BETA PROCESS SPARSE NONNEGATIVE MATRIX FACTORIZATION FOR MUSIC

Dawen Liang
LabROSA, EE Dept.
Columbia University
dl2771@columbia.edu

Matthew D. Hoffman
Adobe Research
Adobe Systems Incorporated
mathoffm@adobe.com

Daniel P. W. Ellis
LabROSA, EE Dept.
Columbia University
dpwe@ee.columbia.edu

ABSTRACT

Nonnegative matrix factorization (NMF) has been widely used for discovering physically meaningful latent components in audio signals to facilitate source separation. Most of the existing NMF algorithms require that the number of latent components is provided *a priori*, which is not always possible. In this paper, we leverage developments from the Bayesian nonparametrics and compressive sensing literature to propose a probabilistic *Beta Process Sparse NMF* (BP-NMF) model, which can automatically infer the proper number of latent components based on the data. Unlike previous models, BP-NMF explicitly assumes that these latent components are often completely silent. We derive a novel mean-field variational inference algorithm for this nonconjugate model and evaluate it on both synthetic data and real recordings on various tasks.

1. INTRODUCTION

Nonnegative matrix factorization (NMF) [9] has been extensively applied to analyze audio signals, since the approximate decomposition of the audio spectrogram into the product of 2 nonnegative matrices $\mathbf{X} \approx \mathbf{WH}$ provides a physically meaningful interpretation. We can view each column of $\mathbf{X}$, which represents the power density across frequencies at a particular time, as a nonnegative linear combination of the columns of $\mathbf{W}$, determined by the column of activation $\mathbf{H}$. Thus $\mathbf{W}$ can be considered as a dictionary, where each column acts as a component. This can be particularly useful for audio source separation, where the goal is to find out the individual sources from mixed signal.

Audio source separation poses a meaningful and challenging problem, which has been actively studied for the last few decades. One of the obstacles which makes source separation difficult is that the number of sources is generally not known. For example, when we listen to a piece of polyphonic music, it is difficult and tedious to figure out how many notes or instruments are being played. However, most existing NMF algorithms require the number of components to be provided as input, based on the assumption that there exists a certain mapping between the learned components and real sources. To address this issue, we propose BP-NMF, a nonparametric Bayesian NMF model that uses a beta process prior. The model automatically determines how many sources it needs to explain the data during posterior inference.

1.1 Related Work

NMF has been applied to many music analysis problems such as music transcription [1, 12], music analysis [5], and music source separation [10, 15].

On the other hand, most of the literature on nonparametric Bayesian latent factor models focuses on conjugate linear Gaussian models, for example, beta process factor analysis [11] which is the main inspiration for BP-NMF. However, such models are not appropriate for audio spectrograms as they do not impose nonnegativity constraints. To address this limitation, [7] proposed a nonparametric Bayesian NMF model based on the gamma process.

BP-NMF extends the standard NMF model in two ways:

- BP-NMF can explicitly and completely silence latent components when they should not be active. This captures the intuition that a note which appears frequently during one phrase may not contribute anything in another phrase, and most notes are silent most of the time.
- The number of latent components, which is difficult to set *a priori*, is inferred by the model.

Both of these issues have been addressed in previous work, but to the authors' knowledge, BP-NMF is the first model to combine them.

2. BP-NMF

We adopt the notational conventions that upper case bold letters (e.g. $\mathbf{X}, \mathbf{D}$, $\mathbf{S}$ and $\mathbf{Z}$) denote matrices and lower case bold letters (e.g. $\boldsymbol{x}$, $\boldsymbol{d}$ $\boldsymbol{s}$, and $\boldsymbol{z}$) denote vectors. $f \in \{1, 2, \cdots, F\}$ is used to index frequency. $t \in \{1, 2, \cdots, T\}$ is used to index time. $k \in \{1, 2, \cdots, K\}$ is used to index dictionary components.

