A Bayesian Approach to Learning Low-Dimensional Signal Models from Incomplete Measurements

[1]Lawrence Carin, [2]Richard Baraniuk, [3]Volkan Cevher, [1]David Dunson, [4]Michael Jordan, [5]Guillermo Sapiro, [6]Michael Wakin

[1]Duke University; [2]Rice University; [3]Ecole Polytechnique Federale de Lausanne; [4]University of California, Berkeley; [5]University of Minnesota; [6]Colorado School of Mines
POC: lcarin@ece.duke.edu

I. Introduction

Sampling, coding and streaming even the most essential data, *e.g.,* in medical imaging and weather monitoring applications, now produce a data deluge that severely stresses the available analog-to-digital converter, communication bandwidth, and digital storage resources. Surprisingly, while the ambient data dimension is large in many problems, the relevant information in the data can reside in a much lower dimensional space. This observation has lead to several important theoretical and algorithmic developments under different low-dimensional modeling frameworks, such as compressive sensing [1], [2], matrix completion [3], [4], and general factor-model representations [5], [6]. These approaches have enabled new measurement systems, tools and methods for information extraction from dimensionality-reduced or incomplete data. A key aspect of maximizing the potential of such techniques is to develop appropriate data models, and in this paper we investigate this challenge from the perspective of nonparametric Bayesian analysis.

Before detailing the Bayesian modeling techniques, we review the form of the measurements. Specifically, we consider measurement systems based on *dimensionality reduction*, where we linearly project the signal of interest into a lower-dimensional space via

$$\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x} + \boldsymbol{\delta}. \tag{1}$$

The *signal* is $\boldsymbol{x} \in \mathbb{R}^d$, the *measurements* are $\boldsymbol{y} \in \mathbb{R}^{d'}$, $\boldsymbol{\Phi}$ is a $d' \times d$ matrix with $d' < d$, and $\boldsymbol{\delta}$ accounts for noise. Such a projection process loses signal information in general, since $\boldsymbol{\Phi}$ has a nontrivial null space. Hence, there has been significant interest over the last few decades in finding dimensionality reductions that preserve as much information as possible in the incomplete measurements $\boldsymbol{y}$ about certain signals $\boldsymbol{x}$. One way to preserve information is for $\boldsymbol{\Phi}$ to provide a *stable embedding* that approximately preserves pairwise distances between all signals in some set of interest. In some cases this property allows recovery of $\boldsymbol{x}$ from its measurements $\boldsymbol{y}$.

Geometric data models, such as sparsity, union-of-subspaces, manifolds, and mixture of factor

analyzers (MFAs) are at the core of low-dimensional modeling frameworks [7]. For instance, given a signal $\boldsymbol{x} \in \mathbb{R}^d$ and an appropriate basis $\boldsymbol{\Psi} \in \mathbb{R}^{d \times d}$, we can transform the signal as $\boldsymbol{x} = \boldsymbol{\Psi}\boldsymbol{\theta}$, where $\boldsymbol{\theta}$ is sparse or can be well-approximated as such, $i.e.$, it has only a few non-zero elements. Compressive sensing exploits this fact to recover signals from their compressive samples $\boldsymbol{y} \in \mathbb{R}^{d'}$, which are dimensionality reducing, non-adaptive random measurements. According to compressive sensing theory, the number of measurements for stable recovery is proportional to the signal sparsity (hence, $d' \ll d$), rather than to its Fourier bandwidth as dictated by the Shannon/Nyquist theorem. While signal recovery at such measurement rates is impressive, significant improvements can be achieved through the generalization of sparsity; for instance, union-of-subspace models encode dependencies among sparse coefficients; manifold models exploit smooth variations in the signals; and mixtures of factor analyzers combine the strength of both models using a mixture of low-rank Gaussians [5], [7], [8].

The existing results in signal recovery from compressive or incomplete measurements are predicated upon the knowledge of the appropriate low-dimensional signal model; a signal recovery algorithm relies on this model to locate the correct signal among all possible signals that can generate the same measurements. In this paper we consider the more difficult but more broadly applicable problem for which we must first learn the signal model from a set of training data. One can use this learned model subsequently to recover the underlying signal from compressive measurements. There are also examples for which we *jointly* learn the underlying model and recover the high-dimensional data, without any *a priori* training data; specifically, this is done when considering the image-interpolation problem (closely related to matrix completion), for which the underlying image is recovered based upon measurement of a small subset of pixels, selected uniformly at random.

The tools and methods we use to tackle the rich problems associated with learning low-dimensional signal models are based on probabilistic, nonparametric Bayesian techniques. By nonparametric, we mean that the number of parameters within the probabilistic models is beforehand unspecified. While it has been historically challenging to find workable prior distributions in the parameter space for such problems, we leverage beta, Bernoulli, Dirichlet, and Indian buffet processes. We observe that these distributions provide a nice scaffold for analytically managing posterior distributions given the set of training samples as well as observations. Additionally, we develop performance bounds for recovering high-dimensional data based upon incomplete measurements. We present several examples of how this technology may be used in practice in compressive sensing, matrix completion (when we recover a full low-rank matrix based upon a small number of randomly sampled matrix elements), and image interpolation based on highly incomplete measurements. These applications are of significant practical importance; for example matrix-completion techniques are of interest for automatic recommendation systems ($e.g.$, for movies, music, books, etc.).

The remainder of the paper is organized as follows. In Section II we provide a review of

dimensionality reduction and low-dimensional signal models. In Section III we review several signal models that enforce low-dimensional latent structure in the signals of interest. In Section IV nonparametric Bayesian statistical tools are reviewed, and it is explained how these may be applied to infer the signal models of interest. Section V summarizes how these models may be employed in practical applications, with example results presented to illustrate concepts. In Section VI theoretical performance guarantees are summarized for recovery of high-dimensional data based on compressive or incomplete measurements (with these bounds linked to the types of models we learn via nonparametric Bayesian analysis). We close with a discussion and conclusions in Section VII.

II. Stable Embeddings

We consider several classes of low-dimensional models for which the dimensionality reduction process (1) is *stable*. This means that we have not only the information preservation guarantee that $\boldsymbol{\Phi}\boldsymbol{x}_1 \neq \boldsymbol{\Phi}\boldsymbol{x}_2$ holds for all signal pairs $\boldsymbol{x}_1, \boldsymbol{x}_2$ belonging to the model set, but also the guarantee that if $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are far apart in $\mathbb{R}^d$ then their respective projections $\boldsymbol{\Phi}\boldsymbol{x}_1$ and $\boldsymbol{\Phi}\boldsymbol{x}_2$ are also far apart in $\mathbb{R}^{d'}$. This latter guarantee ensures robustness of the dimensionality reduction process to noise $\boldsymbol{\delta}$.

A requirement on the matrix $\boldsymbol{\Phi}$ that combines both the information preservation and stability properties for a signal model is the so-called ϵ-*stable embedding property*

$$(1-\epsilon)\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2^2 \leq \|\boldsymbol{\Phi}\boldsymbol{x}_1 - \boldsymbol{\Phi}\boldsymbol{x}_2\|_2^2 \leq (1+\epsilon)\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2^2 \tag{2}$$

which must hold for all $\boldsymbol{x}_1, \boldsymbol{x}_2$ in the model set. The interpretation is simple: a stable embedding approximately preserves the Euclidean distances between all points in a signal model.

A dimensionality reduction $\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x}$ from $\mathbb{R}^d$ down to $\mathbb{R}^{d'}$, $d' < d$, cannot hope to preserve all of the information in all signals $\boldsymbol{x} \in \mathbb{R}^d$, since it is impossible to guarantee that $\boldsymbol{\Phi}\boldsymbol{x}_1 \neq \boldsymbol{\Phi}\boldsymbol{x}_2$ holds for all signal pairs $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^d$. This is because there are infinitely many $\boldsymbol{x} + \boldsymbol{x}'$, with $\boldsymbol{x}'$ from the $(d-d')$-dimensional nullspace of $\boldsymbol{\Phi}$, that yield exactly the same measurement $\boldsymbol{y}$. However, by restricting our attention only to signals from a *low-dimensional model* that occupies a subset of $\mathbb{R}^d$, such an information preservation guarantee becomes possible, meaning that we can uniquely identify/recover any signal $\boldsymbol{x}$ in the model from its measurement $\boldsymbol{y}$.

Let us review three deterministic model classes that have been shown to support stable dimensionality reduction. First, a *sparse* signal $\boldsymbol{x} \in \mathbb{R}^d$ can be represented in terms of just $k \ll d$ nonzero coefficients in the basis expansion $\boldsymbol{x} = \boldsymbol{\Psi}\boldsymbol{\theta}$, where $\boldsymbol{\Psi}$ is a fixed basis. Concisely, we say that $\|\boldsymbol{\theta}\|_{\ell_0} = k$, where ℓ_0 is the pseudo-norm that merely counts the non-zero entries in $\boldsymbol{x}$. The set of all sparse signals Σ_k is the union of the $\binom{d}{k}$, k-dimensional canonical subspaces in $\mathbb{R}^d$ aligned with the coordinate axes of the basis $\boldsymbol{\Psi}$. For sparse signals, the stable embedding property (2) corresponds

to the Restricted Isometry Property (RIP) [9]. While the design of such a stable embedding is an NP-Complete problem, in general, it has been shown that any i.i.d. subgaussian random matrix $\mathbf{\Phi}$ stably embeds Σ_k into $\mathbb{R}^{d'}$ with high probability as long as $d' = O(k \log(d/k))$ [10]. Second, a *structured sparse* signal is not only sparse but also has correlated coefficients such that it lies on one of a subset of the $\binom{d}{k}$ subspaces of Σ_k [8]. As a result, a random dimensionality reduction $\mathbf{\Phi}$ is stable for a commensurately smaller value of d' than for a conventional sparse signal. Third, an ensemble of articulating signals often live on a *manifold*, in particular when the family of signals $\{\boldsymbol{x}_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ is smoothly parameterized by a k-dimensional parameter vector $\boldsymbol{\theta}$ [11]. The manifold dimension k is equal to the number of degrees of freedom of the articulation. It has been shown that a random dimensionality reduction $\mathbf{\Phi}$ stably embeds a k-dimensional smooth manifold from $\mathbb{R}^d$ into $\mathbb{R}^{d'}$ as long $d' = O(k \log(d))$ [12].

