Published in final edited form as:

IEEE Signal Process Mag. 2011 March 1; 28(2): . doi:10.1109/MSP.2010.939733.

Learning Low-Dimensional Signal Models:

A Bayesian approach based on incomplete measurements

Lawrence Carin, Richard G. Baraniuk, Volkan Cevher, David Dunson, Michael I. Jordan, Guillermo Sapiro, and Michael B. Wakin

Sampling, coding, and streaming even the most essential data, e.g., in medical imaging and weather-monitoring applications, 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 led to several important theoretical and algorithmic developments under different low-dimensional modeling frameworks, such as compressive sensing (CS) [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. In this article, we investigate this challenge from the perspective of nonparametric Bayesian analysis.

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

$$\mathbf{y} = \Phi \mathbf{x} + \delta. \quad (1)$$

The signal is $\mathbf{x} \in \mathbb{R}^d$, the measurements are $\mathbf{y} \in \mathbb{R}^{d'}$, Φ is a $d' \times d$ matrix with $d' < d$, and δ accounts for noise.

Such a projection process loses signal information in general, since Φ 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 $\mathbf{y}$ about certain signals $\mathbf{x}$. One way to preserve information is for Φ to provide a stable embedding that approximately preserves pairwise distances between all signals in some set of interest. In some cases, this property allows the recovery of $\mathbf{x}$ from its measurement $\mathbf{y}$.

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 $\mathbf{x} \in \mathbb{R}^d$ and an appropriate basis $\Psi \in \mathbb{R}^{d \times d}$, we can transform the signal as $\mathbf{x} = \Psi \theta$, where θ is sparse or can be well approximated as such, that is, it has only a few nonzero elements. CS exploits this fact to recover signals from its compressive samples $\mathbf{y} \in \mathbb{R}^{d'}$, which are dimensionality-reducing, nonadaptive random measurements. According to CS 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 MFAs 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 of the type discussed in the “Stable Embeddings” section 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 measurement.

In this article 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 to subsequently 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 on the measurement of a small subset of pixels uniformly selected at random.

The tools and methods used 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 unspecified beforehand. While it has been historically challenging to find workable prior distributions in the parameter space for such problems, we leverage the beta process (BP), Bernoulli process, Dirichlet process (DP), and Indian buffet process (IBP). We observe that these distributions provide a nice scaffold for analytically managing posterior distributions, given a set of training samples as well as observations. Additionally, we develop performance bounds for recovering high-dimensional data based on incomplete measurements. We present several examples of how this technology may be used in practice in CS, in matrix completion (when we recover a full low-rank matrix based on a small number of randomly sampled matrix elements), and in 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, and books).

STABLE EMBEDDINGS

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

A requirement on the matrix Φ that combines both information preservation and stability properties for a signal model is the so-called ε -stable embedding property

$$(1-\varepsilon)\|\mathbf{x}_1 - \mathbf{x}_2\|_2^2 \leq \|\Phi\mathbf{x}_1 - \Phi\mathbf{x}_2\|_2^2 \leq (1+\varepsilon)\|\mathbf{x}_1 - \mathbf{x}_2\|_2^2, \quad (2)$$

which must hold for all $\mathbf{x}_1, \mathbf{x}_2$ 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 $\mathbf{y} = \Phi\mathbf{x}$ from $\mathbb{R}^d$ down to $\mathbb{R}^{d'}$, $d' < d$, cannot hope to preserve all of the information in all signals $\mathbf{x} \in \mathbb{R}^d$, since it is impossible to guarantee that $\Phi\mathbf{x}_1 - \Phi\mathbf{x}_2$ holds for all signal pairs $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^d$. This is because there are infinitely many $\mathbf{x} + \mathbf{x}'$, with $\mathbf{x}'$

from the $(d - d')$ -dimensional null space of Φ , which yield exactly the same measurement $\mathbf{y}$. 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 $\mathbf{x}$ in the model from its measurement $\mathbf{y}$.

Let us review the three deterministic model classes that have been shown to support stable dimensionality reduction. First, a sparse signal $\mathbf{x} \in \mathbb{R}^d$ can be represented in terms of just $k \ll d$ nonzero coefficients in the basis expansion $\mathbf{x} = \Psi \theta$, where Ψ is a fixed basis. Concisely, we say that $\|\theta\|_{\ell_0} = k$, where ℓ_0 is the pseudonorm that merely counts the nonzero entries in

$\mathbf{x}$. The set of all sparse signals Σ_k is the union of $\binom{d}{k}$, k -dimensional canonical subspaces in $\mathbb{R}^d$, aligned with the coordinate axes of the basis Ψ . For sparse signals, the stable embedding property (2) corresponds to the restricted isometry property (RIP) [9]. Although the design of such a stable embedding is an nondeterministic polynomial time (NP) complete problem, in general, it has been shown that any independent identically distributed (i.i.d.) sub-Gaussian random matrix Φ 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 Φ 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 $\{\mathbf{x}_\theta, \theta \in \Theta\}$ is smoothly parameterized by a k -dimensional parameter vector θ [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 Φ stably embeds a k -dimensional smooth manifold from $\mathbb{R}^d$ into $\mathbb{R}^{d'}$ as long as $d' = O(k \log(d))$ [12].

Given a stable embedding of the form (1), a number of techniques have been developed to recover a (structured) sparse signal of interest $\mathbf{x}$ from the measurements $\mathbf{y}$, 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 article addresses concerns learning the underlying signal models, particularly union-of-subspace and manifold models, with this learning performed nonparametrically based on the available data. The MFA model discussed later 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 learned, they may be used in algorithms that seek to recover high-dimensional data based on low-dimensional compressive measurements.

LEARNING CONCISE SIGNAL MODELS

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.

