Contributions to probabilistic non-negative matrix factorization

Maximum marginal likelihood estimation and temporal Markovian models

Louis Filstroff
IRIT, Univ. Toulouse, CNRS

Joint work with Cédric Févotte, Alberto Lumbreras, Olivier Gouvert, Olivier Cappé

November 29th, 2019

Outline

Outline

Data as matrices

- In many situations, data are available as matrices

Data as matrices

- In many situations, data are available as matrices
- Consider N samples $\mathbf{v}_n$ in $\mathbb{R}^F$ (i.e., described by F features)

Data as matrices

- In many situations, data are available as matrices
- Consider N samples $\mathbf{v}_n$ in $\mathbb{R}^F$ (i.e., described by F features)
- Samples stored column-wise, yielding an $F \times N$ data matrix $\mathbf{V}$

Data as matrices

- In many situations, data are available as matrices
- Consider N samples $\mathbf{v}_n$ in $\mathbb{R}^F$ (i.e., described by F features)
- Samples stored column-wise, yielding an $F \times N$ data matrix $\mathbf{V}$

$\mathbf{V}$ represents	f	n	Typical F
A corpus of documents	Words	Documents	$10^4 - 10^5$
A collection of grayscale images	Pixels	Images	$10^4 - 10^6$
The spectrogram of an audio signal	Frequencies	Time frames	$10^3 - 10^4$
Ratings	Items	Users	$10^6 - 10^8$

Table 1: Examples of data available as matrices

Matrix factorization (1/3)

Matrix factorization (MF)

MF aims at finding a decomposition of the data matrix $\mathbf{V}$ as the product of two matrices

$$\mathbf{V} \simeq \mathbf{WH}, \tag{1}$$

where $\mathbf{W}$ is of size $F \times K$, and $\mathbf{H}$ is of size $K \times N$

Matrix factorization (1/3)

Matrix factorization (MF)

MF aims at finding a decomposition of the data matrix $\mathbf{V}$ as the product of two matrices

$$\mathbf{V} \simeq \mathbf{WH}, \tag{1}$$

where $\mathbf{W}$ is of size $F \times K$, and $\mathbf{H}$ is of size $K \times N$

- $K \ll \min(F, N)$: low-rank approximation

Matrix factorization (1/3)

Matrix factorization (MF)

MF aims at finding a decomposition of the data matrix $\mathbf{V}$ as the product of two matrices

$$\mathbf{V} \simeq \mathbf{WH}, \tag{1}$$

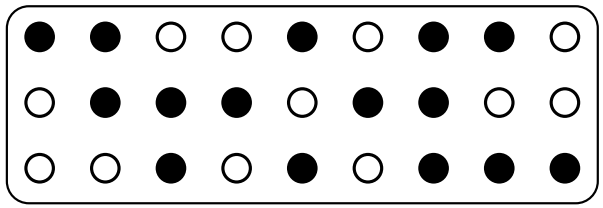
where $\mathbf{W}$ is of size $F \times K$, and $\mathbf{H}$ is of size $K \times N$

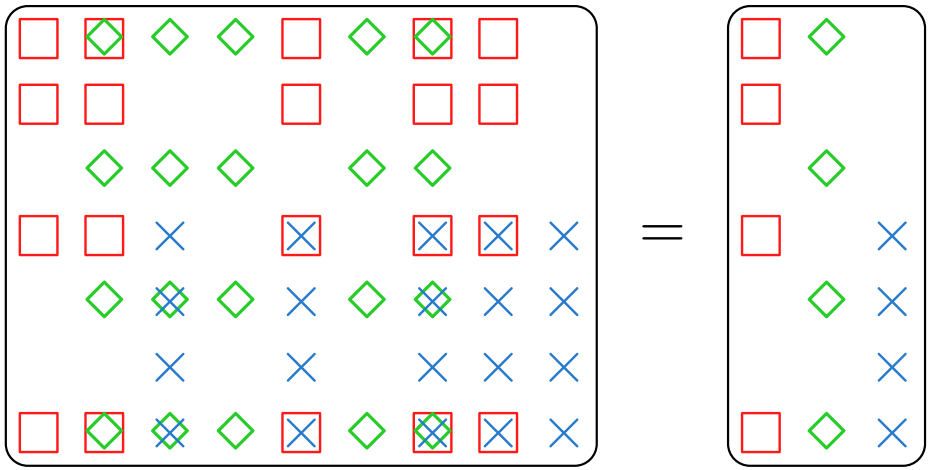
- $K \ll \min(F, N)$: low-rank approximation
- Linear dimensionality reduction technique

$$\mathbf{v}_n \simeq \sum_{k=1}^{K} h_{kn} \mathbf{w}_k \tag{2}$$

Matrix factorization (2/3)

- **W** is called the dictionary. Columns represent characteristic or recurring patterns of the data
- **H** is called the activation coefficients. The n-th column represents how much of each pattern is needed to represent $\mathbf{v}_n$





$=$

$\mathbf{V} = \mathbf{W}$ $\mathbf{H}$

Matrix factorization (3/3)

- MF can be written as an optimization problem

$$\min_{\mathbf{W},\mathbf{H}} D(\mathbf{V}|\mathbf{WH}) = \sum_{f,n} d(v_{fn}|[\mathbf{WH}]_{fn}) \tag{3}$$

  - D is a separable measure of fit ("divergence")
  - Additional constraints over $\mathbf{W}$ and $\mathbf{H}$ for interpretability

Matrix factorization (3/3)

- MF can be written as an optimization problem

$$\min_{\mathbf{W},\mathbf{H}} D(\mathbf{V}|\mathbf{WH}) = \sum_{f,n} d(v_{fn}|[\mathbf{WH}]_{fn}) \tag{3}$$

  - D is a separable measure of fit ("divergence")
  - Additional constraints over $\mathbf{W}$ and $\mathbf{H}$ for interpretability
- Ubiquitous example : principal component analysis (PCA) [Pearson, 1901, Hotelling, 1933]...

Matrix factorization (3/3)

- MF can be written as an optimization problem

$$\min_{\mathbf{W},\mathbf{H}} D(\mathbf{V}|\mathbf{WH}) = \sum_{f,n} d(v_{fn}|[\mathbf{WH}]_{fn}) \tag{3}$$

  - D is a separable measure of fit ("divergence")
  - Additional constraints over $\mathbf{W}$ and $\mathbf{H}$ for interpretability
- Ubiquitous example : principal component analysis (PCA) [Pearson, 1901, Hotelling, 1933]...
- ... but does not take into account the support of the data

Non-negative matrix factorization

Non-negative matrix factorization (NMF)

NMF aims at finding a decomposition of a non-negative data matrix $\mathbf{V}$ as the product of two non-negative matrices

$$\min_{\mathbf{W}\geq 0, \mathbf{H}\geq 0} D(\mathbf{V}|\mathbf{WH}) \tag{4}$$

[Paatero and Tapper, 1994, Lee and Seung, 1999]

Non-negative matrix factorization

Non-negative matrix factorization (NMF)

NMF aims at finding a decomposition of a non-negative data matrix $\mathbf{V}$ as the product of two non-negative matrices

$$\min_{\mathbf{W}\geq 0, \mathbf{H}\geq 0} D(\mathbf{V}|\mathbf{WH}) \tag{4}$$

[Paatero and Tapper, 1994, Lee and Seung, 1999]

- Non-negativity constraints improve interpretability :
  - $\mathbf{W} \geq 0$: direct interpretation of the columns of $\mathbf{W}$
  - $\mathbf{H} \geq 0$: induces part-based representations

Non-negative matrix factorization

Non-negative matrix factorization (NMF)

NMF aims at finding a decomposition of a non-negative data matrix $\mathbf{V}$ as the product of two non-negative matrices

$$\min_{\mathbf{W}\geq 0, \mathbf{H}\geq 0} D(\mathbf{V}|\mathbf{WH}) \tag{4}$$

[Paatero and Tapper, 1994, Lee and Seung, 1999]

- Non-negativity constraints improve interpretability :
  - $\mathbf{W} \geq 0$: direct interpretation of the columns of $\mathbf{W}$
  - $\mathbf{H} \geq 0$: induces part-based representations
- Many application fields :
  - Audio signal processing (source separation [Virtanen, 2007], music transcription [Smaragdis and Brown, 2003])
  - Text information retrieval (topic modeling [Xu et al., 2003])
  - Hyperspectral imaging (unmixing [Bioucas-Dias et al., 2012])

Divergences and statistical models

Name	$d(x\|y)$
Squared Euclidean distance (SED)	$\frac{1}{2}(x-y)^2$
Kullback-Leibler (KL) divergence	$x \log\left(\frac{x}{y}\right) - x + y$
Itakura-Saito (IS) divergence	$\frac{x}{y} - \log\left(\frac{x}{y}\right) + 1$

Table 2: Typical divergences used in NMF

For many usual cost functions, the minimization problem is equivalent to the joint maximum likelihood estimation of $\mathbf{W}$ and $\mathbf{H}$

$$\min_{\mathbf{W}\geq 0, \mathbf{H}\geq 0} D(\mathbf{V}|\mathbf{WH}) \Leftrightarrow \max_{\mathbf{W},\mathbf{H}} p(\mathbf{V}; \mathbf{W}, \mathbf{H}) \tag{5}$$

The probabilistic framework

NMF problem	Equivalent likelihood
SED-NMF	$v_{fn} \sim \mathcal{N}([\mathbf{WH}]_{fn}, \sigma^2)$
KL-NMF	$v_{fn} \sim \text{Poisson}([\mathbf{WH}]_{fn})$
IS-NMF	$v_{fn} \sim \text{Exp}\left(\frac{1}{[\mathbf{WH}]_{fn}}\right)$

Table 3: Equivalences with statistical models

The probabilistic framework

NMF problem	Equivalent likelihood
SED-NMF	$v_{fn} \sim \mathcal{N}([\mathbf{WH}]_{fn}, \sigma^2)$
KL-NMF	$v_{fn} \sim \text{Poisson}([\mathbf{WH}]_{fn})$
IS-NMF	$v_{fn} \sim \text{Exp}\left(\frac{1}{[\mathbf{WH}]_{fn}}\right)$

Table 3: Equivalences with statistical models

Probabilistic NMF

Learning (estimation and/or inference) tasks in statistical models of the form

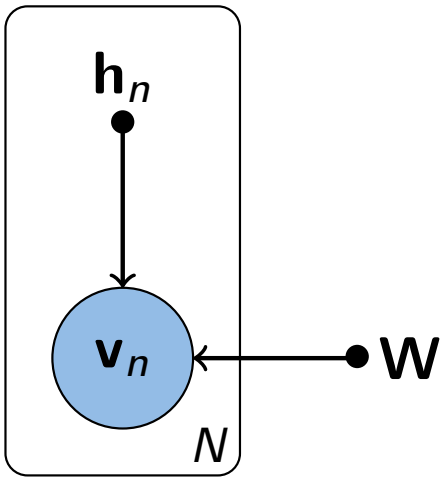
$$\mathbf{v}_n \sim p(.; \mathbf{W}\mathbf{h}_n, \mathbf{\Psi}), \tag{6}$$

i.e., parametrized by the dot product $\mathbf{W}\mathbf{h}_n$ (and $\mathbf{\Psi}$)

NB : Most of the time we have $\mathbb{E}(\mathbf{v_n}) = \mathbf{W}\mathbf{h}_n$

Model variants and learning problems

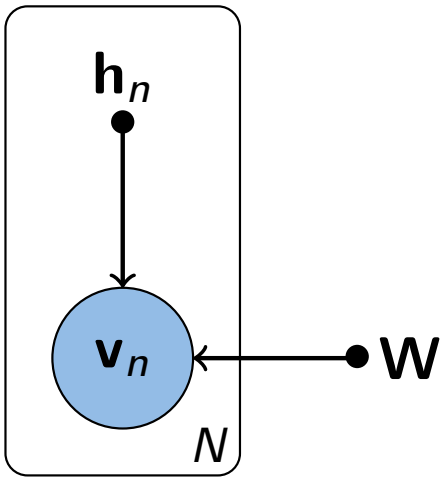
Model variants and learning problems



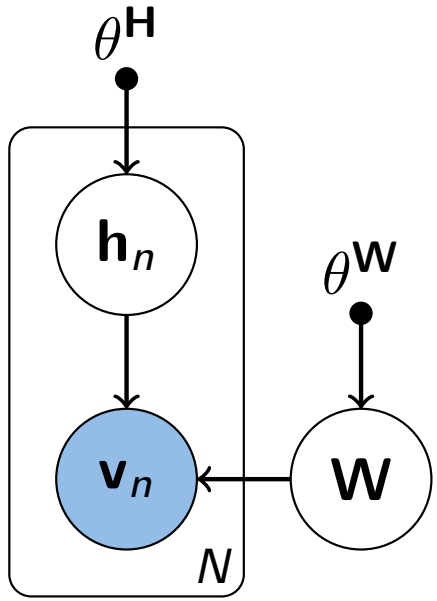
(1)

