POISSON-UNIFORM NONNEGATIVE MATRIX FACTORIZATION

Matthew D. Hoffman

Columbia University
Dept. of Statistics
1255 Amsterdam Avenue
New York, NY 10027

ABSTRACT

Probabilistic models of audio spectrograms used in audio source separation often rely on Poisson or multinomial noise models corresponding to the generalized Kullback-Leibler (GKL) divergence popular in methods using Nonnegative Matrix Factorization (NMF). This noise model works well in practice, but it is difficult to justify since these distributions are technically only applicable to discrete counts data. This issue is particularly problematic in hierarchical and nonparametric Bayesian models where estimates of uncertainty depend strongly on the likelihood model. In this paper, we present a hierarchical Bayesian model that retains the flavor of the Poisson likelihood model but yields a coherent generative process for continuous spectrogram data. This model allows for more principled, accurate, and effective Bayesian inference in probabilistic NMF models based on GKL.

Index Terms— NMF, audio, Bayesian models, variational inference, blind source separation.

1. INTRODUCTION

Nonnegative matrix factorization (NMF) [1] is a popular method that approximately decomposes an $M \times N$ matrix X of nonnegative data into an $M \times K$ matrix W and a $K \times N$ matrix H whose entries are also nonnegative. NMF and its variants are particularly widely used in the audio source separation community, which has found that applying such decompositions to magnitude spectrograms results in W matrices whose columns tend to correspond to the spectra of audio sources present in a mixed recording.

Typically for $K < \min\{M, N\}$ no setting of W and H can satisfy $X = WH$ exactly; the NMF problem is therefore framed in terms of minimizing some cost $D(X, WH)$. [1] presented efficient and simple multiplicative update algorithms for two such costs: Euclidean distance and generalized Kullback-Leibler divergence, the latter being defined as

$$D_{\mathrm{GKL}}(X, Y) \equiv \sum_{m,n} X_{m,n} \log \frac{X_{m,n}}{Y_{m,n}} - X_{m,n} + Y_{m,n}. \quad (1)$$

As shorthand, we will refer to NMF optimizing the cost $D_{\mathrm{GKL}}(X, WH)$ as "KL-NMF." KL-NMF has proven to give better results for audio spectrograms than NMF optimizing Euclidean distance [2].

In the last five years, there has been increasing interest in probabilistic interpretations of and extensions to KL-NMF. In particular, it has been observed that minimizing D_{GKL} is equivalent to finding the maximum-likelihood estimate (MLE) of W and H under the model $X_{m,n} \sim \mathrm{Poisson}([WH]_{m,n})$, where $[WH]_{m,n}$ denotes element m, n of the product of W and H. This interpretation has led to many probabilistic extensions to NMF that use the formalism of hierarchical Bayesian modeling to build additional assumptions and prior knowledge into this simple Poisson likelihood model (e.g. [3, 4, 5, 6, 7]) [1].

There is an obvious issue with the Poisson interpretation of KL-NMF, at least as applied to audio spectrograms: spectrogram data is inherently continuous, and the Poisson distribution is a distribution over discrete counts. Naively treating the Poisson probability mass function as a probability density function results in a density that does not integrate to 1. One can address this issue by quantizing the audio spectrogram data, for example by assuming $\lfloor \nu X_{m,n} \rfloor \sim \mathrm{Poisson}([WH]_{m,n})$, where the scaling factor ν controls the fineness of this quantization. When computing point estimates of W and H one can avoid any loss of resolution by making ν arbitrarily large; in fact, since $D_{\mathrm{GKL}}(X, Y) = D_{\mathrm{GKL}}(\nu X, \nu Y)/\nu$ for any ν and rescaling X therefore only affects the MLE of W and H by a multiplicative constant, one can argue that KL-NMF maximizes $\prod_{m,n} \mathrm{Poisson}(\lfloor \nu X_{m,n} \rfloor; \nu[WH]_{m,n})$ for some very large value of ν. Though technically correct, there is a statistical problem with this interpretation. The ratio of the standard deviation of the Poisson distribution to its mean λ decreases as $1/\sqrt{\lambda}$, so as $\nu \to \infty$, the noise level assumed by the model goes to $-\infty$ dB, meaning that the model is inconsistent with *any* X of rank greater than K. In practice, though, when fitting point estimates of W and H (possibly with regularization as in maximum a posteriori (MAP) estimation) this

[1] Some of this work uses multinomial instead of Poisson likelihoods—most of the derivations and observations in this paper can be adapted to such models with only minor changes.

technical issue does not seem to cause much trouble.