UNION-OF-SUBSPACE MODEL FOR SPARSE SIGNALS

Assume access to a set of N training data $\{\mathbf{x}_n\}_{n=1, N}$. Our goal is to infer a concise model for $\{\mathbf{x}_n\}_{n=1, N}$ appropriate for recovering high-dimensional data from compressive measurements. Further, we would like to learn the model parameters nonparametrically (e.g., without having to set the dimensionality of the subspaces or the number of mixture components). We express each $\mathbf{x}_n \in \mathbb{R}^d$ as

$$\mathbf{x}_n = \mathbf{A}(\mathbf{c}_n \circ \mathbf{b}_n) + \varepsilon_n, \quad (3)$$

where $\mathbf{c}_n \in \mathbb{R}^K$, $\mathbf{b}_n \in \{0, 1\}^K$, 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 $\mathbf{b}_n$ and because $\|\mathbf{b}_n\|_{\ell_0} < d$, each $\mathbf{x}_n$ is represented by a subset of the columns of $\mathbf{A}$ (defining a subspace); ε_n is meant to represent the portion of $\mathbf{x}_n$ not contained within the aforementioned subspace.

If we assume that the components of ε_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 $\mathbf{c}_n \sim \mathcal{N}(0, \alpha^{-1} \mathbf{I}_K)$, then, after integrating out $\mathbf{c}_n$, $\mathbf{x}_n$ is drawn from

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

where $\Lambda_n = \text{diag}(\mathbf{b}_n)$ is a binary diagonal matrix. Therefore, if the columns of $\mathbf{A} \Lambda_n$ are linearly independent, and if r_n represents the number of nonzero components in $\mathbf{b}_n$, then $\mathbf{x}_n$ 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 $\alpha^{-1} \mathbf{A} \Lambda_n \mathbf{A}^T$). Note that priors like $\mathbf{c}_n \sim \mathcal{N}(0, \alpha^{-1} \mathbf{I}_K)$, and other similar priors considered later, 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 k th column of $\mathbf{A}$, $\mathbf{a}_k$, i.i.d. as

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

such that each column has a 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 the n th row defined by $\mathbf{b}_n$; the cumulative set of binary vectors $\{\mathbf{b}_n\}_{n=1, N}$ defines the total number of columns needed from $\mathbf{A}$. The prior we will employ for $\{\mathbf{b}_n\}_{n=1, N}$ is the beta-Bernoulli process, which is closely connected to the IBP [21] developed by Griffiths and Ghahramani; this is discussed in detail in the ‘‘Completely Random Measures’’ section. 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 the ‘‘Applications’’ section, in Figure 1, $\{\mathbf{x}_n\}_{n=1, N}$ correspond to N patches of pixels from a red, green, and blue (RGB) image. In this problem, only 20% of the pixels 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 $\{\mathbf{c}_n \circ \mathbf{b}_n\}_{n=1, N}$. Note that the columns of $\mathbf{A}$ have the same support as $\mathbf{x}_n$, and hence, they may be used to infer missing pixel values via $\mathbf{A}(\mathbf{c}_n \circ \mathbf{b}_n)$. It is important that the pixels are missing at random; if the same pixel is missing in all $\{\mathbf{x}_n\}_{n=1, N}$, 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 the information from a similar patch elsewhere in the image. Hence, the simultaneous (collaborative) analysis of all $\{\mathbf{x}_n\}_{n=1, N}$ allows one to infer information about the missing pixels by exploiting the 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.

MFA MODEL FOR SIGNAL ENSEMBLES

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

Specifically, we generalize (4) as

$$\mathbf{x}_n \sim \sum_{m=1}^M v_m \mathcal{N}(\mu_m, \alpha_m^{-1} \mathbf{A}_m \Lambda_m \mathbf{A}_m^T + \alpha_0^{-1} \mathbf{I}_d), \quad (6)$$

where $\sum_{m=1}^M v_m = 1$, $\Lambda_m = \text{diag}(\mathbf{b}_m)$, with $\mathbf{b}_m \in \{0, 1\}^K$, again a binary vector that selects particular columns of $\mathbf{A}_m$ to define the subspace spanned by the m th mixture component [in (3), there is a separate binary vector $\mathbf{b}_n$ for each sample $\mathbf{x}_n$, and now there is a related binary vector $\mathbf{b}_m$ associated with the mixture component m]. Note that the number of nonzero 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 nonzero components of $\mathbf{b}_m$ is small (i.e., $\|\mathbf{b}_m\|_{\ell_0} \ll d$), then each mixture component $\mathcal{N}(\mu_m, \alpha_m^{-1} \mathbf{A}_m \Lambda_m \mathbf{A}_m^T + \alpha_0^{-1} \mathbf{I}_d)$ defines a relatively low-dimensional pancake in $\mathbb{R}^d$, with the number of principal dimensions in the m th associated subspace defined by $\|\mathbf{b}_m\|_{\ell_0}$. The means μ_m locate the center of each pancake, and these are assumed to be drawn from $\mathcal{N}(0, \beta_m^{-1} \mathbf{I}_d)$, with a gamma prior placed on β_m (again due to conjugacy).

The model (6) is called MFA [22], and the nonzero columns of $\mathbf{A}_m \Lambda_m$ define the factor loadings associated with the m th mixture component. When building an MFA model, a natural question concerns how many mixture components M are appropriate for the training data $\{\mathbf{x}_n\}_{n=1, N}$. 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, later 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 DP [26], as summarized in the ‘‘Completely Random Measures’’ section.

Note that in (3) the $\mathbf{b}_n$ selects a subset of the columns of $\mathbf{A}$ for the representation of $\mathbf{x}_n$, and one may expect that different $\mathbf{x}_n$ will (partially) share the usage of these columns. In the mixture model of (6), $\mathbf{b}_m$ selects which subset of the columns of $\mathbf{A}_m$ are used for the m th

mixture component; $\mathbf{b}_m$ therefore defines the dimensionality and subspace of this mixture component. In general, the sub-spaces spanned by $\mathbf{A}_m\Lambda_m$ and $\mathbf{A}_{m'}\Lambda_{m'}$ are different. Hence, (3) implies that $\mathbf{x}_n$ are drawn from partially overlapping subspaces, without an explicit clustering; (6) explicitly clusters the data, with the data in cluster m spanned by the nonzero columns of $\mathbf{A}_m\Lambda_m$. 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 this is inferred via nonparametric Bayesian inference. In [27], the authors achieve a point estimate of model parameters via expectation maximization (EM), where we estimate a full posterior density function on model parameters.