- (1) Frequentist NMF – Maximum likelihood estimation

Model variants and learning problems



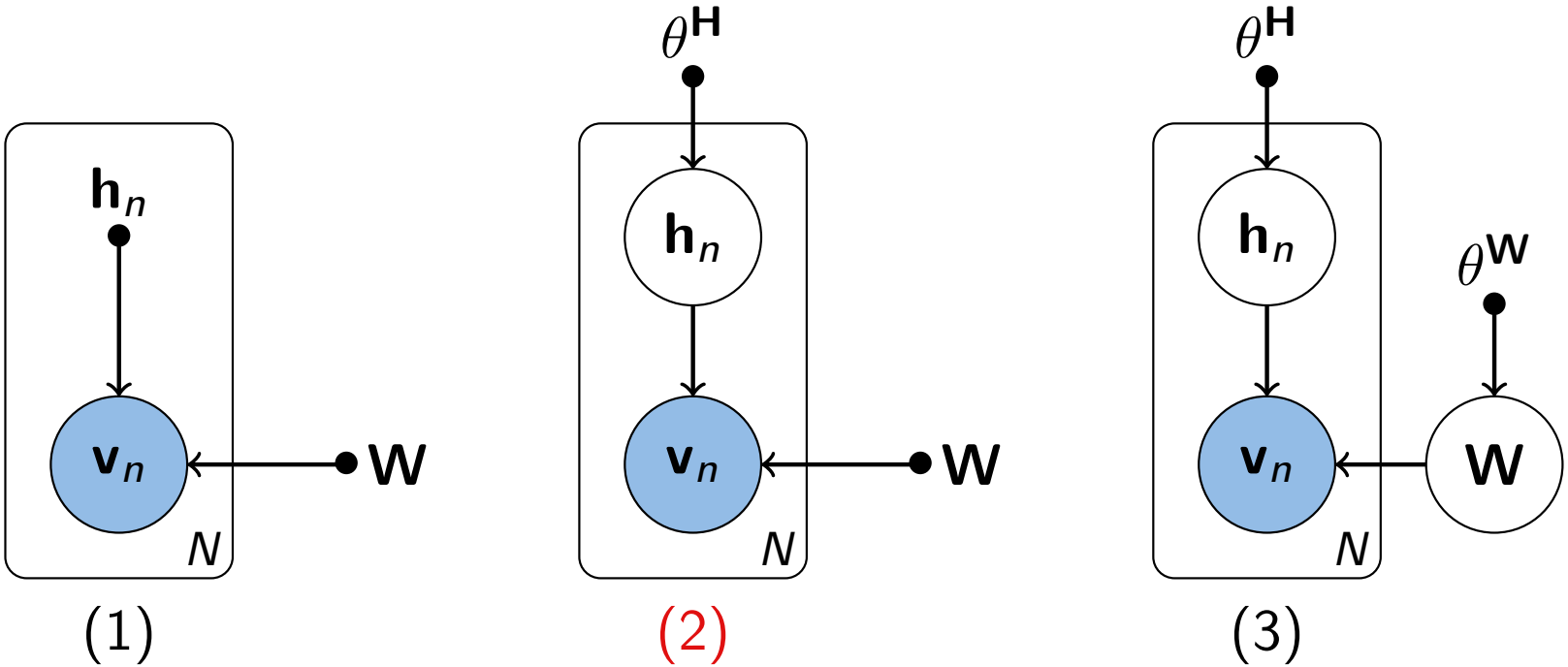
(1)



(3)

- (1) Frequentist NMF – Maximum likelihood estimation
- (3) Bayesian NMF – Infer posterior distribution $p(\mathbf{W}, \mathbf{H}|\mathbf{V})$

Model variants and learning problems



- (1) Frequentist NMF – Maximum likelihood estimation
- (2) Semi-Bayesian NMF – Our setting
- (3) Bayesian NMF – Infer posterior distribution $p(\mathbf{W}, \mathbf{H}|\mathbf{V})$

Inference in semi-Bayesian NMF (1/2)

Two estimation paradigms

Inference in semi-Bayesian NMF (1/2)

Two estimation paradigms

- Maximizing joint likelihood estimation (MJLE) :

$$\max_{\mathbf{W},\mathbf{H}} \log p(\mathbf{V}, \mathbf{H}; \mathbf{W}) = \log p(\mathbf{V}|\mathbf{H}; \mathbf{W}) + \log p(\mathbf{H}) \qquad (7)$$

Estimation of $FK + KN$ parameters

Inference in semi-Bayesian NMF (1/2)

Two estimation paradigms

- Maximizing joint likelihood estimation (MJLE) :

$$\max_{\mathbf{W},\mathbf{H}} \log p(\mathbf{V}, \mathbf{H}; \mathbf{W}) = \log p(\mathbf{V}|\mathbf{H}; \mathbf{W}) + \log p(\mathbf{H}) \tag{7}$$

  Estimation of $FK + KN$ parameters

- Maximizing marginal likelihood estimation (MMLE) :

$$\max_{\mathbf{W}} \log p(\mathbf{V}; \mathbf{W}) = \log \int_{\mathbf{H}} p(\mathbf{V}|\mathbf{H}; \mathbf{W}) p(\mathbf{H}) \mathrm{d}\mathbf{H} \tag{8}$$

  Estimation of FK parameters

Inference in semi-Bayesian NMF (1/2)

Two estimation paradigms

- Maximizing joint likelihood estimation (MJLE) :

$$\max_{\mathbf{W},\mathbf{H}} \log p(\mathbf{V},\mathbf{H};\mathbf{W}) = \log p(\mathbf{V}|\mathbf{H};\mathbf{W}) + \log p(\mathbf{H}) \qquad (7)$$

  Estimation of $FK + KN$ parameters

- Maximizing marginal likelihood estimation (MMLE) :

$$\max_{\mathbf{W}} \log p(\mathbf{V};\mathbf{W}) = \log \int_{\mathbf{H}} p(\mathbf{V}|\mathbf{H};\mathbf{W}) p(\mathbf{H}) \mathrm{d}\mathbf{H} \qquad (8)$$

  Estimation of FK parameters

MMLE is a better-posed approach because the number of parameters to be estimated is fixed w.r.t. the number of samples N

Inference in semi-Bayesian NMF (2/2)

- Empirical comparison of the two paradigms [Dikmen and Févotte, 2011, Dikmen and Févotte, 2012] in the Poisson and Exponential models

Inference in semi-Bayesian NMF (2/2)

- Empirical comparison of the two paradigms [Dikmen and Févotte, 2011, Dikmen and Févotte, 2012] in the Poisson and Exponential models
  - MMLE tends to automatically prune the columns of $\mathbf{W}$, while MJLE makes use of all K columns

Inference in semi-Bayesian NMF (2/2)

- Empirical comparison of the two paradigms [Dikmen and Févotte, 2011, Dikmen and Févotte, 2012] in the Poisson and Exponential models
  - MMLE tends to automatically prune the columns of $\mathbf{W}$, while MJLE makes use of all K columns
  - Favorable behavior that was left unexplained

Inference in semi-Bayesian NMF (2/2)

- Empirical comparison of the two paradigms [Dikmen and Févotte, 2011, Dikmen and Févotte, 2012] in the Poisson and Exponential models
  - MMLE tends to automatically prune the columns of $\mathbf{W}$, while MJLE makes use of all K columns
  - Favorable behavior that was left unexplained
- Related works of the literature

Inference in semi-Bayesian NMF (2/2)

- Empirical comparison of the two paradigms [Dikmen and Févotte, 2011, Dikmen and Févotte, 2012] in the Poisson and Exponential models
  - MMLE tends to automatically prune the columns of $\mathbf{W}$, while MJLE makes use of all K columns
  - Favorable behavior that was left unexplained
- Related works of the literature
  - Integration of nuisance parameters [Berger et al., 1999]

Inference in semi-Bayesian NMF (2/2)

- Empirical comparison of the two paradigms [Dikmen and Févotte, 2011, Dikmen and Févotte, 2012] in the Poisson and Exponential models
  - MMLE tends to automatically prune the columns of $\mathbf{W}$, while MJLE makes use of all K columns
  - Favorable behavior that was left unexplained
- Related works of the literature
  - Integration of nuisance parameters [Berger et al., 1999]
  - Noisy ICA [Moulines et al., 1997]

Inference in semi-Bayesian NMF (2/2)

- Empirical comparison of the two paradigms [Dikmen and Févotte, 2011, Dikmen and Févotte, 2012] in the Poisson and Exponential models
  - MMLE tends to automatically prune the columns of $\mathbf{W}$, while MJLE makes use of all K columns
  - Favorable behavior that was left unexplained
- Related works of the literature
  - Integration of nuisance parameters [Berger et al., 1999]
  - Noisy ICA [Moulines et al., 1997]
  - Latent Dirichlet allocation (LDA) [Blei et al., 2003]

Overview

We will focus on two sub-cases :

Overview

We will focus on two sub-cases :

- Independent priors on $\mathbf{H}$

$$p(\mathbf{H}) = \prod_{n=1}^{N} p(\mathbf{h}_n) \tag{9}$$

  - Poisson likelihood + Gamma prior
  - (Exponential likelihood + Inverse Gamma prior)

Overview

We will focus on two sub-cases :

- Independent priors on $\mathbf{H}$

$$p(\mathbf{H}) = \prod_{n=1}^{N} p(\mathbf{h}_n) \tag{9}$$

  - Poisson likelihood + Gamma prior
  - (Exponential likelihood + Inverse Gamma prior)

- Temporal priors on $\mathbf{H}$

$$p(\mathbf{H}) = p(\mathbf{h}_1) \prod_{n \geq 2} p(\mathbf{h}_n | \mathbf{h}_{n-1}) \tag{10}$$

  - Design of a meaningful prior

Outline

Model and objective

The Gamma-Poisson (GaP) model [Canny, 2004]

$$h_{kn} \sim \text{Gamma}(\alpha_k, \beta_k) \tag{11}$$

$$v_{fn}|\mathbf{h}_n \sim \text{Poisson}([\mathbf{WH}]_{fn}) \tag{12}$$

Model and objective

The Gamma-Poisson (GaP) model [Canny, 2004]

$$h_{kn} \sim \text{Gamma}(\alpha_k, \beta_k) \tag{11}$$

$$v_{fn}|\mathbf{h}_n \sim \text{Poisson}([\mathbf{WH}]_{fn}) \tag{12}$$

- Different application fields for this observation model
  - Text information retrieval (observation model very close to LDA) [Canny, 2004, Buntine and Jakulin, 2006]
  - Recommender systems ("Poisson factorization") [Gopalan et al., 2015]
  - Image processing [Cemgil, 2009]

Model and objective

The Gamma-Poisson (GaP) model [Canny, 2004]

$$h_{kn} \sim \text{Gamma}(\alpha_k, \beta_k) \tag{11}$$

$$v_{fn}|\mathbf{h}_n \sim \text{Poisson}([\mathbf{WH}]_{fn}) \tag{12}$$

- Different application fields for this observation model
  - Text information retrieval (observation model very close to LDA) [Canny, 2004, Buntine and Jakulin, 2006]
  - Recommender systems ("Poisson factorization") [Gopalan et al., 2015]
  - Image processing [Cemgil, 2009]
- MMLE amounts to

$$\min_{\mathbf{W}} \mathcal{L}(\mathbf{W}) = -\log \int p(\mathbf{V}|\mathbf{H};\mathbf{W})p(\mathbf{H})\mathrm{d}\mathbf{H} \tag{13}$$

Hyperparameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ may also be optimized

Data augmentation

- Superposition property of the Poisson distribution

The augmented GaP model

$$h_{kn} \sim \text{Gamma}(\alpha_k, \beta_k) \tag{14}$$

$$c_{fkn} \sim \text{Poisson}(w_{fk} h_{kn}) \tag{15}$$

$$v_{fn} = \sum_k c_{fkn} \tag{16}$$

Data augmentation

- Superposition property of the Poisson distribution

The augmented GaP model

$$h_{kn} \sim \text{Gamma}(\alpha_k, \beta_k) \tag{14}$$

$$c_{fkn} \sim \text{Poisson}(w_{fk} h_{kn}) \tag{15}$$

$$v_{fn} = \sum_k c_{fkn} \tag{16}$$

- $\mathbf{C}$ denotes the $F \times K \times N$ tensor with entries c_{fkn}

Data augmentation

- Superposition property of the Poisson distribution

The augmented GaP model

$$h_{kn} \sim \text{Gamma}(\alpha_k, \beta_k) \tag{14}$$

$$c_{fkn} \sim \text{Poisson}(w_{fk} h_{kn}) \tag{15}$$

$$v_{fn} = \sum_k c_{fkn} \tag{16}$$

- $\mathbf{C}$ denotes the $F \times K \times N$ tensor with entries c_{fkn}
- Thanks to the conjugacy between the Poisson and the Gamma distribution, h_{kn} can be marginalized out from Eqs. (14)-(15)

Marginalizing the activation coefficients

New formulation of GaP

$$\mathbf{c}_{kn} \sim \mathrm{NM}\left(\alpha_k, \left[\frac{w_{1k}}{\sum_f w_{fk} + \beta_k}, \ldots, \frac{w_{Fk}}{\sum_f w_{fk} + \beta_k}\right]^{\mathsf{T}}\right) \tag{17}$$

$$\mathbf{v}_n = \sum_k \mathbf{c}_{kn} \tag{18}$$

where $\mathbf{c}_{kn} = [c_{1kn}, \ldots, c_{Fkn}]^{\mathsf{T}}$ is a vector of size F

Marginalizing the activation coefficients

New formulation of GaP

$$\mathbf{c}_{kn} \sim \mathrm{NM}\left(\alpha_k, \left[\frac{w_{1k}}{\sum_f w_{fk} + \beta_k}, \ldots, \frac{w_{Fk}}{\sum_f w_{fk} + \beta_k}\right]^{\mathsf{T}}\right) \tag{17}$$

$$\mathbf{v}_n = \sum_k \mathbf{c}_{kn} \tag{18}$$

where $\mathbf{c}_{kn} = [c_{1kn}, \ldots, c_{Fkn}]^{\mathsf{T}}$ is a vector of size F