BP-NMF is formulated as:

$$\mathbf{X} = \mathbf{D}(\mathbf{S} \odot \mathbf{Z}) + \mathbf{E} \tag{1}$$

where $\mathbf{X}$ is a $F \times T$ spectrogram and $\mathbf{D}$ is a $F \times K$ dictionary with K components. The activation is the Hadamard product between a nonnegative matrix $\mathbf{S}$ and a binary mask $\mathbf{Z}$, both of which have the shape of $K \times T$. $\mathbf{E}$ is an i.i.d. Gaussian noise matrix which has the same shape as $\mathbf{X}$. We use Gaussian noise model instead of Poisson or exponential mainly for mathematical convenience and extending BP-NMF to more audio-oriented model is part of the future work. Unlike previous models, BP-NMF can explicitly silence some components by turning off the corresponding elements in $\mathbf{Z}$. For example, when a clarinet is playing A3 the model should silence all clarinet notes that are not A3.

We place a beta process [6] prior on the binary mask $\mathbf{Z}$ so that the number of components K can potentially go to infinity and the inference algorithm will choose the proper number to describe the data. We adopt the finite approximation to a beta process in [11]:

$$\begin{aligned} Z_{kt} &\sim \text{Bernoulli}(\pi_k) \\ \pi_k &\sim \text{Beta}(a_0/K, b_0(K-1)/K) \end{aligned} \tag{2}$$

where K is set to a large number. As shown in [4], a finite approximation for Indian buffet process [1] performs comparably well as the infinite model. In this formulation π_k explicitly controls the prevalence of each individual component; the closer π_k is to zero, the less frequently it will contribute to $\mathbf{X}$. The rest of the model is specified as:

$$\begin{aligned} \log(\boldsymbol{d}_k) &\sim \mathcal{N}(0, \mathbf{I}_F) & \boldsymbol{\epsilon}_t &\sim \mathcal{N}(0, \gamma_\epsilon^{-1}\mathbf{I}_F) \\ \boldsymbol{s}_t &\sim \text{Gamma}(\alpha, \beta) & \gamma_\epsilon &\sim \text{Gamma}(c_0, d_0) \end{aligned} \tag{3}$$

The choice of component $\boldsymbol{d}_k$ being lognormal distributed will become natural as we describe our inference algorithm in Section 3.2. For activation $\boldsymbol{s}_t$, being gamma distributed instead of lognormal distributed is easier to extend to a time-dependent gamma chain prior [5]. The full model is summarized in Figure 1.

Exact inference for this model is infeasible, so we instead derive a mean-field [8] variational inference algorithm. Note that this model is nonconjugate between the observation $\mathbf{X}$ and priors $\mathbf{D}$ and $\mathbf{S}$, which makes deriving an inference algorithm more difficult.

3. VARIATIONAL INFERENCE

3.1 Laplace Approximation Variational Inference

Since BP-NMF is nonconjugate, we use Laplace approximation variational inference [16].

Given a probabilistic model $P(X, \Theta)$ where X denotes the observation and Θ denotes hidden variables, mean-field inference approximates the posterior $P(\Theta|X)$ with a fully factorized variational distribution $q(\Theta) = \prod_i q(\theta_i)$ by minimizing the KL-divergence between the variational distribution and the true posterior. Inference can be carried out via coordinate descent for each hidden variable θ_i which is guaranteed to find a local optimum.

[1] It has been shown [13] that beta process is the de Finetti mixing measure for the Indian buffet process.

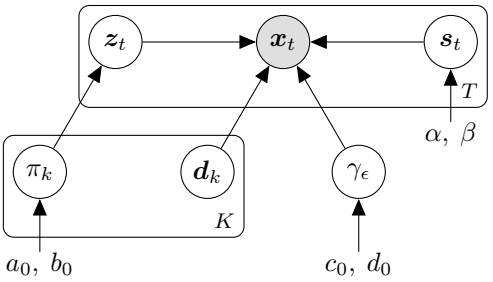


Figure 1: Graphical model representation of BP-NMF. Shaded node represents observed variable (spectrogram). Unshaded nodes represent hidden variables. A directed edge from node a to node b denotes that the variable b depends on the value of variable a. Plates denote replication by the value in the lower right of the plate.

The general mean-field update for a variable θ_i can be shown [3] to be:

$$q(\theta_i) \propto \exp\{\langle \log P(X, \Theta) \rangle_{-\theta_i}\} \tag{4}$$

where $\langle \cdot \rangle_{-\theta_i}$ indicates the expectation with respect to $q(\Theta \setminus \{\theta_i\})$. For simplicity, we will omit the subscript $-\theta_i$ when there is no ambiguity.