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 μ_m roughly correspond to points sampled from the manifold, the columns of $\mathbf{A}_m\Lambda_m$ roughly span the k -dimensional local tangent spaces, the thickness parameter α_0^{-1} depends on the manifold curvature, and the weights v_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 $\mathbf{A}_m\Lambda_m$. This is reminiscent of the classical CS 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 $\mathbf{x}$ 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 that has been studied in CS. An example application of this framework is presented in Figure 2.

MATRIX COMPLETION

As a final model, consider a matrix $\mathbf{M} \in \mathbb{R}^{d \times N}$ with $N \gg d$ (this can always be achieved by matrix transpose). Let the N columns of $\mathbf{M}$ constitute the set of vectors $\{\mathbf{x}_n\}_{n=1}^N$, where $\mathbf{x}_n$ is the n th column manifested with randomly selected missing entries (in this problem $\mathbf{x}_n' = \Phi_n \mathbf{x}_n$, where the rows of Φ_n are randomly selected rows of the $d \times d$ identity matrix, with Φ_n different for each n). If the matrix $\mathbf{M}$ is such that its columns satisfy the properties inherent to (4), specifically that each $\mathbf{x}_n$ resides approximately within a subspace defined by the columns in matrices of the form $\mathbf{A}\Lambda_n$, then the data-recovery technique discussed in the “Pixel/Voxel Recovery Via Union of Subspaces” section 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

$$\mathbf{M} = \sum_{k=1}^d \lambda_k b_k u_k v_k^T + \mathbf{E}, \quad (7)$$

where $\lambda_k \in \mathbb{R}$, $b_k \in \{0, 1\}$, $u_k \in \mathbb{R}^d$, and $v_k \in \mathbb{R}^N$. We may again draw $u_k \sim \mathcal{N}(0, (1/d)\mathbf{I}_d)$, $v_k \sim \mathcal{N}(0, (1/N)\mathbf{I}_N)$, $\lambda_k \sim \mathcal{N}(0, \alpha^{-1})$, and each component of $\mathbf{E}$ drawn i.i.d. from $\mathcal{N}(0, \alpha_0^{-1})$. A binary sparseness-promoting prior (see the ‘‘Completely Random Measures’’ section) may again be employed to define the binary vector $\mathbf{b} = (b_1, \dots, b_d)$, 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 $\mathbf{b}$ (so in this case, we do not use a BP); however, this model may be clearly generalized to the nonlinear case by making $\mathbf{b}$ 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 n th component of v_k . Then the n th column of $\mathbf{M}$ may be expressed as

$$\mathbf{x}_n = \mathbf{U}(\mathbf{c}_n \circ \mathbf{b}) + \varepsilon_n, \quad (8)$$

where ε_n represents the n th column of $\mathbf{E}$, $\mathbf{U} \in \mathbb{R}^{d \times d}$ has columns defined by $\mathbf{u}_k$, and $\mathbf{c}_n = (\lambda_1 v_{1n}, \dots, \lambda_d v_{dn})$. 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 $\mathbf{x}_n$ in general has its own subspace defined by the binary vector $\mathbf{b}_n$.

BAYESIAN NONPARAMETRIC INFERENCE

Based on the earlier 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 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 DP. These nonparametric models represent special cases of a more general concept, the completely random measure. Later, we first review the completely random measure, and then we show three examples for which it may be applied: for the BP, gamma process, and DP. Finally, we explain how BP may be combined with 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 DP may be used to infer the appropriate number of mixture components in MFA. For a thorough discussion of nonparametric Bayesian methods, the interested reader is referred to [33].

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 a general framework of completely random measures, and then we show how to derive some particularly useful stochastic processes—BP and DP—from this framework. When learning MFA, BP is used to infer the number of factor loadings (equivalently the rank) for each mixture component, while DP 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:

$$G = \sum_{i=1}^{\infty} p_i \delta_{\omega_i}, \quad (9)$$

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 Brownian motion).

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 $(0, 1)$

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

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 BP as

$$B \sim \text{BP}(c, B_0), \quad (11)$$

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

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

GAMMA PROCESS—As a second example, let the rate measure be a product of base measure G_0 and an improper gamma distribution

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

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

$$G \sim \text{GaP}(c, G_0) \quad (13)$$

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

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:

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

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

APPLICATION TO DATA MODELS

From the “Learning Concise Signal Models” section, there are two principal modeling objectives: 1) an ability to infer the number of mixture components needed in an MFA and 2) the capacity to infer the number of needed factor loadings and their characteristics. Item 2) is related to inferring the binary matrix $\mathbf{B}$ discussed in the “Union-of-Subspace Model for Sparse Signals” section. First, considering 1), recall from above that a draw from DP 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 the candidate mixture model parameters $(\alpha_m, \mu_m, \mathbf{A}_m, \Lambda_m)$ used in the mixture model of (6). Specifically, we constitute the following generative process for the data $\{\mathbf{x}_n\}_{n=1, N}$ when these data are assumed to be drawn from MFA:

$$\begin{aligned} \mathbf{x}_n &\sim f(\alpha_n, \mu_n, \mathbf{A}_n, \Lambda_n, \alpha_0), \\ (\alpha_n, \mu_n, \mathbf{A}_n, \Lambda_n) &\sim G, \\ G &\sim \text{DP}(\alpha, G_0), \end{aligned} \quad (15)$$

where $f(\cdot)$ represents the Gaussian distribution in (6). In this case, the base measure G_0 from which ω_i is drawn corresponds to a factorized prior for the set of parameters $(\alpha_n, \mu_n, \mathbf{A}_n, \Lambda_n)$, with the individual components of that prior, as defined in the “Mixture of Factor Analyzers Model for Signal Ensembles” section. 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 $\{\alpha_n, \mu_n, \mathbf{A}_n, \Lambda_n\}_{n=1, N}$ are characteristic of being drawn from a mixture model. With probability π_i , any particular set $(\alpha_n, \mu_n, \mathbf{A}_n, \Lambda_n)$ 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 multiple data samples $\{\mathbf{x}_n\}_{n=1, N}$. We therefore manifest a clustering of $\{\mathbf{x}_n\}_{n=1, N}$ (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 DP. 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 DP draw G [33].

