Maximum marginal likelihood estimation for nonnegative dictionary learning

Cédric Févotte

Laboratoire Lagrange, Nice

Xerox Research Center Europe
Grenoble, Jan. 2014

Work in collaboration with Onur Dikmen (Aalto University, Finland)

Outline

Nonnegative data decompositions

Probabilistic latent factor models for nonnegative data

Estimators

Experiments

Nonnegative dictionary learning

Dictionary learning: given data $\mathbf{V} = [\mathbf{v}_1 \dots \mathbf{v}_N]$, learn $\mathbf{W}$ such that

$\mathbf{v}_n$	$\approx$	$\mathbf{W}$	$\mathbf{h}_n$
data vector		"explanatory variables"	"regressors"
		"basis", "dictionary"	"expansion coefficients"
		"patterns"	"activation coefficients"

Nonnegative dictionary learning

Dictionary learning: given data $\mathbf{V} = [\mathbf{v}_1 \dots \mathbf{v}_N]$, learn $\mathbf{W}$ such that

$\mathbf{v}_n$	$\approx$	$\mathbf{W}$	$\mathbf{h}_n$
data vector		"explanatory variables"	"regressors"
		"basis", "dictionary"	"expansion coefficients"
		"patterns"	"activation coefficients"

Nonnegative dictionary learning: $\mathbf{V}$, $\mathbf{W}$ and $\mathbf{H}$ are nonnegative.

- nonneg. of $\mathbf{W}$ ensures *interpretability* of the dictionary (features $\mathbf{w}_k$ and data $\mathbf{v}_n$ belong to same space),
- nonneg. of $\mathbf{H}$ tends to produce *part-based* representations because subtractive combinations are forbidden.

Landmark paper in *Nature* by Lee and Seung (1999).

Nonnegative matrix factorization (NMF)

Obtain $\mathbf{V} \approx \mathbf{WH}$ by minimizing a loss function:

$$\min_{\mathbf{W},\mathbf{H} \geq \mathbf{0}} D(\mathbf{V}|\mathbf{WH}) = \sum_{fn} d([\mathbf{V}]_{fn}|[\mathbf{WH}]_{fn}),$$

where $d(x|y)$ is a scalar measure of fit.

Nonnegative matrix factorization (NMF)

Obtain $\mathbf{V} \approx \mathbf{WH}$ by minimizing a loss function:

$$\min_{\mathbf{W},\mathbf{H} \geq \mathbf{0}} D(\mathbf{V}|\mathbf{WH}) = \sum_{fn} d([\mathbf{V}]_{fn}|[\mathbf{WH}]_{fn}),$$

where $d(x|y)$ is a scalar measure of fit.

Regularization terms are often added to $D(\mathbf{V}|\mathbf{WH})$ to favor certain properties of $\mathbf{W}$ or $\mathbf{H}$ (sparsity, smoothness).

Nonnegative matrix factorization (NMF)

Obtain $\mathbf{V} \approx \mathbf{WH}$ by minimizing a loss function:

$$\min_{\mathbf{W},\mathbf{H} \geq \mathbf{0}} D(\mathbf{V}|\mathbf{WH}) = \sum_{fn} d([\mathbf{V}]_{fn}|[\mathbf{WH}]_{fn}),$$

where $d(x|y)$ is a scalar measure of fit.

Regularization terms are often added to $D(\mathbf{V}|\mathbf{WH})$ to favor certain properties of $\mathbf{W}$ or $\mathbf{H}$ (sparsity, smoothness).

Majorization-minimization algorithms can be derived for large class of loss functions (Euclidean distance, Kullback-Leibler divergence, Itakura-Saito divergence, α-divergences, β-divergences) (Yang and Oja, 2011; Févotte and Idier, 2011)

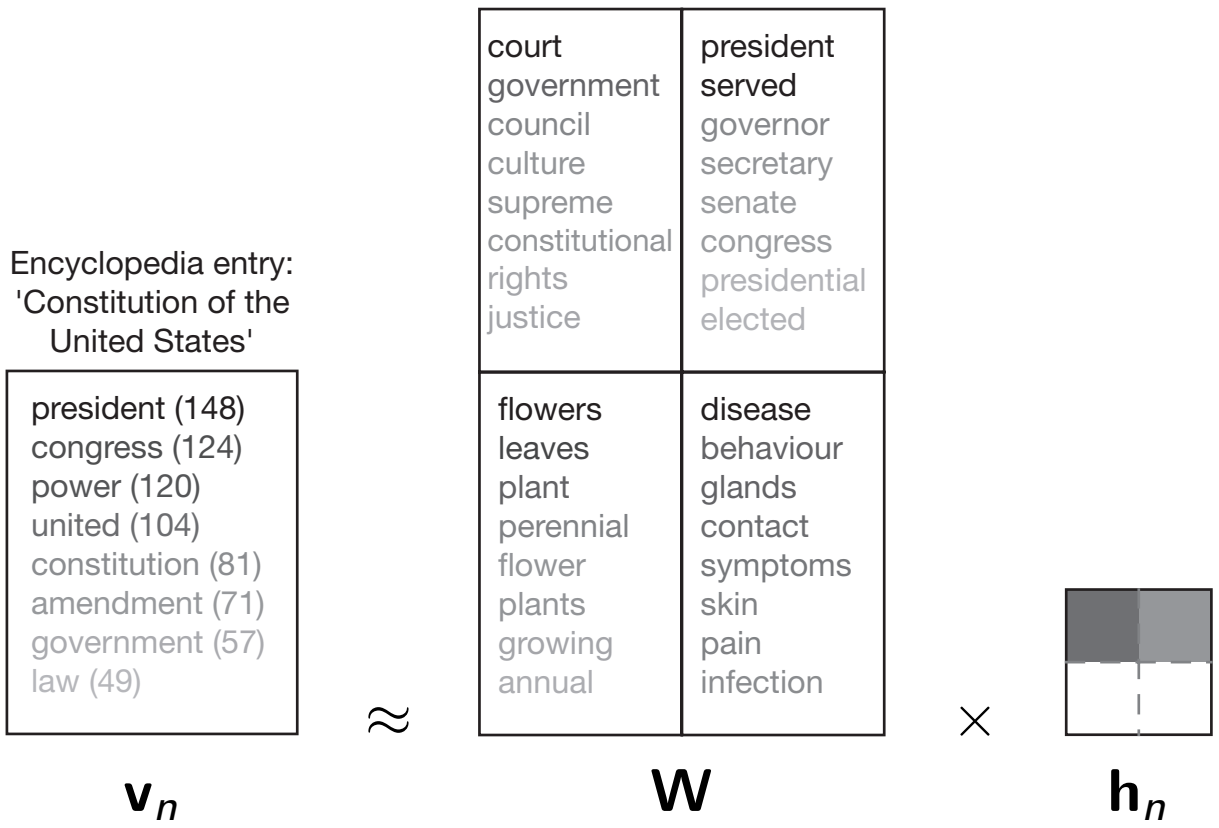
Some applications of nonnegative decompositions

- text analysis (see next)
- music signal processing (see next)
- environmetrics (Paatero and Tapper, 1994)
- video summarization (Cooper and Foote, 2002)
- gene expression analysis (Brunet et al., 2004)
- Scotch whiskies clustering (Young et al., 2006) (!)
- hyperspectral imaging (Berry et al., 2007)
- portfolio diversification (Drakakis et al., 2007)
- clustering of protein interactions (Greene et al., 2008)
- food consumption analysis (Zetlaoui et al., 2010)
- image denoising and inpainting (Mairal et al., 2010)

(selected references)