Given a stable embedding of the form (1), a number of techniques have been developed to recover a (structured) sparse signal of interest $\boldsymbol{x}$ from the measurements $\boldsymbol{y}$ including various sparsity-promoting convex optimizations [1], [2], [13], greedy algorithms [14], [15] and Bayesian approaches [16]–[18]. Recently, algorithms have also been developed that recover signal manifolds from randomized measurements [19]. The challenge this paper addresses concerns *learning* the underlying signal models, particularly for union-of-subspace and manifold models, with this learning performed nonparametrically based upon available data. The mixture-of-factor-analyzer (MFA) model discussed below is a statistical form of the union-of-subspace data model, and the MFA may also be used to approximate a manifold. Once these models are so learned, they may be used in algorithms that seek to recover high-dimensional data based on low-dimensional compressive measurements.

III. Learning Concise Signal Models

The existing results in signal recovery from compressive or incomplete measurements of the type discussed in Section II are predicated upon knowledge of the appropriate low-dimensional signal model. Starting in this section, we assume that we may not have access to the model but instead to *training data* representative of the signals of interest. Our goal is to learn a concise signal model from this data, enabling stable signal recovery. We design these models in a statistical manner, using nonparametric Bayesian techniques.

A. Union-of-subspaces model for sparse signals

Assume access to a set of N training data $\{\boldsymbol{x}_n\}_{n=1,N}$. Our goal is to infer a concise model for $\{\boldsymbol{x}_n\}_{n=1,N}$ appropriate for recovering high-dimensional data from compressive measurements. Further, we would like to learn the model parameters nonparamerically ($e.g.$, without having to

set the dimensionality of the subspaces or the number of mixture components). We express each $\boldsymbol{x}_n \in \mathbb{R}^d$ as

$$\boldsymbol{x}_n = \mathbf{A}(\boldsymbol{c}_n \circ \boldsymbol{b}_n) + \boldsymbol{\epsilon}_n \tag{3}$$

where $\boldsymbol{c}_n \in \mathbb{R}^K$, $\boldsymbol{b}_n \in \{0,1\}^K$, and $\circ$ denotes a pointwise or Hadamard vector product. The columns of the matrix $\mathbf{A} \in \mathbb{R}^{d\times K}$ define a *dictionary*, and in many cases $K > d$, such that $\mathbf{A}$ may be overcomplete. Because of the binary nature of $\boldsymbol{b}_n$ and because $\|\boldsymbol{b}_n\|_{\ell_0} < d$, each $\boldsymbol{x}_n$ is represented by a subset of the columns of $\mathbf{A}$ (defining a subspace); $\boldsymbol{\epsilon}_n$ is meant to represent the portion of $\boldsymbol{x}_n$ not contained within the aforementioned subspace.

If we assume that the components of $\boldsymbol{\epsilon}_n$ are drawn from $\mathcal{N}(0, \alpha_0^{-1}\mathbf{I}_d)$, where $\mathbf{I}_d$ represents the d-dimensional identity matrix, and if $\boldsymbol{c}_n \sim \mathcal{N}(0, \alpha^{-1}\mathbf{I}_K)$, then, after integrating out $\boldsymbol{c}_n$, $\boldsymbol{x}_n$ is drawn from

$$\boldsymbol{x}_n \sim \mathcal{N}(\mathbf{0}, \alpha^{-1}\mathbf{A}\boldsymbol{\Lambda}_n\mathbf{A}^T + \alpha_0^{-1}\mathbf{I}_d) \tag{4}$$

where $\boldsymbol{\Lambda}_n = \text{diag}(\boldsymbol{b}_n)$ is a binary diagonal matrix. Therefore, if the columns of $\mathbf{A}\boldsymbol{\Lambda}_n$ are linearly independent, and if r_n represents the number of nonzero components in $\boldsymbol{b}_n$, then $\boldsymbol{x}_n$ is drawn from a zero-mean Gaussian with approximate rank r_n (*approximate* because $\alpha_0^{-1}\mathbf{I}_d$, with generally small α_0^{-1}, is added to the rank-r_n $\alpha^{-1}\mathbf{A}\boldsymbol{\Lambda}_n\mathbf{A}^T$). Note that priors like $\boldsymbol{c}_n \sim \mathcal{N}(0, \alpha^{-1}\mathbf{I}_K)$, and other similar priors considered below, are typically selected for modeling convenience; the inferred posterior on such model parameters are not in general as simple as the prior ($e.g.$, they are typically not Gaussian).

One of our objectives is to learn the dictionary matrix $\mathbf{A}$, and in a Bayesian setting we place a prior on it. Specifically, a convenient prior is to draw the kth column of $\mathbf{A}$, $\boldsymbol{a}_k$, i.i.d. as

$$\boldsymbol{a}_k \sim \mathcal{N}(\mathbf{0}, \frac{1}{d}\mathbf{I}_d), \quad k = 1, \ldots, K \tag{5}$$

such that each column has unit expected norm and the columns have zero *expected* correlation. One typically also places gamma priors on the precisions α and α_0 (these priors are selected because of model conjugacy [20]).

The final part of the model involves placing a prior on the sparse matrix $\mathbf{B} \in \{0,1\}^{N\times K}$, with nth row defined by $\boldsymbol{b}_n$; the cumulative set of binary vectors $\{\boldsymbol{b}_n\}_{n=1,N}$ defines the total number of columns needed from $\mathbf{A}$. The prior we will employ for $\{\boldsymbol{b}_n\}_{n=1,N}$ is the beta-Bernoulli process, which is closely connected to the Indian buffet process [21] developed by Griffiths and Ghahramani; this is discussed in detail in Section IV-A. At this point we simply assume that an appropriate prior for $\mathbf{B}$ may be constituted.

As a first look at an application, to be discussed further in Section V, in Figure 1 the $\{\boldsymbol{x}_n\}_{n=1,N}$ correspond to N patches of pixels from an RGB image. In this problem only 20% of the pixels

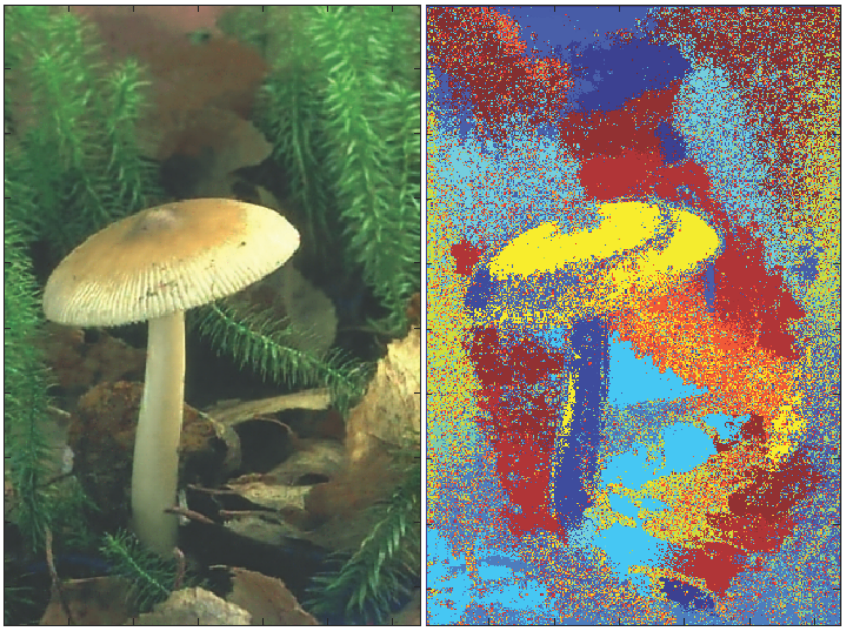

Fig. 1. Recovery of an RGB image based on measuring 20% of the voxels, uniformly at random. (a) Recovered image (PSNR=29.73), (b) local usage of BP dictionary elements, where the color denotes a specific usage of a subset of dictionary elements. [Results courtesy of J. Paisley.]

are observed, selected uniformly at random, and the model is used to infer the missing pixels in the image. In this analysis the incomplete data (image) are used as model inputs, to infer all model parameters, and importantly $\mathbf{A}$ and $\{\boldsymbol{c}_n \circ \boldsymbol{b}_n\}_{n=1,N}$. Note that the columns of $\mathbf{A}$ have the same support as $\boldsymbol{x}_n$, and hence they may be used to infer missing pixel values, via $\mathbf{A}(\boldsymbol{c}_n \circ \boldsymbol{b}_n)$. It is important that the pixels are missing at random; if the same pixel is missing in all $\{\boldsymbol{x}_n\}_{n=1,N}$, then it is impossible to learn the corresponding components in $\mathbf{A}$ (the row of $\mathbf{A}$ corresponding to this missing pixel cannot be inferred). Because the pixels are missing at random, information about a missing pixel may be inferred by using information from a similar patch elsewhere in the image. Hence, the simultaneous ("collaborative") analysis of all $\{\boldsymbol{x}_n\}_{n=1,N}$ allows one to infer information about missing pixels by exploiting observed versions of that pixel from similar patches (we are exploiting "self similarity" between image patches, which is typical of natural imagery). Inferring interrelationships between incomplete patches, with complementary missing pixels, is consequently critical to model success.