We now consider BP as a prior for the binary vectors $\{\mathbf{b}_n\}_{n=1, N}$ in the model (4); these binary vectors are also of interest in the matrix-completion problem in the “Matrix Completion” section of the “Learning Concise Signal Models” section. Recall that a draw from BP may be expressed as $G = \sum_{i=1}^{\infty} \pi_i \delta_{w_i}$, where 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 BP corresponds to the same prior used for the columns $\mathbf{A}$ or $\mathbf{U}$. The random variable π_i defines the probability that the i th column of $\mathbf{A}$ or $\mathbf{U}$ is used to represent the data of interest. Specifically, the data sample $\mathbf{x}_n$ 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, $\mathbf{x}_n$ selects the i th column of $\mathbf{A}$ or $\mathbf{U}$, with the respective column denoted by the atom ω_i . Therefore, π_i defines the parameter of a Bernoulli distribution, with which a particular $\mathbf{x}_n$ 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 $\mathbf{x}_n$ and the columns correspond to specific atoms $\{\omega_i\}_{i=1, \infty}$; hence, the rows of $\mathbf{B}$ are defined by $\{\mathbf{b}_i\}_{i=1, N}$, and sample $\mathbf{x}_n$ selects atom/dish ω_i if the i th component of $\mathbf{b}_n$ 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 nonzero values [21]; in practice, one may truncate the model to K columns/atoms for large K .

The beta-Bernoulli process yields a so-called IBP [33] if the BP 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 $\mathbf{b}_n$ for customer/data n defines which dishes/parameters are selected; if the k th component of $\mathbf{b}_n$ is one, then the k th parameter is used by data n , and if the k th component of $\mathbf{b}_n$ is zero, the k th parameter is not used.

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 conditional distributions. By appropriate design of the model, of the form discussed earlier, these conditional distributions may often be expressed analytically. As an example of such samplers, consider DP in particular.

Working the marginal distribution embodied in 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 the 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 BP, one can work with the marginal distribution embodied in 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 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 BP and for pointers to literature on variational approaches to inference for random measures.

APPLICATIONS

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

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 $\mathcal{D} = \{\mathbf{x}_n\}_{n=1, N}$ are constituted from pixels/voxels in an image. Specifically, each of the N image patches is defined by a set of contiguous pixels, with $\mathbf{x}_n \in \mathbb{R}^d$ representing data from the n th 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 $\mathbf{x}_n$ is defined by the matrix $\mathbf{A} \in \mathbb{R}^{d \times K}$, by the sparse vector $\mathbf{c}_n \circ \mathbf{b}_n$ (with r_n nonzero components), and by $\varepsilon_n \in \mathbb{R}^d$. We assume that ε_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 $\{\mathbf{b}_n\}_{n=1, N}$ are known (i.e., if it is assumed we know which columns of $\mathbf{A}$ are associated with each $\mathbf{x}_n$, 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 $\{\mathbf{x}_n\}_{n=1, N}$ jointly to infer the cumulative set of model parameters (exploiting the fact that $\mathbf{A}$ is shared among all), it appears that we have more data in $\mathcal{D}$ than needed. Further, it would appear that we have enough data to infer, which among the r_n columns of $\mathbf{A}$ are needed for the representation of each $\mathbf{x}_n$ (therefore, we do not need a priori access to $\{\mathbf{b}_n\}_{n=1, N}$).

On the basis of the earlier observations, researchers have recently assumed access to only a subset of components of each $\mathbf{x}_n$, 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 is measured, with the measured subset of pixels

selected uniformly at random. Let $\mathcal{D}' = \{\mathbf{x}'_n\}_{n=1, N}$ represent a modified form of $\mathcal{D}$, with each $\mathbf{x}'_n$ defined by a fraction of the components of each $\mathbf{x}_n$, with the 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 CS [47], [48], in that a small subset of measurements are performed, with the full data recovered based upon the exploitation of properties of the signal (that the signals live in a low-dimensional subspace of $\mathbb{R}^d$). However, in CS, 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 discrete cosine transform (DCT)-type basis, one could use delta-function-like projections (selecting random components of each $\mathbf{x}_n$), like those considered earlier. However, in an earlier collaborative-filtering framework, we not only perform random sampling but 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 the “Matrix Completion” section of the “Applications” section, collaborative filtering for image recovery is closer to the field of matrix completion [3] than it is to CS.

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 $\prod_{n=1}^N \mathcal{N}(\mathbf{x}'_n | \mathbf{0}, \alpha^{-1} \mathbf{A} \Lambda_n \mathbf{A}^T + \alpha_0^{-1} \mathbf{I}_d)$ is simply evaluated at the pixels for which data are observed. As discussed in the “Posterior Inference” section, a Gibbs sampler may be implemented, and from this, one may obtain an approximation to all model parameters. Hence, the posterior probability of each $\mathbf{x}_n$ in $\mathcal{D}$, based on observed $\mathcal{D}'$ and model hyperparameters Θ , may be expressed as

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

with the posterior $p(\mathbf{A}, \alpha, \Lambda_n, \alpha_0 | \mathcal{D}', \Theta)$ 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 $\{\mathbf{x}_n\}_{n=1, N}$. 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 $\{\mathbf{x}_n\}_{n=1, N}$, very distinct images are manifested (recall that each $\mathbf{x}_n$ 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 $\mathbf{x}_n$ and $\mathbf{x}_{n'}$ are spatially proximate, they will likely employ similar factors (manifested by similar binary factor-selection vectors $\mathbf{b}_n$ and $\mathbf{b}_{n'}$).

