Bayesian Nonparametric Matrix Factorization for Recorded Music

Matthew D. Hoffman MDHOFFMA@CS.PRINCETON.EDU
David M. Blei BLEI@CS.PRINCETON.EDU
Perry R. Cook PRC@CS.PRINCETON.EDU
Princeton University, Department of Computer Science, 35 Olden St., Princeton, NJ, 08540 USA

Abstract

Recent research in machine learning has focused on breaking audio spectrograms into separate sources of sound using latent variable decompositions. These methods require that the number of sources be specified in advance, which is not always possible. To address this problem, we develop *Gamma Process Nonnegative Matrix Factorization* (GaP-NMF), a Bayesian nonparametric approach to decomposing spectrograms. The assumptions behind GaP-NMF are based on research in signal processing regarding the expected distributions of spectrogram data, and GaP-NMF automatically discovers the number of latent sources. We derive a mean-field variational inference algorithm and evaluate GaP-NMF on both synthetic data and recorded music.

1. Introduction

Recent research in machine learning has focused on breaking audio spectrograms into separate sources of sound using latent variable decompositions. Such decompositions have been applied to identifying individual instruments and notes, e.g., for music transcription (Smaragdis & Brown, 2003), to predicting hidden or distorted signals (Bansal et al., 2005), and to source separation (Févotte et al., 2009).

A problem with these methods is that the number of sources must be specified in advance, or found with expensive techniques such as cross-validation. This problem is particularly relevant when analyzing music. We want the discovered latent components to correspond to real-world sound sources, and we cannot expect the same number of sources to be present in every recording.

In this article, we develop *Gamma Process Nonnegative Matrix Factorization* (GaP-NMF), a Bayesian nonparametric (BNP) approach to decomposing spectrograms. We posit a generative probabilistic model of spectrogram data where, given an observed audio signal, posterior inference reveals both the latent sources and their number.

The central computational challenge posed by our model is posterior inference. Unlike other BNP factorization methods, our model is not composed of conjugate pairs of distributions—we chose our distributions to be appropriate for spectrogram data, not for computational convenience.

We use variational inference to approximate the posterior, and develop a novel variational approach to inference in nonconjugate models. Variational inference approximates the posterior with a simpler distribution, whose parameters are optimized to be close to the true posterior (Jordan et al., 1999). In mean-field variational inference, each variable is given an independent distribution, usually of the same family as its prior. Where the model is conjugate, optimization proceeds by an elegant coordinate ascent algorithm. Researchers usually appeal to less efficient scalar optimization where conjugacy is absent. We instead use a *bigger* variational family than the model initially asserts. We show that this gives an analytic coordinate ascent algorithm, of the kind usually limited to conjugate models.

We evaluated GaP-NMF on several problems—extracting the sources from music audio, predicting the signal in missing entries of the spectrogram, and classical measures of Bayesian model fit. Our model performs as well as or better than the current state-of-the-art. It finds simpler representations of the data with equal statistical power, without needing to explore many fits over many numbers of sources, and thus with much less computation.

2. GaP-NMF Model

We model the Fourier power spectrogram $\boldsymbol{X}$ of an audio signal. The spectrogram $\boldsymbol{X}$ is an M by N matrix of nonnegative reals; the cell X_{mn} is the power of our input audio signal at time window n and frequency bin m. Each column of the power spectrogram is obtained as follows. First, take the discrete Fourier transform of a window of

Appearing in *Proceedings of the 27^{th} International Conference on Machine Learning*, Haifa, Israel, 2010.

$2(M-1)$ samples. Next, compute the squared magnitude of the complex value in each frequency bin. Finally, keep only the first M bins, since the remaining bins contain only redundant information.

We assume the audio signal is composed of K static sound sources. As a consequence, we can model the observed spectrogram $\boldsymbol{X}$ with the product of two non-negative matrices: an M by K matrix $\boldsymbol{W}$ describing these sources and a K by N matrix $\boldsymbol{H}$ controlling how the amplitude of each source changes over time (Smaragdis & Brown, 2003). Each column of $\boldsymbol{W}$ is the average power spectrum of an audio source; cell W_{mk} is the average amount of energy source k exhibits at frequency m. Each row of $\boldsymbol{H}$ is the time-varying gain of a source; cell H_{kn} is the gain of source k at time n. These matrices are unobserved.

Abdallah & Plumbley (2004) and Févotte et al. (2009) show that (under certain assumptions) mixing K sound sources in the time domain, with average power spectra defined by the columns of $\boldsymbol{W}$ and gains specified by the rows of $\boldsymbol{H}$, yields a mixture whose spectrogram $\boldsymbol{X}$ is distributed

$$X_{mn} \sim \text{Exponential}(\textstyle\sum_k W_{mk} H_{kn}). \tag{1}$$

Previous spectrogram decompositions assume the number of components K is known. In practice, this is rarely true. Our goal is to develop a method that infers both the characters and number of latent audio sources from data. We develop a Bayesian nonparametric model with an infinite number of latent components, a finite number of which are active when conditioned on observed data.

We now describe the Gamma Process Nonnegative Matrix Factorization model (GaP-NMF). As in previous matrix decomposition models, the spectrogram $\boldsymbol{X}$ arises from hidden matrices $\boldsymbol{W}$ and $\boldsymbol{H}$. In addition, the model includes a hidden vector of non-negative values $\boldsymbol{\theta}$, where each element θ_l is the overall gain of the corresponding source l. The key idea is that we allow for the *possibility* of a large number of sources L, but place a sparse prior on $\boldsymbol{\theta}$. During posterior inference, this prior biases the model to use no more sources than it needs.