- The vector $\mathbf{c}_{kn}$ has a so-called negative multinomial (NM) distribution, known in closed form

Marginalizing the activation coefficients

New formulation of GaP

$$\mathbf{c}_{kn} \sim \mathrm{NM}\left(\alpha_k, \left[\frac{w_{1k}}{\sum_f w_{fk} + \beta_k}, \ldots, \frac{w_{Fk}}{\sum_f w_{fk} + \beta_k}\right]^{\mathsf{T}}\right) \tag{17}$$

$$\mathbf{v}_n = \sum_k \mathbf{c}_{kn} \tag{18}$$

where $\mathbf{c}_{kn} = [c_{1kn}, \ldots, c_{Fkn}]^{\mathsf{T}}$ is a vector of size F

- The vector $\mathbf{c}_{kn}$ has a so-called negative multinomial (NM) distribution, known in closed form
- GaP can therefore be seen as a composite NM model

Marginalizing the activation coefficients

New formulation of GaP

$$\mathbf{c}_{kn} \sim \mathrm{NM}\left(\alpha_k, \left[\frac{w_{1k}}{\sum_f w_{fk} + \beta_k}, \ldots, \frac{w_{Fk}}{\sum_f w_{fk} + \beta_k}\right]^{\mathsf{T}}\right) \tag{17}$$

$$\mathbf{v}_n = \sum_k \mathbf{c}_{kn} \tag{18}$$

where $\mathbf{c}_{kn} = [c_{1kn}, \ldots, c_{Fkn}]^{\mathsf{T}}$ is a vector of size F

- The vector $\mathbf{c}_{kn}$ has a so-called negative multinomial (NM) distribution, known in closed form
- GaP can therefore be seen as a composite NM model
- Alternative characterization with the multinomial distribution

Closed-form marginal likelihood

Closed-form marginal likelihood

- Denote by $\mathcal{C}_{\mathbf{V}}$ the set of "admissible components"

$$\mathcal{C}_{\mathbf{V}} = \{\mathbf{C} \in \mathbb{N}^{F \times K \times N} \,|\, \forall (f, n), \sum_{k} c_{fkn} = v_{fn}\}. \qquad (19)$$

Closed-form marginal likelihood

- Denote by $\mathcal{C}_{\mathbf{V}}$ the set of "admissible components"

$$\mathcal{C}_{\mathbf{V}} = \{\mathbf{C} \in \mathbb{N}^{F \times K \times N} \,|\, \forall (f, n), \sum_{k} c_{fkn} = v_{fn}\}. \tag{19}$$

- The marginalization of $\mathbf{C}$ yields

$$p(\mathbf{V}; \mathbf{W}) = \sum_{\mathbf{C} \in \mathcal{C}_{\mathbf{V}}} p(\mathbf{C}; \mathbf{W}) = \sum_{\mathbf{C} \in \mathcal{C}_{\mathbf{V}}} \prod_{k,n} \underbrace{p(\mathbf{c}_{kn}; \mathbf{w}_k)}_{\text{NM}} \tag{20}$$

Exhibiting a regularization term

Exhibiting a regularization term

- After computation, we obtain

$$-\frac{1}{N}\mathcal{L}(\mathbf{W}) = -\frac{1}{N}\log\left(\sum_{\mathbf{C}\in\mathcal{C}_\mathbf{V}} f(\mathbf{C};\mathbf{W})\right) \tag{21}$$

$$+\sum_k \alpha_k \log(||\mathbf{w}_k||_1 + \beta_k) + \text{cst} \tag{22}$$

Exhibiting a regularization term

- After computation, we obtain

$$-\frac{1}{N}\mathcal{L}(\mathbf{W}) = -\frac{1}{N}\log\left(\sum_{\mathbf{C}\in\mathcal{C}_\mathbf{V}} f(\mathbf{C};\mathbf{W})\right) \tag{21}$$

$$+\sum_k \alpha_k \log(||\mathbf{w}_k||_1 + \beta_k) + \text{cst} \tag{22}$$

- "Data-fitting term" + "regularization term"

Exhibiting a regularization term

- After computation, we obtain

$$-\frac{1}{N}\mathcal{L}(\mathbf{W}) = -\frac{1}{N}\log\left(\sum_{\mathbf{C}\in\mathcal{C}_\mathbf{V}} f(\mathbf{C};\mathbf{W})\right) \tag{21}$$

$$+\sum_k \alpha_k \log(||\mathbf{w}_k||_1 + \beta_k) + \text{cst} \tag{22}$$

- "Data-fitting term" + "regularization term"
- Term of the form $R(\mathbf{x}) = \sum_k \log(|x_k| + \epsilon)$ is known to be sparsity-inducing [Candès et al., 2008]

Exhibiting a regularization term

- After computation, we obtain

$$-\frac{1}{N}\mathcal{L}(\mathbf{W}) = -\frac{1}{N}\log\left(\sum_{\mathbf{C}\in\mathcal{C}_\mathbf{V}} f(\mathbf{C};\mathbf{W})\right) \tag{21}$$

$$+\sum_k \alpha_k \log(||\mathbf{w}_k||_1 + \beta_k) + \text{cst} \tag{22}$$

- "Data-fitting term" + "regularization term"
- Term of the form $R(\mathbf{x}) = \sum_k \log(|x_k| + \epsilon)$ is known to be sparsity-inducing [Candès et al., 2008]
- Provides a deeper understanding of the self-regularization phenomenon

Optimization of the marginal likelihood (1/2)

Optimization of the marginal likelihood (1/2)

- EM algorithm [Dempster et al., 1977] iteratively optimizes the following functional

$$Q(\mathbf{W}; \tilde{\mathbf{W}}) = \int_{\mathbf{Z}} \log p(\mathbf{V}, \mathbf{Z}; \mathbf{W}) p(\mathbf{Z}|\mathbf{V}; \tilde{\mathbf{W}}) \tag{23}$$

$\mathbf{Z}$ is the set of latent variables, $\tilde{\mathbf{W}}$ the current value

Optimization of the marginal likelihood (1/2)

- EM algorithm [Dempster et al., 1977] iteratively optimizes the following functional

$$Q(\mathbf{W}; \tilde{\mathbf{W}}) = \int_{\mathbf{Z}} \log p(\mathbf{V}, \mathbf{Z}; \mathbf{W}) p(\mathbf{Z}|\mathbf{V}; \tilde{\mathbf{W}}) \tag{23}$$

  $\mathbf{Z}$ is the set of latent variables, $\tilde{\mathbf{W}}$ the current value

- Three possible choices :
  - $\mathbf{Z} = \{\mathbf{C}, \mathbf{H}\}$: known from [Dikmen and Févotte, 2012]
  - $\mathbf{Z} = \{\mathbf{H}\}$: known from [Dikmen and Févotte, 2012]
  - $\mathbf{Z} = \{\mathbf{C}\}$: novel

Optimization of the marginal likelihood (1/2)

- EM algorithm [Dempster et al., 1977] iteratively optimizes the following functional

$$Q(\mathbf{W}; \tilde{\mathbf{W}}) = \int_{\mathbf{Z}} \log p(\mathbf{V}, \mathbf{Z}; \mathbf{W}) p(\mathbf{Z}|\mathbf{V}; \tilde{\mathbf{W}}) \tag{23}$$

  $\mathbf{Z}$ is the set of latent variables, $\tilde{\mathbf{W}}$ the current value

- Three possible choices :
  - $\mathbf{Z} = \{\mathbf{C}, \mathbf{H}\}$: known from [Dikmen and Févotte, 2012]
  - $\mathbf{Z} = \{\mathbf{H}\}$: known from [Dikmen and Févotte, 2012]
  - $\mathbf{Z} = \{\mathbf{C}\}$: novel
- In all cases the posterior of the latent variables $p(\mathbf{Z}|\mathbf{V}; \tilde{\mathbf{W}})$ is not tractable

Optimization of the marginal likelihood (2/2)

[1] Majorization-Minimization

Optimization of the marginal likelihood (2/2)

- We resort to Monte Carlo EM [Wei and Tanner, 1990]

[1] Majorization-Minimization

Optimization of the marginal likelihood (2/2)

- We resort to Monte Carlo EM [Wei and Tanner, 1990]
- Sampling from $p(\mathbf{C}, \mathbf{H}|\mathbf{V}; \tilde{\mathbf{W}})$ with a Gibbs sampling procedure

[1] Majorization-Minimization

Optimization of the marginal likelihood (2/2)

- We resort to Monte Carlo EM [Wei and Tanner, 1990]
- Sampling from $p(\mathbf{C}, \mathbf{H}|\mathbf{V}; \tilde{\mathbf{W}})$ with a Gibbs sampling procedure
- Optimizes instead

$$\hat{Q}(\mathbf{W}) = \frac{1}{J} \sum_{j} \log p(\mathbf{V}, \mathbf{Z}^{(j)}; \mathbf{W}) \tag{24}$$

Can be carried out in closed form for EM-CH and EM-C, EM-H requires a MM[1]-based procedure

[1] Majorization-Minimization

Experimental work (1/3)

Synthetic dataset $\mathbf{V}_1$ (4×100) generated from the GaP model with $\mathbf{W}_1^\star$

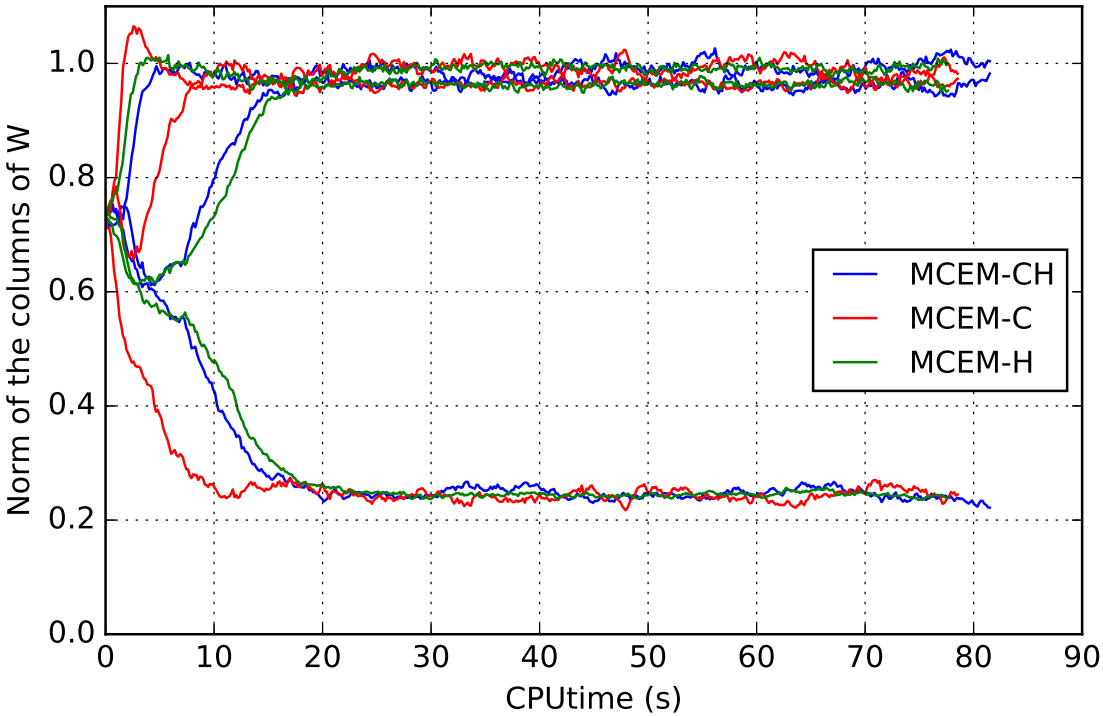


Figure 1: Speed of convergence comparison of the three algorithms on dataset $\mathbf{V}_1$ $(K = 3)$

Experimental work (2/3)

- Synthetic dataset $\mathbf{V}_2$ (4×100) generated from the GaP model with $\mathbf{W}_2^\star = 100 \times \mathbf{W}_1^\star$
- Over-dispersed, non-sparse