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 and 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 on 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 peak signal-to-noise ratio (PSNR) curve shows how the model performance improves as we consider more data in a 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 earlier. For Gibbs rounds 16, 32, and 64, the corresponding PSNR values were 27.64, 28.20, and 28.66 dB. For this example, we used $K = 256$. This example was considered in [50]; the best results reported was 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.

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 $\mathcal{D} = \{\mathbf{x}_n\}_{n=1, N}$. We now wish to measure a single new $\mathbf{x} \in \mathbb{R}^d$, under the assumption that $\mathbf{x}$ 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 $\mathbf{x}$ is drawn from a known manifold [12]. On the basis of this prior knowledge, we wish to measure $\mathbf{y} \in \mathbb{R}^{d'}$, with $d' < d$, and ideally with $d' \ll d$; based on the measured $\mathbf{y}$, $\mathbf{x}$ is recovered.

It is assumed that $\mathbf{y} = \Phi\mathbf{x}$, where $\Phi \in \mathbb{R}^{d' \times d}$ is a projection matrix typically defined randomly. In the “Bayesian Nonparametric Inference” section, we discuss the desired properties of Φ and the connection of such to the characteristics of MFA. In a statistical sense, to recover $\mathbf{x}$ from $\mathbf{y}$, we desire $p(\mathbf{x}|\mathbf{y}) = p(\mathbf{y}|\mathbf{x})p(\mathbf{x})/p(\mathbf{y})$, under the assumption that $p(\mathbf{x})$ is the known MFA of the form in (6). Assuming that the compressive measurements are noisy, we may express $\mathbf{y} = \Phi\mathbf{x} + \delta$, where $\delta \in \mathbb{R}^{d'}$ represents additive noise. If

$\delta \sim \mathcal{N}(\mathbf{0}, \beta_0^{-1} \mathbf{I}_{d'})$, then we have $p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\Phi\mathbf{x}, \beta_0^{-1} \mathbf{I}_{d'})$. If β_0 is known, under the MFA assumption for $p(\mathbf{x})$, the expression $p(\mathbf{x}|\mathbf{y})$ 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 an MFA model for $p(\mathbf{x})$, one may readily constitute a statistical estimate of $\mathbf{x}$ based on the observed $\mathbf{y}$, with performance bounds discussed in the “Performance Guarantees” section. 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.

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 Movie-Lens dataset (10,681 movies by 71,567 users). Table 1 shows the results for both r_a and r_b partitions provided with the data, in which ten 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 root-mean-square error (RMSE) of 0.8740 ± 0.0278 using a ten-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 1, we show the results when the vectors $\mathbf{x}_n$ correspond to the user-dependent rankings of all movies (user profile), and with $\mathbf{x}_n$ 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 the 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, the integer values are simply approximated as real numbers.

These results were computed using a Gibbs sampler, with a truncated BP implementation with $K = 256$ dishes. One Gibbs iteration required 150 s on a 2.53-GHz E5540 Xeon processor, using nonoptimized MATLAB software. The results in Table 1 correspond to 50 burn-in iterations and 100 collection iterations.

PERFORMANCE GUARANTEES

The BP and DP nonparametric methods may be used to infer an MFA based upon the given training data. Once this model is so learned, MFA may be assumed known and can be used in the inversion of subsequent compressive measurements. An example of such MFA learning and the subsequent utilization within CS signal recovery was presented in Figure 2. It is of interest to examine the performance guarantees based on CS measurements and a known MFA model (learned based on training data, using nonparametric techniques of the type discussed earlier). 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 that one learned MFA is used to perform CS inversion, and this model provides an accurate statistical representation of the data; it is for such a learned MFA that the bounds are constituted.

BOUNDS FOR MFAs

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

$$\hat{\mathbf{x}} = \sum_{m=1}^M \hat{v}_m \hat{\mathbf{x}}_m, \quad (17)$$

where

$$\hat{v}_m = \frac{v_m \mathcal{N}(\mathbf{y}; \Phi\mu_m, \beta_0^{-1} \mathbf{I}_{d'} + \Phi\Omega_m\Phi^T)}{\sum_{\ell=1}^M v_\ell \mathcal{N}(\mathbf{y}; \Phi\mu_\ell, \beta_0^{-1} \mathbf{I}_{d'} + \Phi\Omega_\ell\Phi^T)}$$

represents an estimated mixture weight of the m th component,

$$\hat{\mathbf{x}}_m = \Omega_m \Phi^T (\beta_0^{-1} \mathbf{I}_{d'} + \Phi \Omega_m \Phi^T)^{-1} (\mathbf{y} - \Phi \mu_m) + \mu_m$$

equals the signal estimate that would be recovered if only component m was present in the MFA, and $\Omega_m = \alpha_m^{-1} \mathbf{A}_m \Lambda_m \mathbf{A}_m^T + \alpha_0^{-1} \mathbf{I}_d$ represents the covariance matrix of the m th component in the MFA.

We can also consider the situation where $\alpha_0^{-1} \rightarrow 0$ and $\beta_0^{-1} \rightarrow 0$, 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 $\mu_{m_1} - \mu_{m_2}$ for all $1 \leq m_1, m_2 \leq M$ and that (2) also holds for all vectors in $\text{colspan}(\mathbf{A}_m \Lambda_m)$ for all $1 \leq m \leq M$. From the Johnson-Lindenstrauss lemma, we know that when Φ is generated randomly with i.i.d. Gaussian or sub-Gaussian entries, the former property holds with high probability as long as $d' = O(\log(M)\varepsilon^{-2})$, and using similar arguments, the latter property also holds as long as $d' = O((k + \log(M))\varepsilon^{-2})$ [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 $\mathbf{x}$ that is drawn from the manifold.