For nonconjugate model, we cannot write Eq. 4 in closed form. The Laplace method is used to approximate $q(\theta_i)$:

$$\begin{aligned} f(\theta_i) &= \langle \log P(X, \Theta) \rangle_{-\theta_i} \\ &\approx f(\hat{\theta}_i) + \frac{1}{2}(\theta_i - \hat{\theta}_i)^T H(\hat{\theta}_i)(\theta_i - \hat{\theta}_i) \end{aligned} \tag{5}$$

A second-order Taylor expansion is taken in (5) where $\hat{\theta}_i$ is a local maximum of $f(\theta_i)$ and $H(\hat{\theta}_i)$ is the Hessian matrix. This suggests that we use a Gaussian variational distribution for $q(\theta_i)$:

$$q(\theta_i) = \mathcal{N}(\hat{\theta}_i, -H(\hat{\theta}_i)^{-1}) \tag{6}$$

$H(\hat{\theta}_i)$ is guaranteed to be negative definite at any local maximum of $f(\theta)$, provided $f(\theta)$ is smooth. $-H(\hat{\theta}_i)^{-1}$ is therefore a valid covariance matrix. Conjugate gradient or L-BFGS can be used to search for $\hat{\theta}_i$.

3.2 Inference for BP-NMF

Laplace approximation variational inference assumes that the nonconjugate continuous variables are unconstrained, thus we reparametrize $\{\mathbf{D}, \mathbf{S}\}$ as $\{\mathbf{\Phi}, \mathbf{\Psi}\}$, where $\mathbf{\Phi} \in \mathbb{R}^{F \times K}$ with $\Phi_{fk} = \log(D_{fk})$ and $\mathbf{\Psi} \in \mathbb{R}^{K \times T}$ with $\Psi_{kt} = \log(S_{kt})$. The fully-factorized variational distribution is:

$$q(\Theta) = q(\gamma_\epsilon) \prod_{k=1}^{K} q(\pi_k) \Big(\prod_{f=1}^{F} q(\Phi_{fk}) \Big) \prod_{t=1}^{T} q(\Psi_{kt}) q(Z_{kt})$$

where the variational distributions are specified as:

$$\begin{aligned} q(\Phi_{fk}) &= \mathcal{N}(\mu_{fk}^{(\Phi)}, 1/\gamma_{fk}^{(\Phi)}) \\ q(\Psi_{kt}) &= \mathcal{N}(\mu_{kt}^{(\Psi)}, 1/\gamma_{kt}^{(\Psi)}) \\ q(Z_{kt}) &= \text{Bernoulli}(p_{kt}^{(z)}) \\ q(\pi_k) &= \text{Beta}(\alpha_k^{(\pi)}, \beta_k^{(\pi)}) \\ q(\gamma_\epsilon) &= \text{Gamma}(\alpha^{(\epsilon)}, \beta^{(\epsilon)}) \end{aligned} \tag{7}$$

We will briefly describe the mean-field update and a Python implementation is available online [2].

3.2.1 Update $\mathbf{\Phi}$ and $\mathbf{\Psi}$

Following Eq. 4, we can write $q(\Phi_{fk})$ as:

$$\begin{aligned} q(\Phi_{fk}) &\propto \exp\{\langle \log P(\mathbf{X}, \Theta)\rangle_{-\Phi_{fk}}\} \\ &\propto \exp\{\langle \log P(\boldsymbol{x}_f|\boldsymbol{\phi}_f, \mathbf{\Psi}, \mathbf{Z}, \gamma_\epsilon)\rangle + \log P(\Phi_{fk})\} \\ &= \exp\{f(\Phi_{fk})\} \end{aligned} \tag{8}$$

and express $P(\boldsymbol{x}_f|\boldsymbol{\phi}_f, \mathbf{\Psi}, \mathbf{Z}, \gamma_\epsilon)$ in exponential family form:

$$\begin{aligned} &\langle \log P(\boldsymbol{x}_f|\boldsymbol{\phi}_f, \mathbf{\Psi}, \mathbf{Z}, \gamma_\epsilon)\rangle \\ =&\langle \eta(\boldsymbol{\phi}_f, \mathbf{\Psi}, \mathbf{Z}, \gamma_\epsilon)^T\rangle T(\boldsymbol{x}_f) - \langle A(\eta)\rangle. \end{aligned} \tag{9}$$