B. Mixture of factor analyzers model for signal ensembles

In the above model all data share the same dictionary defined by the columns of $\mathbf{A}$, but each sample $\boldsymbol{x}_n$ generally employs a subset of the dictionary elements, defined by the binary vector $\boldsymbol{b}_n$. When the number of samples N is large, it can be expected that many of the $\boldsymbol{b}_n$ will be the same or similar, defining a union of subspaces. This can be represented statistically as a mixture of Gaussians with covariance matrices that are nearly low-rank.

Specifically, we generalize (4) as

$$\boldsymbol{x}_n \sim \sum_{m=1}^{M} \nu_m \mathcal{N}(\boldsymbol{\mu}_m, \alpha_m^{-1}\mathbf{A}_m\boldsymbol{\Lambda}_m\mathbf{A}_m^T + \alpha_0^{-1}\mathbf{I}_d) \tag{6}$$

where $\sum_{m=1}^{M}\nu_m = 1$, $\boldsymbol{\Lambda}_m = \text{diag}(\boldsymbol{b}_m)$, with $\boldsymbol{b}_m \in \{0,1\}^K$ again a binary vector that selects particular columns of $\mathbf{A}_m$ to define the subspace spanned by the mth mixture component (in (3) there is a separate binary vector $\boldsymbol{b}_n$ for each sample $\boldsymbol{x}_n$, and now there is a related binary vector $\boldsymbol{b}_m$ associated with mixture component m). Note that the number of non-zero components in $\mathbf{A}_m$ may vary with m, implying that the dimensionality of the mixture components need not be the same. If for each m the number of non-zero components of $\boldsymbol{b}_m$ is small ($i.e.$, $\|\boldsymbol{b}_m\|_{\ell_0} \ll d$), then each mixture component $\mathcal{N}(\boldsymbol{\mu}_m, \alpha_m^{-1}\mathbf{A}_m\boldsymbol{\Lambda}_m\mathbf{A}_m^T + \alpha_0^{-1}\mathbf{I}_d)$ defines a relatively low-dimensional "pancake" in $\mathbb{R}^d$, with the number of principal dimensions in the mth associated subspace defined by $\|\boldsymbol{b}_m\|_{\ell_0}$. The means $\boldsymbol{\mu}_m$ locate the center of each pancake, and these are assumed drawn from $\mathcal{N}(\mathbf{0}, \beta_m^{-1}\mathbf{I}_d)$, with a gamma prior placed on β_m (again due to conjugacy).

The model (6) is called a *mixture of factor analyzers* (MFA) [22], and the nonzero columns of $\mathbf{A}_m\boldsymbol{\Lambda}_m$ define the factor loadings associated with the mth mixture component. When building an MFA model, a natural question concerns how many mixture components M are appropriate for the training data $\{\boldsymbol{x}_n\}_{n=1,N}$. One may use model-selection techniques to choose a single setting of M. Perhaps the most widely employed approach for choosing M is the Bayesian information criteria (BIC) [23]–[25]. Alternatively, below we consider nonparametric modeling, which yields a posterior distribution on M, and inference essentially performs model averaging across a weighted set of models with different M. This is implemented via the Dirichlet process [26] as summarized below in Section IV-A.

Note that in (3) the $\boldsymbol{b}_n$ select a subset of the columns of $\mathbf{A}$ for representation of $\boldsymbol{x}_n$, and one may expect that different $\boldsymbol{x}_n$ will (partially) share usage of these columns. In the mixture model of (6) the $\boldsymbol{b}_m$ selects which subset of the columns of $\mathbf{A}_m$ are used for the mth mixture component; the $\boldsymbol{b}_m$ therefore defines the dimensionality and the subspace of this mixture component. In general the subspaces spanned by $\mathbf{A}_m\boldsymbol{\Lambda}_m$ and $\mathbf{A}_{m'}\boldsymbol{\Lambda}_{m'}$ are different. Hence, (3) implies that the $\boldsymbol{x}_n$ are drawn from partially overlapping subspaces, without an explicit clustering; (6) explicitly clusters

the data, with the data in cluster m spanned by the non-zero columns of $\mathbf{A}_m\mathbf{\Lambda}_m$. The representation in (6) is of most interest when one wishes to approximate a data manifold as a mixture of low-rank Gaussians, with the number of mixture components and their characteristics ($e.g.$, ranks) inferred by the data.

A related model is the mixture of probabilistic principal component analyzers (MPPCA) framework of Tipping and Bishop [27]; MPPCA is similar to the proposed MFA, but in [27] one must set the dimensionality (rank) of each mixture component as well as the number of mixtures, where here this is inferred via nonparametric Bayesian inference. In [27] the authors achieve a point estimate of model parameters via expectation maximization (EM), where here we estimate a full posterior density function on model parameters.

C. Manifold models for signal ensembles

One intriguing use of the MFA model in (6) is for data living along a nonlinear k-dimensional manifold in $\mathbb{R}^d$. Locally, a k-dimensional manifold can be well approximated by its tangent plane, with the quality of this approximation depending on the local curvature of the manifold. Therefore, an MFA model as in (6) may be considered a candidate for manifold-modeled data, where the mean vectors $\boldsymbol{\mu}_m$ roughly correspond to points sampled from the manifold, the columns of $\mathbf{A}_m\mathbf{\Lambda}_m$ roughly span the k-dimensional local tangent spaces, the thickness parameter α_0^{-1} depends on the manifold curvature, and the weights ν_m reflect the density of the data across the manifold [28].

When an MFA model is used for recovering data of this type from compressive measurements, one will expect the recovered signal to draw only from a small number of MFA components. The recovered signal is therefore an affine combination of the columns of the few active $\mathbf{A}_m\mathbf{\Lambda}_m$. This is reminiscent of the classical compressive sensing problem in which an unknown signal must be recovered as a sparse superposition of vectors from some dictionary. Indeed, one could alternatively formulate the MFA recovery program using CS techniques [5] in which $\widehat{\boldsymbol{x}}$ is recovered as a sparse superposition of the columns of $\mathbf{A}_m$. A key consideration in this formulation, however, is that the set of selected columns may draw from only a few MFA components; this requirement is closely related to the notion of *block sparsity* which has been studied in compressive sensing. An example application of this framework is presented in Figure 2.

D. Matrix completion

As a final model, consider a matrix $\mathbf{M} \in \mathbb{R}^{d\times N}$ with $N \geq d$ (this can always be achieved by matrix transpose). Let the N columns of $\mathbf{M}$ constitute the set of vectors $\{\boldsymbol{x}_n\}_{n=1}^N$, where $\boldsymbol{x}'_n$ is the nth column manifested with randomly selected missing entries (in this problem $\boldsymbol{x}'_n = \mathbf{\Phi}_n\boldsymbol{x}_n$, where

Fig. 2. Sparse signal recovery performed on a mixture of factor analyzers (MFA) model, inferred based on training data using Dirichlet process (DP) and beta process (BP). The left-most images are the original data, and the other columns represent CS recovery, based on random compressive measurements. Results show performance when the number of CS measurements are 5%, ..., 30% of the total number of pixels in the image. In columns two through seven, the left figure employs the CS-recovery algorithm in [29], which does not exploit the MFA, and the right image is based on the learned MFA. [Results courtesy of M. Chen.]

now the rows of $\boldsymbol{\Phi}_n$ are randomly selected rows of the $d \times d$ identity matrix, with $\boldsymbol{\Phi}_n$ different for each n). If the matrix $\mathbf{M}$ is such that its columns satisfy the properties inherent to (4), specifically that each $\boldsymbol{x}_n$ resides approximately within a subspace defined by columns in matrices of the form $\mathbf{A}\boldsymbol{\Lambda}_n$, then the data-recovery technique discussed below in Section V-A may be applied directly to achieve matrix completion.

It is of interest to examine how such a procedure is related to conventional matrix-completion frameworks based on low-rank constructions [3], [30], [31]. In this context, assume that the matrix may be expressed as

$$\mathbf{M} = \sum_{k=1}^{d} \lambda_k b_k \boldsymbol{u}_k \boldsymbol{v}_k^T + \mathbf{E} \tag{7}$$

where $\lambda_k \in \mathbb{R}$, $b_k \in \{0, 1\}$, $\boldsymbol{u}_k \in \mathbb{R}^d$, and $\boldsymbol{v}_k \in \mathbb{R}^N$. We may again draw $\boldsymbol{u}_k \sim \mathcal{N}(\mathbf{0}, \frac{1}{d}\mathbf{I}_d)$, $\boldsymbol{v}_k \sim \mathcal{N}(\mathbf{0}, \frac{1}{N}\mathbf{I}_N)$, $\lambda_k \sim \mathcal{N}(0, \alpha^{-1})$, and each component of $\mathbf{E}$ drawn i.i.d. from $\mathcal{N}(0, \alpha_0^{-1})$. A

binary sparseness-promoting prior (see Section IV-A) may again be employed to define the binary vector $\boldsymbol{b} = (b_1, \ldots, b_d)$, thereby imposing a preference for low-rank constructions. Note that in this case, because it is assumed that each column of $\mathbf{M}$ is drawn from the same linear subspace, there is only a single $\boldsymbol{b}$ (so in this case we do not use a beta *process*); however, this model may clearly be generalized to the nonlinear case by making $\boldsymbol{b}$ a function of index n, as in (3). This shows that the matrix completion problem is closely linked to the inference of missing pixels in images.