The issue of scaling is more serious in the fully Bayesian setting, where we are interested in making inferences about the posterior $p(W, H|X, \nu)$. Bayesian inference methods such as variational Bayes (VB) and Markov chain Monte Carlo (MCMC) have several advantages over point estimation methods such as MLE and MAP: they offer an estimate of the uncertainty of the parameter estimates, they permit automatic selection of the model order K via Bayesian model selection [5] or Bayesian nonparametric modeling [6, 7], and they allow one to automatically fit the hyperparameters that control the prior distributions on W and H rather than (mis)specifying them by hand as MAP methods require (trying to fit hyperparameters in MAP inference often leads to degeneracies) [8]. The scaling parameter ν is critical to realizing these advantages, however. Since ν effectively controls the noise level assumed by the model, a too-small value of ν will lead to a diffuse posterior that is dominated by the prior terms, underfits, and prefers to use too few latent components. Conversely, a too-large value of ν will lead to a posterior that is centered too tightly around the MLE, ignores the prior terms, and prefers to use too many latent components.

These issues motivate the likelihood model presented in this paper, which we call Poisson-uniform NMF (PUNMF). PUNMF is a generative model of continuous nonnegative matrices that retains a Poisson-like likelihood model but fits the scaling parameter ν automatically, permitting the principled use of Bayesian inference methods such as variational Bayes (VB) or Markov chain Monte Carlo (only the former is treated in this paper due to space limitations). In the following sections, we will derive the PUNMF model and a corresponding VB inference algorithm, and evaluate this VB algorithm's ability to automatically tune its prior distributions to give good performance on a blind source separation task. These experiments will demonstrate the importance of the scaling parameter ν to this task.

2. POISSON-UNIFORM NMF

Poisson-uniform NMF (PUNMF) assumes that the following two-step stochastic process generated the spectrogram X given the component matrix W and the activation matrix H:

$$\begin{aligned} \tilde{X}_{m,n} &\sim \text{Poisson}(\nu[WH]_{m,n}); \\ \nu X_{m,n} &\sim \begin{cases} \text{Uniform}([\tilde{X}_{m,n}, \tilde{X}_{m,n}+1)) & \text{if} \tilde{X}_{m,n} > 0; \\ \text{Beta}(\alpha, \beta)) & \text{if} \tilde{X}_{m,n} = 0. \end{cases} \end{aligned} \tag{2}$$

We first draw a discrete variable $\tilde{X}_{m,n}$ from a Poisson distribution with mean $[WH]_{m,n}$. Then, if $\tilde{X}_{m,n}$ is greater than 0, we sample the observed value $X_{m,n}$ from a uniform distribution varying between $\frac{\tilde{X}_{m,n}}{\nu}$ and $\frac{\tilde{X}_{m,n}+1}{\nu}$. If $X_{m,n} = 0$, we sample $X_{m,n}$ from a rescaled beta distribution with parameters α and β. Note that given $X_{m,n}$ we can infer that $\tilde{X}_{m,n} = \lfloor \nu X_{m,n} \rfloor$; this property dramatically simplifies inference for W and H. The additional complexity of modeling some elements of X as coming from a beta distribution is necessary to capture the nonuniform shape of the noise floor of audio magnitude spectra, which is illustrated in figure 1.

