Gamma Hidden Markov Model as a Probabilistic Nonnegative Matrix Factorization

Nasser Mohammadiha, W. Bastiaan Kleijn, Arne Leijon

Department of Electrical Engineering, KTH Royal Institute of Technology, Stockholm, Sweden

September 10, 2013
EUSIPCO, Marrakech
Outline

1 Background
   - Introduction to NMF
   - Dynamic NMF

2 Proposed Method: Gamma NHMM
   - Gamma HMM
   - NMF Representation

3 Conclusions
Nonnegative Matrix Factorization

\[ x \approx vw \]

For speech signals:

Discrete Fourier Transform

Nasser Mohammadiha
Gamma HMM as a Probabilistic NMF
Nonnegative Matrix Factorization

\[ x \approx vw \]

- For speech signals:

  - Speech Signal in time domain
  - Discrete Fourier Transform
  - Speech Spectrogram
Deterministic Decomposition

$$(v, w) = \arg\min_{v, w} D(x \parallel vw),$$
subject to $v \geq 0, w \geq 0$

- $D(x \parallel \hat{x})$ is a cost function: e.g., Euclidean distance
- Non-convex problem
- **Solution**: iterative gradient decent algorithms
What is the motivation?

- Providing a probabilistic generative model of the data
- The signals are stochastic, like speech signals
- Applying prior information is more intuitive
- Deriving optimal estimators is possible

Examples:

- EUC-NMF corresponds to \( x_{kt} \sim \mathcal{N}(\hat{x}_{kt}, \sigma^2) \)
- KL-NMF corresponds to \( x_{kt} \sim \mathcal{P}(\hat{x}_{kt}) \)

where \( x = \{x_{kt}\} \), \( \hat{x} = vw \), and \( \mathcal{N} \) and \( \mathcal{P} \) denote the normal and Poisson distributions, respectively.
Using Temporal Dependencies in NMF

- Regularizing the cost function, e.g., [Virtanen 2007]

\[
(v, w) = \arg \min_{v,w} D(x \| vw) + \mu h(w),
\]

subject to \( v \geq 0, w \geq 0 \)
Using Temporal Dependencies in NMF

- Regularizing the cost function, e.g., [Virtanen 2007]

\[
(v, w) = \arg\min_{v, w} D(x \| vw) + \mu h(w),
\]

subject to \( v \geq 0, w \geq 0 \)

- Bayesian NMF: recursively update the prior distributions over time [Mohammadiha et al. 2013a]
Using Temporal Dependencies in NMF

- Regularizing the cost function, e.g., [Virtanen 2007]
  \[(v, w) = \arg \min_{v, w} D(x\|vw) + \mu h(w),\]
  subject to \(v \geq 0, w \geq 0\)

- Bayesian NMF: recursively update the prior distributions over time [Mohammadiha et al. 2013a]

- Continuous state-space formulations [Mohammadiha et al. 2013b, Févotte et al. 2013]

- Discrete state-space formulations or nonnegative hidden Markov models (NHMM)
  - Factorial scaled hidden Markov model [Ozerov et al. 2009]
  - Multinomial NHMM [Mysore et al. 2010]
  - Proposed method: Gamma NHMM
Nonnegative Hidden Markov Models (NHMM)

- Factorial scaled hidden Markov model [Ozerov et al. 2009]
  - Generalizes the Itakura-Saito NMF and Gaussian Scaled Mixture Model
  - Models each sample of the complex data as a sum of independent complex-valued components
  - Assumes an HMM for each component with Gaussian output density functions and structured diagonal covariance matrices
Nonnegative Hidden Markov Models (NHMM)

- Factorial scaled hidden Markov model [Ozerov et al. 2009]
  - Generalizes the Itakura-Saito NMF and Gaussian Scaled Mixture Model
  - Models each sample of the complex data as a sum of independent complex-valued components
  - Assumes an HMM for each component with Gaussian output density functions and structured diagonal covariance matrices

- Multinomial NHMM [Mysore et al. 2010]
  - Models the observed nonnegative data with an HMM
  - Each state-dependent distribution is given by a mixture of multinomial density functions. That is, given a state, the data is assumed to be generated by a linear combination of the nonnegative basis vectors corresponding to that state.
Gamma HMM With Gain Modeling

- Model each observation vector $x$ using an $l$-state HMM:

$$f(x_{kt} \mid S_t = i, G_t = g_t) = \text{Gamma}(a_{ki}, g_t b_{ki})$$

where

- $k$: dimension index
- $t$: time index
- $S$: State variable, $G$: Gain variable
- $\text{Gamma}(a, b)$ is a gamma distribution with $a$ and $b$ as the shape and scale parameters.
**Gamma HMM With Gain Modeling**

- Model each observation vector $x$ using an $I$-state HMM:

  $f(x_{kt} \mid S_t = i, G_t = g_t) = \text{Gamma}(a_{ki}, g_t b_{ki})$

  where
  - $k$: dimension index
  - $t$: time index
  - $S$: State variable, $G$: Gain variable
  - $\text{Gamma}(a, b)$ is a gamma distribution with $a$ and $b$ as the shape and scale parameters.