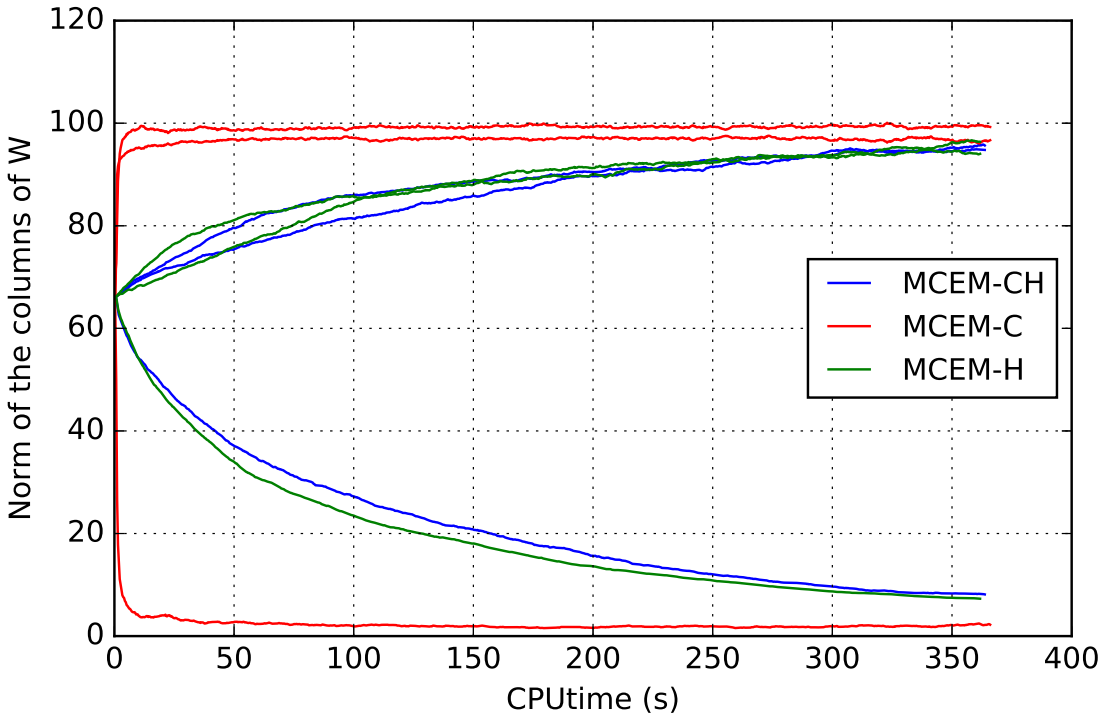


Figure 2: Speed of convergence comparison of the three algorithms on dataset $\mathbf{V}_2$ $(K = 3)$

Experimental work (3/3)

- `Taste Profile` dataset (1509×805)
- Over-dispersed, sparse

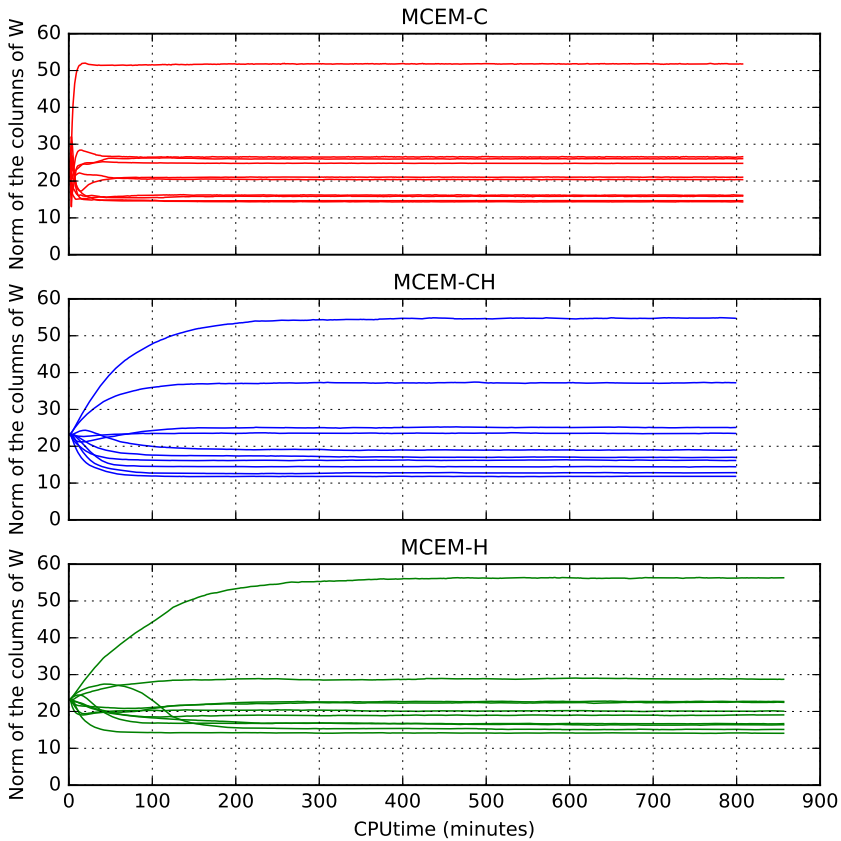


Figure 3: Speed of convergence comparison of the three algorithms the `Taste Profile` dataset ($K = 10$)

Outline

Temporal prior

- Matrices $\mathbf{V}$ where the samples are correlated

Temporal prior

- Matrices $\mathbf{V}$ where the samples are correlated
- Add correlation to the models by lifting the independence assumption on the columns of $\mathbf{H}$

Temporal prior

- Matrices $\mathbf{V}$ where the samples are correlated
- Add correlation to the models by lifting the independence assumption on the columns of $\mathbf{H}$
- Markov structure on the columns + independence of the rows

$$p(\mathbf{H}) = \prod_k p(h_{k1}) \prod_{n \geq 2} p(h_{kn} | h_{k(n-1)}) \tag{25}$$

So-called temporal models

Temporal prior

- Matrices $\mathbf{V}$ where the samples are correlated
- Add correlation to the models by lifting the independence assumption on the columns of $\mathbf{H}$
- Markov structure on the columns + independence of the rows

$$p(\mathbf{H}) = \prod_k p(h_{k1}) \prod_{n \geq 2} p(h_{kn} | h_{k(n-1)}) \tag{25}$$

  So-called temporal models
- Non-negative Markov chains, in relation with the Gamma distribution

Gamma Markov Chains of the literature (1/3)

Chaining on the rate parameter

$$h_{kn}|h_{k(n-1)} \sim \text{Gamma}\left(\alpha, \frac{\beta}{h_{k(n-1)}}\right) \tag{26}$$

[Févotte et al., 2009, Févotte, 2011]

Gamma Markov Chains of the literature (1/3)

Chaining on the rate parameter

$$h_{kn}|h_{k(n-1)} \sim \text{Gamma}\left(\alpha, \frac{\beta}{h_{k(n-1)}}\right) \tag{26}$$

[Févotte et al., 2009, Févotte, 2011]

- $\mathbb{E}(h_{kn}|h_{k(n-1)}) = \frac{\alpha}{\beta} h_{k(n-1)}$
- No well-defined stationary distribution

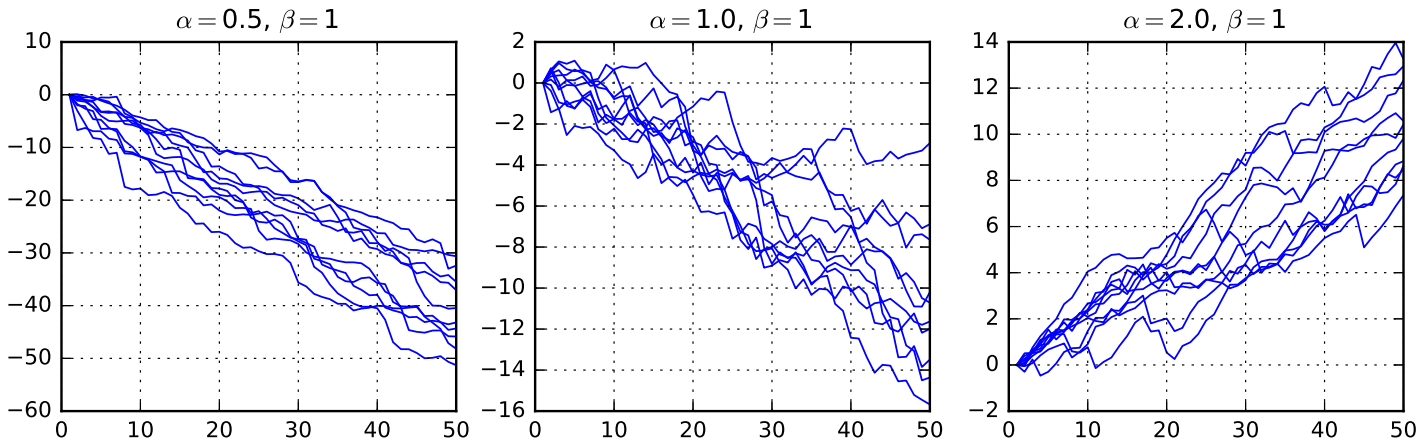
Gamma Markov Chains of the literature (1/3)

Chaining on the rate parameter

$$h_{kn}|h_{k(n-1)} \sim \text{Gamma}\left(\alpha, \frac{\beta}{h_{k(n-1)}}\right) \tag{26}$$

[Févotte et al., 2009, Févotte, 2011]

- $\mathbb{E}(h_{kn}|h_{k(n-1)}) = \frac{\alpha}{\beta} h_{k(n-1)}$
- No well-defined stationary distribution



Gamma Markov Chains of the literature (2/3)

Chaining on the rate parameter with an auxiliary variable

$$z_{kn}|h_{k(n-1)} \sim \text{Gamma}\left(\alpha_z, \beta_z h_{k(n-1)}\right) \tag{27}$$

$$h_{kn}|z_{kn} \sim \text{Gamma}\left(\alpha_h, \beta_h z_{kn}\right) \tag{28}$$

[Cemgil and Dikmen, 2007]

Gamma Markov Chains of the literature (2/3)

Chaining on the rate parameter with an auxiliary variable

$$z_{kn}|h_{k(n-1)} \sim \text{Gamma}\left(\alpha_z, \beta_z h_{k(n-1)}\right) \tag{27}$$

$$h_{kn}|z_{kn} \sim \text{Gamma}\left(\alpha_h, \beta_h z_{kn}\right) \tag{28}$$

[Cemgil and Dikmen, 2007]

- $\mathbb{E}(h_{kn}|h_{k(n-1)}) = \frac{\beta_z \alpha_h}{\beta_h(\alpha_z - 1)} h_{k(n-1)}$
- No well-defined stationary distribution

Gamma Markov Chains of the literature (2/3)

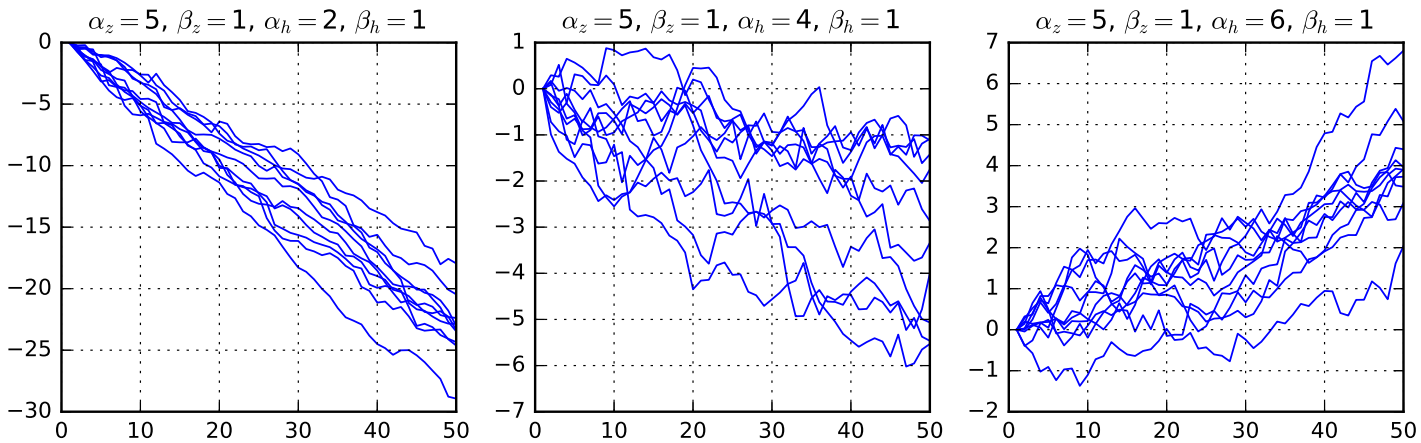
Chaining on the rate parameter with an auxiliary variable

$$z_{kn}|h_{k(n-1)} \sim \text{Gamma}\left(\alpha_z, \beta_z h_{k(n-1)}\right) \tag{27}$$

$$h_{kn}|z_{kn} \sim \text{Gamma}\left(\alpha_h, \beta_h z_{kn}\right) \tag{28}$$

[Cemgil and Dikmen, 2007]

- $\mathbb{E}(h_{kn}|h_{k(n-1)}) = \frac{\beta_z \alpha_h}{\beta_h(\alpha_z - 1)} h_{k(n-1)}$
- No well-defined stationary distribution



Gamma Markov Chains of the literature (3/3)

Chaining on the shape parameter

$$h_{kn} |\ h_{k(n-1)} \sim \text{Gamma}\left(\alpha h_{k(n-1)}, \beta\right) \tag{29}$$

[Acharya et al., 2015, Schein et al., 2016]

Gamma Markov Chains of the literature (3/3)

Chaining on the shape parameter

$$h_{kn} |\ h_{k(n-1)} \sim \text{Gamma}\left(\alpha h_{k(n-1)}, \beta\right) \tag{29}$$

[Acharya et al., 2015, Schein et al., 2016]

- $\mathbb{E}(h_{kn} | h_{k(n-1)}) = \frac{\alpha}{\beta} h_{k(n-1)}$
- No well-defined stationary distribution