Considering (7), let v_{kn} represent the nth component of $\boldsymbol{v}_k$. Then the nth column of $\mathbf{M}$ may be expressed as

$$\boldsymbol{x}_n = \mathbf{U}(\boldsymbol{c}_n \circ \boldsymbol{b}) + \boldsymbol{\epsilon}_n \tag{8}$$

where $\boldsymbol{\epsilon}_n$ represents the nth column of $\mathbf{E}$, $\mathbf{U} \in \mathbb{R}^{d \times d}$ has columns defined by $\boldsymbol{u}_k$, and $\boldsymbol{c}_n = (\lambda_1 v_{1n}, \ldots, \lambda_d v_{dn})$. Therefore, matrix completion based on a sparseness constraint of the form in (7) represents each column of $\mathbf{M}$ as being drawn from a *single* linear subspace spanned by the columns of $\mathbf{U}$, while the union-of-subspace construction [32] in (3), applied to the N columns of $\mathbf{M}$, is nonlinear in that each column $\boldsymbol{x}_n$ in general has its own subspace defined by the binary vector $\boldsymbol{b}_n$.

IV. Bayesian Nonparametric Inference

Based upon the above discussions, learning a model for concise representation of high-dimensional data requires the ability to infer the dimensionality of the subspace data reside in, with this defined by the number of columns needed in $\mathbf{A}$ and $\mathbf{U}$. Further, in the context of the MFA model, we require a means of inferring an appropriate number of mixture components. The former problem will be addressed using the beta-Bernoulli process. The latter will be addressed via the Dirichlet process. These nonparametric models represent special cases of a more general concept, the completely random measure. In the section below we first review the completely random measure, and then we show three examples for which it may be applied: for the beta process, the gamma process and the Dirichlet process. Finally, we explain how the beta process may be combined with a Bernoulli process to place a prior on the aforementioned matrix $\mathbf{B}$ (to infer the dimensionality of the subspace in which a signal resides), and how the Dirichlet process may be used to infer the appropriate number of mixture components in the MFA. For a thorough discussion of nonparametric Bayesian methods, the interested reader is referred to [33].

A. Completely random measures

The key idea of Bayesian nonparametrics is easily stated: one replaces classical finite-dimensional prior distributions with general stochastic processes. Recall that a stochastic process is an indexed

collection of random variables, where the index set may be infinite; thus, by using stochastic processes as priors we introduce an open-ended number of degrees of freedom in a model. For this idea to be useful in practical models, it is necessary for these stochastic processes to have simplifying properties, and in particular it is necessary that they combine in simple ways with the likelihoods that arise in common statistical models, so that posterior inference is feasible. One general approach to designing such stochastic processes is to make use of the notion of *completely random measures*, a class of objects that embody a simplifying independence assumption. We begin by presenting the general framework of completely random measures and then we show how to derive some particularly useful stochastic processes—the beta process and the Dirichlet process—from this framework. When learning the MFA, the beta process is used to infer the number of factor loadings (equivalently the rank) for each mixture component, while the Dirichlet process is used to infer the number of mixture components.

Letting Ω denote a measurable space endowed with a sigma algebra $\mathcal{A}$, a *random measure* G is a stochastic process whose index set is $\mathcal{A}$. That is, $G(A)$ is a random variable for each set A in the sigma algebra. A *completely random measure* G is defined by the additional requirement that whenever A_1 and A_2 are disjoint sets in $\mathcal{A}$, the corresponding random variables $G(A_1)$ and $G(A_2)$ are independent [34].

Kingman [34] presented a way to construct completely random measures based on the nonhomogeneous Poisson process. The construction runs as follows (see Figure 3 for a graphical depiction). Consider the product space $\Omega \otimes \mathbb{R}$, and place a product measure η on this space. Treating η as the rate measure for a nonhomogeneous Poisson process, draw a sample $\{(\omega_i, p_i)\}$ from this Poisson process. From this sample, form a measure on Ω in the following way:

$$G = \sum_{i=1}^{\infty} p_i \delta_{\omega_i}. \tag{9}$$

where δ_{ω_i} corresponds to a unit point measure concentrated at the parameter/"atom" ω_i. We refer to $\{\omega_i\}$ as the *atoms* of the measure G and $\{p_i\}$ as the *weights*.

Clearly the random measure defined in (9) is completely random because the Poisson process assigns independent mass to disjoint sets. The interesting fact is that all completely random processes can be obtained this way (up to a deterministic component and a Brownian motion).

1) Beta process: The beta process (BP) is an example of a completely random measure. In this case we define the rate measure η as a product of an arbitrary measure B_0 on Ω and an "improper" beta distribution on $(0, 1)$:

$$\eta(d\omega, dp) = cp^{-1}(1-p)^{c-1} dp\, B_0(d\omega), \tag{10}$$

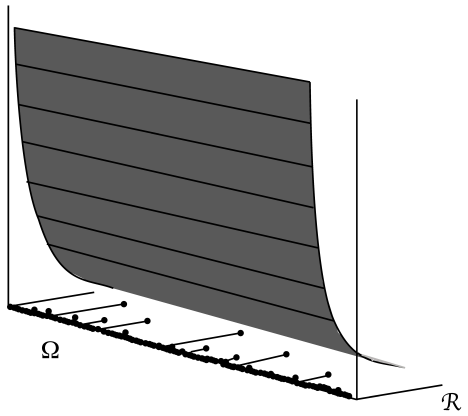

Fig. 3. Construction of a completely random measure on Ω from a nonhomogeneous Poisson process on $\Omega \otimes \mathbb{R}$.

where $c > 0$. Note that the expression $cp^{-1}(1-p)^{c-1}$ integrates to infinity; this has the consequence that a countably infinite number of points are obtained from the Poisson process.

We denote a draw from the beta process as follows:

$$B \sim \mathrm{BP}(c, B_0), \tag{11}$$

where $c > 0$ is referred to as a *concentration parameter* and where B_0 is the *base measure*.

For further details on this derivation of the beta process, see [35]. For an alternative derivation that does not make use of the framework of completely random measures, see [36]. Additional work on applications of the beta process can be found in [37]–[39].

2) Gamma process: As a second example, let the rate measure be a product of a base measure G_0 and an improper gamma distribution

$$\eta(d\omega, dp) = cp^{-1}e^{-cp}dp\, G_0(d\omega). \tag{12}$$

Again the density on p integrates to infinity, yielding a countably infinite number of atoms. The resulting completely random measure is known as the gamma process. We write:

$$G \sim \mathrm{GaP}(c, G_0) \tag{13}$$

to denote a draw from the gamma process. Note that the weights $\{p_i\}$ lie in $(0, \infty)$ and their sum is again finite.

3) Dirichlet process: It is also of interest to consider random measures that are obtained from completely random measures by normalization. For example, returning to the rate measure defining the gamma process in (12), let $\{(\omega_i, p_i)\}$ denote the points obtained from the corresponding Poisson

process. Form a random probability measure as follows:

$$G = \sum_{i=1}^{\infty} \pi_i \delta_{\omega_i}, \tag{14}$$

where $\pi_i = p_i / \sum_{j=1}^{\infty} p_j$. This is the Dirichlet process (DP) [26]. We denote a draw from the DP as $G \sim \text{DP}(\alpha, H_0)$, where $\alpha = G_0(\Omega)$ and $H_0 = G_0/\alpha$.

B. Application to data models

From Section III there are two principal modeling objectives: (i) an ability to infer the number of mixture components needed in an MFA, and (ii) the capacity to infer the number of needed factor loadings and their characteristics. Item (ii) is related to inferring the binary matrix $\mathbf{B}$ discussed in Section III-A. First considering (i), recall from above that a draw from a Dirichlet process may be expressed as $G = \sum_{i=1}^{\infty} \pi_i \delta_{\omega_i}$, where because of the aforementioned normalization $\sum_{i=1}^{\infty} \pi_i = 1$, and $\pi_i \geq 0$. In this application the atoms $\{\omega_i\}_{i=1,\infty}$ correspond to candidate mixture model parameters $(\alpha_m, \boldsymbol{\mu}_m, \mathbf{A}_m, \boldsymbol{\Lambda}_m)$ used in the mixture model of (6). Specifically, we constitute the following generative process for the data $\{\boldsymbol{x}_n\}_{n=1,N}$, when these data are assumed drawn from an MFA:

$$\begin{aligned} \boldsymbol{x}_n &\sim f(\alpha_n, \boldsymbol{\mu}_n, \mathbf{A}_n, \boldsymbol{\Lambda}_n, \alpha_0) \\ (\alpha_n, \boldsymbol{\mu}_n, \mathbf{A}_n, \boldsymbol{\Lambda}_n) &\sim G \\ G &\sim \text{DP}(\alpha, G_0) \end{aligned} \tag{15}$$

where $f(\cdot)$ represents the Gaussian distribution in (6). In this case the base measure G_0 from which the ω_i are drawn corresponds to a factorized prior for the set of parameters $(\alpha_n, \boldsymbol{\mu}_n, \mathbf{A}_n, \boldsymbol{\Lambda}_n)$, with the individual components of that prior as defined in Section III-B. A gamma prior is also placed on α_0. Note that because of the form of $G = \sum_{i=1}^{\infty} \pi_i \delta_{\omega_i}$ with $\sum_{i=1}^{\infty} \pi_i = 1$ and $\pi_i \geq 0$, the set of parameters $\{\alpha_n, \boldsymbol{\mu}_n, \mathbf{A}_n, \boldsymbol{\Lambda}_n\}_{n=1,N}$ are characteristic of being drawn from a mixture model. With probability π_i any particular set $(\alpha_n, \boldsymbol{\mu}_n, \mathbf{A}_n, \boldsymbol{\Lambda}_n)$ corresponds to ω_i. Therefore, although there are an infinite set of atoms in $G = \sum_{i=1}^{\infty} \pi_i \delta_{\omega_i}$, at most N of them will be used in the generative process, and typically fewer than N are needed, as often the same atom ω_i is shared among multple data samples $\{\boldsymbol{x}_n\}_{n=1,N}$. We therefore manifest a clustering of $\{\boldsymbol{x}_n\}_{n=1,N}$ (with data in the same cluster sharing a particular model parameter ω_i), and the model posterior density function allows inference of the number of mixture components needed. Therefore in principle the number of mixture components is unbounded, while in practice the model allows one to infer the finite number of mixture components needed to represent the data.