Specifically, GaP-NMF assumes that $\boldsymbol{X}$ is drawn according to the following generative process:

$$\begin{aligned} W_{ml} &\sim \text{Gamma}(a, a) \\ H_{ln} &\sim \text{Gamma}(b, b) \\ \theta_l &\sim \text{Gamma}(\alpha/L, \alpha c) \\ X_{mn} &\sim \text{Exponential}(\textstyle\sum_l \theta_l W_{ml} H_{ln}). \end{aligned} \tag{2}$$

As the truncation level L increases towards infinity, the vector $\boldsymbol{\theta}$ approximates an infinite sequence drawn from a gamma process with shape parameter α and inverse-scale parameter αc (Kingman, 1993). A property of this sequence is that the number of elements K greater than some number $\epsilon > 0$ is finite almost surely. Specifically:

$$K \sim \text{Poisson}\left(\frac{1}{c}\int_{\epsilon}^{\infty} x^{-1} e^{-x\alpha c} dx\right). \tag{3}$$

For truncation levels L that are sufficiently large relative to the shape parameter α, we likewise expect that only a few of the L elements of $\boldsymbol{\theta}$ will be substantially greater than 0. During posterior inference, this property leads to a preference for explanations that use relatively few components.

Note that the expected value of X_{mn} under this model is constant with respect to L, α, a, and b:

$$\mathbb{E}_p[X_{mn}] = \textstyle\sum_l \mathbb{E}_p[\theta_l]\mathbb{E}_p[W_{ml}]\mathbb{E}_p[H_{ln}] = \frac{1}{c}. \tag{4}$$

This equation suggests the heuristic of setting the expected mean of the spectrogram $\boldsymbol{X}$ under the prior equal to its empirical mean $\bar{X}$ by setting $c = 1/\bar{X}$.

3. Variational Inference

Posterior inference is the central computational problem for analyzing data with the GaP-NMF model. Given an observed spectrogram $\boldsymbol{X}$, we want to compute the posterior distribution $p(\boldsymbol{\theta}, \boldsymbol{W}, \boldsymbol{H} | \boldsymbol{X}, \alpha, a, b, c)$. Exact Bayesian inference is intractable. We appeal to mean-field variational inference (Jordan et al., 1999).

Variational inference is a deterministic alternative to Markov Chain Monte Carlo (MCMC) methods that replaces sampling with optimization. It has permitted efficient large-scale inference for several Bayesian nonparametric models (e.g. Blei & Jordan, 2004; Doshi-Velez et al., 2009; Paisley & Carin, 2009). Variational inference algorithms approximate the true posterior distribution with a simpler variational distribution controlled by free parameters. These parameters are optimized to make the variational distribution close (in Kullback-Leibler divergence) to the true posterior of interest. Mean-field variational inference uses a fully factorized variational distribution—i.e., under the variational distribution all variables are independent. In conjugate models this permits easy coordinate ascent updates using variational distributions of the same families as the prior distributions.

Less frequently, variational methods are applied to non-conjugate models, which allow increased model expressivity at the price of greater algorithmic challenges. Our model is such a model. The usual strategy is to use a factorized variational distribution with the same families as the priors, bound or approximate the objective function, and use numerical techniques to optimize difficult parameters (Blei & Lafferty, 2006; Braun & McAuliffe, 2008).

We use a different strategy. We adopt an expanded family for our variational distributions, one that generalizes the

priors' family. This allows us to derive analytic coordinate ascent updates for the variational parameters, eliminating the need for numerical optimization.

3.1. Variational Objective Function

It is standard in mean-field variational inference to give each variable a variational distribution from the same family as its prior distribution (Jordan et al., 1999). We instead use the more flexible Generalized Inverse-Gaussian (GIG) family (Jørgenson, 1982):

$$\begin{aligned} q(W_{ml}) &= \text{GIG}(\gamma_{ml}^{(W)}, \rho_{ml}^{(W)}, \tau_{ml}^{(W)}) \\ q(H_{ln}) &= \text{GIG}(\gamma_{ln}^{(H)}, \rho_{ln}^{(H)}, \tau_{ln}^{(H)}) \\ q(\theta_l) &= \text{GIG}(\gamma_l^{(\theta)}, \rho_l^{(\theta)}, \tau_l^{(\theta)}). \end{aligned} \tag{5}$$

The GIG distribution is an exponential family distribution with sufficient statistics x, $1/x$, and $\log x$, and its PDF (in canonical exponential family form) is

$$\text{GIG}(y; \gamma, \rho, \tau) = \frac{\exp\{(\gamma-1)\log y - \rho y - \tau/y\}\rho^{\gamma/2}}{2\tau^{\gamma/2}\mathcal{K}_\gamma(2\sqrt{\rho\tau})}, \tag{6}$$

for $x \geq 0$, $\rho \geq 0$, and $\tau \geq 0$. ($\mathcal{K}_\nu(x)$ denotes a modified Bessel function of the second kind.)

Note that the GIG family's sufficient statistics (y, $1/y$, and $\log y$) are a superset of those of the gamma family (y and $\log y$), and so the gamma family is a special case of the GIG family where $\gamma > 0, \tau \to 0$.

