{x}_2^{(C_2=i)}\}$, and $K_1 = K_2 = 5$ for this example. These Poisson rates can be considered as "topics" of the images within the same class and we assume that one image class corresponds to a unique topic. Therefore, the class relationship between the image and side information is modeled as $p(C_1, C_2) = \delta(C_1 = C_2)$.

We randomly select 100 images per class for training, and test the classification performance on the rest. For side information of training images, we consider $\mathbf{\Phi}_2 = \mathbf{I}$ and use the maximal likelihood (ML) estimator to obtain the Poisson rates $\{\mathbf{x}_2^{(C_2)}\}_{C_2=1}^{5}$.

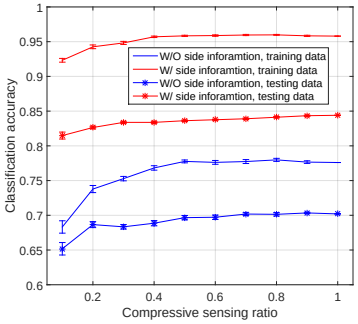

Fig. 3: Classification accuray against compression ratio, average of 10 trials.

The image patches are compressed via aforementioned random $\mathbf{\Phi}_1 \in \mathbb{R}^{m_1 \times n_1}$. The compression sensing ratio (CSr) is defined as $\text{CSr} = m_1/n_1$. By increasing the number of measurements, m_1, the classification accuracies with and without (non-compressed) side information are plotted in Figure 3. It can be seen that the classification accuracies improve significantly ($\sim 20\%$ for training and $> 15\%$ for testing) due to the presence of side information at various compression ratios, which is consistent with conclusion of Theorem 2.

6. CONCLUSION

We have developed a framework for compressive linear-projection measurements in the presence of side information. Classification and signal-recovery tasks have been addressed, based on the proposed framework. Motivated by real applications, two special examples of the general model have been considered. A joint GMM on the signal and side information has been utilized in the first example, and necessary and sufficient conditions for perfect signal recovery in the asymptotically low-noise regime have been derived. Further, a GMM signal with side information drawn from a mixture in the exponential family has been considered in the second example, and theoretical results quantifying misclassification probability have been presented. It has been demonstrated that the presence of side information yields improved performance relative to the cases for which such is absent.

7. REFERENCES

[1] M.W. Seeger and H. Nickisch, "Compressed sensing and Bayesian experimental design," in *International Conference on Machine Learning*, 2008.

[2] L. Wang, D. Carlson, M. Rodrigues, D. Wilcox, R. Calderbank, and L. Carin, "Designed measurements for vector count data," in *Neural Information Processing Systems (NIPS)*, 2013.

[3] L. Wang, D. Carlson, M. Rodrigues, R. Calderbank, and L. Carin, "A Bregman matrix and the gradient of mutual information for vector Poisson and Gaussian channels," *IEEE Transactions on Information Theory*, vol. 60, no. 6, 2014.

[4] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, 2001.

[5] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Jounral of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[6] Y. C. Eldar and M. Mishali, "Robust recovery of signals from a structured union of subspaces," *IEEE Transactions on Information Theory*, vol. 55, no. 11, pp. 5302–5316, 2009.

[7] T. Blumensath and M.E. Davies, "Sampling theorems for signals from the union of finite-dimensional linear subspaces," *IEEE Transactions on Information Theory*, vol. 55, no. 4, pp. 1872–1882, 2009.

[8] R. G. Baraniuk and M. B. Wakin, "Random projections of smooth manifolds," *Foundations of Computational Mathematics*, vol. 9, no. 1, pp. 51–77, 2009.

[9] M. Chen, J. Silva, J. Paisley, C. Wang, D. Dunson, and L. Carin, "Compressive sensing on manifolds using a nonparametric mixture of factor analyzers: Algorithm and performance bounds," *IEEE Transactions on Signal Processing*, vol. 58, no. 12, pp. 6140–6155, 2010.

[10] A.D. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Transactions on Information Theory*, vol. 22, no. 1, pp. 1–10, 1976.

[11] E.J. Candès, J. Romberg, and T. Tao, "Stable signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics*, vol. 59, no. 8, pp. 1207–1223, 2006.

[12] R. Ahlswede and J. Körner, "Source coding with side information and a converse for degraded broadcast channels," *IEEE Transactions on Information Theory*, vol. 21, no. 6, pp. 629–637, 1975.

[13] D. Slepian and J.K. Wolf, "Noiseless coding of correlated information sources," *IEEE Transactions on Information Theory*, vol. 19, no. 4, pp. 471–480, 1973.

[14] X. Wang and J. Liang, "Side information-aided compressed sensing reconstruction via approximate message passing," *arXiv preprint arXiv:1311.0576*, 2013.

[15] G. Chen, J. Tang, and S. Leng, "Prior image constrained compressed sensing (PICCS): a method to accurately reconstruct dynamic CT images from highly undersampled projection data sets," *Medical Physics*, vol. 35, no. 2, pp. 660–663, 2008.

[16] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification (2nd Edition)*, Wiley-Interscience, New York, NY, 2000.

[17] W.R. Carson, M. Chen, M.R.D. Rodrigues, R. Calderbank, and L. Carin, "Communications-inspired projection design with application to compressive sensing," *SIAM Journal Imaging Sciences*, vol. 5, no. 4, 2012.

[18] C. Bishop and N. Nasrabadi, *Pattern Recognition and Machine Learning*, vol. 1, Springer New York, 2006.

[19] F. Renna, R. Calderbank, L. Carin, and M. R. D. Rodrigues, "Reconstruction of signals drawn from a Gaussian mixture via noisy compressive measurements," *IEEE Transactions on Signal Processing*, vol. 62, no. 9, pp. 2265–2277, 2014.

[20] H. Reboredo, F. Renna, R. Calderbank, and M. R. D. Rodrigues, "Compressive classification of a mixture of Gaussians: analysis, designs and geometrical interpretation," *arXiv preprint arXiv:1401.6962v1*, 2014.

[21] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories," in *CVPR Workshop on Generative-Model Based Vision*, 2004.

[22] E. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 489–509, 2006.

[23] C. Wang, D. Blei, and L Fei-Fei, "Simultaneous image classification and annotation," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[24] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Jounral of Machine Learning Research*, vol. 3, pp. 993–1022, Mar. 2003.