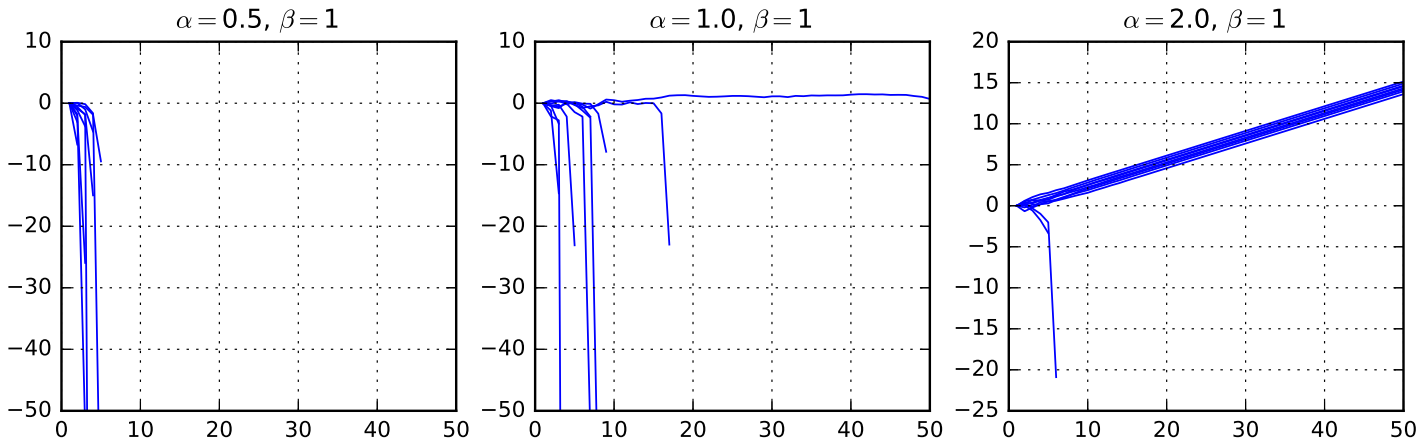
Gamma Markov Chains of the literature (3/3)

Chaining on the shape parameter

$$h_{kn} |\ h_{k(n-1)} \sim \text{Gamma}\left(\alpha h_{k(n-1)}, \beta\right) \tag{29}$$

[Acharya et al., 2015, Schein et al., 2016]

- $\mathbb{E}(h_{kn} | h_{k(n-1)}) = \frac{\alpha}{\beta} h_{k(n-1)}$
- No well-defined stationary distribution



Limitations

Limitations

- All the chains proposed in the NMF literature are based on $\mathbb{E}(h_{kn}|h_{k(n-1)}) \propto h_{k(n-1)}$

Limitations

- All the chains proposed in the NMF literature are based on $\mathbb{E}(h_{kn}|h_{k(n-1)}) \propto h_{k(n-1)}$
- All share the same drawback : the absence of a well-defined stationary distribution

Limitations

- All the chains proposed in the NMF literature are based on $\mathbb{E}(h_{kn}|h_{k(n-1)}) \propto h_{k(n-1)}$
- All share the same drawback : the absence of a well-defined stationary distribution
- Leads to degenerate realizations of the chain
- Difficult to interpret from a generative perspective

Limitations

- All the chains proposed in the NMF literature are based on $\mathbb{E}(h_{kn}|h_{k(n-1)}) \propto h_{k(n-1)}$
- All share the same drawback : the absence of a well-defined stationary distribution
- Leads to degenerate realizations of the chain
- Difficult to interpret from a generative perspective
- We propose to use Markov chains with a well-defined stationary distribution

The BGAR(1) process (1/2)

First-order autoregressive Beta-Gamma process – BGAR(1)

$$h_{k1} \sim \text{Gamma}(\alpha, \beta) \tag{30}$$

$$h_{kn} = b_{kn} h_{k(n-1)} + \epsilon_{kn} \tag{31}$$

where $b_{kn} \in [0, 1]$ and $\epsilon_{kn} \geq 0$ are i.i.d. r.v. such that

$$b_{kn} \sim \text{Beta}(\alpha\rho, \alpha(1-\rho)) \tag{32}$$

$$\epsilon_{kn} \sim \text{Gamma}(\alpha(1-\rho), \beta) \tag{33}$$

[Lewis et al., 1989]

The BGAR(1) process (1/2)

First-order autoregressive Beta-Gamma process – BGAR(1)

$$h_{k1} \sim \text{Gamma}(\alpha, \beta) \tag{30}$$

$$h_{kn} = b_{kn} h_{k(n-1)} + \epsilon_{kn} \tag{31}$$

where $b_{kn} \in [0, 1]$ and $\epsilon_{kn} \geq 0$ are i.i.d. r.v. such that

$$b_{kn} \sim \text{Beta}(\alpha\rho, \alpha(1-\rho)) \tag{32}$$

$$\epsilon_{kn} \sim \text{Gamma}(\alpha(1-\rho), \beta) \tag{33}$$

[Lewis et al., 1989]

- We have

$$\mathbb{E}(h_{kn} | h_{k(n-1)}) = \rho h_{k(n-1)} + \frac{\alpha(1-\rho)}{\beta} \tag{34}$$

The BGAR(1) process (1/2)

First-order autoregressive Beta-Gamma process – BGAR(1)

$$h_{k1} \sim \text{Gamma}(\alpha, \beta) \tag{30}$$

$$h_{kn} = b_{kn} h_{k(n-1)} + \epsilon_{kn} \tag{31}$$

where $b_{kn} \in [0, 1]$ and $\epsilon_{kn} \geq 0$ are i.i.d. r.v. such that

$$b_{kn} \sim \text{Beta}(\alpha\rho, \alpha(1-\rho)) \tag{32}$$

$$\epsilon_{kn} \sim \text{Gamma}(\alpha(1-\rho), \beta) \tag{33}$$

[Lewis et al., 1989]

- We have

$$\mathbb{E}(h_{kn} | h_{k(n-1)}) = \rho h_{k(n-1)} + \frac{\alpha(1-\rho)}{\beta} \tag{34}$$

- h_{kn} is marginally Gamma distributed

The BGAR(1) process (2/2)

More precisely...

The BGAR(1) process (2/2)

More precisely...

- α and β control the marginal distribution

The BGAR(1) process (2/2)

More precisely...

- α and β control the marginal distribution
- ρ controls the correlation between two successive values

The BGAR(1) process (2/2)

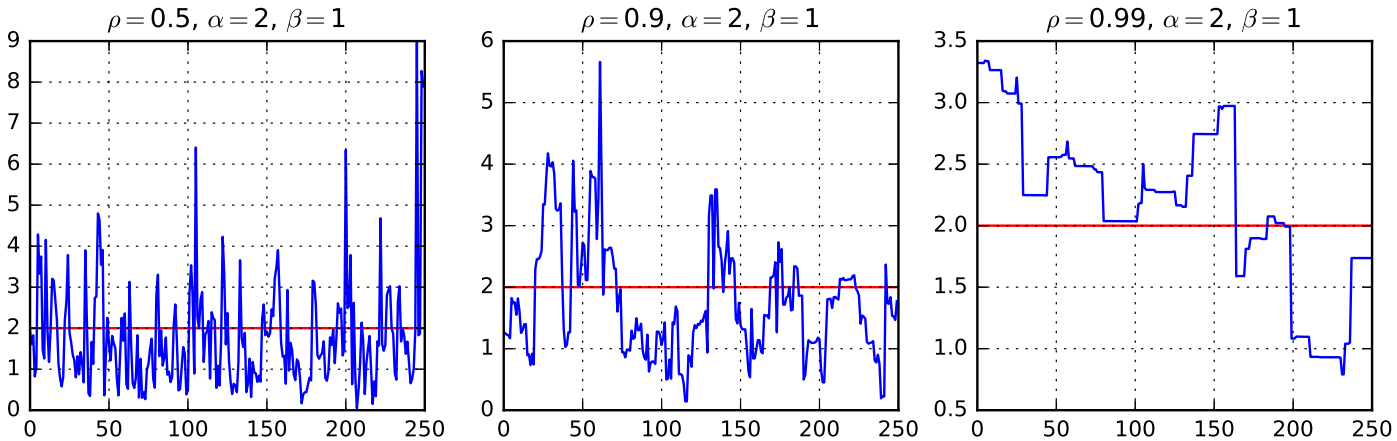
More precisely...

- α and β control the marginal distribution
- ρ controls the correlation between two successive values
- $\rho \to 0$: i.i.d. random variables, $\rho \to 1$: deterministic process

The BGAR(1) process (2/2)

More precisely...

- α and β control the marginal distribution
- ρ controls the correlation between two successive values
- $\rho \to 0$: i.i.d. random variables, $\rho \to 1$: deterministic process



A novel temporal NMF model

The BGAR-NMF model

$$\underline{\mathbf{h}}_k \sim \text{BGAR}(\rho_k, \alpha_k, \beta_k) \tag{35}$$

$$v_{fn} | \ \mathbf{h}_n \sim \text{Poisson}([\mathbf{WH}]_{fn}) \tag{36}$$

A novel temporal NMF model

The BGAR-NMF model

$$\underline{\mathbf{h}}_k \sim \text{BGAR}(\rho_k, \alpha_k, \beta_k) \tag{35}$$

$$v_{fn} | \, \mathbf{h}_n \sim \text{Poisson}([\mathbf{WH}]_{fn}) \tag{36}$$

- $\mathbf{W}$ is left to be a deterministic variable to be estimated

A novel temporal NMF model

The BGAR-NMF model

$$\underline{\mathbf{h}}_k \sim \text{BGAR}(\rho_k, \alpha_k, \beta_k) \tag{35}$$

$$v_{fn} | \, \mathbf{h}_n \sim \text{Poisson}([\mathbf{WH}]_{fn}) \tag{36}$$

- $\mathbf{W}$ is left to be a deterministic variable to be estimated
- $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, $\boldsymbol{\rho}$ are treated as fixed hyperparameters

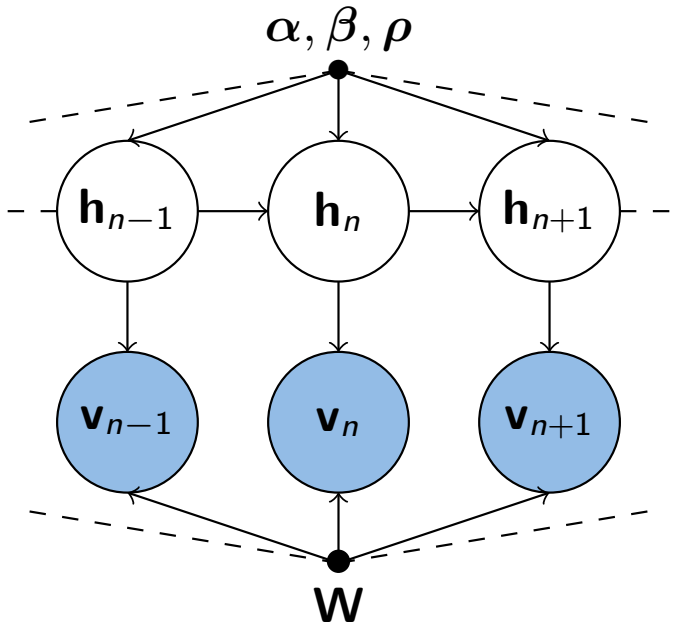
A novel temporal NMF model

The BGAR-NMF model

$$\underline{\mathbf{h}}_k \sim \text{BGAR}(\rho_k, \alpha_k, \beta_k) \tag{35}$$

$$v_{fn} | \, \mathbf{h}_n \sim \text{Poisson}([\mathbf{WH}]_{fn}) \tag{36}$$

- $\mathbf{W}$ is left to be a deterministic variable to be estimated
- α, β, ρ are treated as fixed hyperparameters
- $\mathbf{V}$ and $\mathbf{H}$ define a hidden Markov model [Cappé et al., 2005]



MMLE in the BGAR-NMF model (1/2)

MMLE in the BGAR-NMF model (1/2)

- Maximize the marginal likelihood

$$\max_{\mathbf{W}} p(\mathbf{V};\mathbf{W}) = \int_{\mathbf{H},\mathbf{B}} p(\mathbf{V},\mathbf{H},\mathbf{B};\mathbf{W})\mathrm{d}\mathbf{H}\mathrm{d}\mathbf{B} \tag{37}$$

  Using $\mathbf{B}$ as auxiliary variables

MMLE in the BGAR-NMF model (1/2)

- Maximize the marginal likelihood

$$\max_{\mathbf{W}} p(\mathbf{V};\mathbf{W}) = \int_{\mathbf{H},\mathbf{B}} p(\mathbf{V},\mathbf{H},\mathbf{B};\mathbf{W})\mathrm{d}\mathbf{H}\mathrm{d}\mathbf{B} \tag{37}$$

  Using $\mathbf{B}$ as auxiliary variables

- Amounts to estimating the static parameters of the HMM [Kantas et al., 2015]

MMLE in the BGAR-NMF model (1/2)

- Maximize the marginal likelihood

$$\max_{\mathbf{W}} p(\mathbf{V}; \mathbf{W}) = \int_{\mathbf{H},\mathbf{B}} p(\mathbf{V}, \mathbf{H}, \mathbf{B}; \mathbf{W}) \mathrm{d}\mathbf{H} \mathrm{d}\mathbf{B} \tag{37}$$

  Using $\mathbf{B}$ as auxiliary variables
- Amounts to estimating the static parameters of the HMM [Kantas et al., 2015]
- MCEM algorithm whose sampling step is carried out with sequential Monte Carlo (SMC)

MMLE in the BGAR-NMF model (2/2)

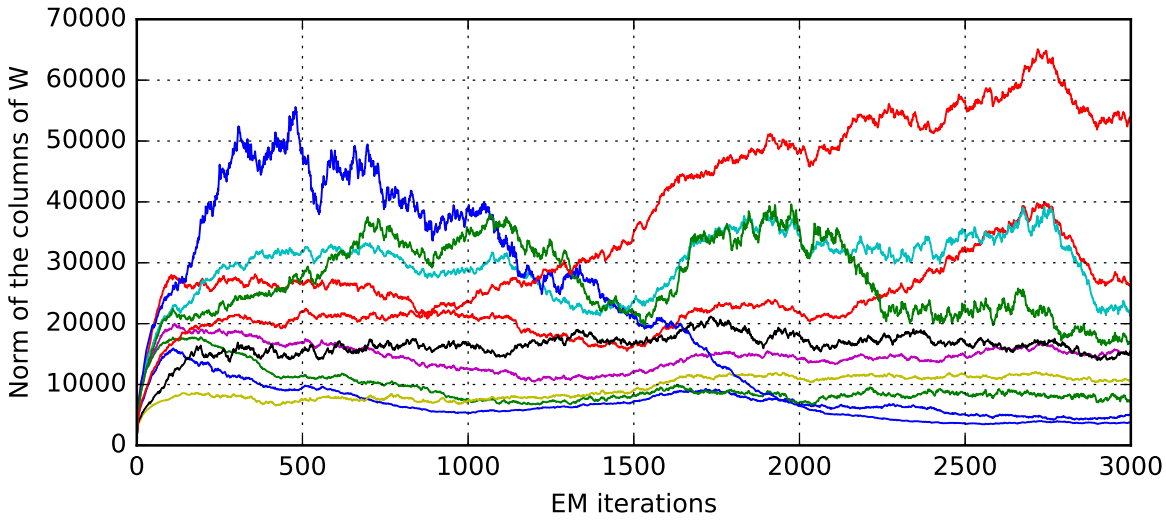


Figure 4: Evolution of the norm of the columns of $\mathbf{W}$ w.r.t. the number of EM iterations on the `NIPS` dataset (11463×29).