To compute the bound in equation 8, we will need the expected values of each W_{ml}, H_{ln}, and θ_l and of their reciprocals under our variational GIG distributions. For a variable $y \sim \text{GIG}(\gamma, \rho, \tau)$ these expectations are

$$\mathbb{E}[y] = \frac{\mathcal{K}_{\gamma+1}(2\sqrt{\rho\tau})\sqrt{\tau}}{\mathcal{K}_\gamma(2\sqrt{\rho\tau})\sqrt{\rho}}; \quad \mathbb{E}\left[\frac{1}{y}\right] = \frac{\mathcal{K}_{\gamma-1}(2\sqrt{\rho\tau})\sqrt{\rho}}{\mathcal{K}_\gamma(2\sqrt{\rho\tau})\sqrt{\tau}}. \tag{7}$$

Having chosen a fully factorized variational family, we can lower bound the marginal likelihood of the input spectrogram under the GaP-NMF model (Jordan et al., 1999):

$$\begin{aligned} \log p(\boldsymbol{X}|\alpha, a, b, c) \geq\ & \mathbb{E}_q[\log p(\boldsymbol{X}|\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{\theta})] \\ &+ \mathbb{E}_q[\log p(\boldsymbol{W}|a)] - \mathbb{E}_q[\log q(\boldsymbol{W})] \\ &+ \mathbb{E}_q[\log p(\boldsymbol{H}|b)] - \mathbb{E}_q[\log q(\boldsymbol{H})] \\ &+ \mathbb{E}_q[\log p(\boldsymbol{\theta}|\alpha, c)] - \mathbb{E}_q[\log q(\boldsymbol{\theta})]. \end{aligned} \tag{8}$$

The difference between the left and right sides of equation 8 is the Kullback-Leibler (KL) divergence between the true posterior and the variational distribution q. Thus, maximizing this bound with respect to q minimizes the KL divergence between q and our posterior distribution of interest.

The second, third, and fourth lines of equation 8 can be computed using the expectations in equation 7.

The likelihood term in equation 8 expands to

$$\begin{aligned} &\mathbb{E}_q[\log p(\boldsymbol{X}|\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{\theta})] = \\ &\sum_{m,n} \mathbb{E}_q\left[\frac{-X_{mn}}{\sum_l \theta_l W_{ml} H_{ln}}\right] - \mathbb{E}_q\left[\log \sum_l \theta_l W_{ml} H_{ln}\right]. \end{aligned} \tag{9}$$

We cannot compute either of the expectations on the right. However, we can compute lower bounds on both of them.

First, the function $-x^{-1}$ is concave. Jensen's inequality says that for any vector $\boldsymbol{\phi}$ such that $\phi_l \geq 0$ and $\sum_l \phi_l = 1$

$$-\frac{1}{\sum_l x_l} = -\frac{1}{\sum_l \phi_l \frac{x_l}{\phi_l}} \geq -\sum_l \phi_l \frac{1}{\frac{x_l}{\phi_l}} = -\sum_l \phi_l^2 \frac{1}{x_l}. \tag{10}$$

We use this inequality to derive a bound on the first expectation in equation 9:

$$\mathbb{E}_q\left[\frac{-X_{mn}}{\sum_l \theta_l W_{ml} H_{ln}}\right] \geq \sum_l \phi_{lmn}^2 \mathbb{E}_q\left[\frac{-X_{mn}}{\theta_l W_{ml} H_{ln}}\right] \tag{11}$$

Second, the function $-\log x$ is convex. We can therefore bound the second expectation in equation 9 using a first-order Taylor approximation about an arbitrary (positive) point ω_{mn} as in (Blei & Lafferty, 2006) [1]:

$$\begin{aligned} -\mathbb{E}_q\left[\log \sum_l \theta_l W_{ml} H_{ln}\right] \geq & \\ -\log(\omega_{mn}) + 1 - \frac{1}{\omega_{mn}} \sum_l \mathbb{E}_q\left[\theta_l W_{ml} H_{ln}\right]. \end{aligned} \tag{12}$$

We use equations 11 and 12 to bound equation 9:

$$\begin{aligned} &\mathbb{E}_q[\log p(\boldsymbol{X}|\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{\theta})] \geq \\ &\sum_{m,n} -X_{mn} \sum_l \phi_{lmn}^2 \mathbb{E}_q\left[\frac{1}{\theta_l W_{ml} H_{ln}}\right] \\ &\quad -\log(\omega_{mn}) + 1 - \frac{1}{\omega_{mn}} \sum_l \mathbb{E}_q\left[\theta_l W_{ml} H_{ln}\right]. \end{aligned} \tag{13}$$

Note that this bound involves the expectations both of the model parameters *and* of their reciprocals under the variational distribution q. Since both y and $1/y$ are sufficient statistics of $\text{GIG}(y; \gamma, \rho, \tau)$, this will not pose a problem during inference, as it would if we were to use variational distributions from the gamma family.

We denote as $\mathcal{L}$ the sum of the likelihood bound in equation 13 and the second, third, and fourth lines of equation

[1] Braun & McAuliffe (2008) observe that this bound is maximized when the Taylor approximation is taken around the expected value of the argument to the logarithm function, which corresponds to the 0th-order delta method. However, retaining the "redundant" parameter ω_{mn} permits faster and simpler updates for our other parameters.

8. $\mathcal{L}$ lower bounds the likelihood $p(\boldsymbol{X}|\alpha, a, b, c)$. Our variational inference algorithm maximizes this bound over the free parameters, yielding an approximation $q(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{\theta})$ to the true posterior $p(\boldsymbol{W}, \boldsymbol{H}, \boldsymbol{\theta}|\boldsymbol{X}, \alpha, a, b, c)$.