For BP-NMF, both $\langle \eta(\boldsymbol{\phi}_f, \boldsymbol{\psi}_t, \boldsymbol{z}_t, \gamma_\epsilon)\rangle$ and $\langle A(\eta)\rangle$ can be computed in closed form. Thus, we can search for a local maximum $\hat{\Phi}_{fk}$. The mean-field update following Eq. 6 is:

$$\begin{aligned} \mu_{fk}^{(\Phi)} &\leftarrow \hat{\Phi}_{fk}, \\ \gamma_{fk}^{(\Phi)} &\leftarrow -\frac{\partial^2 f}{\partial \Phi_{fk}^2}(\hat{\Phi}_{fk}). \end{aligned} \tag{10}$$

The update for Ψ_{kt} is basically the same as Φ_{fk}, except that $P(\Psi_{kt})$ is a log-gamma distribution:

$$P(\Psi_{kt}) \propto \exp\{\alpha\Psi_{kt} - \beta\exp\{\Psi_{kt}\}\}. \tag{11}$$

3.2.2 Update $\mathbf{Z}$

Similarly, we can follow Eq. 4:

$$\begin{aligned} q(Z_{kt}) &\propto \exp\{\langle \log P(\mathbf{X}, \Theta)\rangle_{-Z_{kt}}\} \\ &\propto \exp\{\langle \log P(\boldsymbol{x}_t|\mathbf{\Phi}, \boldsymbol{\psi}_t, \mathbf{z}_t, \gamma_\epsilon)\rangle + \langle \log P(Z_{kt}|\pi_k)\rangle\} \end{aligned} \tag{12}$$

Since Z_{kt} is Bernoulli distributed, we can explicitly compute $P_0 \propto q(Z_{kt} = 0)$ and $P_1 \propto q(Z_{kt} = 1)$, respectively. Then the update for Z_{kt} can be carried out:

$$p_{kt}^{(z)} \leftarrow \frac{P_1}{P_0 + P_1} \tag{13}$$

3.2.3 Update $\boldsymbol{\pi}$ and γ_ϵ

In BP-NMF, $\boldsymbol{\pi}$ and $\mathbf{Z}$ are conjugate, therefore we can directly derive the mean-field update for $\boldsymbol{\pi}$ in closed form:

$$\begin{aligned} \alpha_k^{(\pi)} &\leftarrow \frac{a_0}{K} + \sum_{t=1}^{T}\langle Z_{kt}\rangle \\ \beta_k^{(\pi)} &\leftarrow \frac{b_0(K-1)}{K} + T - \sum_{t=1}^{T}\langle Z_{kt}\rangle \end{aligned} \tag{14}$$

Similarly, γ_ϵ can also be updated in closed form:

$$\begin{aligned} \alpha^{(\epsilon)} &\leftarrow c_0 + \frac{1}{2}FT \\ \beta^{(\epsilon)} &\leftarrow d_0 + \frac{1}{2}\sum_{t=1}^{T} \parallel \boldsymbol{x}_t - \langle \mathbf{D}(\boldsymbol{s}_t \odot \boldsymbol{z}_t)\rangle \parallel_2^2 \end{aligned} \tag{15}$$

[2] https://github.com/dawenl/bp_nmf

3.3 Accelerating inference

Both [11] and [7] proposed heuristics to speed up the inference. In general, we want to set number of dictionary components K to be large so that it can better approximate the infinite functional prior. On the other hand, a large value of K will dramatically increase the time for inference. This can be compensated by setting K initially to a large value and truncating the rarely-used dictionary components as the inference proceeds. The heuristic applied for BP-NMF is that, for dictionary component $\boldsymbol{d}_k$, if the corresponding π_k drops below 10^{-3} of the maximal $\boldsymbol{\pi}$, we skip it during the inference. The first few iterations may be slow, but inference accelerates as more elements of $\boldsymbol{\pi}$ are driven towards 0.