- Method seemingly works on small dimensioned datasets
- Fails to produce exploitable results on real datasets
  - Samples of poor quality ?
  - Label switching

MMLE in the BGAR-NMF model (2/2)

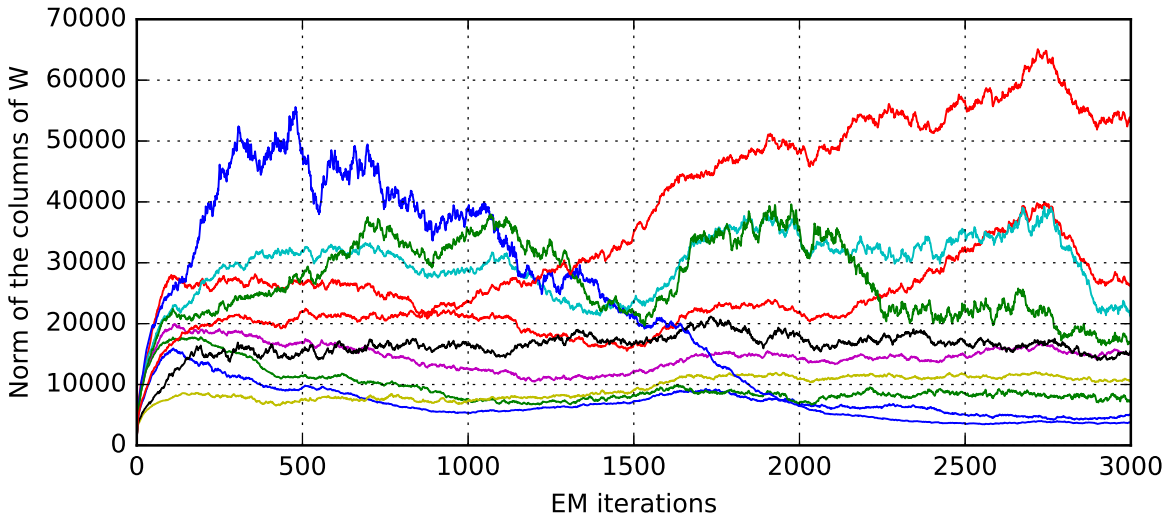


Figure 4: Evolution of the norm of the columns of $\mathbf{W}$ w.r.t. the number of EM iterations on the `NIPS` dataset (11463×29).

- Method seemingly works on small dimensioned datasets
- Fails to produce exploitable results on real datasets
  - Samples of poor quality ?
  - Label switching
- Need for an alternative estimation paradigm in the BGAR-NMF model

MAP estimation in the BGAR-NMF model (1/3)

- MAP estimation amounts to the minimization of the following function

$$C(\mathbf{W}, \mathbf{H}, \mathbf{B}) = -\log p(\mathbf{H}, \mathbf{B}|\mathbf{V}; \mathbf{W}) \tag{38}$$

$$= -\log p(\mathbf{V}|\mathbf{H}; \mathbf{W}) - \log p(\mathbf{H}, \mathbf{B}) + \text{cst} \tag{39}$$

MAP estimation in the BGAR-NMF model (1/3)

- MAP estimation amounts to the minimization of the following function

$$C(\mathbf{W}, \mathbf{H}, \mathbf{B}) = -\log p(\mathbf{H}, \mathbf{B}|\mathbf{V}; \mathbf{W}) \tag{38}$$
$$= -\log p(\mathbf{V}|\mathbf{H}; \mathbf{W}) - \log p(\mathbf{H}, \mathbf{B}) + \text{cst} \tag{39}$$

- We resort to a MM-based scheme. Only $-\log p(\mathbf{V}|\mathbf{H}; \mathbf{W})$ needs to be majorized. Standard scheme in the NMF literature [Lee and Seung, 2000, Févotte and Idier, 2011]

MAP estimation in the BGAR-NMF model (1/3)

- MAP estimation amounts to the minimization of the following function

$$C(\mathbf{W}, \mathbf{H}, \mathbf{B}) = -\log p(\mathbf{H}, \mathbf{B}|\mathbf{V}; \mathbf{W}) \tag{38}$$
$$= -\log p(\mathbf{V}|\mathbf{H}; \mathbf{W}) - \log p(\mathbf{H}, \mathbf{B}) + \text{cst} \tag{39}$$

- We resort to a MM-based scheme. Only $-\log p(\mathbf{V}|\mathbf{H}; \mathbf{W})$ needs to be majorized. Standard scheme in the NMF literature [Lee and Seung, 2000, Févotte and Idier, 2011]
- For $\mathbf{H}$ and $\mathbf{B}$, leads to order-3 polynomial equations to solve
- We have to restrict ourselves to certain values of hyperparameters

MAP estimation in the BGAR-NMF model (2/3)

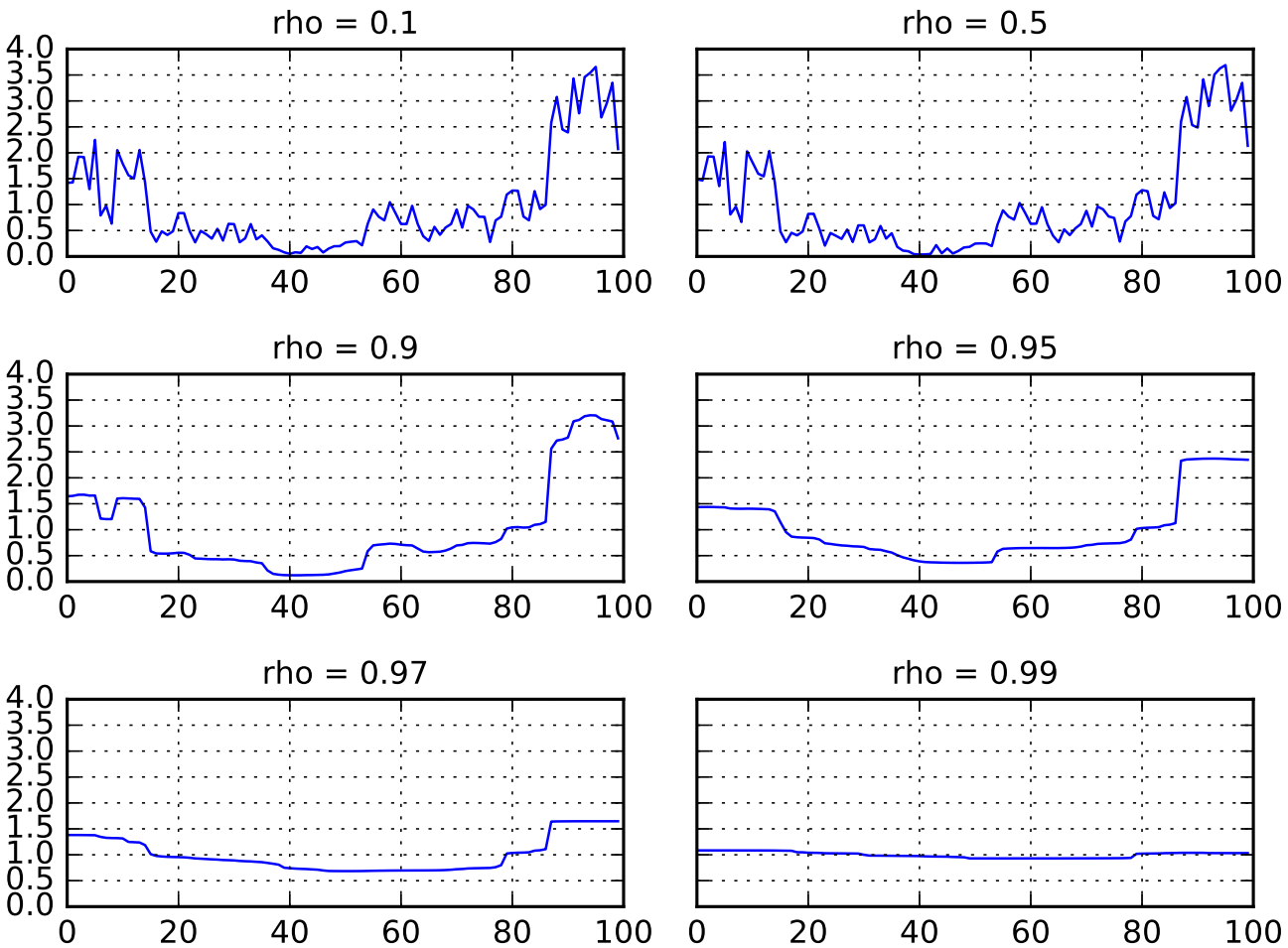


Figure 5: Evolution of $\underline{\mathbf{h}}$ w.r.t. ρ on a synthetic dataset

MAP estimation in BGAR-NMF model (3/3)

MAP estimation in BGAR-NMF model (3/3)

- Comparison of all the presented models in a MAP framework

MAP estimation in BGAR-NMF model (3/3)

- Comparison of all the presented models in a MAP framework
- Prediction problem on the `NIPS` dataset. 80/10/10 split

MAP estimation in BGAR-NMF model (3/3)

- Comparison of all the presented models in a MAP framework
- Prediction problem on the `NIPS` dataset. 80/10/10 split

Method	ℓ_1 error	ℓ_2 error	KL error
GaP	17.14 ± 0.47	4011 ± 446	267031 ± 9859
Rate	13.07 ± 0.58	4652 ± 1907	206574 ± 12839
Rate + Aux	9.34 ± 0.29	912 ± 225	$\mathbf{136412 \pm 6212}$
Shape	12.63 ± 0.3	1946 ± 240	192849 ± 5753
BGAR	$\mathbf{9.3 \pm 0.17}$	$\mathbf{839 \pm 119}$	138351 ± 6262

Table 4: Prediction results on the `NIPS` dataset

Outline

Conclusions

We tackled maximum marginal likelihood estimation in semi-Bayesian NMF models

Conclusions

We tackled maximum marginal likelihood estimation in semi-Bayesian NMF models

- Study of two particular instances : the GaP model and the IGCN model

Conclusions

We tackled maximum marginal likelihood estimation in semi-Bayesian NMF models

- Study of two particular instances : the GaP model and the IGCN model
  - Rewriting of the models free of $\mathbf{H}$, which led to an expression of the marginal likelihood

Conclusions

We tackled maximum marginal likelihood estimation in semi-Bayesian NMF models

- Study of two particular instances : the GaP model and the IGCN model
  - Rewriting of the models free of $\mathbf{H}$, which led to an expression of the marginal likelihood
  - The expression revealed a penalty term on $\mathbf{W}$

Conclusions

We tackled maximum marginal likelihood estimation in semi-Bayesian NMF models

- Study of two particular instances : the GaP model and the IGCN model
  - Rewriting of the models free of $\mathbf{H}$, which led to an expression of the marginal likelihood
  - The expression revealed a penalty term on $\mathbf{W}$
  - We tackled the optimization of the likelihood with (MC)-EM algorithms

Conclusions

We tackled maximum marginal likelihood estimation in semi-Bayesian NMF models

- Study of two particular instances : the GaP model and the IGCN model
  - Rewriting of the models free of $\mathbf{H}$, which led to an expression of the marginal likelihood
  - The expression revealed a penalty term on $\mathbf{W}$
  - We tackled the optimization of the likelihood with (MC)-EM algorithms
- Study of temporal Markovian NMF models