For example, supposing that $\beta_0^{-1} \rightarrow 0$, the isometry property for $\text{colspan}(\mathbf{A}_m \Lambda_m)$ discussed essentially guarantees that $\|\mathbf{x} - \hat{\mathbf{x}}_m\|_2$ is a combination of the two error terms, one depending on the size of $\mathbf{x} - \mu_m$ when projected onto $\text{colspan}(\mathbf{A}_m \Lambda_m)$, and one depending on the size of $\mathbf{x} - \mu_m$ when projected orthogonal to $\text{colspan}(\mathbf{A}_m \Lambda_m)$.

The size of α_0^{-1} controls the balance between these two terms, and by choosing sufficiently small α_0^{-1} , we can ensure that the dependence on the first of these terms is small; this allows us to guarantee that $\|\mathbf{x} - \hat{\mathbf{x}}_m\|_2$ is small for any signal $\mathbf{x}$ living near mixture component m .

Analysis of the recovery error $\|\mathbf{x} - \hat{\mathbf{x}}\|_2$ for a multicomponent mean estimator (17) is more involved but can proceed based on the observation that for any $m_0 \in \{1, 2, \dots, M\}$, we can write $\|\mathbf{x} - \hat{\mathbf{x}}\|_2 \leq \|\mathbf{x} - \hat{\mathbf{x}}_{m_0}\|_2 + \sum_{m \neq m_0} v_m \|\mathbf{x} - \hat{\mathbf{x}}_m\|_2$. One conclusion that can be drawn from this is that if the mixture centers $\{\mu_m\}$ are well separated in $\mathbb{R}^d$ and remain well separated in $\mathbb{R}^{d'}$ (as discussed earlier), then for a signal $\mathbf{x}$ living near mixture component m_0 , all v_m will be small for $m \neq m_0$, thus $\|\mathbf{x} - \hat{\mathbf{x}}\|_2$ will be small.

We refer the interested readers to [28] for a more detailed analysis.

BOUNDS ON MATRIX COMPLETION

Earlier, we considered bounds for CS measurements and an underlying MFA model, with the example results; in Figure 2, the CS projection matrix Φ 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 the experimental results in the ‘‘Applications’’ section 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 the ‘‘Matrix Completion’’ section of the ‘‘Learning Concise Signal Models’’ section, let us consider a matrix $\mathbf{M} \in \mathbb{R}^{d \times N}$ with $N \gg d$ 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, 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 RIP from CS, which is discussed in the “Stable Embeddings” section. 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 the “Matrix Completion” section of the “Learning Concise Signal Models” section, 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, 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.

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 article, we have examined such learning from a nonparametric Bayesian viewpoint, with the example results presented for CS of signals that reside on a union of subspaces, image interpolation, and matrix completion. We have also reviewed the theoretical results on the accuracy of data recovery for such problems.

Concerning future research, note that the discussion in the “Bayesian Nonparametric Inference” section on completely random measures is quite general, with the DP and beta-Bernoulli processes considered here as special cases. It is of interest to consider more general nonparametric models. For example, such models may be replaced by generalized forms, which yield power-law behavior in the number of clusters and 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].

Acknowledgments

Many graduate students contributed to the ideas and results reviewed in this article. The authors particularly acknowledge the contributions of Minhua Chen, Armin Eftekhar, John Paisley, and Mingyuan Zhou. The authors

also thank the reviewers for a careful reading of the original version of this article and suggestions that led to a significantly improved final article.

References

1. Candès E, Romberg J, Tao T. Signal recovery from incomplete and inaccurate measurements. *Commun Pure Appl Math.* 2005; 59:1207–1223.
2. Donoho DL. Compressed sensing. *IEEE Trans Inform Theory.* 2006; 52:1289–1306.
3. Candès EJ, Recht B. Exact matrix completion via convex optimization. *Found Comput Math.* 2008; 9:717–772.
4. Recht B, Fazel M, Parrilo P. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. submitted for publication.
5. Chen M, Silva J, Paisley J, Wang C, Dunson D, Carin L. Compressive sensing on manifolds using a nonparametric mixture of factor analyzers: Algorithm and performance bounds. *IEEE Trans Signal Processing.* 2010
6. Paisley, J.; Carin, L. Nonparametric factor analysis with beta process priors. *Proc. Int. Conf. Machine Learning*; 2009.
7. Baraniuk, R.; Cevher, V.; Wakin, M. Low-dimensional models for dimensionality reduction and signal recovery: A geometric perspective. *Proc. IEEE*; to be published
8. Baraniuk, RG.; Cevher, V.; Duarte, MF.; Hegde, C. Model-based compressive sensing Preprint [Online]. 2008. Available: <http://dsp.rice.edu/cs>
9. Candès E, Tao T. Decoding by linear programming. *IEEE Trans Inform Theory.* 2005; 51:4203–4215.
10. Baraniuk R, Davenport M, DeVore R, Wakin M. A simple proof of the restricted isometry property for random matrices. *Construct Approx.* 2008; 28:253–263.
11. Donoho DL, Grimes C. Image manifolds which are isometric to Euclidean space. *J Math Imag Comput Vision.* Jul; 2005 23(1):5–24.
12. Baraniuk R, Wakin M. Random projections of smooth manifolds. *Foundat Comput Math.* 2009; 9(1):51–77.
13. Chen SS, Donoho DL, Saunders MA. Atomic decomposition by basis pursuit. *SIAM J Scient Comput.* 1998; 20:33.
14. Pati, Y.; Rezaifar, R.; Krishnaprasad, P. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. *Proc. 27th Ann. Asilomar Conf. Signals, Systems, and Computers*; 1993.
15. Needell D, Tropp J. CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Appl Comput Harmon Anal.* to be published.
16. Wipf DP, Rao BD. Sparse Bayesian learning for basis selection. *IEEE Trans Signal Processing.* 2004; 52(8):2153–2164.
17. Ji S, Xue Y, Carin L. Bayesian compressive sensing. *IEEE Trans Signal Processing.* 2008; 56(6): 2346–2356.
18. Schniter, P.; Potter, LC.; Ziniel, J. Fast Bayesian matching pursuit. *Proc. Information Theory and Applications Workshop*; 2008. p. 326–333.
19. Wakin MB. Manifold-based signal recovery and parameter estimation from compressive measurements. 2008 Preprint.
20. Bernardo, J.; Smith, A. *Bayesian Theory*. New York: Wiley; 2004.
21. Griffiths, TL.; Ghahramani, Z. Infinite latent feature models and the Indian buffet process. *Proc. Advances in Neural Information Processing Systems*; 2005.
22. Ghahramani, Z.; Beal, M. Variational inference for Bayesian mixtures of factor analysers. *Proc. Conf. Advances in Neural Information Processing Systems, NIPS*; 2000.
23. Berger J, Ghosh J, Mukhopadhyay N. Approximation and consistency of Bayes factors as model dimension grows. *J Statist Plan Infer.* 2003; 112:241–258.
24. Press S, Shigemasu K. A note on choosing the number of factors. *Commun Statist Theory Methods.* 1999; 28:1653–1670.