Text analysis

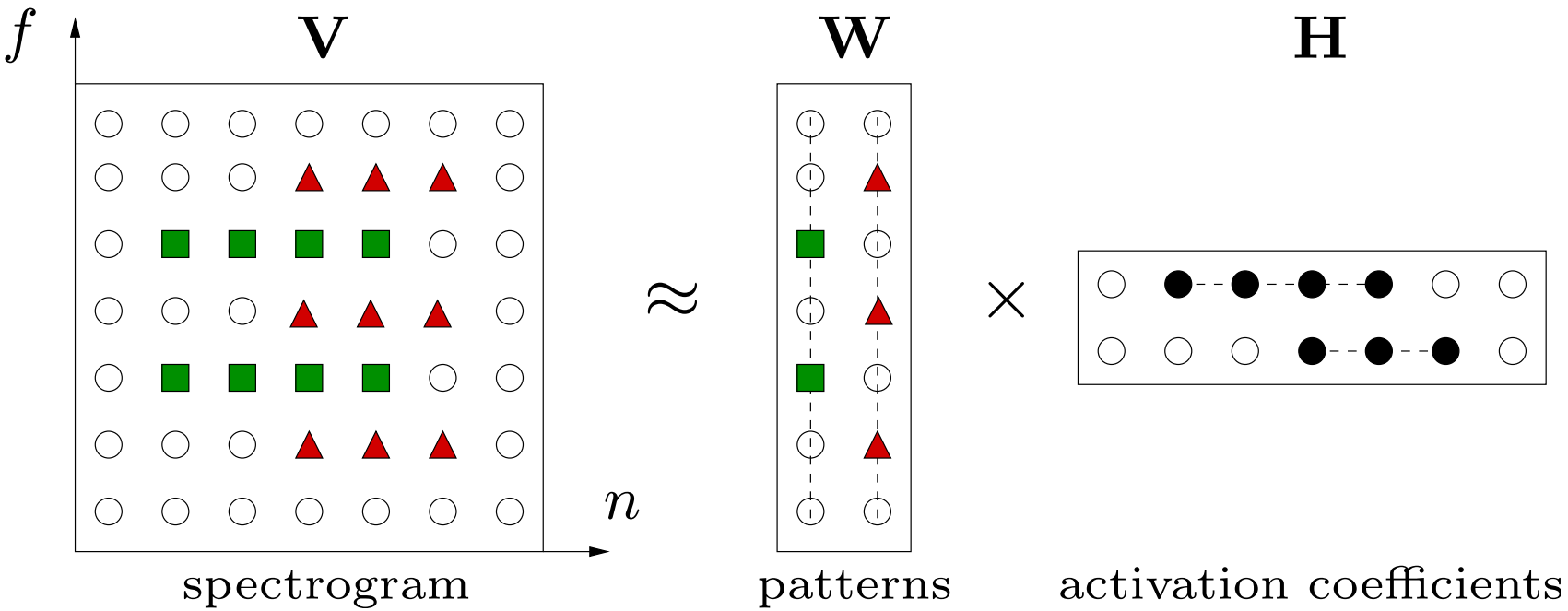
(Lee and Seung, 1999; Hofmann, 1999; Blei et al., 2003; Gaussier and Goutte, 2005; Buntine and Jakulin, 2006)



reproduced from Lee and Seung (1999)

Music signal processing

(Smaragdis and Brown, 2003; Virtanen, 2007; Févotte et al., 2009)



Outline

Nonnegative data decompositions

Probabilistic latent factor models for nonnegative data

Estimators

Experiments

Why probabilistic models ?

NMF is intrinsically deterministic.

- how to choose the loss function $D(\mathbf{V}|\mathbf{WH})$?
- do we have statistical guarantees about the estimates of $\mathbf{W}$ and $\mathbf{H}$ returned by NMF ?
- how to choose the rank of the factorization ?

The probabilistic setting may bring answers to these questions.

Review of probabilistic nonnegative factor models

Featuring the additive Gaussian, Poisson, multinomial, multiplicative Gamma and Tweedie models.

- probabilistic model $\mathbf{V} \sim p(\mathbf{V}|\mathbf{WH})$
- conditional independence of the observations

$$p(\mathbf{V}|\mathbf{WH}) = \prod_n p(\mathbf{v}_n|\mathbf{W}\mathbf{h}_n)$$

- conditional independence of the features (except for the multinomial model)

$$p(\mathbf{v}_n|\mathbf{W}\mathbf{h}_n) = \prod_f p(v_{fn}|[\mathbf{WH}]_{fn})$$

- linear conditional expectation

$$\mathrm{E}[\mathbf{V}|\mathbf{WH}] = \mathbf{WH}$$

Additive Gaussian model

(Schmidt et al., 2009; Zhong and Girolami, 2009)

Generative model:

$$v_{fn} = [\mathbf{WH}]_{fn} + \epsilon_{fn}$$
$$\epsilon_{fn} \sim N(0, \sigma^2)$$

Anti log-likelihood:

$$-\log p(\mathbf{V}|\mathbf{WH}) = \frac{1}{\sigma^2} D_{EUC}(\mathbf{V}|\mathbf{WH}) + cst$$

with $D_{EUC}(\mathbf{X}|\mathbf{Y}) = \|\mathbf{X} - \mathbf{Y}\|_F^2$.

Ill-posed model for nonnegative data as it may generate negative values in large variance settings.

Poisson model

(Canny, 2004; Buntine and Jakulin, 2006; Cemgil, 2009)

Generative model:

$$v_{fn} \sim Pois([\mathbf{WH}]_{fn})$$

Domain: $v_{fn} \in \mathbb{N}$
Anti log-likelihood:

$$-\log p(\mathbf{V}|\mathbf{WH}) = D_{GKL}(\mathbf{V}|\mathbf{WH}) + cst$$

where $D_{GKL}(\mathbf{X}|\mathbf{Y}) = \sum_{ij} x_{ij} \log \frac{x_{ij}}{y_{ij}} - x_{ij} + y_{ij}$ is the generalized Kullback-Leibler divergence.

Application: relevant model for counts, long history in photon tomography, text analysis.

Multinomial model

(Hofmann, 1999; Blei et al., 2003)

Generative model

$$\mathbf{v}_n \sim Mult(\sum_f v_{fn}, \mathbf{W}\mathbf{h}_n)$$

where the columns of $\mathbf{W}$ and $\mathbf{h}_n$ sum to 1.
Domain: $v_{fn} \in \mathbb{N}$
Anti log-likelihood:

$$-\log p(\mathbf{V}|\mathbf{WH}) = \sum_n \|\mathbf{v}_n\|_1 D_{KL}(\bar{\mathbf{v}}_n|\mathbf{W}\mathbf{h}_n) + cst$$

where $\bar{\mathbf{v}}_n$ is the normalized data and $D_{KL}(\mathbf{x}|\mathbf{y}) = \sum_i x_i \log \frac{x_i}{y_i}$ is the Kullback-Leibler divergence between normalized vectors.

Application: relevant model for counts, popular in text analysis.

Multiplicative Gamma model

(Févotte, Bertin, and Durrieu, 2009)

Generative model:

$$v_{fn} = [\mathbf{WH}]_{fn} \cdot \epsilon_{fn}$$
$$\epsilon_{fn} \sim G(\alpha, \alpha) \quad (= \text{Gamma distribution with expectation 1})$$

Domain: $v_{fn} \in \mathbb{R}^+$

Anti log-likelihood:

$$-\log p(\mathbf{V}|\mathbf{WH}) = \alpha\, D_{IS}(\mathbf{V}|\mathbf{WH}) + cst$$

where $D_{IS}(\mathbf{X}|\mathbf{Y}) = \sum_{ij} \frac{x_{ij}}{y_{ij}} - \log \frac{x_{ij}}{y_{ij}} - 1$ is the Itakura-Saito divergence.

Application: decomposition of spectrograms.

Tweedie model

(Yilmaz, 2012; Tan and Févotte, 2013)

Additive Gaussian, Poisson and multiplicative Gamma models are special cases of

$$v_{fn} \sim T([\mathbf{WH}]_{fn}, \phi, \beta)$$

where $T(\mu, \phi, \beta)$ refers the Tweedie distribution (Tweedie, 1984; Jørgensen, 1987) defined by

$$T(x|\mu, \phi, \beta) = h(x, \phi) \exp\left[\frac{1}{\phi}\left(\frac{1}{\beta - 1} x \mu^{\beta-1} - \frac{1}{\beta}\mu^{\beta}\right)\right]$$

with expectation μ, dispersion ϕ and shape β.

Tweedie model

(Yilmaz, 2012; Tan and Févotte, 2013)

Additive Gaussian, Poisson and multiplicative Gamma models are special cases of

$$v_{fn} \sim T([\mathbf{WH}]_{fn}, \phi, \beta)$$

where $T(\mu, \phi, \beta)$ refers the Tweedie distribution (Tweedie, 1984; Jørgensen, 1987) defined by

$$T(x|\mu, \phi, \beta) = h(x, \phi) \exp\left[\frac{1}{\phi}\left(\frac{1}{\beta - 1} x \mu^{\beta-1} - \frac{1}{\beta}\mu^{\beta}\right)\right]$$

with expectation μ, dispersion ϕ and shape β.