4. EXPERIMENTS

We conducted a set of experiments to evaluate if BP-NMF can effectively capture the latent components from music recordings. First, we performed a sanity check on a synthetic example. Then we tested BP-NMF on 2 different tasks: bandwidth expansion and blind source separation. We also designed a transcription-based mechanism to evaluate the quality of the learned dictionary.

All experiments were done on magnitude spectrum with hyperparameters $\alpha = \beta = 2$, $a_0 = b_0 = 1$, and $c_0 = d_0 = 10^{-6}$. All the variational parameters were randomly initialized. The initial K was set to 512. All recordings were sampled at 22.05 kHz.

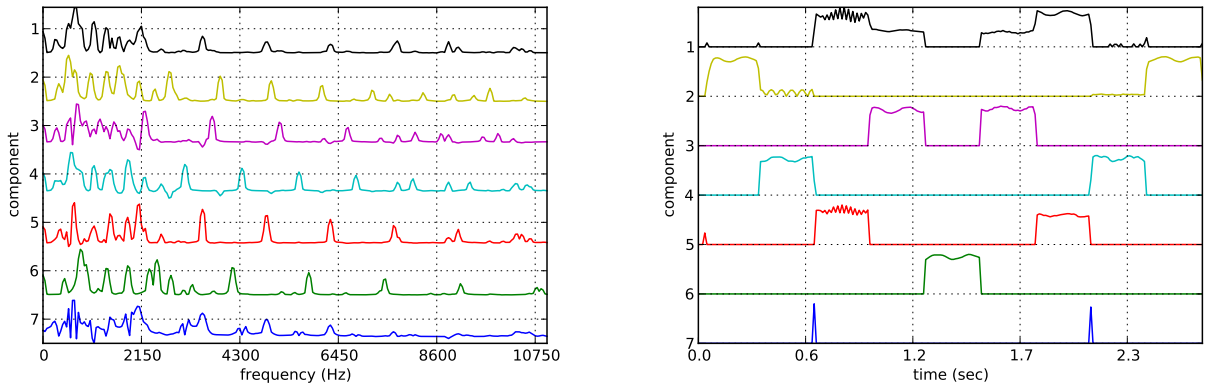
We compared with three other NMF algorithms: GaP-NMF [7] which is another nonparametric Bayesian NMF based on the gamma process, IS-NMF [5] which uses the audio-oriented Itakura-Saito divergence as loss function, and EUC-NMF [9] which minimizes the sum of the squared Euclidean distance and can be considered as a finite version of BP-NMF.

4.1 Synthetic Data

We synthesized a short clip of audio with 5 distinct piano notes and 5 distinct clarinet notes using *ChucK* [3] which is based on physical models for the instruments. At any given time, one piano note and one clarinet note are played simultaneously at different pitches.

DFTs of 512 samples (23.2 ms) were computed with 50% overlap. K quickly converged to 7 after a few iterations of variational inference. Ideally, there should be 10 components, but since some of the notes only appear with some others, they were grouped together by BP-NMF. The learned dictionary components (in log scale) and activations are shown in the Figure 2, from top to bottom in descending order of π_k. As we can see, there are clear harmonic structures in the learned dictionary and the activation does reasonably reflect the location where note combinations are played. Most importantly, the binary mask $\boldsymbol{Z}$ succeeds in explicitly controlling the appearance and disappearance of the components, which is not reflected when we test on GaP-NMF, IS-NMF, and EUC-NMF.

[3] http://chuck.stanford.edu/



(a) Dictionary components $\mathbf{D}$ in log scale.

(b) Activations $\mathbf{S} \odot \mathbf{Z}$.

Figure 2: The learned dictionary components and activations from BP-NMF on synthetic data. Both of them are listed in descending order of π_k.

4.2 Bandwidth Expansion

The basic idea of bandwidth expansion [2] is to infer the high-frequency content of a signal given only the low frequency part of the spectrum.