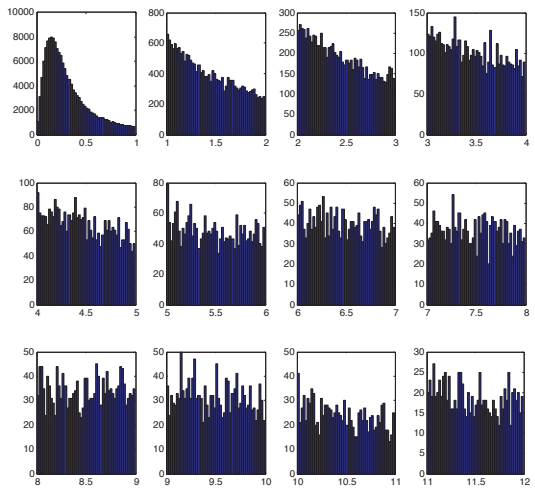

Fig. 1. Histograms of spectrogram amplitudes within various ranges. Each histogram shows the relative frequencies of amplitudes of a scaled spectrogram νX that fall within the ranges $[0, 1)$, $[1, 2)$, etc. The range $[0, 1)$ is much less uniformly distributed than the higher ranges.

The PUNMF likelihood model is compatible with any prior specification for W and H. In this paper, we use a simple independent gamma prior for W and the gamma chain prior of [4] to enforce smoothness on H:

$$\begin{aligned} W_{m,k} &\sim \text{Gamma}(a, a); \quad H_{k,n} \sim \text{Gamma}(b, Z_{k,n}); \\ Z_{k,1} &\sim \text{Gamma}(b, bc); \quad Z_{k,n>1} \sim \text{Gamma}(b, H_{k,n-1}). \end{aligned}$$

Given W and H and letting $p^0_{m,n} = \text{Poisson}(0; \nu[WH]_{m,n})$, the expected value of the observed spectrogram X is

$$\mathbb{E}_p[X_{m,n}|W, H] = p^0_{m,n} \tfrac{\alpha}{\nu\beta} + (1 - p^0_{m,n})([WH]_{m,n} + \tfrac{1}{2\nu}). \tag{3}$$

Thus, PUNMF defines a coherent generative process that assumes that the elements of X are produced by corrupting WH with both Poisson and uniform or beta noise.

3. VARIATIONAL INFERENCE FOR PUNMF

In this section we derive a variational Bayesian inference algorithm for the PUNMF model. We will fit a variational distribution $q(W, H)$ of the form

$$\begin{aligned} q(W, H) = &(\textstyle\prod_{m,k} \text{Gamma}(W_{m,k}; \gamma^W_{m,k}, \rho^W_{m,k})) \\ &\times (\textstyle\prod_{k,n} \text{Gamma}(H_{k,n}; \gamma^H_{k,n}, \rho^H_{k,n})) \end{aligned} \tag{4}$$

to minimize the Kullback-Leibler divergence (KLD) $D_{\mathrm{KL}}(q(W,H)||p(W,H|X,a,b,c,\nu,\alpha,\beta))$ between q and the posterior over W and H given the data X and the hyperparameters a, b, c, ν, α and β. We fit point estimates of these hyperparameters by maximum marginal likelihood.

We approximately minimize the KLD between q and the target posterior by maximizing the following Evidence Lower BOund (ELBO) on the marginal probability of the data [9]:

$$\begin{aligned}
&\log p(X) \geq \mathbb{E}_q[\log p(X,W,H)] - \mathbb{E}_q[\log q(W,H)] \\
&\geq \textstyle\sum_{m,n} \mathbb{I}[\lfloor \nu X_{m,n} \rfloor = 0] \log \mathrm{Beta}(\nu X_{m,n}; \alpha, \beta) \\
&\qquad + \lfloor \nu X_{m,n} \rfloor \log(\nu \textstyle\sum_k e^{\mathbb{E}_q[\log W_{m,k}] + \mathbb{E}_q[\log H_{k,n}]}) \\
&\qquad - \nu \textstyle\sum_k \mathbb{E}_q[W_{m,k} H_{k,n}] - \log \Gamma(\lfloor \nu X_{m,n} \rfloor + 1)) \\
&\quad + \textstyle\sum_{m,k} (a - \gamma^W_{m,k}) \mathbb{E}_q[\log W_{m,k}] - (a - \rho^W_{m,k}) \mathbb{E}_q[W_{m,k}] \\
&\qquad - \gamma^W_{m,k} \log \rho^W_{m,k} + \log \Gamma(\gamma^W_{m,k}) \\
&\quad + \textstyle\sum_{k,n} (b - \gamma^Z_{k,n}) \mathbb{E}_q[\log Z_{k,n}] - \sum_k (bc - \rho^Z_{k,1}) \mathbb{E}_q[Z_{k,1}] \\
&\quad - \textstyle\sum_k \sum_{n=2}^N (b \mathbb{E}_q[H_{k,n-1}] - \rho^Z_{k,n}) \mathbb{E}_q[Z_{k,n}] \\
&\quad + \textstyle\sum_{k,n} (b - \gamma^H_{k,n}) \mathbb{E}_q[\log H_{k,n}] \\
&\quad - \textstyle\sum_{k,n} (b \mathbb{E}_q[Z_{k,n}] - \rho^H_{k,n}) \mathbb{E}_q[H_{k,n}] \\
&\quad + Kb \log c + b \textstyle\sum_k \sum_{n=1}^{N-1} \mathbb{E}_q[\log H_{k,n}] \\
&\quad + b(\textstyle\sum_{k,n} \mathbb{E}_q[\log Z_{k,n}]) + 2KN(b \log b - \log \Gamma(b)) \\
&\quad + MK(a \log a - \log \Gamma(a)) + MN \log \nu,
\end{aligned} \tag{5}$$

where $\mathbb{I}[\cdot]$ is 1 if its argument is true and 0 otherwise and $\Gamma(\cdot)$ denotes the gamma function. As in [5], we lower bound the intractable expectation $\mathbb{E}_q[\log \sum_k W_{m,k} H_{k,n}]$ with $\log \sum_k e^{\mathbb{E}_q[\log W_{m,k}] + \mathbb{E}_q[\log H_{k,n}]}$ using the convexity of the log-sum-exp function. The necessary expectations are

$$\begin{aligned}
\mathbb{E}_q[\log W_{m,k}] &= \psi(\gamma^W_{m,k}) - \log \rho^W_{m,k};\ \mathbb{E}_q[W_{m,k}] = \tfrac{\gamma^W_{m,k}}{\rho^W_{m,k}}; \\
\mathbb{E}_q[\log H_{k,n}] &= \psi(\gamma^H_{k,n}) - \log \rho^H_{k,n};\ \mathbb{E}_q[H_{k,n}] = \tfrac{\gamma^H_{k,n}}{\rho^H_{k,n}}
\end{aligned} \tag{6}$$

We maximize the ELBO using coordinate ascent. Using Jensen's inequality as in [1] or an equivalent argument based on latent variables as in [5], one can show that the updates

$$\begin{aligned}
\gamma^W_{m,k} &= a + e^{\mathbb{E}_q[\log W_{m,k}]} \textstyle\sum_n \lfloor \nu X_{m,n} \rfloor e^{\mathbb{E}_q[\log H_{k,n}]}; \\
\rho^W_{m,k} &= a + \nu \textstyle\sum_n \mathbb{E}_q[H_{k,n}],
\end{aligned} \tag{7}$$

increase the ELBO (unless it is already at a maximum). The updates

$$\begin{aligned}
\gamma^H_{k,n<N} &= 2b + e^{\mathbb{E}_q[\log H_{k,n}]} \textstyle\sum_m \lfloor \nu X_{m,n} \rfloor e^{\mathbb{E}_q[\log W_{m,k}]}; \\
\gamma^H_{k,N} &= b + e^{\mathbb{E}_q[\log H_{k,N}]} \textstyle\sum_m \lfloor \nu X_{m,n} \rfloor e^{\mathbb{E}_q[\log W_{m,k}]}; \\
\rho^H_{k,n<N} &= b(\mathbb{E}_q[Z_{k,n}] + \mathbb{E}_q[Z_{k,n+1}]) + \nu \textstyle\sum_m \mathbb{E}_q[W_{m,k}]; \\
\rho^H_{k,N} &= b \mathbb{E}_q[Z_{k,N}] + \nu \textstyle\sum_m \mathbb{E}_q[W_{m,k}]
\end{aligned} \tag{8}$$