25. Lee S, Song X. Bayesian selection on the number of factors in a factor analysis model. *Behaviormetrika*. 2002; 29:23–39.
26. Ferguson T. A Bayesian analysis of some nonparametric problems. *Ann Statistics*. 1973; 1:209–230.
27. Tipping M, Bishop C. Mixture of principal component analyzers. *Neural Comput*. 1999; 11:443–482. [PubMed: 9950739]
28. Eftekhari A, Wakin MB. Performance bounds for compressive sensing with a mixture of factor analyzers. School of Mines, Tech Rep. 2010
29. He L, Carin L. Exploiting structure in wavelet-based Bayesian compressive sensing. *IEEE Trans Signal Processing*. 2009; 57:3488–3497.
30. Salakhutdinov, R.; Mnih, A. Bayesian probabilistic matrix factorization with MCMC. *Proc. Advances in Neural Information Processing Systems*; 2008.
31. Salakhutdinov, R.; Mnih, A. Probabilistic matrix factorization. *Proc. Advances in Neural Information Processing Systems*; 2008.
32. Eldar YC, Mishali M. Robust recovery of signals from a union of subspaces. *IEEE Trans Inform Theory*. 2008; 12:1338–1351.
33. Hjort, N.; Holmes, C.; Muller, P.; Walker, S. *Bayesian Nonparametrics*. Cambridge, MA: Cambridge Univ. Press; 2010.
34. Kingman JFC. Completely random measures. *Pacific J Math*. 1967; 21(1):59–78.
35. Thibaux, R.; Jordan, MI. Hierarchical beta processes and the Indian buffet process. *Proc. Int. Workshop Artificial Intelligence and Statistics*; 2007.
36. Hjort NL. Nonparametric Bayes estimators based on beta processes in models for life history data. *Ann Stat*. 1990; 18(3):1259–1294.
37. Paisley, J.; Carin, L. Nonparametric factor analysis with beta process priors. *Proc. Int. Conf. Machine Learning (ICML)*; 2009.
38. Zhou, M.; Chen, H.; Paisley, J.; Ren, L.; Sapiro, G.; Carin, L. Non-parametric Bayesian dictionary learning for sparse image representations. *Proc. Conf. Advances in Neural Information Processing Systems, NIPS*; 2009.
39. Fox, E.; Sudderth, E.; Jordan, MI; Willsky, A. Sharing features among dynamical systems with beta processes. *Proc. Advances in Neural Information Processing (NIPS)*; Cambridge, MA: MIT Press; 2010.
40. Escobar MD. Estimating normal means with a Dirichlet process prior. *J Amer Stat Assoc*. 1994; 89:268–277.
41. Neal RM. Markov chain sampling methods for Dirichlet process mixture models. *J Comput Graph Stat*. 2000; 9:249–265.
42. Ishwaran H, James LF. Gibbs sampling methods for stick-breaking priors. *J Amer Stat Assoc Theory Methods*. 2001; 96(453):161–173.
43. Walker SG. Sampling the Dirichlet mixture model with slices. *Commun Stat Simulat Comput*. 2007; 36:45.
44. Teh, YW.; Jordan, MI. *Bayesian Nonparametrics: Principles and Practice*. Cambridge: U.K: Cambridge Univ. Press; 2010. Hierarchical Bayesian nonparametric models with applications.
45. Mairal, J.; Bach, F.; Ponce, J.; Sapiro, G.; Zisserman, A. Supervised dictionary learning. *Proc. Conf. Advances in Neural Information Processing Systems, NIPS*; 2008.
46. Mairal, J.; Bach, F.; Ponce, J.; Sapiro, G. Online dictionary learning for sparse coding. *Proc. Int. Conf. Machine Learning (ICML)*; 2009.
47. Candès E, Romberg J, Tao T. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans Inform Theory*. 2006; 52:489–509.
48. Candès EJ, Wakin MB. An introduction to compressive sampling. *IEEE Signal Processing Mag*. 2008; 25:21–30.
49. Candès EJ. The restricted isometry property and its implications for compressed sensing. *Compte Rendus de l'Academie des Sciences, Paris, Serie I*. 2008; 346:589–592.
50. Mairal J, Elad M, Sapiro G. Sparse representation for color image restoration. *IEEE Trans Image Processing*. 2008; 17

51. Lawrence, N.; Urtasun, R. Non-linear matrix factorization with Gaussian processes. Proc. Int. Conf. Machine Learning; 2009.
52. Candès EJ, Tao T. The power of convex relaxation: Near-optimal matrix completion. IEEE Trans Inform Theory. 2009; 56:2053–2080.
53. Candès, EJ.; Plan, Y. Matrix completion with noise. Proc. IEEE; 2010.
54. Goldwater, S.; Griffiths, T.; Johnson, M. Interpolating between types and tokens by estimating power-law generators. Proc. Conf. Advances in Neural Information Processing Systems, NIPS; 2006.
55. Teh, Y.; Gorur, D. Indian buffet processes with power-law behavior. Proc. Conf. Advances in Neural Information Processing Systems, NIPS; 2009.