There is a so-called "Chinese restaurant process" (CRP) viewpoint of the Dirichlet process. The data are viewed as "customers", and the clusters are "tables", with the "dish" associated with a

given table manifested by the associated model parameters. One may explicitly draw from this CRP, by marginalizing out the Dirichlet process draw G [33].

We now consider the beta process as a prior for the binary vectors $\{\boldsymbol{b}_n\}_{n=1,N}$ in the model (4); these binary vectors are also of interest in the matrix-completion problem of Section III-D. Recall that a draw from a beta process may be expressed as $G = \sum_{i=1}^{\infty} \pi_i \delta_{\omega_i}$, where now each $\pi_i \in (0, 1)$. In this case each atom ω_i corresponds to a potential column of the matrix $\mathbf{A}$ in (4) or a potential column of $\mathbf{U}$ in (8). Therefore, in this case the base measure B_0 in the beta process corresponds to the same prior used for the columns of $\mathbf{A}$ or $\mathbf{U}$. The random variable π_i defines the probability that the ith column of $\mathbf{A}$ or $\mathbf{U}$ is used to represent the data of interest. Specifically, data sample $\boldsymbol{x}_n$ selects from among "dishes" in a "buffet", where the dishes correspond to the columns of $\mathbf{A}$ or $\mathbf{U}$. With probability π_i data $\boldsymbol{x}_n$ selects the ith column of $\mathbf{A}$ or $\mathbf{U}$, with this respective column denoted by the atom ω_i. Therefore the π_i defines the parameter of a Bernoulli distribution, with which a particular $\boldsymbol{x}_n$ decides which ω_i to use for data representation. This beta-Bernoulli process therefore defines a binary matrix $\mathbf{B} \in \{0, 1\}^{N \times \infty}$, where each row corresponds to a particular data sample $\boldsymbol{x}_n$ and the columns correspond to specific atoms $\{\omega_i\}_{i=1,\infty}$; hence, the rows of $\mathbf{B}$ are defined by $\{\boldsymbol{b}_i\}_{i=1,N}$, and sample $\boldsymbol{x}_n$ selects atom/"dish" ω_i if the ith component of $\boldsymbol{b}_n$ is equal to one. While $\mathbf{B}$ has an infinite number of columns in principle, it can be shown that only a finite number of columns in each row will have non-zero values [21]; in practice one may truncate the model to K columns/atoms, for large K.

The beta-Bernoulli process yields a so-called "Indian buffet process" (IBP) [33] if the beta-process draw is marginalized out. In this construction the data are again customers, and the model parameters are dishes at a buffet. Each customer sequentially selects parameters from the buffet, where the binary vector $\boldsymbol{b}_n$ for customer/data n defines which dishes/parameters are selected; if the kth component of $\boldsymbol{b}_n$ is one, then the kth parameter is used by data n, and if the kth component of $\boldsymbol{b}_n$ is zero the kth parameter is not used.

C. Posterior inference

Markov chain Monte Carlo (MCMC) procedures provide the dominant approach to inference with random measures. In such methods one approximates the posterior distribution of all model parameters in terms of a set of parameter-vector samples. These samples yield an ensemble of models, and the relative frequency of samples approximates the posterior distribution. In this manner one need not explicitly compute the high-dimensional integrals that would be required of a direct evaluation of the posterior distribution.

A special case of MCMC is Gibbs sampling, for which samples from the posterior distribution are drawn by sequentially sampling from conditional distributions. By appropriate design of the model,

of the form discussed above, these conditional distributions may often be expressed analytically. As an example of such samplers, consider the DP in particular. Working the marginal distribution embodied in the CRP, the core problem is to sample the seating assignment of a single customer conditioning on the seating assignments of the remaining customers. By exchangeability, one can pretend that this customer is the last to arrive in the restaurant, and the contribution of the prior to the seating assignment becomes the following rule: the customer sits at a table with probability proportional to the number of customers at that table. Multiplying this prior by a likelihood term one obtains a conditional probability that can be sampled. Similarly, in models based on the BP, one can work with the marginal distribution embodied in the IBP, and sampling the sparse binary vector associated with a data point by pretending that that data point is the last to arrive in the restaurant.

The insight of exploiting exchangeability in inference for random measures is due to Escobar [40], and a large literature has emerged. See Neal [41] for a thorough discussion in the case of the DP. Another direction of research on inference has involved working directly with the random measures rather than the marginals obtained from these random measures. There have been two main approaches: (1) truncate the random measure by limiting the random measure to a fixed number of atoms that is larger than any value expected to arise during sampling [42]; and (2) use slice sampling to adaptively truncate the random measure [43]. See [44] for a discussion of these methods in the setting of the BP, and for pointers to literature on variational approaches to inference for random measures.

V. Applications

To illustrate the broad applicability and high performance of the above approach to learning concise signal models, we consider several representative examples.

A. Pixel/voxel recovery via union of subspaces

We first consider an application of the union of subspaces model from (4), in which it is assumed that the data $\mathcal{D} = \{\boldsymbol{x}_n\}_{n=1,N}$ are constituted from pixels/voxels in an image. Specifically, each of the N image "patches" is defined by a set of *contiguous* pixels, with $\boldsymbol{x}_n \in \mathbb{R}^d$ representing data from the nth patch (it is possible that patches may overlap). For a color image one often considers $d = 8 \cdot 8 \cdot 3 = 192$, corresponding to the RGB components of an 8×8 image patch.

From (3), note that $\boldsymbol{x}_n$ is defined by the matrix $\mathbf{A} \in \mathbb{R}^{d \times K}$, by the *sparse* vector $\boldsymbol{c}_n \circ \boldsymbol{b}_n$ (with r_n non-zero components), and by $\boldsymbol{\epsilon}_n \in \mathbb{R}^d$. We assume that $\boldsymbol{\epsilon}_n$ may be made negligibly small, which implies $\alpha_0 \gg \alpha$ (to be demonstrated in the experiments). Since $\mathbf{A}$ is shared for all vectors

in $\mathcal{D}$, the total number of real model components needed is $dK + \sum_{n=1}^{N} r_n$, *if* the $\{\boldsymbol{b}_n\}_{n=1,N}$ are known ($i.e.$, if it is assumed we know *which* columns of $\mathbf{A}$ are associated with each $\boldsymbol{x}_n$, with this clearly impossible in practice). Nevertheless, under these assumptions, note that we have Nd real numbers in $\mathcal{D}$ available for computation of $dK + \sum_{n=1}^{N} r_n$ real model parameters. Therefore, if $N \gg K$ and $r_n \ll d$, and if we process all $\{\boldsymbol{x}_n\}_{n=1,N}$ jointly to infer the cumulative set of model parameters (exploiting the fact that $\mathbf{A}$ is shared among all), it appears we have more data in $\mathcal{D}$ than needed. Further, it would appear that we have enough data to also infer which among the r_n columns of $\mathbf{A}$ are needed for representation of each $\boldsymbol{x}_n$ (and therefore we do *not* need *a priori* access to $\{\boldsymbol{b}_n\}_{n=1,N}$).

Based upon the above observations, researchers have recently assumed access to only a subset of the components of each $\boldsymbol{x}_n$, with the observed components selected uniformly at random [38], [45], [46]. For example, rather than measuring all $d = 192$ contiguous pixels in an $8 \times 8 \times 3$ patch, a fraction of the pixels are measured, with the measured subset of pixels selected uniformly at random. Let $\mathcal{D}' = \{\boldsymbol{x}'_n\}_{n=1,N}$ represent a *modified* form of $\mathcal{D}$, with each $\boldsymbol{x}'_n$ defined by a fraction of the components of each $\boldsymbol{x}_n$, with observed samples selected uniformly at random. Processing all of the data in $\mathcal{D}'$ jointly ("collaboratively"), it has been demonstrated that for real, natural images one may indeed recover the missing data accurately, even when downsampling $\mathcal{D}$ significantly. Further, the compressive measurements may be performed very simply: by just randomly sampling/measuring the pixels/voxels in *existing* cameras (no need to develop new compressive-sampling cameras). An example is shown in Figure 1.

Note that this looks *like* compressive sensing [47], [48], in that a small subset of measurements are performed, with the full data recovered based upon exploitation of properties of the signal (that the signals live in a low-dimensional subspace of $\mathbb{R}^d$). However, in compressive sensing it is typically assumed that projection-type measurements are performed, and that the signal is sparse in an underlying *known* basis or frame. The projections should be incoherent with the basis vectors [49], and for a DCT-type basis one could use delta-function-like projections (selecting random components of each $\boldsymbol{x}_n$), like those considered above. However, in the above collaborative-filtering framework, not only do we perform random sampling, we also *infer* the underlying union of subspaces in which the signals reside, as defined by the columns of $\mathbf{A}$, thereby matching the signal subspace to the observed data, adaptively. In fact, as discussed in Section V-C, collaborative filtering for image recovery is closer to the field of matrix completion [3] than it is to compressive sensing.