Underlies the β-divergence $D_\beta(\mathbf{V}|\mathbf{WH})$, a common divergence in NMF, see, e.g., (Févotte and Idier, 2011).

Outline

Nonnegative data decompositions

Probabilistic latent factor models for nonnegative data

Estimators

Experiments

Maximum likelihood estimation

At this stage we are equipped with a probabilistic model

$$p(\mathbf{V}|\mathbf{WH})$$

and we established a correspondence between NMF and maximum likelihood on $\mathbf{W}$ and $\mathbf{H}$

$$-\log p(\mathbf{V}|\mathbf{WH}) \iff D(\mathbf{V}|\mathbf{WH})$$

(with possibly restrictions on the domain of $\mathbf{V}$)

Maximum likelihood estimation

At this stage we are equipped with a probabilistic model

$$p(\mathbf{V}|\mathbf{WH})$$

and we established a correspondence between NMF and maximum likelihood on $\mathbf{W}$ and $\mathbf{H}$

$$-\log p(\mathbf{V}|\mathbf{WH}) \iff D(\mathbf{V}|\mathbf{WH})$$

(with possibly restrictions on the domain of $\mathbf{V}$)

⚠Statistical optimality of ML is in question because the number of parameters grows with the number of data points (because of "nuisance" parameter $\mathbf{H}$).

Bayesian inference

Set priors $p(\mathbf{W})$, $p(\mathbf{H})$ and characterize the posterior $p(\mathbf{W}, \mathbf{H}|\mathbf{V})$.

Bayesian inference

Set priors $p(\mathbf{W})$, $p(\mathbf{H})$ and characterize the posterior $p(\mathbf{W}, \mathbf{H}|\mathbf{V})$.

Maximum a posteriori (penalized NMF)

$$\min_{\mathbf{W},\mathbf{H}\geq 0} \; -\log p(\mathbf{V}|\mathbf{WH}) - \log p(\mathbf{W}) - \log p(\mathbf{H}) + cst$$

Bayesian inference

Set priors $p(\mathbf{W})$, $p(\mathbf{H})$ and characterize the posterior $p(\mathbf{W}, \mathbf{H}|\mathbf{V})$.

Maximum a posteriori (penalized NMF)

$$\min_{\mathbf{W},\mathbf{H}\geq 0} -\log p(\mathbf{V}|\mathbf{WH}) - \log p(\mathbf{W}) - \log p(\mathbf{H}) + cst$$

Monte-Carlo approaches, e.g., (Buntine and Jakulin, 2006; Schmidt et al., 2009; Zhong and Girolami, 2009)

$$\mathbf{W}^{(i)}, \mathbf{H}^{(i)} \sim p(\mathbf{W}, \mathbf{H}|\mathbf{V})$$

Bayesian inference

Set priors $p(\mathbf{W})$, $p(\mathbf{H})$ and characterize the posterior $p(\mathbf{W}, \mathbf{H}|\mathbf{V})$.

Maximum a posteriori (penalized NMF)

$$\min_{\mathbf{W},\mathbf{H}\geq 0} -\log p(\mathbf{V}|\mathbf{WH}) - \log p(\mathbf{W}) - \log p(\mathbf{H}) + cst$$

Monte-Carlo approaches, e.g., (Buntine and Jakulin, 2006; Schmidt et al., 2009; Zhong and Girolami, 2009)

$$\mathbf{W}^{(i)}, \mathbf{H}^{(i)} \sim p(\mathbf{W}, \mathbf{H}|\mathbf{V})$$

Variational approaches, e.g., (Blei et al., 2003; Buntine and Jakulin, 2006; Hoffman et al., 2010; Seeger and Bouchard, 2012)

$$p(\mathbf{W}, \mathbf{H}|\mathbf{V}) \approx q(\mathbf{W}, \mathbf{H})$$

Maximum *marginal* likelihood estimation

Defined as of (Dikmen and Févotte, 2011, 2012)

- treat $\mathbf{W}$ as a *deterministic* variable.

Maximum *marginal* likelihood estimation

Defined as of (Dikmen and Févotte, 2011, 2012)

- treat $\mathbf{W}$ as a *deterministic* variable.
- treat $\mathbf{H}$ as a *random* latent variable with prior $p(\mathbf{H})$.

Maximum *marginal* likelihood estimation

Defined as of (Dikmen and Févotte, 2011, 2012)

- treat $\mathbf{W}$ as a *deterministic* variable.
- treat $\mathbf{H}$ as a *random* latent variable with prior $p(\mathbf{H})$.
- optimize the *marginal* likelihood of $\mathbf{V}$ and $\mathbf{W}$:

$$\min_{\mathbf{W} \geq 0} -\log p(\mathbf{V}|\mathbf{W}) = -\log \int_{\mathbf{H}} p(\mathbf{V}|\mathbf{W}\mathbf{H}) p(\mathbf{H}) d\mathbf{H}.$$

Maximum *marginal* likelihood estimation

Defined as of (Dikmen and Févotte, 2011, 2012)

- treat $\mathbf{W}$ as a *deterministic* variable.
- treat $\mathbf{H}$ as a *random* latent variable with prior $p(\mathbf{H})$.
- optimize the *marginal* likelihood of $\mathbf{V}$ and $\mathbf{W}$:

$$\min_{\mathbf{W} \geq 0} -\log p(\mathbf{V}|\mathbf{W}) = -\log \int_{\mathbf{H}} p(\mathbf{V}|\mathbf{W}\mathbf{H}) p(\mathbf{H}) d\mathbf{H}.$$

+ no need to set a prior on $\mathbf{W}$
+ better posed than NMF (fixed number of parameters)
+ better-behaved with respect to scales
+ self-regularization of the rank

☹ involves complex optimization and integration problems

Maximum *marginal* likelihood estimation

Origins

General idea can be tracked back to

- statistics literature about estimation with nuisance parameters
- **independent component analysis**
  E.g., additive Gaussian model and Laplacian activations in Lewicki and Sejnowski (2000)
- **latent dirichlet allocation** (LDA) (Blei et al., 2003)
  multinomial model and Dirichlet activations
- **discrete component analysis** (DCA) (Buntine and Jakulin, 2006)
  Poisson model and Gamma activations

Maximum *marginal* likelihood estimation

EM algorithms

Generic EM: complete data $\mathbf{V}$ with $\mathbf{H}$ and optimize

$$Q(\mathbf{W}|\tilde{\mathbf{W}}) = -\int_{\mathbf{H}} \log p(\mathbf{V},\mathbf{H}|\mathbf{W}) p(\mathbf{H}|\mathbf{V},\tilde{\mathbf{W}}) d\mathbf{H}$$

$p(\mathbf{H}|\mathbf{V},\tilde{\mathbf{W}})$ is not available in most models.

Maximum *marginal* likelihood estimation

EM algorithms

Generic EM: complete data $\mathbf{V}$ with $\mathbf{H}$ and optimize

$$Q(\mathbf{W}|\tilde{\mathbf{W}}) = -\int_{\mathbf{H}} \log p(\mathbf{V},\mathbf{H}|\mathbf{W}) p(\mathbf{H}|\mathbf{V},\tilde{\mathbf{W}}) d\mathbf{H}$$

$p(\mathbf{H}|\mathbf{V},\tilde{\mathbf{W}})$ is not available in most models.

Resort to

- variational EM: $q(\mathbf{H}) \approx p(\mathbf{H}|\mathbf{V},\tilde{\mathbf{W}})$

$$Q^{\text{VB}}(\mathbf{W}|\tilde{\mathbf{W}}) = -\int_{\mathbf{H}} \log p(\mathbf{V},\mathbf{H}|\mathbf{W}) q(\mathbf{H}) d\mathbf{H}$$

Maximum *marginal* likelihood estimation

EM algorithms

Generic EM: complete data $\mathbf{V}$ with $\mathbf{H}$ and optimize

$$Q(\mathbf{W}|\tilde{\mathbf{W}}) = -\int_{\mathbf{H}} \log p(\mathbf{V},\mathbf{H}|\mathbf{W}) p(\mathbf{H}|\mathbf{V},\tilde{\mathbf{W}}) d\mathbf{H}$$

$p(\mathbf{H}|\mathbf{V},\tilde{\mathbf{W}})$ is not available in most models.