We use 2 pieces of music: *Pink Moon* by Nick Drake and *Funky Kingston* by Toots and the Maytals, both of which are also used in [7] for bandwidth expansion evaluation. DFTs of 512 samples are computed with no overlap. We take the middle 4000 frames of each piece and do a 5-fold cross-validation: 4/5 of the data is used to learn the dictionary. For the remaining 1/5, the top 2 octaves (192 frequency bins) are removed as a held-out set. We encode the low-frequency content with the corresponding part of the learned dictionary and predict the high-frequency content by reconstructing the full frequency band with the whole dictionary on the encoded activation.

Here we use predictive likelihood as a metric. We compare BP-NMF with GaP-NMF and EUC-NMF. The reason for not including IS-NMF is that it has been compared on the exactly same task with GaP-NMF in [7] and GaP-NMF has comparably better performance.

Unlike BP-NMF and GaP-NMF, EUC-NMF needs to specify the number of components K. Given the relationship between BP-NMF and EUC-NMF, we set K to the average number of dictionary components inferred by BP-NMF. Since both BP-NMF and EUC-NMF assume the noise is Gaussian distributed while GaP-NMF assumes the noise is exponential distributed, the predictive likelihood should be computed differently. However, this would give the exponential model an advantage, as the Gaussian distribution assigns very low probability to outcomes far from its mean, while the exponential distribution can give moderately high likelihood to values close to 0 even if they are far from the mean. To adjust for this, we evaluate the predictive likelihood under an exponential distribution for all three models. This may arguably still favor GaP-NMF as it is trained using the exponential model that it is tested on.

Geometric mean of predictive likelihood with standard error under exponential model is reported in Figure 3. Contrary to the results in [7], GaP-NMF does not dominate. This is partially due to the adjustment for Gaussian models. But the lower sampling rate of 22.05 kHz also helps the Gaussian model, since a fixed-variance Gaussian distribution has trouble modeling low-energy signals such as those that tend to appear at very high frequencies.

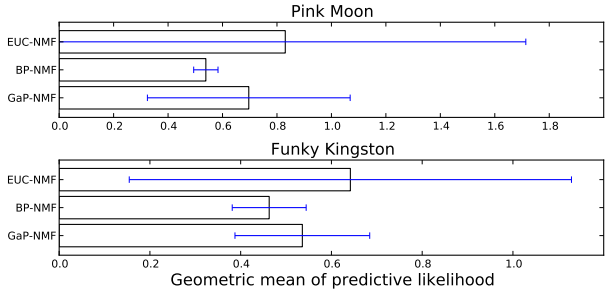


Figure 3: The geometric mean of the predictive likelihood under an exponential model for *Pink Moon* and *Funky Kingston* on a 5-fold cross-validation with standard error. In general, there is no significant difference among EUC-NMF, BP-NMF, and GaP-NMF. However, BP-NMF gives more stable results with smaller standard error.

The results in Figure 3 show that there is no significant difference among EUC-NMF, BP-NMF, and GaP-NMF. However, BP-NMF gives more stable results with smaller standard error, while its parametric counterpart EUC-NMF has much larger standard error compared with both BP-NMF and GaP-NMF.

4.3 Blind Source Separation

As with GaP-NMF, BP-NMF can also be applied to blind source separation. The model formulation of BP-NMF can be directly adopted for blind source separation, where each dictionary component can be considered as all or part of the source.

We evaluate BP-NMF for blind source separation and compare with GaP-NMF using MIREX F_0 estimation data,

Table 1: Instrument-level `bss_eval` results. The last column lists the number of components K inferred by the models.

	SDR	SIR	SAR	K
BP-NMF	0.65	7.46	4.81	46
GaP-NMF	-1.86	3.89	6.12	31

which is a woodwind quintet recording, consisting of bassoon, clarinet, flute, horn, and oboe. This piece has rich content across frequencies and various sound textures. The goal is to separate the signal on the instrument level. We compute DFTs of 1024 samples with 50% overlap.