The model in (4) is well suited for recovering the missing components of $\mathcal{D}$ from $\mathcal{D}'$ [38]. Specifically, when performing computations for the posterior distribution of the model parameters, the likelihood function represented by $\prod_{n=1}^{N} \mathcal{N}(\boldsymbol{x}'_n|\mathbf{0}, \alpha^{-1}\mathbf{A}\boldsymbol{\Lambda}_n\mathbf{A}^T + \alpha_0^{-1}\mathbf{I}_d)$ is simply evaluated at the pixels for which data are observed. As discussed in Section IV-C, a Gibbs sampler may be implemented, and from this one may obtain an approximation to all model parameters. Hence, the

posterior probability of each $\boldsymbol{x}_n$ in $\mathcal{D}$, based on observed $\mathcal{D}'$ and model hyperparameters Θ, may be expressed as

$$p(\boldsymbol{x}_n|\mathcal{D}',\Theta)=\int_{\mathbf{A}}\int_{\alpha}\int_{\boldsymbol{\Lambda}_n}\int_{\alpha_0}\mathcal{N}(\boldsymbol{x}_n|\mathbf{0},\alpha^{-1}\mathbf{A}\boldsymbol{\Lambda}_n\mathbf{A}^T+\alpha_0^{-1}\mathbf{I}_d)p(\mathbf{A},\alpha,\boldsymbol{\Lambda}_n,\alpha_0|\mathcal{D}',\Theta) \tag{16}$$

with the posterior $p(\mathbf{A},\alpha,\boldsymbol{\Lambda}_n,\alpha_0|\mathcal{D}',\Theta)$ approximated via samples from the Gibbs computations, and the integrals are approximated as sums.

A model like that in (3) or (6) has a posterior on model parameters that is invariant to exchanging the order of the data $\{\boldsymbol{x}_n\}_{n=1,N}$. In other words, any permutation of the order of the data will yield exactly the same inferred model parameters. This implies that the model is not utilizing all available prior information, since if one reconstituted an image after permuting the order of $\{\boldsymbol{x}_n\}_{n=1,N}$, very distinct images are manifested (recall that each $\boldsymbol{x}_n$ corresponds to an $8\times 8\times 3$ patch of contiguous pixels, and reordering these patches causes significant changes to the overall image). It is therefore desirable to impose within the model that if $\boldsymbol{x}_n$ and $\boldsymbol{x}_{n'}$ are spatially proximate, that they will likely employ similar factors (manifested by similar binary factor-selection vectors $\boldsymbol{b}_n$ and $\boldsymbol{b}_{n'}$).

Toward this end, we utilize DP in an additional manner (beyond within the MFA), to exploit spatial information. Specifically, we cluster the image patches spatially using a DP and impose that if two patches are spatially proximate, they are likely to be drawn from the same Gaussian mixture component, from the spatial mixture component. Figure 1(b) uses a different color to represent each Gaussian mixture component, and effective spatial segmentation is realized. One may therefore envision extending this framework for simultaneous image recovery and segmentation based upon randomly subsampled images.

As another example of this type, consider Figure 4. In this example, rather than processing all possible (overlapping) image patches at once, we select a subset of them for analysis; the approximate posterior on model parameters so inferred is used as a prior for the next randomly selected subset of patches for analysis. In Figure 4 the PSNR curve shows how the model performance improves as we consider more data, in this sequential manner. Each analysis of a subset of the image patches is termed a "learning round".

Theoretically, one would expect to need thousands of Gibbs iterations to achieve convergence. However, our experience is that even a *single* iteration in each of the above B^2 rounds yields good results. In Figure 4 we show the PSNR as a function of each of the 64 rounds discussed above. For Gibbs rounds 16, 32 and 64 the corresponding PSNR values were 27.64 dB, 28.20 dB and 28.66 dB. For this example we used $K=256$. This example was considered in [50]; the best results reported there were a PSNR of 29.65 dB. However, to achieve those results a training data set was employed for initialization [50]; the BP results are achieved with no *a priori* training data.

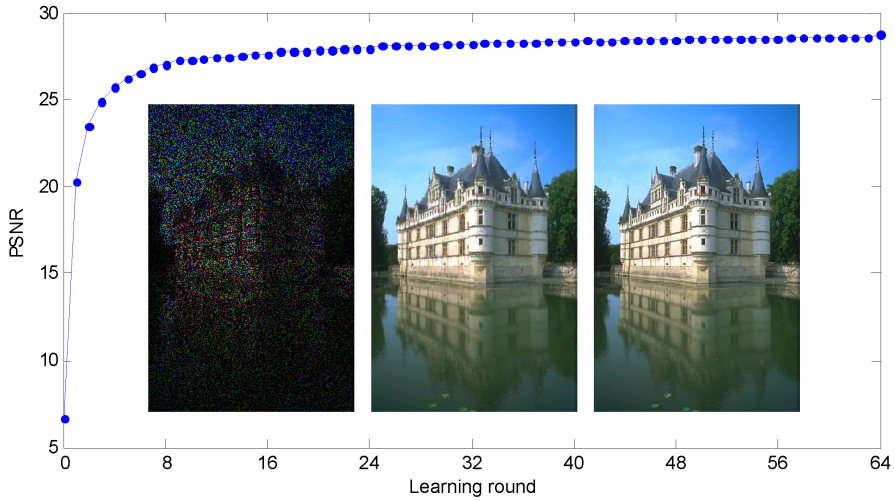

Fig. 4. Inpainting results. The curve shows the PSNR as a function of the 64 Gibbs learning rounds. The left figure is the test image, with 80% of the RGB pixels missing, the middle figure is the result after 64 after Gibbs rounds (final result), and the right figure is the original uncontaminated image. [Results courtesy of M. Zhou.]

B. Signal recovery from MFAs and manifolds

Assume that it is known *a priori* that the data of interest are drawn from an MFA of the form in (6), with the MFA learned "offline" based upon training data $\mathcal{D} = \{\boldsymbol{x}_n\}_{n=1,N}$. We now wish to measure a *single* new $\boldsymbol{x} \in \mathbb{R}^d$, under the assumption that $\boldsymbol{x}$ is drawn from the same MFA [5]. Since the MFA may be used to approximate a manifold, we may also consider the case for which $\boldsymbol{x}$ is drawn from a known manifold [12]. Based upon this prior knowledge, we wish to measure $\boldsymbol{y} \in \mathbb{R}^{d'}$, with $d' < d$, and ideally with $d' \ll d$; based upon the measured $\boldsymbol{y}$ we wish to recover $\boldsymbol{x}$.

It is assumed that $\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x}$, where $\boldsymbol{\Phi} \in \mathbb{R}^{d' \times d}$ is a projection matrix, typically defined randomly. In Section VI we discuss the desired properties of $\boldsymbol{\Phi}$, and the connection of such to the characteristics of the MFA. In a statistical sense, to recover $\boldsymbol{x}$ from $\boldsymbol{y}$ we desire $p(\boldsymbol{x}|\boldsymbol{y}) = p(\boldsymbol{y}|\boldsymbol{x})p(\boldsymbol{x})/p(\boldsymbol{y})$, under the assumption that $p(\boldsymbol{x})$ is the known MFA of the form in (6). Assuming that the compressive measurements are noisy, we may express $\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x} + \boldsymbol{\delta}$, where $\boldsymbol{\delta} \in \mathbb{R}^{d'}$ represents additive noise. If $\boldsymbol{\delta} \sim \mathcal{N}(\boldsymbol{0}, \beta_0^{-1}\mathbf{I}_{d'})$, then we have $p(\boldsymbol{y}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{\Phi}\boldsymbol{x}, \beta_0^{-1}\mathbf{I}_{d'})$. If β_0 is known, under the MFA assumption for $p(\boldsymbol{x})$, the expression $p(\boldsymbol{x}|\boldsymbol{y})$ may also be expressed *analytically* in terms of a mixture of Gaussians. If needed we may also infer β_0, by placing a (conjugate) gamma prior on it. Therefore under the assumption of the MFA model for $p(\boldsymbol{x})$, one may readily constitute a statistical estimate of $\boldsymbol{x}$ based on observing $\boldsymbol{y}$, with performance bounds discussed in Section VI. An example of CS recovery for images that live on a union of subspaces is shown in Figure 2; we are unaware of such CS inversion being performed by any previous method.

TABLE I
RMSE OF UNION-OF-SUBSPACE MODEL ON 10M MOVIELENS DATA. [RESULTS COURTESY OF M. ZHOU.]

Methods	r_a partition	r_b partition
User Profile	0.8749 ± 0.0009	0.8328 ± 0.0004
Movie Profile	0.8676 ± 0.0006	0.8323 ± 0.0002

C. Matrix completion

As our final example we consider the problem of matrix completion, as applied to movie-rating matrices. We tested the union-of-subspace construction on the widely employed 10M MovieLens dataset (10,681 movies by 71,567 users). Table I shows the results for both the r_a and r_b partitions provided with the data, in which 10 ratings per user are held out for testing. One of the best competing algorithms is the Gaussian process latent-variable model (GP-LVM) [51]. Averaged over both partitions, the GP-LVM reports the RMSE of 0.8740 ± 0.0278 using a 10 dimensional latent space, while the baselines of our approaches achieve average RMSEs of 0.8539 ± 0.0298 and 0.8499 ± 0.0250. In this example we employed the model in (3), in two constructions. In Table I we show results when the vectors $\boldsymbol{x}_n$ correspond to the user-dependent rankings of all movies (User Profile), and with $\boldsymbol{x}_n$ corresponding to the movie-dependent rankings manifested by all people (Movie Profile). In other words, one construction is in terms of the rows of the ranking matrix, and the other construction is in terms of the columns, with state-of-the-art results manifested in each case. While the Bayesian models may readily be extended to integer observed matrices via a probit or logistic link function, here the integer values are simply approximated as real numbers.