and

$$\begin{aligned}
\gamma^Z_{k,n} &= 2b; \quad \rho^Z_{k,1} = b(c + \mathbb{E}_q[H_{k,1}]); \\
\rho^Z_{k,n>1} &= b(\mathbb{E}_q[H_{k,n-1}] + \mathbb{E}_q[Z_{k,n}])
\end{aligned} \tag{9}$$

likewise increase the ELBO with respect to γ^H and ρ^H and γ^Z and ρ^Z, respectively.

Closed-form updates for the hyperparameters a, b, α, β, and ν are not available. We optimize the hyperparameters a, b and c, and α and β via Newton's method. The optimal values of these hyperparameters only depend on summary statistics of the data and the variational parameters γ and ρ, so they can be fit efficiently. The ELBO is discontinuous with respect to ν, but the local optima are quite shallow. We can therefore find a local optimum at or near the global optimum using a bisection search—in practice, this approach seems to produce solutions near the global optimum. The terms in the ELBO that depend on ν cannot be efficiently summarized, so this step is more expensive; we amortize this expense by only updating ν once for every ten times we update the other parameters.

4. EXPERIMENTAL EVALUATION

The primary difference between PUNMF and previous hierarchical Bayesian NMF models with Poisson emission models is that in previous work the scaling parameter ν is fixed at some arbitrary value (or is implicitly set at $\nu = 1$), while in PUNMF ν is fit based on the data. In this section we show the practical importance of selecting an appropriate value for ν.

The task is to decompose a single-channel mixed audio signal into several tracks, each of which contains only the audio generated by a single instrument at a single pitch. In our experiment, we used a synthesized recording consisting of a mixture of 10 seconds of randomly generated clarinet and organ music. The samping rate is 44.1KHz. At all times there are two clarinet tones and two organ tones playing, the tones being randomly selected from a single octave of a C pentatonic scale starting an octave below middle C. There are thus a total of 10 unique tones, and the task is to isolate the signals associated with each of them. We do so by fitting an NMF model with $K = 10$ latent sources to the audio magnitude spectrogram (generated with no overlap and a Hamming window of 1024 samples), estimating the contribution of latent source k at time n and frequency bin m as $X_{m,n} W_{m,k} H_{k,n} / [WH_{m,n}]$, and using Wiener filtering to isolate the energy associated with each latent source k. We then used the `BSS_EVAL` toolbox [10] to evaluate the signal-artifact-ratios (SAR), signal-distortion-ratios (SDR), and signal-interference-ratios (SIR) for each signal, which each measure a different dimension of the quality of the separation of each signal. The numbers we report are the SAR, SDR, and SIR averaged across the 10 separated signals. The separated signals are matched to the source

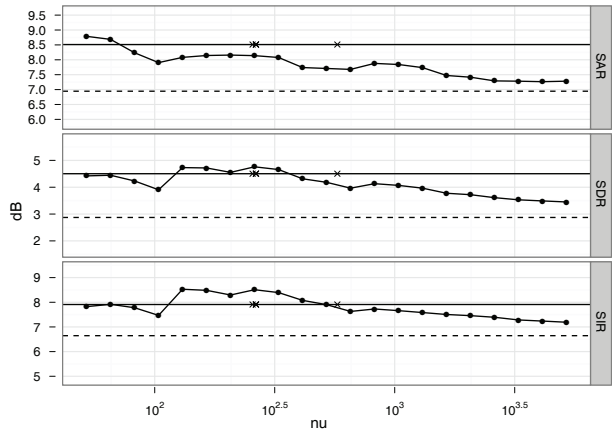

Fig. 2. Summary of blind source separation experiments. The dashed horizontal lines show the performance of classic KL-NMF, the solid horizontal lines show the performance of PUNMF with the scaling parameter ν fit automatically, and the solid curves show the performance of PUNMF for various fixed values of ν. Each "x" denotes a learned value for ν from a different run.

signals by the toolbox. The evaluation code is available at `http://www.cs.princeton.edu/~mdhoffma`.