3.2. Coordinate Ascent Optimization

We maximize the bound $\mathcal{L}$ using coordinate ascent, iteratively optimizing each parameter while holding all other parameters fixed. There are two sets of parameters to optimize: those used to bound the likelihood term in equation 9 and those that control the variational distribution q.

3.2.1. Tightening the likelihood bound

In equations 11 and 12, we derived bounds on the intractable expectations in equation 9. After updating the variational distributions on each set of parameters $\boldsymbol{W}$, $\boldsymbol{H}$, and $\boldsymbol{\theta}$, we update $\boldsymbol{\phi}$ and $\boldsymbol{\omega}$ to re-tighten these bounds.

Using Lagrange multipliers, we find that the optimal $\boldsymbol{\phi}$ is

$$\phi_{lmn} \propto \mathbb{E}_q\left[\frac{1}{\theta_l W_{ml} H_{ln}}\right]^{-1}. \tag{14}$$

The bound in equation 12 is tightest when

$$\omega_{mn} = \textstyle\sum_l \mathbb{E}_q\left[\theta_l W_{ml} H_{ln}\right]. \tag{15}$$

I.e., this bound is tightest when we take the Taylor approximation about the expected value of the function's argument.

3.2.2. Optimizing the variational distributions

The derivative of $\mathcal{L}$ with respect to any of $\gamma_{ml}^{(W)}$, $\rho_{ml}^{(W)}$, or $\tau_{ml}^{(W)}$ equals 0 when

$$\begin{aligned}\gamma_{ml}^{(W)} = a; \quad \rho_{ml}^{(W)} = a + \mathbb{E}_q[\theta_l] \textstyle\sum_n \frac{\mathbb{E}_q[H_{ln}]}{\omega_{mn}};\\ \tau_{ml}^{(W)} = \mathbb{E}_q\left[\tfrac{1}{\theta_l}\right] \textstyle\sum_n X_{mn} \phi_{lmn}^2 \mathbb{E}_q\left[\tfrac{1}{H_{ln}}\right].\end{aligned} \tag{16}$$

Simultaneously updating the parameters $\boldsymbol{\gamma}^{(\boldsymbol{W})}$, $\boldsymbol{\rho}^{(\boldsymbol{W})}$, and $\boldsymbol{\tau}^{(\boldsymbol{W})}$ according to equation 16 will maximize $\mathcal{L}$ with respect to those parameters.

Similarly, the derivative of $\mathcal{L}$ with respect to any of $\gamma_{ln}^{(H)}$, $\rho_{ln}^{(H)}$, or $\tau_{ln}^{(H)}$ equals 0 and $\mathcal{L}$ is maximized when

$$\begin{aligned}\gamma_{ln}^{(H)} = b; \quad \rho_{ln}^{(H)} = b + \mathbb{E}_q[\theta_l] \textstyle\sum_m \frac{\mathbb{E}_q[W_{ml}]}{\omega_{mn}};\\ \tau_{ln}^{(H)} = \mathbb{E}_q\left[\tfrac{1}{\theta_l}\right] \textstyle\sum_m X_{mn} \phi_{lmn}^2 \mathbb{E}_q\left[\tfrac{1}{W_{ml}}\right].\end{aligned} \tag{17}$$

Finally, the derivative of $\mathcal{L}$ with respect to any of $\gamma_l^{(\theta)}$, $\rho_l^{(\theta)}$, or $\tau_l^{(\theta)}$ equals 0 and $\mathcal{L}$ is maximized when

$$\begin{aligned}\gamma_l^{(\theta)} = \tfrac{\alpha}{L}; \quad \rho_l^{(\theta)} = \alpha c + \textstyle\sum_m \sum_n \frac{\mathbb{E}_q[W_{ml} H_{ln}]}{\omega_{mn}};\\ \tau_l^{(\theta)} = \textstyle\sum_m \sum_n X_{mn} \phi_{lmn}^2 \mathbb{E}_q\left[\tfrac{1}{W_{ml} H_{ln}}\right].\end{aligned} \tag{18}$$

We iterate between updating bound parameters and variational parameters according to equations 14, 15, 16, 17, and 18. Each update tightens the variational bound on $\log p(\boldsymbol{X}|\alpha, a, b, c)$, ultimately reaching a local optimum.

3.3. Accelerating Inference

Paisley & Carin (2009) observed that if $\mathbb{E}_q[\theta_l]$ becomes small for some component l, then we can safely skip the updates for the variational parameters associated with that component. (In our experiments we used 60 dB below $\sum_l \mathbb{E}_q[\theta_l]$ as a threshold.) This heuristic allows the use of large truncation levels L (yielding a better approximation to an infinite gamma process) without incurring too severe a performance penalty. The first few iterations will be expensive, but the algorithm will require less time per iteration as it becomes clear that only a small number of components (relative to L) are needed to explain the data.

4. Evaluation

We conducted several experiments to assess the decompositions provided by the GaP-NMF model. We tested GaP-NMF's ability to recover the true parameters used to generate a synthetic spectrogram, compared the marginal likelihoods of real songs under GaP-NMF to the marginal likelihoods of those songs under a simpler version of the model, evaluated GaP-NMF's ability to predict held-out data with a bandwidth expansion task, and evaluated GaP-NMF's ability to separate individual notes from mixed recordings.