Resort to

- variational EM: $q(\mathbf{H}) \approx p(\mathbf{H}|\mathbf{V},\tilde{\mathbf{W}})$

$$Q^{\mathrm{VB}}(\mathbf{W}|\tilde{\mathbf{W}}) = -\int_{\mathbf{H}} \log p(\mathbf{V},\mathbf{H}|\mathbf{W}) q(\mathbf{H}) d\mathbf{H}$$

- Monte-Carlo EM: $\mathbf{H}^{(i)} \sim p(\mathbf{H}|\mathbf{V},\tilde{\mathbf{H}})$

$$Q^{\mathrm{MC}}(\mathbf{W}|\tilde{\mathbf{W}}) = -\sum_i \log p(\mathbf{V},\mathbf{H}^{(i)}|\mathbf{W})$$

Outline

Nonnegative data decompositions

Probabilistic latent factor models for nonnegative data

Estimators

Experiments

Two models considered

	Gamma-Poisson	Gamma-Exponential
$p(\mathbf{V}\|\mathbf{WH})$	$\prod_{fn} Pois(v_{fn}\|[\mathbf{WH}]_{fn})$	$\prod_{fn} Exp(v_{fn}\|[\mathbf{WH}]_{fn})$ (mult. Gamma model with $\alpha = 1$)
$p(\mathbf{H}\|\beta)$	$\prod_{kn} G(h_{kn}\|\alpha_k, \beta_k)$	
Data	Word counts	Spectrogram
Reference	(Dikmen and Févotte, 2012)	(Dikmen and Févotte, 2011)

MMLE vs MJLE

Comparison of two estimators of **W** (and **H**)

Maximum joint likelihood estimation (MJLE)

$$\begin{aligned} C_{JL}(\mathbf{W}, \mathbf{H}, \beta) &= -\log p(\mathbf{V}, \mathbf{H}|\mathbf{W}, \beta) \\ &= -\log p(\mathbf{V}|\mathbf{WH}) - \log p(\mathbf{H}|\beta) \end{aligned}$$

Optimization with majorization-minimization.
Equivalent to penalized NMF.

MMLE vs MJLE

Comparison of two estimators of $\mathbf{W}$ (and $\mathbf{H}$)

Maximum joint likelihood estimation (MJLE)

$$\begin{aligned} C_{JL}(\mathbf{W},\mathbf{H},\boldsymbol{\beta}) &= -\log p(\mathbf{V},\mathbf{H}|\mathbf{W},\boldsymbol{\beta}) \\ &= -\log p(\mathbf{V}|\mathbf{WH}) - \log p(\mathbf{H}|\boldsymbol{\beta}) \end{aligned}$$

Optimization with majorization-minimization.
Equivalent to penalized NMF.

Maximum marginal likelihood estimation (MMLE)

$$\begin{aligned} C_{ML}(\mathbf{W},\boldsymbol{\beta}) &= -\log p(\mathbf{V}|\mathbf{W},\boldsymbol{\beta}) \\ &= -\log \int_{\mathbf{H}} p(\mathbf{V}|\mathbf{WH}) p(\mathbf{H}|\boldsymbol{\beta}) d\mathbf{H} \end{aligned}$$

Optimization with variational EM or MC-EM.
Estimation of $\mathbf{H}$ given $\hat{\mathbf{W}}$ in a second step by MAP.

MMLE vs MJLE

Scales

Let $\mathbf{\Lambda}$ be a nonnegative diagonal matrix.

MMLE is scale-invariant

$$C_{ML}(\mathbf{W}\mathbf{\Lambda}^{-1}, \mathbf{\Lambda}\boldsymbol{\beta}) = C_{ML}(\mathbf{W}, \boldsymbol{\beta})$$

We may set $\beta_k = 1$ and let $\mathbf{W}$ free.

MMLE vs MJLE

Scales

Let $\mathbf{\Lambda}$ be a nonnegative diagonal matrix.

MMLE is scale-invariant

$$C_{ML}(\mathbf{W}\mathbf{\Lambda}^{-1}, \mathbf{\Lambda}\boldsymbol{\beta}) = C_{ML}(\mathbf{W}, \boldsymbol{\beta})$$

We may set $\beta_k = 1$ and let $\mathbf{W}$ free.

MJLE is not

$$C_{JL}(\mathbf{W}\mathbf{\Lambda}^{-1}, \mathbf{\Lambda}\mathbf{H}, \mathbf{\Lambda}\boldsymbol{\beta}) = C_{JL}(\mathbf{W}, \mathbf{H}, \boldsymbol{\beta}) + N \sum_k \log \lambda_k$$

$\Rightarrow$ degenerate solutions $\|\mathbf{W}\| \to \infty$, $\|\mathbf{H}\| \to 0$, $\|\boldsymbol{\beta}\| \to 0$.
$\Rightarrow$ the norm of $\mathbf{W}$ needs to be controlled.

Synthetical data

Gamma-Poisson model

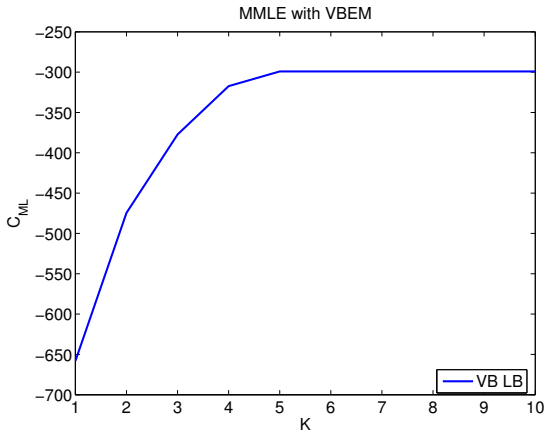
- True matrix $\mathbf{W}^\star$ of size 10×5 composed of zeros and tens.
- $\mathbf{H}^\star \sim Exp(1)$
- $\mathbf{V} \sim Pois(\mathbf{WH})$ with $N = 50$.

Synthetical data

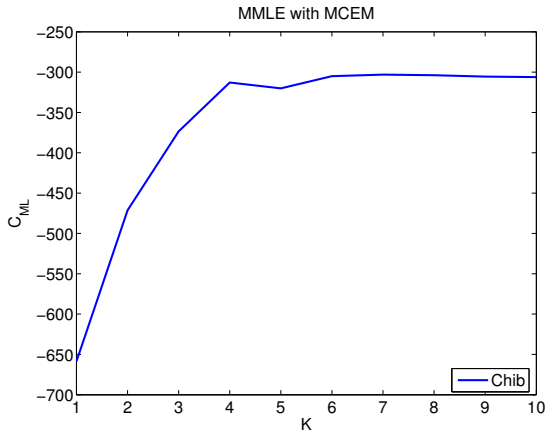
Gamma-Poisson model

- True matrix $\mathbf{W}^\star$ of size 10×5 composed of zeros and tens.
- $\mathbf{H}^\star \sim Exp(1)$
- $\mathbf{V} \sim Pois(\mathbf{WH})$ with $N = 50$.