To separate out different instruments, we need to filter out the audio signals belonging to different dictionary components. As in [5], given the complex spectrogram $\mathbf{X}^c$, to reconstruct the estimated complex spectrogram for the kth component $\hat{\mathbf{X}}^{(k)}$, we can apply Wiener filtering:

$$\hat{X}_{ft}^{(k)} = X_{ft}^c \frac{D_{fk}S_{kt}Z_{kt}}{\sum_{l=1}^K D_{fl}S_{lt}Z_{lt}} \tag{16}$$

There is no direct information to determine how the sources and instruments correspond. The heuristic in [7] is adopted: for each instrument, we pick the single component whose corresponding activation $\boldsymbol{s}_k \odot \boldsymbol{z}_k$ has the largest correlation with the power envelope of the single-track instrument signal. Note that the number of components inferred is larger than the number of instruments, thus the selected components only represent part of sources.

`Bss_eval` [14] is used to quantitatively evaluate the blind source separation. Table 1 lists the average SDR (Source to Distortion Ratio), SIR (Source to Interferences Ratio), and SAR (Sources to Artifacts Ratio) across instruments for BP-NMF and GaP-NMF (higher ratios are better). BP-NMF performs comparably well. BP-NMF decomposes the piece into 46 components, and GaP-NMF decomposes the piece into 31 components. We attribute this to the sparsity induced by BP-NMF's binary mask $\mathbf{Z}$; one needs a richer dictionary to explain the data with a sparse activation matrix.

4.4 Dictionary Quality Evaluation

The evaluation of latent component discovery and source separation is always difficult. We propose an evaluation mechanism similar to music transcription and provide statistically significant results.

To evaluate the model's ability to discover latent components from mixed signals, we can instead work on monophonic signals, which is a substantially simpler problem. We can then compare the results with those from mixed signal. If there is significant similarity, it indicates that the model can do equally well as it would have even if the problem were made artificially easier.

Since we have the single-track recordings for each instrument in the woodwind quintet recording from Section 4.3, we can apply BP-NMF to each of them separately and we will expect the learned dictionaries to be of high quality. We compute DFTs of 512 samples with no overlap. The number of learned components from each instrument is larger than the number of distinct notes V, thus only the top V components are selected according to the corresponding importance π_v. The selected components are shown in Figure 4a. The blocks are grouped according to instruments and sorted by approximated fundamental frequency. In the original piece, the bassoon is mostly playing low-pitch notes, while flute is playing high-pitch notes, both of which are reflected in the learned dictionaries.

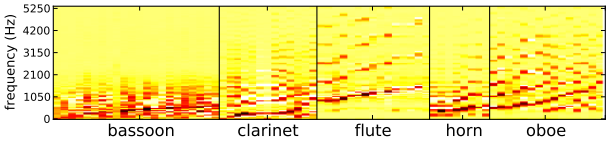
Now we would like to see if the results from BP-NMF on the mixed signal are similar to those from single-track recordings. Again there is no explicit information about the correspondence between the components learned from the mixed signal and the single-track signals. Thus, we adopt a greedy search which tries to match the dictionary components based on their correlation. BP-NMF discovers 29 components to describe the data [4], which is less than the number of distinct notes. Thus we only match the top 29 from components in Figure 4a.

After obtaining a one-to-one matching between dictionaries, we compute the correlations between the corresponding activations $\boldsymbol{s}_k \odot \boldsymbol{z}_k$. A box-and-whisker plot of correlations is shown in Figure 4b. As comparison, we also show the correlations from random matchings. Random matching has correlation close to 0, indicating there is no linear dependence. The minimum of the correlation from BP-NMF matching is close to 0 due to the fact that a few of the activations on the mixed signal are fairly sparse. But the overall quantiles do not overlap.

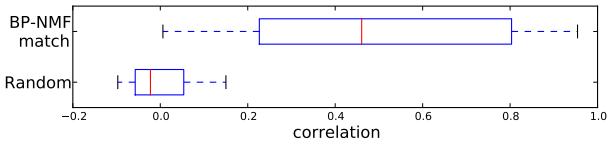
To formally test if the results from BP-NMF matching and random matching are significantly different, we apply hypothesis testing. Since we do not assume the correlations are normally distributed, a paired Wilcoxon signed-rank test [17], instead of a Student's t-test, is performed between the correlations from BP-NMF matching and random matching. The null hypothesis is that the correlations from BP-NMF matching and random matching come from the same population and we get p-value less than 0.01, which indicates that their difference is statistically significant. This gives a solid evidence that BP-NMF is able to learn dictionary components equally well in mixed signal, when compared with dictionaries learned from single-track instrument recordings.