We compared GaP-NMF to two variations on the same model:

Finite Bayesian model. This is a finite version of the GaP-NMF model fit using the same variational updates but without the top-level gain parameters $\boldsymbol{\theta}$. This simpler model's generative process is

$$\begin{aligned}W_{mk} \sim \text{Gamma}(a, ac); \quad H_{kn} \sim \text{Gamma}(b, b);\\ X_{mn} \sim \text{Exponential}(\textstyle\sum_k W_{mk} H_{kn}),\end{aligned} \tag{19}$$

where $k \in \{1, \ldots, K\}$ and the model order K is chosen a priori. The hyperparameters a, b, and c are set to the same values as in the GaP-NMF model in all experiments. We will refer to this model as GIG-NMF, for Generalized Inverse-Gaussian Nonnegative Matrix Factorization.

Finite non-Bayesian model. This model fits $\boldsymbol{W}$ and $\boldsymbol{H}$ to maximize the likelihood in equation 1. Févotte et al. (2009) derive iterative multiplicative updates to maximize this likelihood, calling the resulting algorithm Itakura-Saito Nonnegative Matrix Factorization (IS-NMF).

We also compared GaP-NMF to the two nonnegative matrix factorization (NMF) algorithms described by Lee & Seung (2001). Both of these algorithms also attempt to ap-

proximately decompose the spectrogram $\boldsymbol{X}$ into an M by K matrix $\boldsymbol{W}$ and a K by N matrix $\boldsymbol{H}$ so that $\boldsymbol{X} \approx \boldsymbol{W}\boldsymbol{H}$. The first algorithm, which we refer to as EU-NMF, minimizes the sum of the squared Euclidean distances between the elements of $\boldsymbol{X}$ and $\boldsymbol{W}\boldsymbol{H}$. The second algorithm, which we refer to as KL-NMF, minimizes the generalized KL-divergence between $\boldsymbol{X}$ and $\boldsymbol{W}\boldsymbol{H}$. KL-NMF (and its extensions) in particular is widely used to analyze audio spectrograms (e.g. Smaragdis & Brown, 2003; Bansal et al., 2005).

We focus on approaches that explain power spectrograms in terms of components that can be interpreted as audio power spectra. Other approaches may be useful for some tasks, but they do not decompose mixed audio signals into their component sources. This requirement excludes, for example, standard linear Gaussian factor models, whose latent factors cannot be interpreted as audio spectra unless audio signals are allowed to have negative power.

We normalized all spectrograms to have a maximum value of 1.0. (The high probability densities in our experiments result from low-power bins in the spectrograms.) To avoid numerical issues, we forced the values of the spectrograms to be at least 10^{-8}, 80 dB below the peak value of 1.0.

In all experiments, we initialized the variational parameters $\boldsymbol{\rho}$ for each W, H, and θ with random draws from a gamma distribution with shape parameter 100 and inverse-scale parameter 1000, the variational parameters $\boldsymbol{\tau}$ to 0.1, and each $\gamma_{mk}^{(W)} = a$, $\gamma_{kn}^{(H)} = b$, and $\gamma_k^{(\theta)} = \alpha/K$. This yields a diffuse and smooth initial variational posterior, which helped avoid local optima. We ran variational inference until the variational bound increased by less than 0.001%. The GIG-NMF and IS-NMF algorithms were optimized to the same criterion. KL-NMF and EU-NMF were iterated until their cost functions decreased by less than 0.01 and 0.001, respectively. (We found no gains in performance from letting EU-NMF or KL-NMF run longer.) All algorithms were implemented in MATLAB[2].

We found GaP-NMF to be insensitive to the choice of α, and so we set $\alpha = 1$ in all reported experiments.

4.1. Synthetic Data

We evaluated the GaP-NMF model's ability to correctly discover the latent bases that generated a matrix $\boldsymbol{X}$, and how many such bases exist. To test this, we fit GaP-NMF to random matrices $\boldsymbol{X}$ drawn according to the process:

$$\begin{aligned} W_{mk} &\sim \text{Gamma}(0.1, 0.1); \\ H_{kn} &\sim \text{Gamma}(0.1, 0.1); \\ X_{mn} &\sim \text{Exponential}(\textstyle\sum_k W_{mk} H_{kn}), \end{aligned} \tag{20}$$

[2]MATLAB code for GaP-NMF is available at http://www.cs.princeton.edu/~mdhoffma

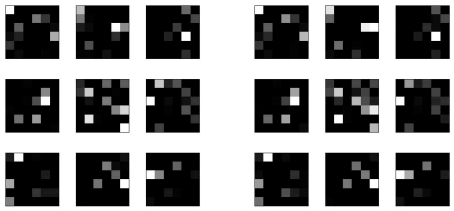

Figure 1. True synthetic bases (left) and expected values under the variational posterior of the nine bases found by the model (right). Brighter denotes more active. The 36-dimensional basis vectors are presented in 6×6 blocks for visual clarity.

where $m \in \{1, \ldots, M = 36\}$, $n \in \{1, \ldots, N = 300\}$, $k \in \{1, \ldots, K\}$ for $K = 9$.

We ran variational inference with the truncation level L set to 50, and hyperparameters $\alpha = 1$, $a = b = 0.1$, $c = 1/\bar{X}$ (where $\bar{X}$ is the mean of $\boldsymbol{X}$). After convergence, only nine of these components were associated with the observed data. (The smallest element of θ associated with one of these nine components was 0.06, while the next largest element was 2.4×10^{-8}). Figure 1 shows that the latent components discovered by the model correspond closely to those used to generate the data.

4.2. Marginal Likelihood

We want to evaluate the ability of GaP-NMF to choose a good number of components to model recorded music. To determine a "good" number of components, we use variational inference to fit GIG-NMF with various orders K and examine the resulting variational bounds on the marginal log-likelihood $\log p(\boldsymbol{X}|a, b, c)$.