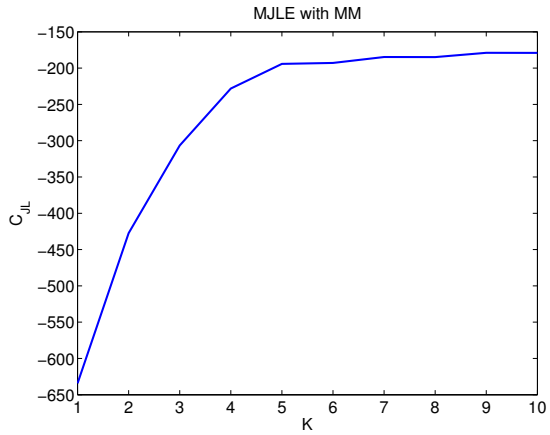
MMLE with VB-EM



MMLE with MC-EM



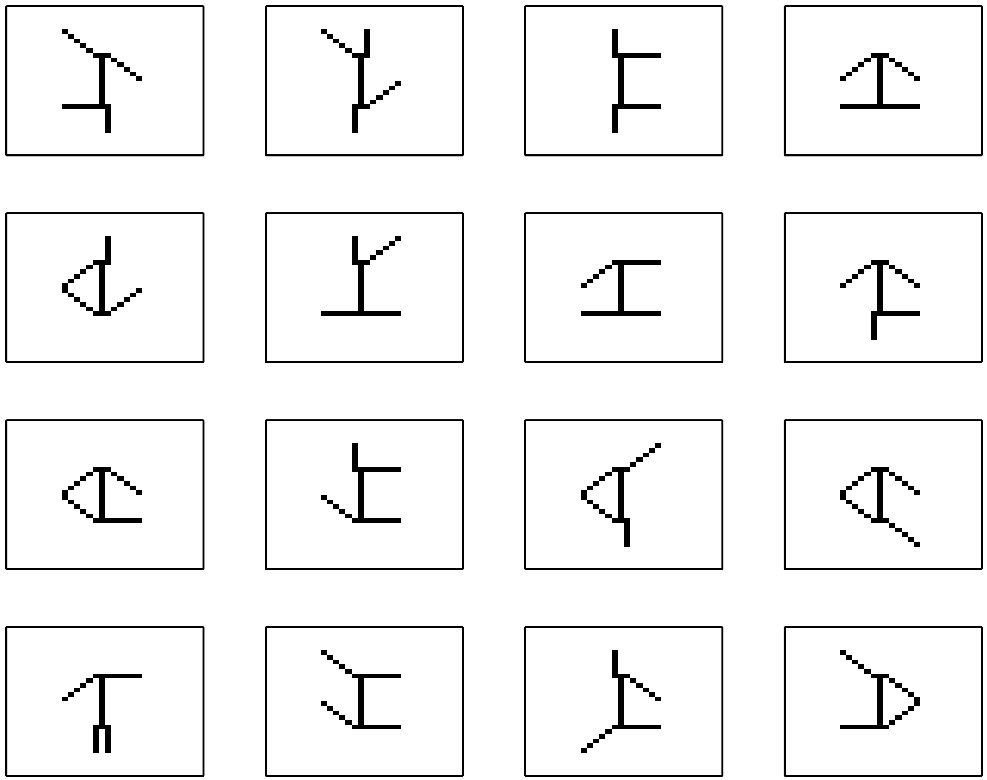
MJLE with MM



Synthetical data

Gamma-Exponential model

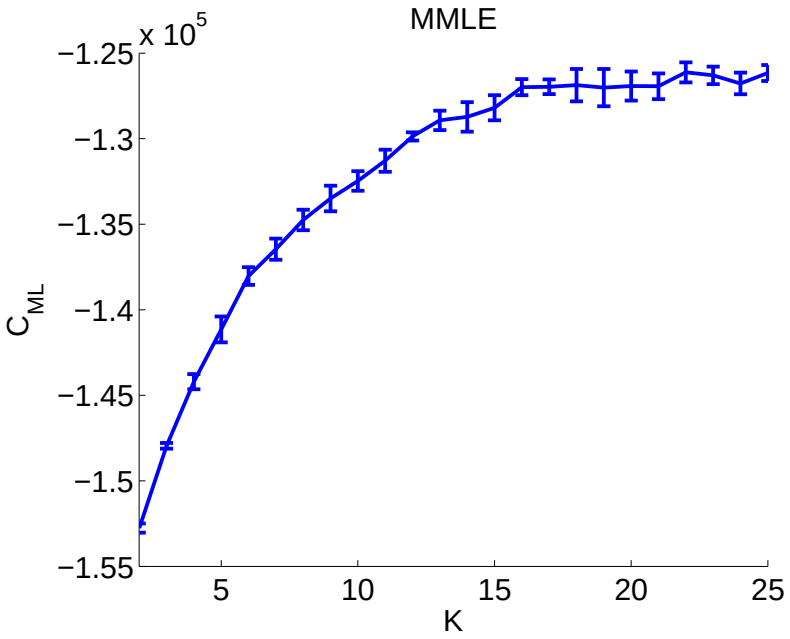
Swimmer dataset corrupted by multiplicative exponential noise.

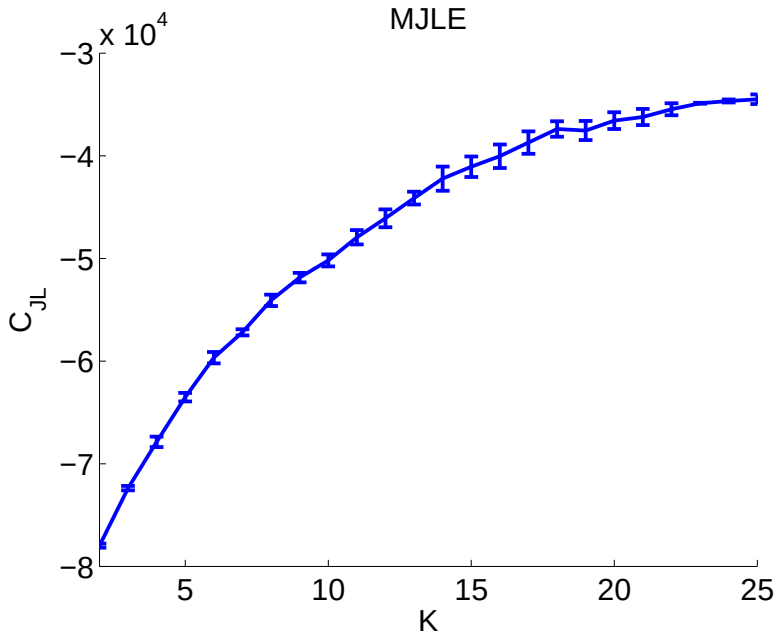


(clean samples)