Biographies

Lawrence Carin (lcarin@ece.duke.edu) earned his B.S., M.S., and Ph.D. degrees in electrical engineering from the University of Maryland, College Park, in 1985, 1986, and 1989, respectively. He is the William H. Younger Professor in the Electrical and Computer Engineering Department at Duke University, a cofounder of Signal Innovations Group, where he is the director of technology, and a Fellow of the IEEE. His recent research addresses applications of nonparametric Bayesian methods to general signal analysis, with an emphasis on sensing.

Richard G. Baraniuk (richb@rice.edu) is the Victor E. Cameron Professor of Electrical and Computer Engineering at Rice University. His work on the Rice single-pixel compressive camera was selected by *MIT Technology Review* as a “TR10 Top 10 Emerging Technology” for 2007. He is a Fellow of the IEEE and the American Association for the Advancement of Science and has received national young investigator Awards from the National Science Foundation and the Office of Naval Research, the Rosenbaum Fellowship from the Isaac Newton Institute of Cambridge University, the ECE Young Alumni Achievement Award from the University of Illinois, and the Wavelet Pioneer Award from SPIE. He is the founder of Connexions. His research interests lie in new theory, algorithms, and hardware for sensing and signal processing.

Volkan Cevher (volkan.cevher@epfl.ch) received his B.Sc. degree (valedictorian) in electrical engineering from Bilkent University in 1999, and his Ph.D. degree in electrical and computer engineering from the Georgia Institute of Technology in 2005. He was a research scientist at the University of Maryland, College Park (2006–2007) and at Rice University (2008–2009). Currently, he is an assistant professor at Ecole Polytechnique Federale de Lausanne, with a joint appointment at Idiap Research Institute and a faculty fellow at Rice University. His research interests include signal processing theory, machine learning, graphical models, and information theory.

David Dunson (dunson@stat.duke.edu) is the professor of statistical science at Duke University. He is a fellow of the American Statistical Association and the Institute of Mathematical Statistics. He is the winner of the 2007 Mortimer Spiegelman Award for the top public health statistician, the 2010 Myrto Lefkopoulou Distinguished Lectureship at Harvard University, and the 2010 COPSS Presidents’ Award for the top statistician under 41. His research focuses on developing Bayesian statistical methods motivated by high-dimensional and complex data sets.

Michael I. Jordan (jordan@eecs.Berkeley.edu) is the Pehong Chen Distinguished Professor in the Department of Electrical Engineering and Computer Science and the Department of Statistics at the University of California, Berkeley. He received the IEEE Neural Networks Pioneer Award in 2006, the SIAM Activity Group on Optimization Prize in 2008, and the

ACM/ AAAI Allen Newell Award in 2009. He was named as Neyman Lecturer and Medallion Lecturer by IMS. He is a Fellow of AAAS, IMS, IEEE, AAAI, and ASA. In 2010, he was named to the National Academy of Engineering and the National Academy of Sciences. His research has focused on Bayesian nonparametric analysis, probabilistic graphical models, spectral methods, kernel machines, and applications to problems in computational biology, information retrieval, signal processing, and speech recognition.

Guillermo Sapiro (guille@umn.edu) received his B.Sc. (summa cum laude), M.Sc., and Ph.D. degrees from the Department of Electrical Engineering at Technion, Israel Institute of Technology, in 1989, 1991, and 1993, respectively. After postdoctoral research at MIT, he became a member of technical staff at the research facilities of HP Labs in Palo Alto, California. He is currently with the Department of Electrical and Computer Engineering at the University of Minnesota, where he holds the position of Distinguished McKnight University Professor and Vincentine Hermes-Luh Chair in Electrical and Computer Engineering. He received the Gutwirth Scholarship for Special Excellence in Graduate Studies in 1991, the Ollendorff Fellowship for Excellence in Vision and Image Understanding Work in 1992, the Rothschild Fellowship for postdoctoral studies in 1993, the Office of Naval Research Young Investigator Award in 1998, the Presidential Early Career Awards for Scientist and Engineers (PECASE) in 1998, the NSF Career Award in 1999, and the National Security Science and Engineering Faculty Fellowship in 2010. He is the funding editor-in-chief of *SIAM Journal on Imaging Sciences*.

Michael B. Wakin (mwakin@mines.edu) earned his Ph.D. degree in electrical engineering from Rice University in 2007. He was an NSF mathematical sciences postdoctoral research fellow at the California Institute of Technology. Before joining the Division of Engineering at the Colorado School of Mines, he was an assistant professor at the University of Michigan. His research interests include sparse, geometric, and manifold-based models for signal and image processing, approximation, compression, CS, and dimensionality reduction. He is a recipient of the Hershel M. Rich Invention Award from Rice University and DARPA Young Faculty Award.

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.)

FIG 2.

Sparse signal recovery performed on an MFA model, inferred based on the training data using DP and BP. Column (a) shows the original data, and columns (b)–(g) represent CS recovery based on random compressive measurements. The results show the performance when the number of CS measurements are 5%, ..., 30% of the total number of pixels in the image. In columns (b)–(g), 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.)

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

FIG 4. Inpainting results. The curve shows the PSNR as a function of the 64 Gibbs learning rounds. Part (a) shows is the test image, with 80% of the RGB pixels missing, (b) is the result 64 after Gibbs rounds (final result), and (c) is the original uncontaminated image. (Results courtesy of M. Zhou.)

TABLE 1

RMSE OF UNION-OF-SUBSPACE MODEL ON 10 MILLION MOVIELENS DATA.

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

RESULTS COURTESY OF M. ZHOU.