As above, we set the prior parameters for the GaP-NMF model to $\alpha = 1$, $a = b = 0.1$, and $c = 1/\bar{X}$. We set the prior parameters for the simplified model to $a = b = 0.1$ and $c = 1/\bar{X}$. The value of 0.1 for a and b was chosen because it gave slightly better bounds than higher or lower values. The results were not very sensitive to α.

We computed power spectrograms from three songs: *Pink Moon* by Nick Drake, *Funky Kingston* by Toots and the Maytals, and a clip from the *Kreutzer Sonata* by Ludwig van Beethoven. These analyses used 2048-sample (46 ms) Hann windows with no overlap, yielding spectrograms of 1025 frequency bins by 2731, 6322, and 2584 time windows, respectively. We fit variational posteriors for GaP-NMF and GIG-NMF, conditioning on these spectrograms. We used a truncation level L of 100 for the nonparametric model, and values of K ranging from 1 to 100 for the finite GIG-NMF model.

The computational cost of fitting the GaP-NMF model was lower than the cost of fitting GIG-NMF with $K = 100$ (thanks to the accelerated inference trick in section 3.3),

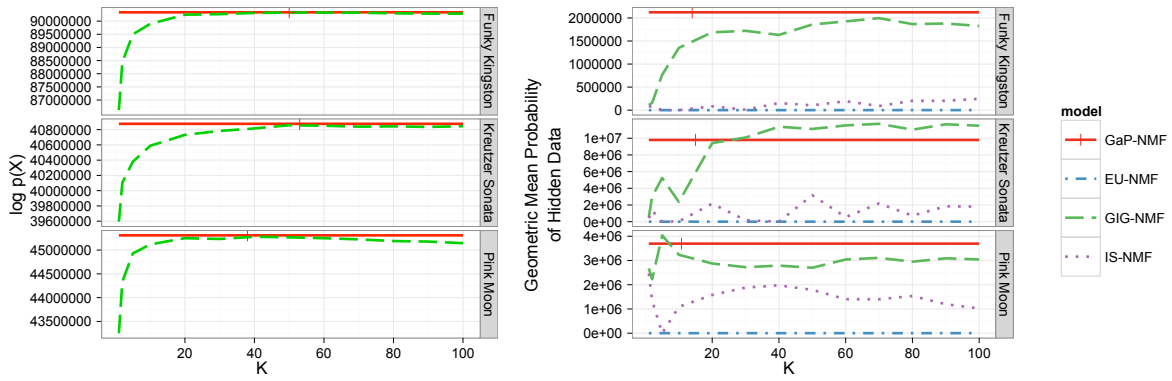

Figure 2. Left: Bounds on $\log p(\boldsymbol{X}|\text{prior})$ for the nonparametric GaP-NMF model and its parametric counterpart GIG-NMF with different numbers of latent components K. Ticks on the horizontal lines showing the bound for the GaP-NMF model indicate the number of components K used to explain the data. For all three songs the values of K chosen by GaP-NMF are close to the optimal value of K for the parametric model. Right: Geometric mean of the likelihood assigned to each censored observation by the nonparametric, finite, and unregularized models. Ticks again indicate the number of components K used to explain the data. The unregularized models overfit. EU-NMF performs badly, with likelihoods orders of magnitude lower than the other models.

and much lower than the cost of repeatedly fitting GIG-NMF with different values of K. For example, on a single core of a 2.3 GHz AMD Opteron 2356 Quad-Core Processor, fitting the 100-component GIG-NMF model to *Pink Moon* took 857 seconds, while fitting the GaP-NMF model to the same song took 645 seconds.

The results are summarized in figure 2 (left). The GaP-NMF model used 50, 53, and 38 components to explain the spectrograms of *Funky Kingston*, the *Kreutzer Sonata*, and *Pink Moon* respectively. In each case the value of K chosen by GaP-NMF was close to the best value of K tested for the GIG-NMF model. This suggests that GaP-NMF performs automatic order selection as well as the more expensive approach of fitting multiple finite-order models.

4.3. Bandwidth Expansion

One application of statistical spectral analysis is bandwidth expansion, the problem of inferring what the high-frequency content of a signal is likely to be given only the low-frequency content of the signal (Bansal et al., 2005). This task has applications to restoration of low-bandwidth audio and lossy audio compression. This is a missing data problem. We compared the ability of different models and inference algorithms to predict the held-out data.

We computed a power spectrogram from 4000 1024-sample (23 ms) Hann windows taken from the middles of the same three songs used to evaluate marginal likelihoods: *Funky Kingston*, the *Kreutzer Sonata*, and *Pink Moon*. For each song, this yielded a 513×4000 spectrogram $\boldsymbol{X}$ describing 93 seconds of the song. We ran five-fold cross-validation to compare GaP-NMF's predictions of the missing high-frequency content to those of GIG-NMF, EU-NMF, and IS-NMF. (It is more difficult to evaluate KL-NMF's ability to predict missing data, since it does not correspond to a probabilistic model of continuous data.) We divided each spectrogram into five contiguous 800-frame sections. For each fold, we censored the top two octaves (i.e., the top 384 out of 513 frequency bins) of one of those sections. We then predicted the values in the censored bins based on the data in the uncensored bins.

The prior hyperparameters for the Bayesian models were set to $a = b = 1$, $c = 1/\bar{X}$, and $\alpha = 1$ (for GaP-NMF). We chose a higher value for a and b for this experiment since stronger smoothing can improve the models' ability to generalize to held-out data.