Synthetical data

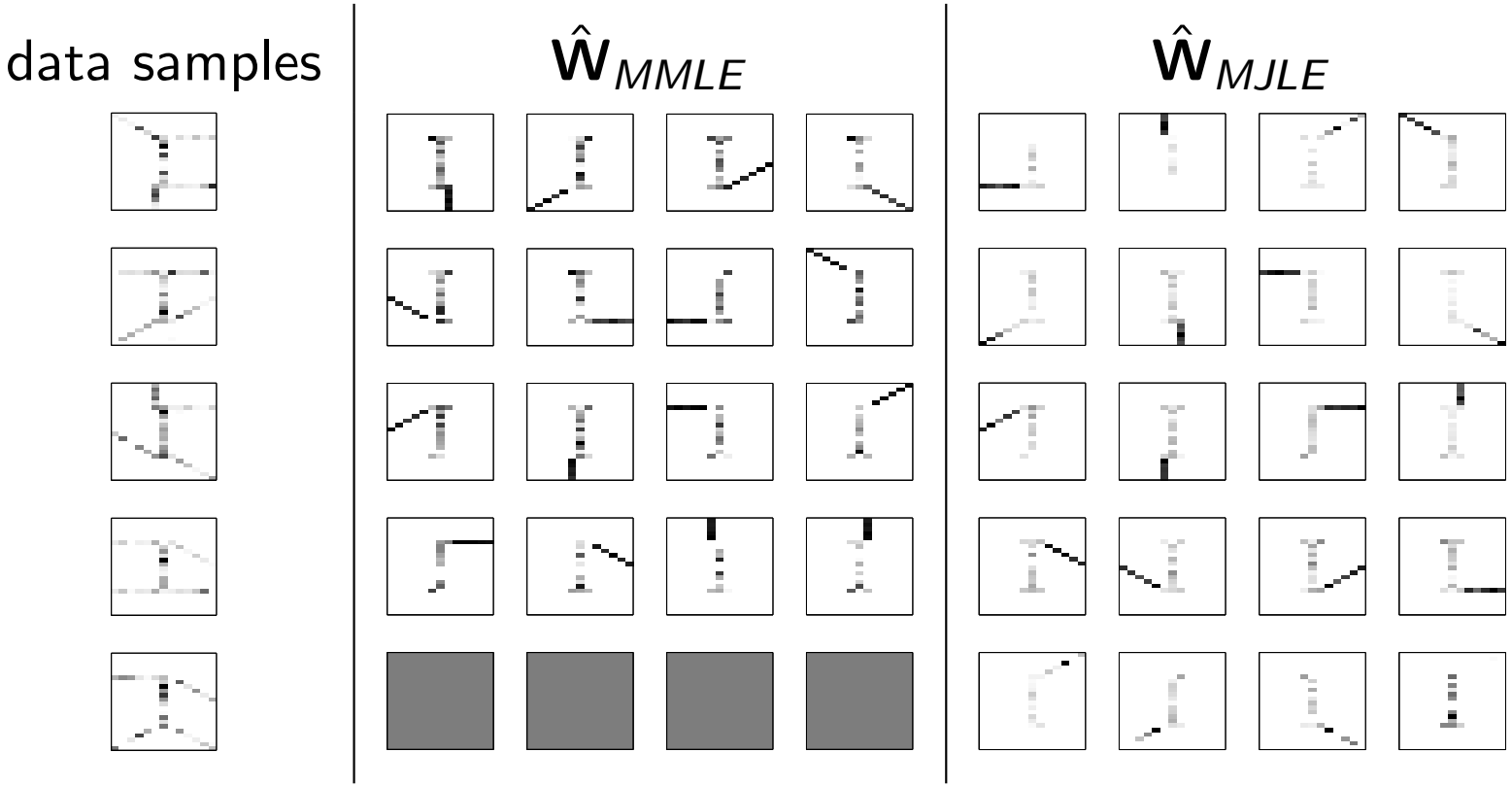
Gamma-Exponential model





Synthetical data

Gamma-Exponential model



MMLE returns four null columns in $\hat{\mathbf{W}}$ while MJLE overfits.

Text data

Gamma-Poisson model

- MusiXmatch public lyrics database with more than $230,000$ songs.
- Bag-of-words representation of each song using $5,000$ most frequent (stemmed) words.
- Analysis of $N = 10,000$ random songs with $K = 200$ with MMLE, MJLE and LDA.
- # occurrences of word f from topic k in song n is reconstructed its posterior mean:

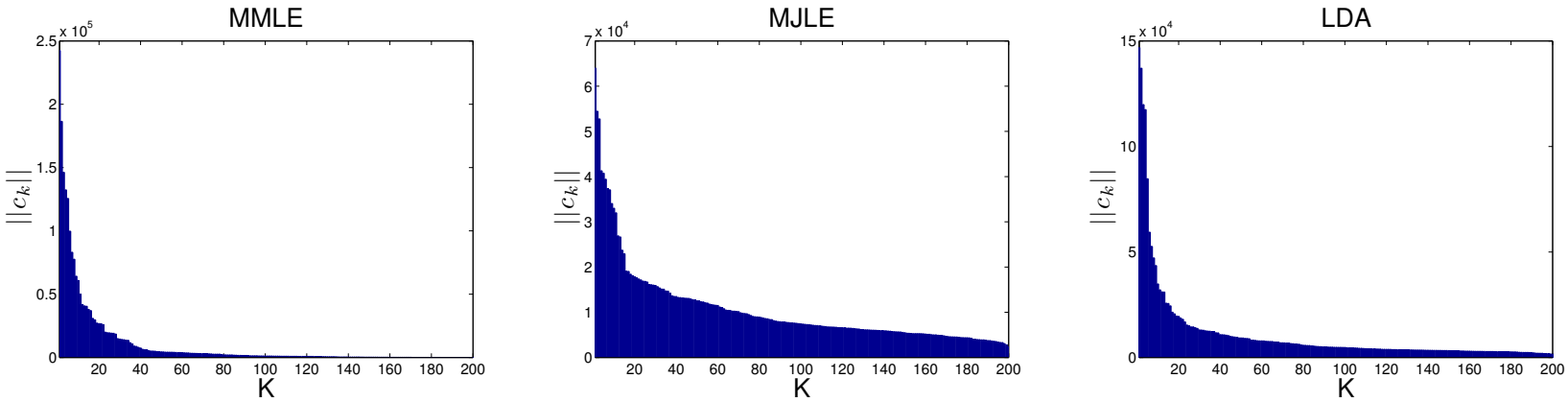
$$\hat{c}_{k,fn} = \frac{\hat{w}_{fk}\hat{h}_{kn}}{[\mathbf{WH}]_{fn}} v_{fn}$$

It follows that $\mathbf{V} = \sum_k \hat{\mathbf{C}}_k$.

Text data

Gamma-Poisson model

Norms $\|\hat{\mathbf{C}}_k\|$ of the components from the 3 approaches.



In contrast with MJLE and LDA, MMLE cancels out about 50 of the components.

Text data

Gamma-Poisson model

4 topics extracted by MMLE and their 5 most representative songs.

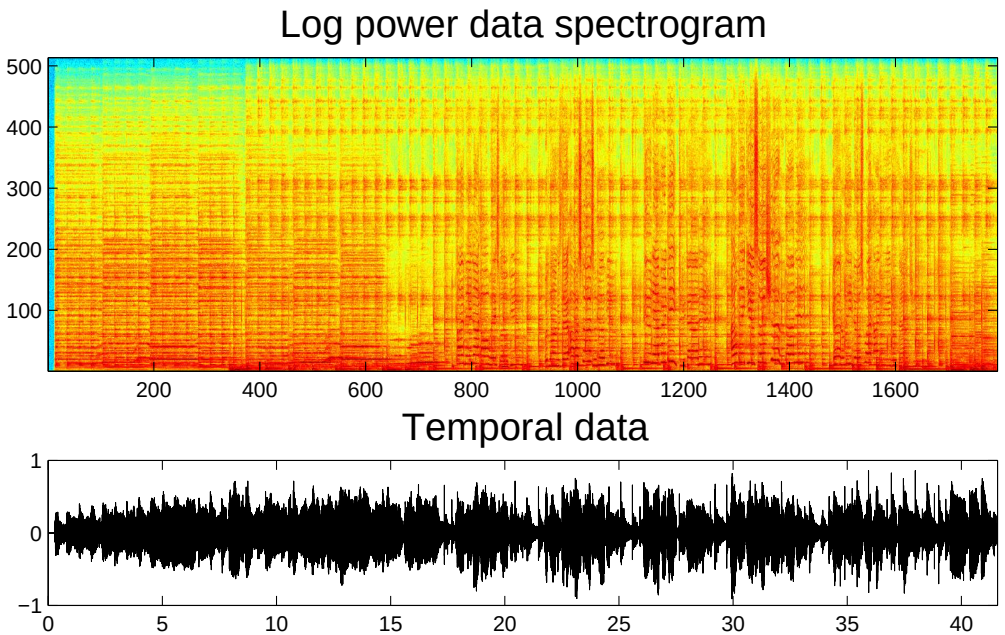
($k = 2$) get nigga the ya shit like fuck em got hit bitch up off yall ass they that cmon money and	
UGK (Underground Kingz) - Murder	i the to nigga my a you got murder and it is am from we so with they yo cuz
Big Punisher - Nigga Shit	shit that nigga the i and my what to out am in on for love me with gettin you do
E-40 - Turf Drop [Clean]	gasolin the my i hey to a it on you some fuck spit of what one ride nigga sick gold
Cam'Ron - Sports Drugs & Entertainment	a the you i got yo stop shot is caus or street jump short wick either to on but in
Foxy Brown - Chyna Whyte	the nigga and you shit i not yall to a on with bitch no fuck uh it money white huh
($k = 8$) god of blood soul death die fear pain hell power within shall earth blind human bleed scream evil holi peac	
Demolition Hammer - Epidemic Of Violence	of pain death reign violenc and a kill rage vicious the to in down blue dead cold
Disgorge - Parallels Of Infinite Torture	of the tortur by their within upon flow throne infinit are no they see life eye befor
Tacere - Beyond Silence	silenc beyond a dark beauti i the you to and me it not in my is of your that do
Cannibal Corpse - Perverse Suffering	to my pain of i me for agoni in by and from way etern lust tortur crave the not be
Showbread - Sampsa Meets Kafka	to of no one die death loneli starv i the you and a me it not in my is your
($k = 26$) she her girl beauti woman & queen sex sexi cloth herself doll shes pink gypsi bodi midnight callin dress hair	
Headhunter - Sex & Drugs & Rock'N Roll	& sex drug rock roll n is good veri inde and not my are all need dead bodi brain i
Holy Barbarians - She	she of kind girl my is the a littl woman like world and gone destroy tiger me on an
X - Devil Doll	devil doll her she and a the in is of eye bone & shoe rag batter you to on no
Kittie - Paperdoll	her she you i now soul pain to is down want eat fit size and not in all dead bodi
Ottawan - D.I.S.C.O.	is she oh disco i o s d c super incred a crazi such desir sexi complic special candi
($k = 13$) je et les le pas dan pour des cest qui de tout mon moi au comm ne sur jai	
Veronique Sanson - Feminin	cest comm le car de bien se les mai a fait devant heur du et une quon quelqu etre
Nevrotic Explosion - Heritage	quon faut mieux pour nous qui nos ceux de la un plus tous honor parent ami oui
Kells - Sans teint	de la se le san des est loin peur reve pour sa sang corp lumier larm
Stille Volk - Corps Magicien	de les ell dan la se le du pass est sa par mond leur corp vivr lair voyag feu
Florent Pagny - Tue-Moi	si plus que un tu mon mes jour souvenir parc