These results were computed using a Gibbs sampler, with a truncated beta-process implementation with $K = 256$ "dishes." One Gibbs iteration required 150 seconds on a 2.53GHz E5540 Xeon processor, using non-optimized Matlab software. The results in Table I correspond to 50 burn-in iterations and 100 collection iterations.

VI. PERFORMANCE GUARANTEES

The BP and DP nonparametric methods may be used to infer an MFA based upon given training data. Once this model is so learned, the MFA may be assumed known, and can be used in the inversion of subsequent compressive measurements. An example of such MFA learning, and subsequent utilization within compressive-sensing (CS) signal recovery, was presented in Figure 2. It is of interest to examine performance guarantees based on CS measurements and a known MFA model (learned based on training data, using nonparametric techniques of the type discussed above). It should be emphasized that the underlying MFA for general data is typically not identifiable, or unique. This implies that multiple MFAs may provide similar generative models for the underlying

data of interest. For the following bounds we assume one learned MFA is used to perform CS inversion, and that this model provides an accurate statistical representation for the data; it is for such a learned MFA that the bounds are constituted.

A. Bounds for MFAs

Recall our expression for compressive measurements $\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x} + \boldsymbol{\delta} \in \mathbb{R}^{d'}$ of a signal $\boldsymbol{x} \in \mathbb{R}^d$ as in (1). As discussed in Section V-B, if we assume that $\boldsymbol{x}$ is drawn from an MFA of the form in (6) whose parameters are known (based on training data), then the posterior distribution $p(\boldsymbol{x}|\boldsymbol{y})$ can be expressed as a mixture of Gaussians [5]. Using this model, we obtain an analytical expression for the mean estimate of $\boldsymbol{x}$:

$$\widehat{\boldsymbol{x}} = \sum_{m=1}^{M} \widehat{\nu}_m \widehat{\boldsymbol{x}}_m, \tag{17}$$

where

$$\widehat{\nu}_m = \frac{\nu_m \mathcal{N}(\boldsymbol{y}; \boldsymbol{\Phi}\boldsymbol{\mu}_m, \beta_0^{-1}\mathbf{I}_{d'} + \boldsymbol{\Phi}\boldsymbol{\Omega}_m\boldsymbol{\Phi}^T)}{\sum_{\ell=1}^{M} \nu_\ell \mathcal{N}(\boldsymbol{y}; \boldsymbol{\Phi}\boldsymbol{\mu}_\ell, \beta_0^{-1}\mathbf{I}_{d'} + \boldsymbol{\Phi}\boldsymbol{\Omega}_\ell\boldsymbol{\Phi}^T)}$$

represents the estimated mixture weight of the mth component,

$$\widehat{\boldsymbol{x}}_m = \boldsymbol{\Omega}_m\boldsymbol{\Phi}^T(\beta_0^{-1}\mathbf{I}_{d'} + \boldsymbol{\Phi}\boldsymbol{\Omega}_m\boldsymbol{\Phi}^T)^{-1}(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\mu}_m) + \boldsymbol{\mu}_m$$

equals the signal estimate that would be recovered if only component m were present in the MFA, and $\boldsymbol{\Omega}_m = \alpha_m^{-1}\mathbf{A}_m\boldsymbol{\Lambda}_m\mathbf{A}_m^T + \alpha_0^{-1}\mathbf{I}_d$ represents the covariance matrix of the mth component in the MFA. We can also consider the situation where $\alpha_0^{-1} \to 0$ and $\beta_0^{-1} \to 0$, in which case the matrix inverse in (17) should be treated as a pseudoinverse.

For an MFA being used for manifold-modeled data, an analogous requirement to the stable embedding is that (2) holds for $\boldsymbol{\mu}_{m_1} - \boldsymbol{\mu}_{m_2}$ for all $1 \leq m_1, m_2 \leq M$ and that (2) also holds for all vectors in $\mathrm{colspan}(\mathbf{A}_m\boldsymbol{\Lambda}_m)$ for all $1 \leq m \leq M$. From the Johnson-Lindenstrauss Lemma, we know that when $\boldsymbol{\Phi}$ is generated randomly with i.i.d. Gaussian or subgaussian entries, the former property holds with high probability as long as $d' = O(\log(M)\epsilon^{-2})$, and using similar arguments, the latter property also holds as long as $d' = O((k + \log(M))\epsilon^{-2})$ [10].

Under the assumption that these two conditions are met, we can establish certain guarantees [28] about the performance of the mean estimator (17) when recovering a signal $\boldsymbol{x}$ that is drawn from the manifold. For example, supposing that $\beta_0^{-1} \to 0$, the isometry property for $\mathrm{colspan}(\mathbf{A}_m\boldsymbol{\Lambda}_m)$ discussed above essentially guarantees that $\|\boldsymbol{x} - \widehat{\boldsymbol{x}}_m\|_2$ is a combination of two error terms, one depending on the size of $\boldsymbol{x} - \boldsymbol{\mu}_m$ when projected onto $\mathrm{colspan}(\mathbf{A}_m\boldsymbol{\Lambda}_m)$, and one depending on the size of $\boldsymbol{x} - \boldsymbol{\mu}_m$ when projected orthogonal to $\mathrm{colspan}(\mathbf{A}_m\boldsymbol{\Lambda}_m)$. The size of α_0^{-1} controls the balance between these two terms, and by choosing α_0^{-1} sufficiently small, we can ensure that the

dependence on the first of these terms is small; this allows us to guarantee that $\|\boldsymbol{x} - \widehat{\boldsymbol{x}}_m\|_2$ is small for any signal $\boldsymbol{x}$ living near mixture component m. Analysis of the recovery error $\|\boldsymbol{x} - \widehat{\boldsymbol{x}}\|_2$ for a multi-component mean estimator (17) is more involved but can proceed based on the observation that for any $m_0 \in \{1, 2, \ldots, M\}$, we can write $\|\boldsymbol{x} - \widehat{\boldsymbol{x}}\|_2 \leq \|\boldsymbol{x} - \widehat{\boldsymbol{x}}_{m_0}\|_2 + \sum_{m \neq m_0} \widehat{\nu}_m \|\boldsymbol{x} - \widehat{\boldsymbol{x}}_m\|_2$. One conclusion that can be drawn from this is that if the mixture centers $\{\boldsymbol{\mu}_m\}$ are well separated in $\mathbb{R}^d$ and remain well separated in $\mathbb{R}^{d'}$ (as discussed above), then for a signal $\boldsymbol{x}$ living near mixture component m_0, all $\widehat{\nu}_m$ will be small for $m \neq m_0$, and thus $\|\boldsymbol{x} - \widehat{\boldsymbol{x}}\|_2$ will be small. We refer the interested reader to [28] for a more detailed analysis.

B. Bounds on matrix completion

Above we have considered bounds for CS measurements and an underlying MFA model, with example results in Figure 2; in Figure 2 the CS projection matrix $\boldsymbol{\Phi}$ was defined by draws from a Gaussian distribution. In Figure 1 rather than taking such random-projection measurements, we observed a small subset of pixels, with these selected at random. This is closely related to the matrix-completion problem, for which we briefly review theoretical guarantees. We also showed experimental results in Section V for the matrix-completion problem, using nonparametric Bayesian techniques.

Several recent papers have examined the problem of recovering a low-rank matrix from just a fraction of its entries. As in Section III-D, let us consider a matrix $\mathbf{M} \in \mathbb{R}^{d \times N}$ with $N \geq d$, and let us suppose that $\mathbf{M}$ has rank r. Because such a matrix has only $(d+N-r)r$ degrees of freedom [4], it seems natural that one may be able to recover the matrix when observing far less than all of its dN entries. We denote by $m \ll dN$ the number of available entries.

A recent approach for recovering the missing entries of $\mathbf{M}$ involves solving a convex optimization problem, wherein one seeks the matrix $\mathbf{M}'$ having the smallest *nuclear norm* such that $\mathbf{M}'$ agrees with $\mathbf{M}$ at the m observed entries. (The nuclear norm of a matrix equals the sum of its singular values.) As an example result, it has been shown that with high probability nuclear norm minimization recovers the matrix $\mathbf{M}$ exactly supposing that $m \sim CNr \log^6 N$ and that the locations of the m observed positions are drawn uniformly at random [52]; the constant C in this expression depends on the "coherence" of the singular vectors of $\mathbf{M}$, implying that some matrices are easier to recover than others. Similar statements [4] have also been made for matrix recovery in terms of a generalization of the RIP from CS, which is discussed above in Section II. Like signal recovery in CS, matrix completion has also been shown to be robust to noise in the observed entries [53].

To the best of our knowledge, there do not exist bounds available for the alternative form of matrix completion discussed in Section III-D, in which each column of the matrix is defined by a unique subspace and thus conventional rank minimization techniques will not be appropriate. This problem

includes conventional rank-based matrix recovery as a special case (when the columns happen to share a common subspace), however, and so it is likely to be more difficult to solve in general, both in terms of the requisite number of observations and in terms of algorithmic complexity.