- $G_t$ is a stochastic gain parameter to model long-term level changes:

  $f(g_t \mid S_t = i) = \text{Gamma}(\phi, \theta_t)$
Model each observation vector $x$ using an $I$-state HMM:

$$f(x_{kt} \mid S_t = i, G_t = g_t) = \text{Gamma}(a_{ki}, g_t b_{ki})$$

where

- $k$: dimension index
- $t$: time index
- $S$: State variable, $G$: Gain variable
- $\text{Gamma}(a, b)$ is a gamma distribution with $a$ and $b$ as the shape and scale parameters.

$G_t$ is a stochastic gain parameter to model long-term level changes:

$$f(g_t \mid S_t = i) = \text{Gamma}(\phi, \theta_t)$$

For speech data, the gamma assumption is in line with super-Gaussianity of the speech DFT coefficients.
- $q$: Transition probability matrix with $q_{ij} = f(S_t = j \mid S_{t-1} = i)$
- Only $\theta_t$ is time-varying and has to be updated over time.
Approximate a nonnegative vector by its expected value under the model assumptions, $x_t \approx \hat{x}_t$:

$$\hat{x}_t = \sum_{i=1}^{l} \int E(X_t \mid S_t = i, g_t) f(S_t = i, g_t \mid x) dg_t,$$

where the posterior distributions of the state and gain variables conditioned on the entire sequence of the observed signal are used.
Big Picture

- Approximate a nonnegative vector by its expected value under the model assumptions, $x_t \approx \hat{x}_t$:

$$\hat{x}_t = \sum_{i=1}^{I} \int E(X_t \mid S_t = i, g_t) f(S_t = i, g_t \mid x) \, dg_t,$$

where the posterior distributions of the state and gain variables conditioned on the entire sequence of the observed signal are used.

- Next, we show that $\hat{x}_t$ is factorized into a product of two nonnegative factors: $\hat{x}_t = vw_t$. 
NMF Derivation

- Define a basis matrix \( \mathbf{v} = \{v_{ki}\} \) with \( v_{ki} = a_{ki}b_{ki} \)
- Number of bases = Number of states = \( I \)
NMF Derivation

- Define a basis matrix $\mathbf{v} = \{v_{ki}\}$ with $v_{ki} = a_{ki}b_{ki}$
  - Number of bases = Number of states = $I$
- Define an indicator column vector $\mathbf{s}_t = 1; \text{ with } s_{it} = 1$

We can show that:

$$x_t \approx ^{x}_t = \mathbf{w}_t \mathbf{v}_t$$

where

$$\mathbf{w}_t = \sum \mathbf{s}_t f(s_t | x_t)$$

$E(g_t | s_t, x_t)$ is a generalized inverse Gaussian (GIG) distribution and $E(g_t | s_t, x_t)$ is available in a closed form in terms of the modified Bessel function of the second kind.

We use EM algorithm to estimate the parameters $\{p, q, a, b, \phi, \theta\}$.
Define a basis matrix \( \mathbf{v} = \{v_{ki}\} \) with \( v_{ki} = a_{ki} b_{ki} \)

Number of bases = Number of states = 1

Define an indicator column vector \( \mathbf{s}_t = 1 \) with \( s_{it} = 1 \)

We can show that:

\[
\mathbf{x}_t \approx \hat{\mathbf{x}}_t = \mathbf{vw}_t \text{ where } \\
\mathbf{w}_t = \sum_{\mathbf{s}_t} s_{t} f(s_{t} | \mathbf{x}) E(g_{t} | s_{t}, \mathbf{x})
\]
NMF Derivation

- Define a basis matrix \( v = \{v_{ki}\} \) with \( v_{ki} = a_{ki} b_{ki} \)
  - Number of bases = Number of states = \( I \)
- Define an indicator column vector \( s_t = 1 \) with \( s_{it} = 1 \)
- We can show that:

  \[ x_t \approx \hat{x}_t = vv_t \quad \text{where} \]
  \[ w_t = \sum_{s_t} s_t f(s_t \mid x) E(g_t \mid s_t, x) \]

- \( f(g_t \mid s_t, x) \) is a generalized inverse Gaussian (GIG) distribution
  and \( E(g_t \mid s_t, x) \) is available in a closed form in terms of the modified Bessel function of the second kind
NMF Derivation

- Define a basis matrix \( \mathbf{v} = \{v_{ki}\} \) with \( v_{ki} = a_{ki} b_{ki} \)
  - Number of bases = Number of states = \( I \)
- Define an indicator column vector \( \mathbf{s}_t = 1; \) with \( s_{it} = 1 \)
- We can show that:

\[
\mathbf{x}_t \approx \hat{\mathbf{x}}_t = \mathbf{v} \mathbf{w}_t \quad \text{where}
\]

\[
\mathbf{w}_t = \sum_{\mathbf{s}_t} s_{t} f \left( s_{t} \mid \mathbf{x} \right) E \left( g_{t} \mid s_{t}, \mathbf{x} \right)
\]