Spectral data

Gamma-Exponential model

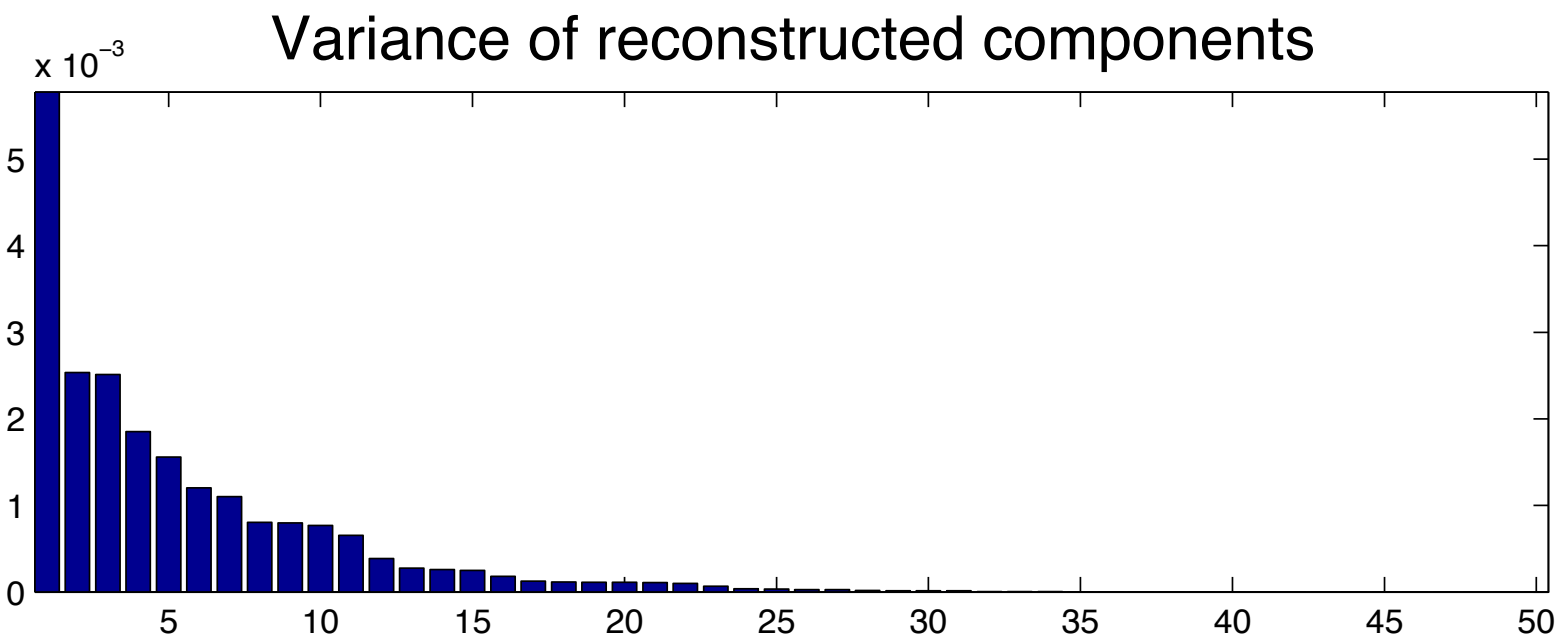
- 40 seconds of *God Only Knows* by the Beach Boys.
- MMLE decomposition of the spectrogram $v_{fn} = |x_{fn}|^2$ with $K = 50$ components.

  Note that the Gamma-Exponential model is a valid generative model of the STFT (Févotte et al., 2009).
- component reconstruction $\hat{c}_{k,fn} = \frac{\hat{w}_{fk}\hat{h}_{kn}}{[\mathbf{WH}]_{fn}} x_{fn}$.