Conclusions

We tackled maximum marginal likelihood estimation in semi-Bayesian NMF models

- Study of two particular instances : the GaP model and the IGCN model
  - Rewriting of the models free of $\mathbf{H}$, which led to an expression of the marginal likelihood
  - The expression revealed a penalty term on $\mathbf{W}$
  - We tackled the optimization of the likelihood with (MC)-EM algorithms
- Study of temporal Markovian NMF models
  - Thorough review of the literature, which revealed that all considered Markov chains shared the same drawback

Conclusions

We tackled maximum marginal likelihood estimation in semi-Bayesian NMF models

- Study of two particular instances : the GaP model and the IGCN model
  - Rewriting of the models free of $\mathbf{H}$, which led to an expression of the marginal likelihood
  - The expression revealed a penalty term on $\mathbf{W}$
  - We tackled the optimization of the likelihood with (MC)-EM algorithms
- Study of temporal Markovian NMF models
  - Thorough review of the literature, which revealed that all considered Markov chains shared the same drawback
  - We proposed the use of an overlooked model from the time series literature, BGAR(1)

Conclusions

We tackled maximum marginal likelihood estimation in semi-Bayesian NMF models

- Study of two particular instances : the GaP model and the IGCN model
  - Rewriting of the models free of $\mathbf{H}$, which led to an expression of the marginal likelihood
  - The expression revealed a penalty term on $\mathbf{W}$
  - We tackled the optimization of the likelihood with (MC)-EM algorithms
- Study of temporal Markovian NMF models
  - Thorough review of the literature, which revealed that all considered Markov chains shared the same drawback
  - We proposed the use of an overlooked model from the time series literature, BGAR(1)
  - MMLE tackled with SMC : not satisfying

Conclusions

We tackled maximum marginal likelihood estimation in semi-Bayesian NMF models

- Study of two particular instances : the GaP model and the IGCN model
  - Rewriting of the models free of $\mathbf{H}$, which led to an expression of the marginal likelihood
  - The expression revealed a penalty term on $\mathbf{W}$
  - We tackled the optimization of the likelihood with (MC)-EM algorithms
- Study of temporal Markovian NMF models
  - Thorough review of the literature, which revealed that all considered Markov chains shared the same drawback
  - We proposed the use of an overlooked model from the time series literature, BGAR(1)
  - MMLE tackled with SMC : not satisfying
  - MAP estimation tackled with an MM-based algorithm showed better performance on prediction tasks

Perspectives

Perspectives

- Model aspects

Perspectives

- Model aspects
  - Break out of conjugate prior distributions

Perspectives

- Model aspects
  - Break out of conjugate prior distributions
  - Break out of composite models

Perspectives

- Model aspects
  - Break out of conjugate prior distributions
  - Break out of composite models
  - Carry out the analysis in a family of distributions

Perspectives

- Model aspects
  - Break out of conjugate prior distributions
  - Break out of composite models
  - Carry out the analysis in a family of distributions
  - Other Markov chains with stationary Gamma distribution

Perspectives

- Model aspects
  - Break out of conjugate prior distributions
  - Break out of composite models
  - Carry out the analysis in a family of distributions
  - Other Markov chains with stationary Gamma distribution
- Optimization aspects

Perspectives

- Model aspects
  - Break out of conjugate prior distributions
  - Break out of composite models
  - Carry out the analysis in a family of distributions
  - Other Markov chains with stationary Gamma distribution
- Optimization aspects
  - Break out of MC-EM : alternative schemes ?

Perspectives

- Model aspects
  - Break out of conjugate prior distributions
  - Break out of composite models
  - Carry out the analysis in a family of distributions
  - Other Markov chains with stationary Gamma distribution
- Optimization aspects
  - Break out of MC-EM : alternative schemes ?
  - Direct optimization of the likelihood

Associated publications

- Filstroff, L., Lumbreras, A., and Févotte, C. (2018).
  Closed-form Marginal Likelihood in Gamma-Poisson Matrix Factorization.
  In Proceedings of the International Conference of Machine Learning (ICML).
- Filstroff, L., and others (2019)
  Temporal Non-negative Matrix Factorization with Gamma Markov Chains.
  In preparation. IEEE Transactions on Signal Processing.
- Xia, R., Tan, V.Y.F., Filstroff, L., and Févotte, C. (2019).
  A Ranking Model Motivated by Nonnegative Matrix Factorization with Applications to Tennis Tournaments.
  In Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD).
- Lumbreras, A., Filstroff, L., and Févotte, C. (2018).
  Bayesian mean-parameterized nonnegative binary matrix factorization.
  In revision. Data Mining and Knowledge Discovery.

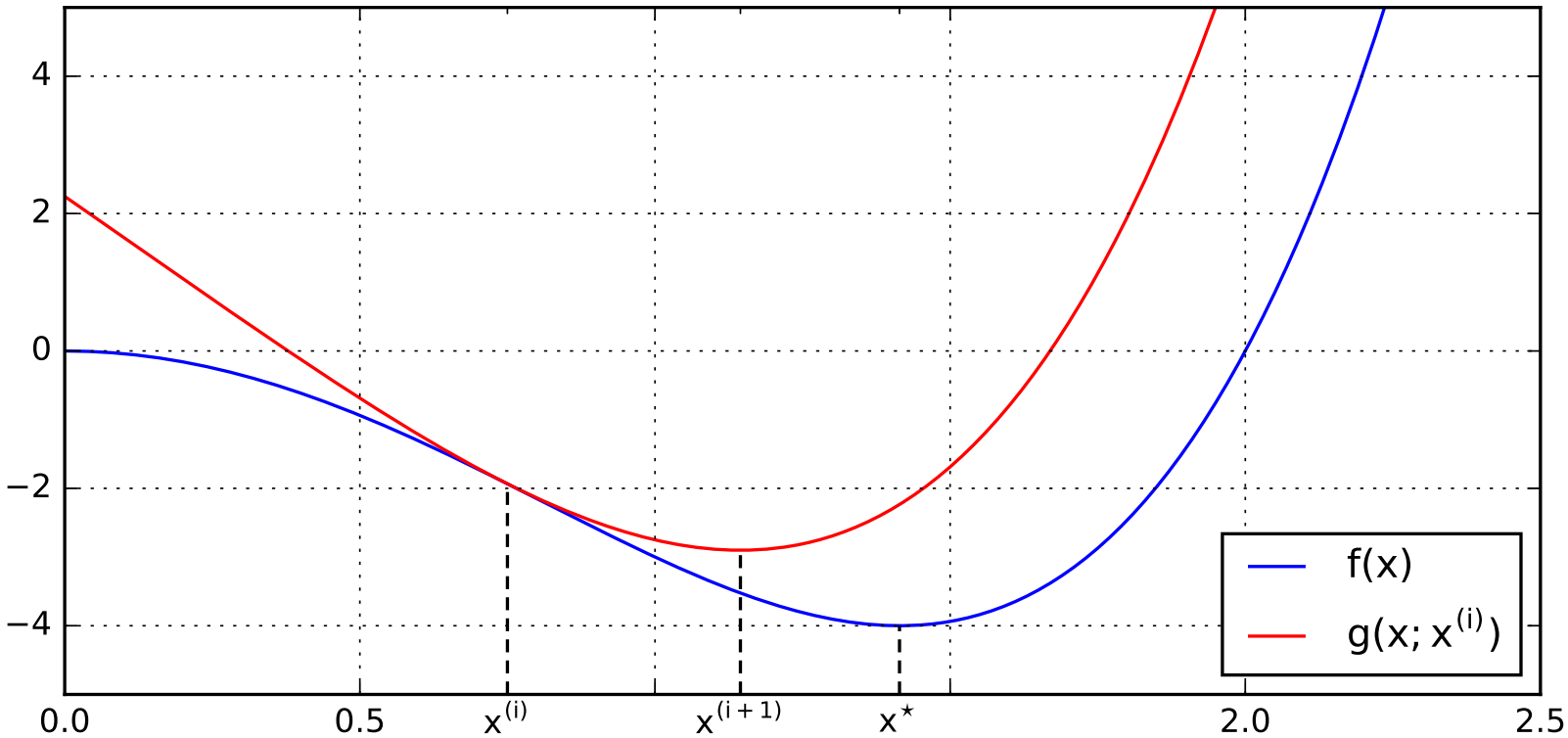
Thank you for your attention.

Additional slides

- The MM framework
- Variants of the EM algorithm
- Links between PF and LDA
- NB and NM distributions
- Examples of self-reg.
- The IGCN model
- MM algorithm for BGAR-NMF
- MAP estimation

The Majorization-Minimization (MM) framework

- Majorize the function f by an auxiliary function g
- Minimize g instead
- g is such that $g(x; \tilde{x}) \geq f(x)$ and $g(\tilde{x}; \tilde{x}) = f(\tilde{x})$



Go back

Variants of the EM algorithm

- MCEM : converges as $N_{iter} \to +\infty$ AND $J \to +\infty$ At iteration t, uses J samples for maximization
- SAEM [Delyon et al., 1999, Kuhn and Lavielle, 2004]
  Converges as $N_{iter} \to +\infty$

$$\hat{Q}_t(\boldsymbol{\theta}) = (1 - \gamma_t)\hat{Q}_{t-1}(\boldsymbol{\theta}) + \frac{\gamma_t}{J_t} \sum_j \log p(\mathbf{V}, \mathbf{Z}^{(j)}; \boldsymbol{\theta}) \qquad (40)$$

  $(\gamma_l)_{l \geq 1}$ is a sequence of positive step sizes decreasing to 0
  At iteration t, uses all past samples for maximization
- On-line EM [Cappé et al., 2005]
  Stochastic approximation "one sample at a time"

◂ Go back

Equivalence between Poisson Factorization and LDA

$$v_{fn} \sim \text{Poisson}([\mathbf{WH}]_{fn}) \tag{41}$$

is equivalent to

$$L_n \sim \text{Poisson}\left(\sum_f w_{fk} \sum_k h_{kn}\right) \tag{42}$$

$$\mathbf{v}_n | L_n \sim \text{Mult}\left(L_n, \boldsymbol{\lambda}_n\right) \tag{43}$$

with

$$\lambda_{fn} = \frac{\sum_k w_{fk} h_{kn}}{\sum_f w_{fk} \sum_k h_{kn}} \tag{44}$$

Imposing $\sum_f w_{fk} = 1$ and $\sum_k h_{kn} = 1$ leads to the observation model of LDA (difference in budget)

◂ Go back

NB and NM distributions

Negative Binomial (NB) distribution

With $\alpha > 0$ and $p \in [0, 1]$. For all $c \in \mathbb{N}$

$$\mathbb{P}(X = c) = \frac{\Gamma(\alpha + c)}{\Gamma(\alpha) c!}(1 - p)^{\alpha} p^{c} \tag{45}$$

$$\begin{cases} \lambda \sim \mathcal{G}(\alpha, \beta) \\ X|\lambda \sim \text{Poisson}(\lambda) \end{cases} \Leftrightarrow X \sim \text{NB}\left(\alpha, \frac{1}{\beta + 1}\right) \tag{46}$$

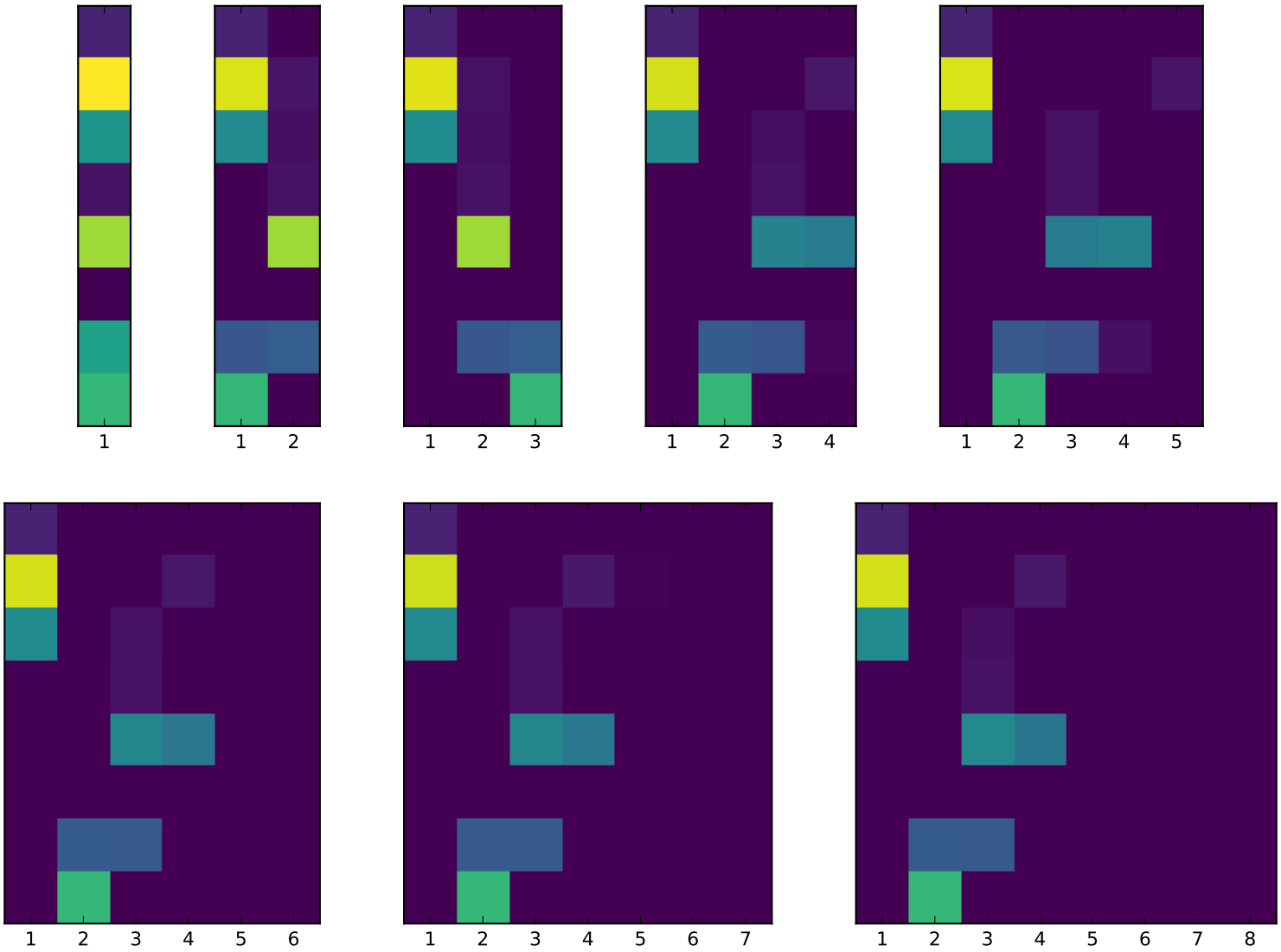
Negative Multinomial (NM) distribution

With $\alpha > 0$ and $p_f \in [0, 1]$ and $\sum_f p_f \leq 1$. For all $c_1, \ldots, c_F \in \mathbb{N}^F$

$$\mathbb{P}(X_1 = c_1, \ldots, X_n = c_F) = \frac{\Gamma(\alpha + \sum_f c_f)}{\Gamma(\alpha) \prod_f c_f!} p_0^{\alpha} \prod_f p_f^{c_f} \tag{47}$$