- \( f \left( g_{t} \mid s_{t}, \mathbf{x} \right) \) is a generalized inverse Gaussian (GIG) distribution and \( E \left( g_{t} \mid s_{t}, \mathbf{x} \right) \) is available in a closed form in terms of the modified Bessel function of the second kind
- We use EM algorithm to estimate the parameters \( \{p, q, a, b, \phi, \theta\} \)
Demonstration

Proposed Method: Gamma NHMM

Nasser Mohammadiha  
Gamma HMM as a Probabilistic NMF
Comparison With State of the Art

- **Algorithms:**
  - Sparse and non-sparse multinomial NHMM
  - Sparse and non-sparse KL-NMF

- **Objective Measures**
  - Compare sparsity levels using $l_0$ norm of the activation vectors $w_t$
  - Compare accuracy of fit using log-spectral distortion (log SD)
  - Compare modeling of temporal dynamics by measuring correlations between columns of $w_t (P_{rep})$
## Results

<table>
<thead>
<tr>
<th>Method</th>
<th>$l_0$ norm</th>
<th>log SD (dB)</th>
<th>$P_{rep}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed NMF (55 states)</td>
<td>1</td>
<td>9.4 ± 0.7</td>
<td>0.59 ± 0.06</td>
</tr>
<tr>
<td>Sparse NHMM (60 states, 1 comp. each)</td>
<td>1</td>
<td>16.4 ± 1.8</td>
<td>0.6 ± 0.05</td>
</tr>
<tr>
<td>Non-sparse NHMM (40 states, 10 comp. each)</td>
<td>10</td>
<td>14.3 ± 1.9</td>
<td>0.66 ± 0.05</td>
</tr>
<tr>
<td>Sparse KL-NMF (55 basis vectors)</td>
<td>1</td>
<td>12.6 ± 1.5</td>
<td>0.53 ± 0.07</td>
</tr>
<tr>
<td>Non-sparse KL-NMF (55 basis vectors)</td>
<td>10</td>
<td>10.3 ± 1.1</td>
<td>0.53 ± 0.08</td>
</tr>
<tr>
<td>Non-sparse KL-NMF (55 basis vectors)</td>
<td>55</td>
<td>7.8 ± 1</td>
<td>0.51 ± 0.05</td>
</tr>
</tbody>
</table>
A Real Application [Mohammadiha & Leijon 2013]

- Nonnegative HMM for Babble Noise Derived From Speech
  HMM: Application to Speech Enhancement
A Real Application [Mohammadiha & Leijon 2013]

- Nonnegative HMM for Babble Noise Derived From Speech
  HMM: Application to Speech Enhancement
  

- Motivation: The basis matrices of speech and babble are quite similar so signals get confused in NMF
Proposed Method: Gamma NHMM

Conclusions

Gamma HMM

NMF Representation

A Real Application [Mohammadiha & Leijon 2013]

- Nonnegative HMM for Babble Noise Derived From Speech
  HMM: Application to Speech Enhancement
  
  N. Mohammadiha and A. Leijon, “Nonnegative HMM for
  babble noise derived from speech HMM: Application to speech

- Motivation: The basis matrices of speech and babble are quite
  similar so signals get confused in NMF

- Solution:
  
  Drive a gamma HMM for speech
  
  Formulate gamma-HMM as a probabilistic sparse NMF (as
  explained here)
  
  To model babble, keep the basis matrix fixed and learn the
  activations by relaxing the sparsity constraint
  
  Combine two models to obtain a factorial HMM to enhance
  the noisy signal
We proposed an NHMM to employ temporal dynamics in NMF

- An HMM with gamma output density functions was elaborated
- The gamma HMM was shown to be equivalent to a sparse NMF
We proposed an NHMM to employ temporal dynamics in NMF

- An HMM with gamma output density functions was elaborated
- The gamma HMM was shown to be equivalent to a sparse NMF

Our experiments using speech signals shows that the proposed approach leads to a better compromise between sparsity, goodness of fit, and temporal modeling compared to state-of-the-art.
We proposed an NHMM to employ temporal dynamics in NMF

- An HMM with gamma output density functions was elaborated
- The gamma HMM was shown to be equivalent to a sparse NMF

Our experiments using speech signals shows that the proposed approach leads to a better compromise between sparsity, goodness of fit, and temporal modeling compared to state-of-the-art.

This work forms a basis for many applications, e.g., the NHMM for babble noise derived from speech HMM.
References (1/2)


Mohammadiha et al. 2013b  N. Mohammadiha, P. Smaragdis, and A. Leijon, “Prediction based filtering and smoothing to exploit temporal dependencies in NMF,” in Proc ICASSP, may 2013.