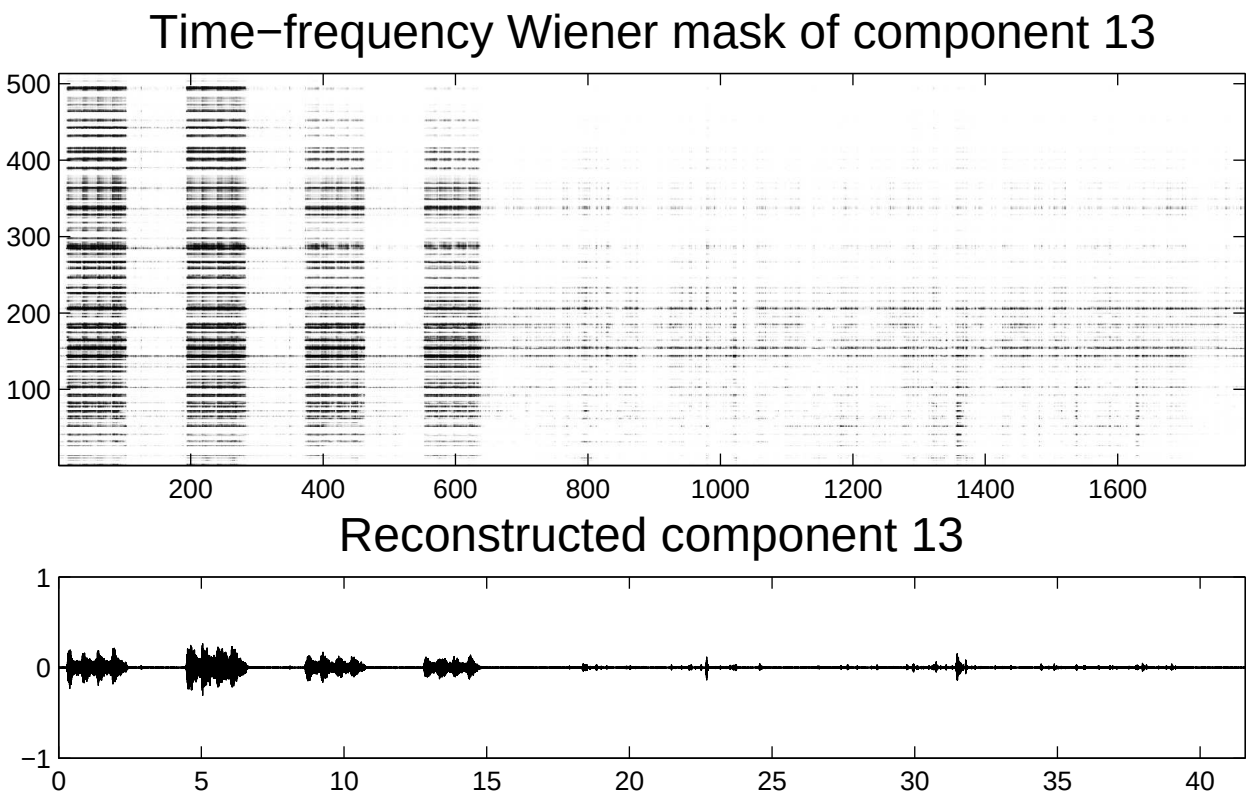
Spectral data

Gamma-Exponential model



Spectral data

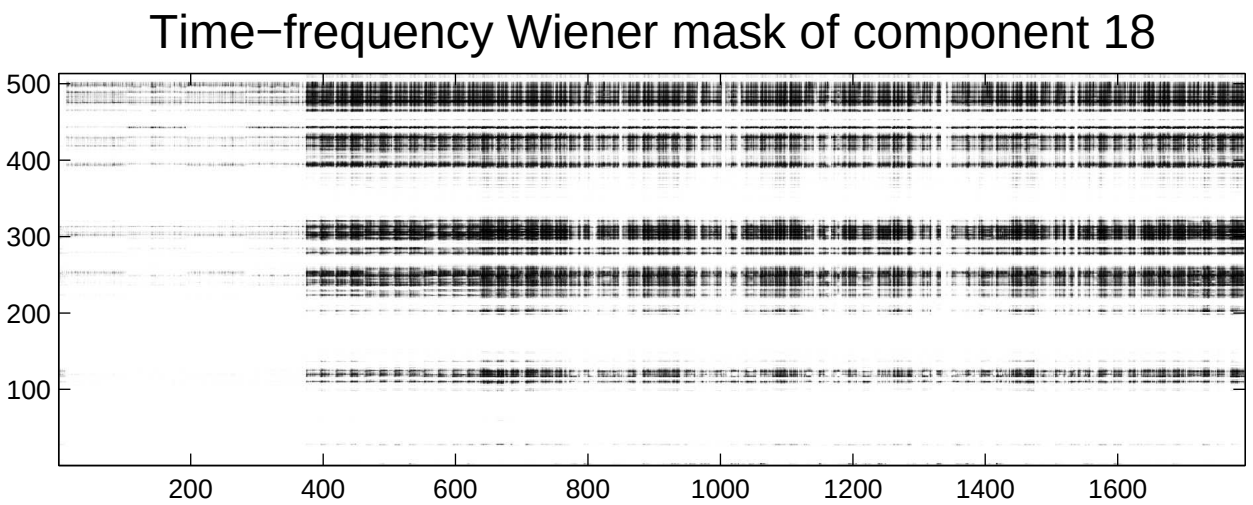
Gamma-Exponential model



audio

Spectral data

Gamma-Exponential model



audio

Conclusion

- Review of probabilistic latent factor models for nonnegative data.
- Review of estimators.
- MMLE leads to a better-posed estimator than MAP/MJLE
  - statistically well-posed (finite number of parameters)
  - scale-invariant
- For the two models considered, MMLE is found empirically to self-regularize the rank of the dictionary.
  - surprising and very appealing result
  - Laplace approximation of the marginal likelihood provides a start to explain this phenomenon, see (Dikmen and Févotte, 2012)

References I

M. W. Berry, M. Browne, A. N. Langville, V. P. Pauca, and R. J. Plemmons. Algorithms and applications for approximate nonnegative matrix factorization. *Computational Statistics & Data Analysis*, 52(1):155–173, Sep. 2007.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, Jan. 2003.

J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov. Metagenes and molecular pattern discovery using matrix factorization. In *Proceedings of the National Academy of Sciences*, pages 4164–4169, Mar. 2004.

W. L. Buntine and A. Jakulin. Discrete component analysis. In *Lecture Notes in Computer Science*, volume 3940, pages 1–33. Springer, 2006. URL `http://www.springerlink.com/content/d53027666542q3v7/`.

J. F. Canny. GaP: A factor model for discrete data. In *Proceedings of the 27th ACM international Conference on Research and Development of Information Retrieval (SIGIR)*, pages 122–129, 2004.

A. T. Cemgil. Bayesian inference for nonnegative matrix factorisation models. *Computational Intelligence and Neuroscience*, 2009(Article ID 785152):17 pages, 2009. doi:10.1155/2009/785152.

M. Cooper and J. Foote. Summarizing video using non-negative similarity matrix factorization. In *Proc. IEEE Workshop on Multimedia Signal Processing*, 2002.

References II

O. Dikmen and C. Févotte. Nonnegative dictionary learning in the exponential noise model for adaptive music signal representation. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 24 (NIPS)*, pages 2267–2275, Granada, Spain, Dec. 2011. MIT Press. URL `http://www.unice.fr/cfevotte/publications/proceedings/nips11.pdf`.

O. Dikmen and C. Févotte. Maximum marginal likelihood estimation for nonnegative dictionary learning in the Gamma-Poisson model. *IEEE Transactions on Signal Processing*, 60(10):5163–5175, Oct. 2012. doi: http://dx.doi.org/10.1109/TSP.2012.2207117. URL `http://www.unice.fr/cfevotte/publications/journals/ieee_sp_mmle.pdf`.

K. Drakakis, S. Rickard, R. de Frein, and A. Cichocki. Analysis of financial data using non-negative matrix factorization. *International Journal of Mathematical Sciences*, 6(2), June 2007.

C. Févotte and J. Idier. Algorithms for nonnegative matrix factorization with the beta-divergence. *Neural Computation*, 23(9):2421–2456, Sep. 2011. doi: 10.1162/NECO_a_00168. URL `http://www.unice.fr/cfevotte/publications/journals/neco11.pdf`.

References III

C. Févotte, N. Bertin, and J.-L. Durrieu. Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis. *Neural Computation*, 21(3):793–830, Mar. 2009. doi: 10.1162/neco.2008.04-08-771. URL `http://www.unice.fr/cfevotte/publications/journals/neco09_is-nmf.pdf`.

E. Gaussier and C. Goutte. Relation between PLSA and NMF and implications. In *Proc. 28th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'05)*, pages 601–602, New York, NY, USA, 2005. ACM.

D. Greene, G. Cagney, N. Krogan, and P. Cunningham. Ensemble non-negative matrix factorization methods for clustering protein-protein interactions. *Bioinformatics*, 24 (15):1722–1728, 2008.

M. Hoffman, D. Blei, and P. Cook. Bayesian nonparametric matrix factorization for recorded music. In *Proc. 27th International Conference on Machine Learning (ICML)*, Haifa, Israel, 2010.

T. Hofmann. Probabilistic latent semantic indexing. In *Proc. 22nd International Conference on Research and Development in Information Retrieval (SIGIR)*, 1999. URL `http://www.cs.brown.edu/~th/papers/Hofmann-SIGIR99.pdf`.

B. Jørgensen. Exponential dispersion models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 49(2):127–162, 1987.

References IV

D. D. Lee and H. S. Seung. Learning the parts of objects with nonnegative matrix factorization. *Nature*, 401:788–791, 1999.

M. S. Lewicki and T. J. Sejnowski. Learning overcomplete representations. *Neural Computation*, 12:337–365, 2000.

J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11:10–60, 2010.

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

M. N. Schmidt, O. Winther, and L. K. Hansen. Bayesian non-negative matrix factorization. In *In Proc. 8th Internation conference on Independent Component Analysis and Signal Separation (ICA)*, Paraty, Brazil, Mar. 2009.

M. Seeger and G. Bouchard. Fast variational bayesian inference for non-conjugate matrix factorization models. In *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2012.

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 (WASPAA'03)*, Oct. 2003.

References V

V. Y. F. Tan and C. Févotte. Automatic relevance determination in nonnegative matrix factorization with the beta-divergence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1592 – 1605, July 2013. URL `http://www.unice.fr/cfevotte/publications/journals/pami13_ardnmf.pdf`.

M. Tweedie. An index which distinguishes between some important exponential families. In J. K. Ghosh and J. Roy, editors, *Proc. of the Indian Statistical Institute Golden Jubilee International Conference*, Statistics: Applications and New Directions, pages 579–604, Calcutta, 1984.

T. Virtanen. Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria. *IEEE Transactions on Audio, Speech and Language Processing*, 15(3):1066–1074, Mar. 2007.

Z. Yang and E. Oja. Unified development of multiplicative algorithms for linear and quadratic nonnegative matrix factorization. *IEEE Transactions on Neural Networks*, 22:1878 – 1891, Dec. 2011. doi: http://dx.doi.org/10.1109/TNN.2011.2170094.

Y. K. Yilmaz. *Generalized tensor factorization*. PhD thesis, Boğaziçi University, Istanbul, Turkey, 2012.

S. S. Young, P. Fogel, and D. Hawkins. Clustering Scotch whiskies using non-negative matrix factorization. *Joint Newsletter for the Section on Physical and Engineering Sciences and the Quality and Productivity Section of the American Statistical Association*, 14(1):11–13, June 2006.

References VI

M. Zetlaoui, M. Feinberg, P. Verger, and S. Clémençon. Extraction of food consumption systems by non-negative matrix factorization (nmf) for the assessment of food choices. Technical report, Arxiv, 2010. URL `http://hal.archives-ouvertes.fr/docs/00/48/47/94/PDF/NMF_food.pdf`.

M. Zhong and M. Girolami. Reversible jump MCMC for non-negative matrix factorization. In *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, page 8, 2009.