◂ Go back

Examples of self-regularization



◂ Go back

The IGCN model and objective

The IGCN model

$$h_{kn} \sim \mathcal{IG}(\alpha_k, \beta_k) \tag{48}$$

$$x_{fn}|\mathbf{h}_n \sim \mathcal{CN}(0, [\mathbf{WH}]_{fn}) \tag{49}$$

- $\mathcal{CN}$ denotes the complex normal distribution
- Standard STFT model in audio signal processing ("Gaussian composite model")
  [Févotte et al., 2009, Hoffman et al., 2010]
- Inverse Gamma prior for practical reasons : conjugacy with the normal distribution of known mean
- MMLE amounts to

$$\min_{\mathbf{W}} \mathcal{L}(\mathbf{W}) = -\log \int p(\mathbf{X}|\mathbf{H};\mathbf{W})p(\mathbf{H})\mathrm{d}\mathbf{H} \tag{50}$$

Go back

Marginalization of the activation coefficients

- Same augmentation with variables $\mathbf{C}$ as in the GaP model (superposition property of the normal distribution)
- IGCN can be written as a composite complex Student's t model
- Does not lead to a closed-form expression of the likelihood, because we end up with an intractable integral

$$p(\mathbf{X}; \mathbf{W}) = \int_{\mathbf{C} \in \mathcal{C}} p(\mathbf{C}; \mathbf{W}) \mathrm{d}\mathbf{C} = \int_{\mathbf{C} \in \mathcal{C}} \prod_{k,n} p(\mathbf{c}_{kn}; \mathbf{w}_k) \mathrm{d}\mathbf{C} \quad (51)$$

- Still exhibits a term of the form $\sum_{f,k} \log(w_{fk})$, which promotes "local" sparsity

Optimization of the marginal likelihood

- In [Dikmen and Févotte, 2011], the optimization was tackled with a variational algorithm
- We have proposed three novel EM algorithms based on three choices of the latent variables
- E-step based on a shared Gibbs sampling procedure
- M-step in closed form (EM-CH), or tackled with MM-based schemes (EM-H, EM-C)

Experimental results

- Real audio decomposition task on a short piano sequence
- Performance compared with the standard IS-NMF
- No obvious advantage in this case (similar audio accuracy, dictionary not especially sparse, computationally prohibitive)
- Conceptually interesting, but shows the limitations of the method

◂ Go back

MM : Constraints

In BGAR, we have

$$h_{kn} = b_{kn} h_{k(n-1)} + \epsilon_{kn} \tag{52}$$

$$h_{k(n+1)} = b_{k(n+1)} h_{kn} + \epsilon_{k(n+1)} \tag{53}$$

Leads to

$$b_{kn} h_{k(n-1)} \leq h_{kn} \leq \frac{h_{k(n+1)}}{b_{k(n+1)}} \tag{54}$$

and

$$0 \leq b_{kn} \leq \min\left(1, \frac{h_{kn}}{h_{k(n-1)}}\right) \tag{55}$$

◂ Go back

MM : Hyperparameter constraints

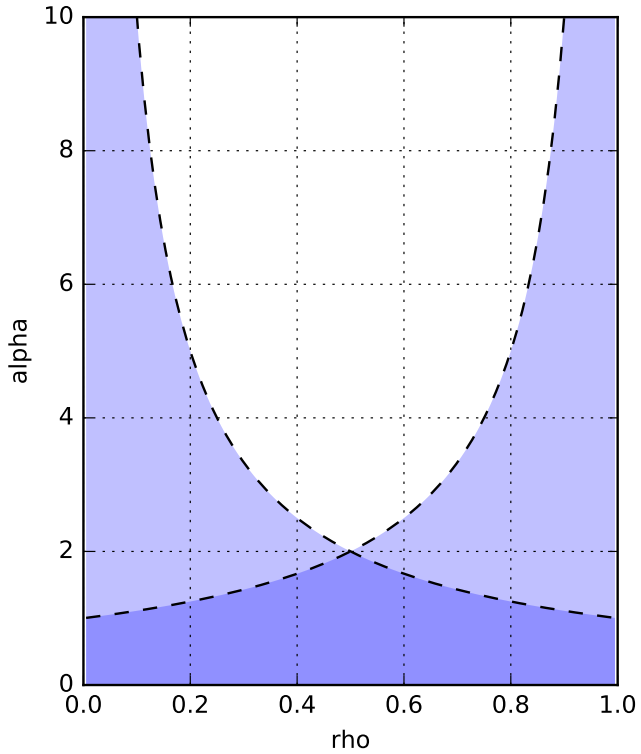


Figure 6: Hyperparameter values of the parameters α_k and ρ_k ensuring a well-posed MAP estimation in the BGAR-NMF model

◂ Go back

MAP Estimation

Objective function

$$C(\mathbf{W}, \mathbf{H}) = \underbrace{-\log p(\mathbf{V}|\mathbf{H}; \mathbf{W})}_{\text{Majorize}} - \log p(\mathbf{H}) \tag{56}$$

Standard majorization in the Poisson case

$$G(\mathbf{H}; \tilde{\mathbf{H}}) = -\sum_{k,n} p_{kn} \log h_{kn} + \sum_{k,n} q_k h_{kn} \tag{57}$$

- Rate : order-2 polynomials
- Rate + Aux : order-1 polynomials
- Shape : Newton's method
- BGAR : order-3 polynomials

Go back

References I

[Acharya et al., 2015] Acharya, A., Ghosh, J., and Zhou, M. (2015).
Nonparametric Bayesian Factor Analysis for Dynamic Count Matrices.
In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1–9.

[Berger et al., 1999] Berger, J. O., Liseo, B., and Wolpert, R. L. (1999).
Integrated Likelihood Methods for Eliminating Nuisance Parameters.
Statistical Science, 14(1):1–28.

[Bioucas-Dias et al., 2012] Bioucas-Dias, J. M., Plaza, A., Dobigeon, N., Parente, M., Du, Q., Gader, P., and Chanussot, J. (2012).
Hyperspectral Unmixing Overview: Geometrical, Statistical, and Sparse Regression-Based Approaches.
IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 5(2):354–379.

[Blei et al., 2003] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003).
Latent Dirichlet Allocation.
Journal of Machine Learning Research, 3(Jan):993–1022.

[Buntine and Jakulin, 2006] Buntine, W. and Jakulin, A. (2006).
Discrete component analysis.
In *Subspace, Latent Structure and Feature Selection*, pages 1–33. Springer.

References II

[Candès et al., 2008] Candès, E. J., Wakin, M. B., and Boyd, S. P. (2008).
Enhancing sparsity by reweighted $\ell 1$ minimization.
Journal of Fourier Analysis and Applications, 14(5-6):877–905.

[Canny, 2004] Canny, J. (2004).
GaP: A Factor Model for Discrete Data.
In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 122–129.

[Cappé et al., 2005] Cappé, O., Moulines, E., and Rydén, T. (2005).
Inference in Hidden Markov Models.
Springer.

[Cemgil, 2009] Cemgil, A. T. (2009).
Bayesian Inference for Nonnegative Matrix Factorisation Models.
Computational Intelligence and Neuroscience, (Article ID 785152).

[Cemgil and Dikmen, 2007] Cemgil, A. T. and Dikmen, O. (2007).
Conjugate Gamma Markov random fields for modelling nonstationary sources.
In *Proceedings of the International Conference on Independent Component Analysis and Signal Separation (ICA)*, pages 697–705.

[Delyon et al., 1999] Delyon, B., Lavielle, M., and Moulines, E. (1999).
Convergence of a stochastic approximation version of the em algorithm.
The Annals of Statistics, 27(1):94–128.

References III

[Dempster et al., 1977] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977).
Maximum likelihood from incomplete data via the EM algorithm.
Journal of the Royal Statistical Society, Series B (Methodological), 39(1):1–38.

[Dikmen and Févotte, 2011] Dikmen, O. and Févotte, C. (2011).
Nonnegative dictionary learning in the exponential noise model for adaptive music signal representation.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 2267–2275.

[Dikmen and Févotte, 2012] Dikmen, O. and Févotte, C. (2012).
Maximum marginal likelihood estimation for nonnegative dictionary learning in the Gamma-Poisson model.
IEEE Transactions on Signal Processing, 60(10):5163–5175.

[Févotte, 2011] Févotte, C. (2011).
Majorization-Minimization Algorithm for Smooth Itakura-Saito Nonnegative Matrix Factorization.
In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1980–1983.

[Févotte et al., 2009] Févotte, C., Bertin, N., and Durrieu, J.-L. (2009).
Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis.
Neural Computation, 21(3):793–830.

References IV

[Févotte and Idier, 2011] Févotte, C. and Idier, J. (2011).
Algorithms for nonnegative matrix factorization with the β-divergence.
Neural Computation, 23(9):2421–2456.

[Gopalan et al., 2015] Gopalan, P., Hofman, J. M., and Blei, D. M. (2015).
Scalable recommendation with hierarchical poisson factorization.
In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 326–335.

[Hoffman et al., 2010] Hoffman, M. D., Blei, D. M., and Cook, P. R. (2010).
Bayesian Nonparametric Matrix Factorization for Recorded Music.
In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 439–446.

[Hotelling, 1933] Hotelling, H. (1933).
Analysis of a complex of statistical variables into principal components.
Journal of Educational Psychology, 24(6):417–441.

[Kantas et al., 2015] Kantas, N., Doucet, A., Singh, S. S., Maciejowski, J., and Chopin, N. (2015).
On Particle Methods for Parameter Estimation in State-space Models.
Statistical Science, 30(3):328–351.

References V

[Kuhn and Lavielle, 2004] Kuhn, E. and Lavielle, M. (2004).
Coupling a stochastic approximation version of em with an mcmc procedure.
ESAIM: Probability and Statistics, 8:115–131.

[Lee and Seung, 1999] Lee, D. D. and Seung, H. S. (1999).
Learning the parts of objects by non-negative matrix factorization.
Nature, 401(6755):788–791.

[Lee and Seung, 2000] Lee, D. D. and Seung, H. S. (2000).
Algorithms for Non-negative Matrix Factorization.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 556–562.

[Lewis et al., 1989] Lewis, P., McKenzie, E., and Hugus, D. K. (1989).
Gamma processes.
Communications in Statistics. Stochastic Models, 5(1):1–30.

[Moulines et al., 1997] Moulines, E., Cardoso, J.-F., and Gassiat, E. (1997).
Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models.
In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3617–3620.

References VI

[Paatero and Tapper, 1994] Paatero, P. and Tapper, U. (1994).
Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values.
Environmetrics, 5(2):111–126.

[Pearson, 1901] Pearson, K. (1901).
On lines and planes of closest fit to systems of points in space.
The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2(11):559–572.

[Schein et al., 2016] Schein, A., Wallach, H. M., and Zhou, M. (2016).
Poisson-Gamma Dynamical Systems.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 5005–5013.

[Smaragdis and Brown, 2003] Smaragdis, P. and Brown, J. C. (2003).
Non-negative matrix factorization for polyphonic music transcription.
In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 177–180.

[Virtanen, 2007] Virtanen, T. (2007).
Monaural Sound Source Separation by NonnegativeMatrix Factorization With Temporal Continuityand Sparseness Criteria.
IEEE Transactions on Audio, Speech, and Language Processing, 15(3):1066–1074.

References VII

[Wei and Tanner, 1990] Wei, G. C. G. and Tanner, M. A. (1990).
A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms.
Journal of the American Statistical Association, 85(411):699–704.

[Xu et al., 2003] Xu, W., Liu, X., and Gong, Y. (2003).
Document Clustering Based On Non-negative MatrixFactorization.
In *Proceedings of the International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pages 267–273.