We also apply the same procedures to IS-NMF and GaP-NMF. For IS-NMF, as we need to specify the number of components K, each single-track recording is decomposed with K equals the number of distinct notes. For the mixed signal, we set $K \in \{5, 10, 20, \cdots, 70\}$. When K is between 10 and 30, the Wilcoxon signed-rank test shows that the difference is significant at $p = 0.05$ level. For the rest of the K's, we get larger p-values and cannot reject the null hypothesis. GaP-NMF decomposes the data into 20 components and the test results show significant difference between GaP-NMF matching and random matching.

[4] This number is smaller than that in Section 4.3 because a smaller DFT size with no overlap is used, which leads to less data.



(a) The selected components learned from single-track instrument. For each instrument, the components are sorted by approximated fundamental frequency. The dictionary is cut off above 5512.5 Hz for visualization purposes.



(b) The box-and-whisker plot for the correlations from both BP-NMF matching and random matching. A paired Wilcoxon signed-rank test shows that they are significantly different.

Figure 4: The results from the proposed evaluation.

Therefore, this evaluation mechanism can also be applied to determine a range for the "proper" number of components to describe the data.

5. CONCLUSION

In this paper, we propose BP-NMF, a Bayesian nonparametric extension of nonnegative matrix factorization, which can automatically infer the number of latent components. BP-NMF explicitly assumes that some of the components are often completely silent. BP-NMF performs well under existing metrics and under a novel evaluation mechanism.

6. ACKNOWLEDGMENTS

The authors thank the reviewers for comments and the helpful discussion with Brian McFee and Colin Raffel. This work was supported by the NSF under grant IIS-1117015.

7. REFERENCES

[1] Samer A. Abdallah and Mark D. Plumbley. Polyphonic music transcription by non-negative sparse coding of power spectra. In *Proceedings of the 5th International Society for Music Information Retrieval Conference*, pages 10–14, 2004.

[2] Dhananjay Bansal, Bhiksha Raj, and Paris Smaragdis. Bandwidth expansion of narrowband speech using non negative matrix factorization. In *9th European Conference on Speech Communication (Eurospeech)*, 2005.

[3] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., 2006.

[4] Finale Doshi-Velez, Kurt T. Miller, Jurgen Van Gael, and Yee Whye Teh. Variational inference for the Indian buffet process. In *International Conference on Artificial Intelligence and Statistics*, 2009.

[5] Cédric Févotte, Nancy Bertin, and Jean-Louis Durrieu. Nonnegative matrix factorization with the Itakura-Saito divergence: with application to music analysis. *Neural Computation*, 21(3):793–830, 2009.

[6] Nils Lid Hjort. Nonparametric bayes estimators based on beta processes in models for life history data. *The Annals of Statistics*, pages 1259–1294, 1990.

[7] Matthew D. Hoffman, David M. Blei, and Perry R. Cook. Bayesian nonparametric matrix factorization for recorded music. In *Proceedings of the 27th Annual International Conference on Machine Learning*, pages 439–446, 2010.

[8] Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

[9] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems*, 13:556–562, 2001.

[10] Alexey Ozerov and Cédric Févotte. Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(3):550–563, 2010.

[11] John Paisley and Lawrence Carin. Nonparametric factor analysis with beta process priors. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 777–784, 2009.

[12] Paris Smaragdis and Judith C. Brown. Non-negative matrix factorization for polyphonic music transcription. In *Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on.*, pages 177–180. IEEE, 2003.

[13] Romain Thibaux and Michael I. Jordan. Hierarchical beta processes and the Indian buffet process. In *International Conference on Artificial Intelligence and Statistics*, 2007.

[14] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte. Performance measurement in blind audio source separation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(4):1462–1469, 2006.

[15] Tuomas Virtanen. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(3):1066–1074, 2007.

[16] Chong Wang and David M. Blei. Variational inference in nonconjugate models. *Journal of Machine Learning Research*, 14:899–925, 2013.

[17] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics bulletin*, 1(6):80–83, 1945.