We compared three approaches: PUNMF with gamma chain priors on H, fitting all hyperparameters; PUNMF with gamma chain priors on H, fitting all hyperparameters except ν and holding ν fixed at a variety of settings; and maximum-likelihood estimation (MLE) of W and H, implemented via the original multiplicative algorithm of [1]. For each run, we applied each approach with five different random initializations and used the decomposition that gave the best ELBO (for PUNMF) or D_{GKL} score (for MLE). We repeated this process five times, averaging the mean SAR, SDR, and SIR across the five super-runs. Figure 2 plots the average SAR, SDR, and SIR as a function of the scaling parameter ν for the PUNMF models with fixed ν. The solid horizontal line shows the average performance of PUNMF with optimized ν, and the dashed horizontal line shows the performance of the classic MLE method. The five "x" marks on the solid horizontal line show the values of ν that were chosen by PUNMF.

PUNMF with the gamma chain prior on H generally outperforms the simple Poisson MLE on this task, but its performance is dependent on the scaling parameter ν, which affects the analysis mostly by determining the relative influence of the prior and the likelihood. Settings between about $\nu = 125$ and $\nu = 325$ give the best performance; as ν increases beyond this region overfitting becomes a problem (since the data are given too much weight), while as ν decreases underfitting and quantization error become more of a concern. In four out of five cases, PUNMF finds a value for ν that gives near-optimal performance, indicating that fitting this hyperparameter automatically is viable.

5. DISCUSSION

We have presented PUNMF, a likelihood model for hierarchical Bayesian NMF models to audio spectra that addresses the modeling issues associated with trying to fit continuous audio spectrograms with a fundamentally discrete noise model. PUNMF's likelihood model can be swapped in for the basic Poisson (or, with minor adjustments, multinomial) likelihood models used by the various probabilistic extensions to KL-NMF, allowing for more principled and effective Bayesian inference in these models.

6. REFERENCES

[1] D.D. Lee and H.S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems 13 (NIPS)*, 2001, pp. 556–562.

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

[3] P. Smaragdis, B. Raj, and M. Shashanka, "Sparse and shift-invariant feature extraction from non-negative data," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE Int'l Conf. on*, 2008, pp. 2069–2072.

[4] T. Virtanen, A.T. Cemgil, and S. Godsill, "Bayesian extensions to non-negative matrix factorisation for audio signal modelling," in *Proc. IEEE Conf. on Acoustics, Speech and Signal Processing, 2008*, 2008, pp. 1825–1828.

[5] A. Cemgil, "Bayesian inference for nonnegative matrix factorisation models," *Computational Intelligence and Neuroscience*, vol. 2009, pp. 4:1–4:17, January 2009.

[6] M.D. Hoffman, D.M. Blei, and P.R. Cook, "Finding latent sources in recorded music with a shift-invariant HDP," in *Proc. Digital Audio Effects (DAFx-09)*, 2009.

[7] K. Yoshii and M. Goto, "A nonparametric Bayesian multipitch analyzer based on infinite latent harmonic allocation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. PP, no. 99, pp. 1, 2011.

[8] A. Gelman, J. Carlin, H. Stern, and D. Rubin, *Bayesian Data Analysis*, CRC press, 2004.

[9] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul, "Introduction to variational methods for graphical models," *Machine Learning*, vol. 37, pp. 183–233, 1999.

[10] C. Févotte, R. Gribonval, and E. Vincent, "Bss eval toolbox user guide," *IRISA, Rennes, France, Tech. Rep*, vol. 1706, 2005.