VII. Conclusions

While the dimensionality of data used for visualization by humans ($e.g.$, imagery and video) may be very large, the underlying information content in the data may be relatively low. We have reviewed addressing this problem through the representation of data in terms of the underlying (low-dimensional) manifold or union of subspaces on which it resides. By exploiting this low-dimensional representation, one may significantly reduce the quantity of data that need be measured from a given scene (or needed within a general data matrix), manifesting compressive or incomplete measurements. There are several technical challenges that must be addressed, including development of models to learn the underlying low-dimensional latent space. In this paper we have examined such learning from a nonparametric Bayesian viewpoint, with example results presented for compressive sensing of signals that reside on a union of subspaces, image interpolation, and matrix completion. We have also reviewed theoretical results on the accuracy of data recovery for such problems.

Concerning future research, note that the discussion in Section IV on completely random measures is quite general, with the Dirichlet process and beta-Bernoulli processes considered here special cases. It is of interest to consider more general nonparametric models. For example, such models may be replaced by generalized forms, that yield power-law behavior in the number of clusters and in the number of dictionary elements, as a function of the quantity of data. Such power-law behavior may be better matched to the properties of real data, such as images, video and general matrices. There are early and promising studies that have examined this power-law construction [54], [55].

Acknowledgements

Many graduate students contributed toward the ideas and results reviewed in this paper. The authors would particularly like to acknowledge the contributions of Minhua Chen, Armin Eftekhari, John Paisley, and Mingyuan Zhou. The authors also thank the reviewers for a careful reading of the original version of this paper, and suggestions that led to a significantly improved final paper.

References

[1] E. Candès, J. Romberg, and T. Tao, "Signal recovery from incomplete and inaccurate measurements," *Communications on Pure and Applied Mathematics*, vol. 59, pp. 1207–1223, 2005.

[2] D. L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, pp. 1289–1306, 2006.

[3] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Preprint*.

[4] B. Recht, M. Fazel, and P. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Review*, to appear.

[5] 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 Trans. Signal Processing*, 2010.

[6] J. Paisley and L. Carin, "Nonparametric factor analysis with beta process priors," in *Proc. Int. Conf. Machine Learning*, 2009.

[7] R. Baraniuk, V. Cevher, and M. Wakin, "Low-dimensional models for dimensionality reduction and signal recovery: A geometric perspective," accepted to Proceedings of the IEEE.

[8] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde, "Model-based compressive sensing," 2008, preprint. Available at http://dsp.rice.edu/cs.

[9] E. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. on Inform. Theory*, vol. 51, pp. 4203–4215, 2005.

[10] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, "A simple proof of the restricted isometry property for random matrices," *Constructive Approximation*, vol. 28, pp. 253–263, 2008.

[11] D. L. Donoho and C. Grimes, "Image manifolds which are isometric to Euclidean space," *J. Math. Imaging Comp. Vision*, vol. 23, no. 1, pp. 5–24, July 2005.

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

[13] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic Decomposition by Basis Pursuit," *SIAM Journal on Scientific Computing*, vol. 20, p. 33, 1998.

[14] Y. Pati, R. Rezaiifar, and P. Krishnaprasad, "Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition," in *Proc. 27th Ann. Asilomar Conf. Signals, Systems, and Computers*, 1993.

[15] D. Needell and J. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Applied and Computational Harmonic Analysis*, June 2008, to be published.

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

[17] S. Ji, Y. Xue, and L. Carin, "Bayesian compressive sensing," *IEEE Trans. Signal Process.*, vol. 56, no. 6, pp. 2346–2356, 2008.

[18] P. Schniter, L. C. Potter, and J. Ziniel, "Fast bayesian matching pursuit," in *Information Theory and Applications Workshop*, 2008, pp. 326–333.

[19] M. B. Wakin, "Manifold-based signal recovery and parameter estimation from compressive measurements," *Preprint*, 2008.

[20] J. Bernardo and A. Smith, *Bayesian Theory*. Wiley, 2004.

[21] T. L. Griffiths and Z. Ghahramani, "Infinite latent feature models and the Indian buffet process," in *Advances in Neural Information Processing Systems*, 2005.

[22] Z. Ghahramani and M. Beal, "Variational inference for Bayesian mixtures of factor analysers," in *Proc. Conf. Advances in Neural Information Processing Systems, NIPS*, 2000.

[23] J. Berger, J. Ghosh, and N. Mukhopadhyay, "Approximation and consistency of Bayes factors as model dimension grows," *J. Statist. Plann. Inference*, vol. 112, p. 241258, 2003.

[24] S. Press and K. Shigemasu, "A note on choosing the number of factors," *Comm. Statist. Theory Methods*, vol. 28, p. 16531670, 1999.

[25] S. Lee and X. Song, "Bayesian selection on the number of factors in a factor analysis model," *Behaviormetrika*, vol. 29, p. 2339, 2002.

[26] T. Ferguson, "A Bayesian analysis of some nonparametric problems," *The Annals of Statistics*, vol. 1, pp. 209–230, 1973.

[27] M. Tipping and C. Bishop, "Mixture of principal component analyzers," *Neural Computation*, vol. 11, pp. 443–482, 1999.

[28] A. Eftekhari and M. B. Wakin, "Performance bounds for compressive sensing with a mixture of factor analyzers," *Technical Report*, 2010.

[29] L. He and L. Carin, "Exploiting structure in wavelet-based Bayesian compressive sensing," *IEEE Trans. Signal Processing*, vol. 57, pp. 3488–3497, 2009.

[30] R. Salakhutdinov and A. Mnih, "Bayesian probabilistic matrix factorization with MCMC," in *Advances in Neural Information Processing Systems*, 2008.

[31] ——, "Probabilistic matrix factorization," in *Advances in Neural Information Processing Systems*, 2008.

[32] Y. C. Eldar and M. Mishali, "Robust recovery of signals from a union of subspaces," *IEEE Trans. Information Theory*, vol. 12, pp. 1338–1351, 2008.

[33] N. Hjort, C. Holmes, P. Muller, and S. Walker, *Bayesian nonparametrics*. Cambridge University Press, 2010.

[34] J. F. C. Kingman, "Completely random measures," *Pacific Journal of Mathematics*, vol. 21, no. 1, pp. 59–78, 1967.

[35] R. Thibaux and M. I. Jordan, "Hierarchical beta processes and the Indian buffet process," in *Proceedings of the International Workshop on Artificial Intelligence and Statistics*, vol. 11, 2007.

[36] N. L. Hjort, "Nonparametric Bayes estimators based on beta processes in models for life history data," *Annals of Statistics*, vol. 18, no. 3, pp. 1259–1294, 1990.

[37] J. Paisley and L. Carin, "Nonparametric factor analysis with beta process priors," in *Proc. Int. Conf. Machine Learning (ICML)*, 2009.

[38] M. Zhou, H. Chen, J. Paisley, L. Ren, G. Sapiro, and L. Carin, "Non-parametric Bayesian dictionary learning for sparse image representations," in *Proc. Conf. Advances in Neural Information Processing Systems, NIPS*, 2009.

[39] E. Fox, E. Sudderth, M. I. Jordan, and A. Willsky, "Sharing features among dynamical systems with beta processes," in *Advances in Neural Information Processing (NIPS) 22*. Cambridge, MA: MIT Press, 2010.

[40] M. D. Escobar, "Estimating normal means with a Dirichlet process prior," *Journal of the American Statistical Association*, vol. 89, pp. 268–277, 1994.

[41] R. M. Neal, "Markov chain sampling methods for Dirichlet process mixture models," *Journal of Computational and Graphical Statistics*, vol. 9, pp. 249–265, 2000.

[42] H. Ishwaran and L. F. James, "Gibbs sampling methods for stick-breaking priros," *Journal of the American Statistical Association, Theory and Methods*, vol. 96, no. 453, pp. 161–173, 2001.

[43] S. G. Walker, "Sampling the Dirichlet mixture model with slices," *Communications in Statistics–Simulation and Computation*, vol. 36, p. 45, 2007.

[44] Y. W. Teh and M. I. Jordan, "Hierarchical Bayesian nonparametric models with applications," in *Bayesian Nonparametrics: Principles and Practice*. Cambridge, UK: Cambridge University Press, 2010.

[45] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Supervised dictionary learning," in *Proc. Conf. Advances in Neural Information Processing Systems, NIPS*, 2008.

[46] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *Proc. Int. Conf. Machine Learning (ICML)*, 2009.

[47] 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, pp. 489–509, 2006.

[48] E. J. Candès and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Processing Magazine*, vol. 25, pp. 21–30, 2008.

[49] E. J. Candès, "The restricted isometry property and its implications for compressed sensing," *Compte Rendus de l'Academie des Sciences, Paris, Serie I*, vol. 346, pp. 589–592, 2008.

[50] J. Mairal, M. Elad, and G. Sapiro, "Sparse representation for color image restoration," *IEEE Transactions on Image Processing*, vol. 17, 2008.

[51] N. Lawrence and R. Urtasun, "Non-linear matrix factorization with Gaussian processes," in *Proc. Int. Conf. Machine Learning*, 2009.

[52] E. J. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *Preprint*, 2009.

[53] E. J. Candès and Y. Plan, "Matrix completion with noise," *Preprint*, 2009.

[54] S. Goldwater, T. Griffiths, and M. Johnson, "Interpolating between types and tokens by estimating power-law generators," in *Proc. Conf. Advances in Neural Information Processing Systems, NIPS*, 2006.

[55] Y. Teh and D. Gorur, "Indian buffet processes with power-law behavior," in *Proc. Conf. Advances in Neural Information Processing Systems, NIPS*, 2009.