For each fit model, we computed an estimate X_{mn}^{pred} of each missing value X_{mn}^{miss}. For the models fit using variational inference, we used the expected value of the missing data under the variational posterior q, $\mathbb{E}_q[\mathbb{E}_p[X_{mn}^{\text{miss}}]]$. For the GaP-NMF model, this expectation is

$$X_{mn}^{\text{pred}} = \mathbb{E}_q[\mathbb{E}_p[X_{mn}^{\text{miss}}]] = \textstyle\sum_k \mathbb{E}_q[\theta_k W_{mk} H_{kn}],$$

and for the GIG-NMF model it is

$$X_{mn}^{\text{pred}} = \mathbb{E}_q[\mathbb{E}_p[X_{mn}^{\text{miss}}]] = \textstyle\sum_k \mathbb{E}_q[W_{mk} H_{kn}].$$

For IS-NMF and EU-NMF we predicted $X_{mn}^{\text{pred}} = \sum_k W_{mk} H_{kn}$.

To evaluate the quality of fit for IS-NMF, GaP-NMF, and GIG-NMF, we compute the likelihood of each unobserved element X_{mn}^{miss} under an exponential distribution with mean X_{mn}^{pred}. To evaluate EU-NMF, we first compute the mean squared error of the estimate of the observed data $\sigma^2 = \text{Mean}[(X^{\text{obs}} - [\boldsymbol{WH}]^{\text{obs}})^2]$. We then compute the likelihood of each unobserved element X_{mn}^{miss} under a normal distribution with mean X_{mn}^{pred} and variance σ^2.

Figure 2 (right) plots the geometric mean of the likelihood of each unobserved element of $\boldsymbol{X}$ for the nonparametric model and for models fit with different numbers of components K. The Bayesian models do very well compared with the unregularized models, which overfit badly for any number of components K greater than 1. GaP-NMF used fewer components to explain the songs than in the previous experiment, which we attribute to the stronger smoothing, smaller number of observations, and smaller window size.

4.4. Blind Monophonic Source Separation

GaP-NMF can also be applied to blind source separation, where the goal is to recover a set of audio signals that combined to produce a mixed audio signal. For example, we may want to separate a polyphonic piece of music into notes to facilitate transcription (Smaragdis & Brown, 2003), denoising, or upmixing (Févotte et al., 2009).

The GaP-NMF model assumes that the audio signal is a linear combination of L sources (some of which have extremely low gain). Given the complex magnitude spectrogram $\boldsymbol{X}^c$ of the original audio and an estimate of the model parameters $\boldsymbol{W}$, $\boldsymbol{H}$, and $\boldsymbol{\theta}$, we can compute maximum-likelihood estimates of the spectrograms of the L unmixed sources using Wiener filtering (Févotte et al., 2009):

$$\hat{X}_{lmn} = X^c_{mn} \frac{\theta_l W_{ml} H_{ln}}{\sum_{i \in \{1,\dots,L\}} \theta_i W_{mi} H_{in}}, \qquad (21)$$

where $\hat{X}_{lmn}$ is the estimate of the complex magnitude spectrum of the lth source at time n and frequency n. We can invert these spectrograms to obtain estimates of the audio signals that are combined in the mixed audio signal.

We evaluated GaP-NMF's ability to separate signals from two synthesized pieces of music. We used synthetic music rather than live performances so that we could easily isolate each note. The pieces we used were a randomly generated four-voice clarinet piece using 21 unique notes, and a two-part Bach invention synthesized using a physical model of a vibraphone using 36 unique notes.

We compared the Signal-to-Noise Ratios (SNRs) of the separated tracks for the GaP-NMF model with those obtained using GIG-NMF, IS-NMF, EU-NMF, and KL-NMF. For the finite models we also used Wiener filtering to separate the tracks, dropping θ from equation 21.

The models do not provide any explicit information about the correspondence between sources and notes. To decide which separated signal to associate with which note, we adopt the heuristic of assigning each note to the component k whose gain signal $\boldsymbol{H}_k$ has the highest correlation with the power envelope of the true note signal. We only consider the V components that make the largest contribution to the mixed signal, where V is the true number of notes.

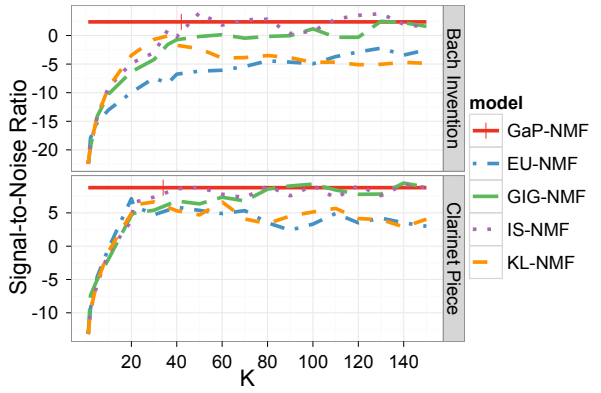

Figure 3. Average Signal-to-Noise Ratios (SNRs) across notes in the source separation task. Approaches based on the exponential likelihood model do well, EU-NMF and KL-NMF do less well. Ticks on the horizontal lines showing GaP-NMF's performance denote the final number of components K used to explain the data.

Figure 3 shows the average SNRs of the tracks corresponding to individual notes for each piece. The approaches based on the exponential likelihood model do comparably well. The KL-NMF and EU-NMF models perform considerably worse, and are sensitive to the model order K. GaP-NMF decomposed the clarinet piece into 34 components, and the Bach invention into 42 components. In both cases, some of these components were used to model the temporal evolution of the instrument sounds.

5. Related Work

GaP-NMF is closely related to recent work in Bayesian nonparametrics and probabilistic interpretations of NMF.

Bayesian Nonparametrics Most of the literature on Bayesian nonparametric latent factor models focuses on conjugate linear Gaussian models in conjunction with the Indian Buffet Process (IBP) (Griffiths & Ghahramani, 2005) or the Beta Process (BP) (Thibaux & Jordan, 2007), using either MCMC or variational methods for posterior inference (e.g. Doshi-Velez et al., 2009; Paisley & Carin, 2009)). (An exception that uses MCMC in non-conjugate infinite latent factor models is (Teh et al., 2007).)

A standard linear Gaussian likelihood model is not appropriate for audio spectrogram data, whereas the exponential likelihood model has theoretical justification and gives interpretable components. The nonlinearity and lack of conjugacy of the exponential likelihood model make inference using an IBP or BP difficult. Our use of a gamma process prior allows us to derive an infinite latent factor model that is appropriate for audio spectrograms and permits a simple and efficient variational inference algorithm.

Probabilistic NMF Févotte & Cemgil (2009) sug-

gest the outline of a variational inference algorithm for Itakura-Saito NMF based on the space-alternating generalized expectation-maximization algorithm in Févotte et al. (2009). This approach introduces $K \times M \times N$ complex hidden variables whose posteriors must be estimated. In our informal experiments, this gave a much looser variational bound, much longer convergence times, and a less flexible approximate posterior than the variational inference algorithm presented in this paper.

Our approach of weighting the contribution of each component k by a parameter θ_k resembles the strategy of Automatic Relevance Determination (ARD), which has been used in a Maximum A Posteriori (MAP) estimation algorithm for a different NMF cost function (Tan & Févotte, 2009). Though similar in spirit, this ARD approach is less amenable to fully Bayesian inference.

6. Discussion

We developed the GaP-NMF model, a Bayesian nonparametric model capable of determining the number of latent sources needed to explain an audio spectrogram. We demonstrated the effectiveness of the GaP-NMF model on several problems in analyzing and processing recorded music. Although this paper has focused on analyzing music, GaP-NMF is equally applicable to other types of audio, such as speech or environmental sounds.

Acknowledgments

We thank the reviewers for their helpful observations and suggestions. David M. Blei is supported by ONR 175-6343, NSF CAREER 0745520.

References

Abdallah, S.A. and Plumbley, M.D. Polyphonic music transcription by non-negative sparse coding of power spectra. In *Proc. 5th Int'l Conf. on Music Information Retrieval (ISMIR)*, pp. 10–14, 2004.

Bansal, D., Raj, B., and Smaragdis, P. Bandwidth expansion of narrowband speech using non-negative matrix factorization. In *Proc. 9th European Conf. on Speech Communication and Technology*, 2005.

Blei, D. and Jordan, M. Variational methods for the Dirichlet process. In *Proc. 21st Int'l Conf. on Machine Learning*, 2004.

Blei, D. and Lafferty, J. Correlated topic models. In *Advances in Neural Information Processing Systems 18 (NIPS) 18*, pp. 147–154. MIT Press, 2006.

Braun, M. and McAuliffe, J. Variational inference for large-scale models of discrete choice. *arXiv*, (0712.2526), 2008.

Doshi-Velez, F., Miller, K.T., Van Gael, J., and Teh, Y.W. Variational inference for the indian buffet process. In *Proc. 13th Int'l Conf. on Artificial Intelligence and Statistics*, pp. 137–144, 2009.

Févotte, C. and Cemgil, A.T. Nonnegative matrix factorizations as probabilistic inference in composite models. In *Proc. 17th European Signal Processing Conf. (EUSIPCO), Glasgow, Scotland*, 2009.

Févotte, C., Bertin, N., and Durrieu, J.L. Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis. *Neural Computation*, 21 (3):793–830, 2009.

Griffiths, T.L. and Ghahramani, Z. Infinite latent feature models and the indian buffet process. In *Advances in Neural Information Processing Systems 17 (NIPS)*, pp. 475–482. MIT Press, 2005.

Jordan, M., Ghahramani, Z., Jaakkola, T., and Saul, L. Introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999.

Jørgenson, Bent. *Statistical properties of the generalized inverse-Gaussian distribution*. Springer-Verlag, New York, 1982.

Kingman, J.F.C. *Poisson processes*. Oxford University Press, USA, 1993.

Lee, D.D. and Seung, H.S. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems 13 (NIPS)*, pp. 556–562. MIT; 1998, 2001.

Paisley, J. and Carin, L. Nonparametric factor analysis with beta process priors. In *Proc. 26th Int'l Conf. on Machine Learning*, 2009.

Smaragdis, P. and Brown, J.C. Non-negative matrix factorization for polyphonic music transcription. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 177–180, 2003.

Tan, V.Y.F. and Févotte, C. Automatic relevance determination in nonnegative matrix factorization. In *Proc. Workshop on Signal Processing with Adaptative Sparse Structured Representations (SPARS09)*, 2009.

Teh, Y., Gorur, D., and Ghahramani, Z. Stick-breaking construction for the Indian buffet process. In *11th Conf. on Artificial Intelligence and Statistics*, 2007.

Thibaux, R. and Jordan, M. Hierarchical beta processes and the Indian buffet process. In *11th Conf. on Artificial Intelligence and Statistics*, 2007.