Maximum marginal likelihood estimation for nonnegative dictionary learning

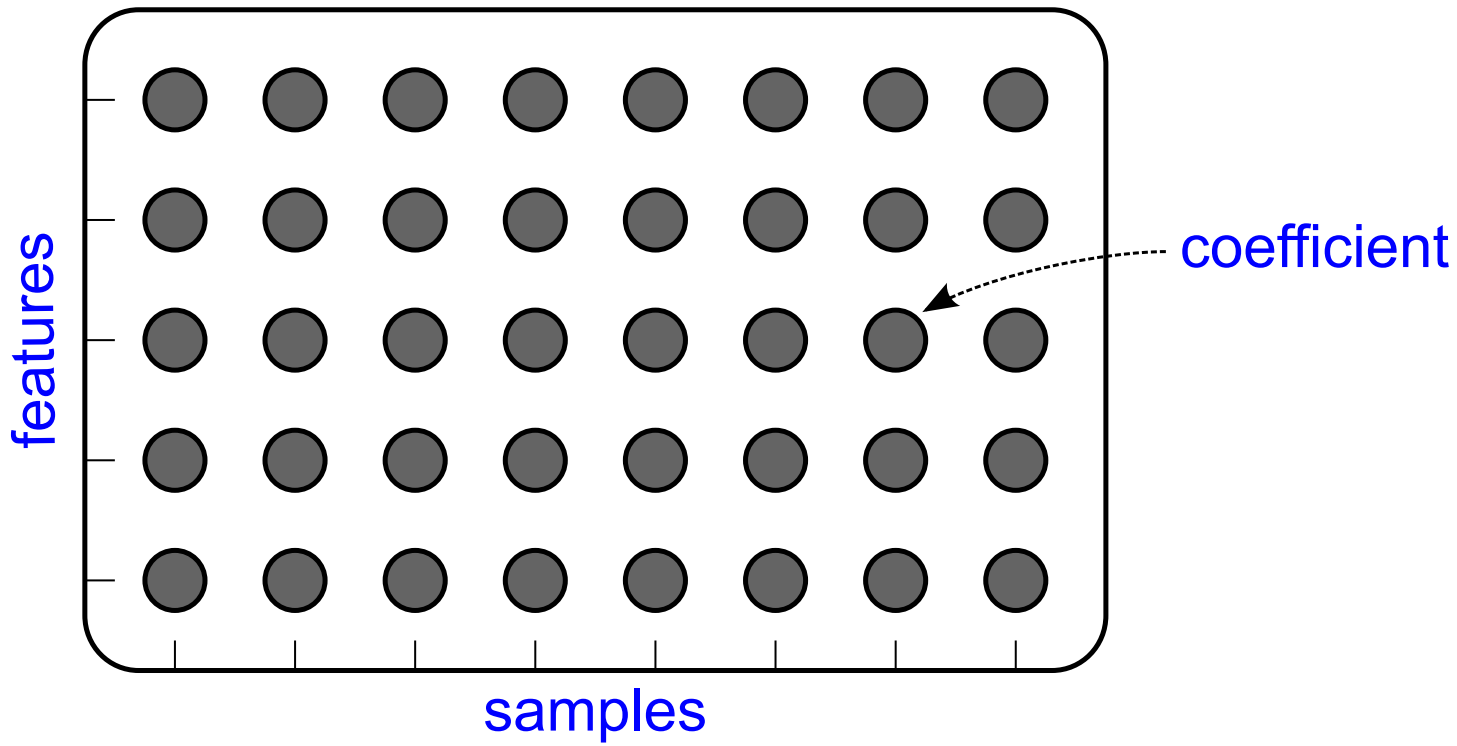
Cédric Févotte
Institut de Recherche en Informatique de Toulouse (IRIT)

Unité MIAT, INRA
Toulouse, March 2017

Outline

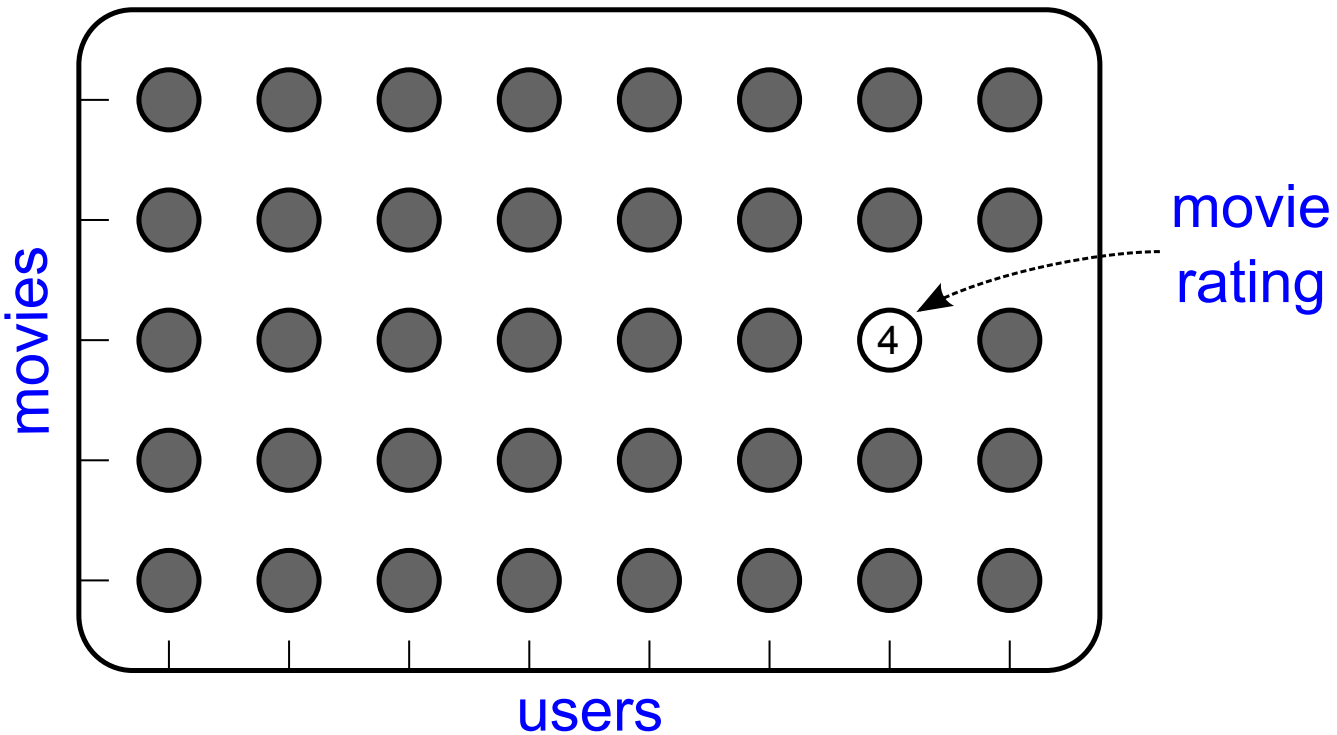
Matrix factorisation in data processing

Data often available in matrix form.



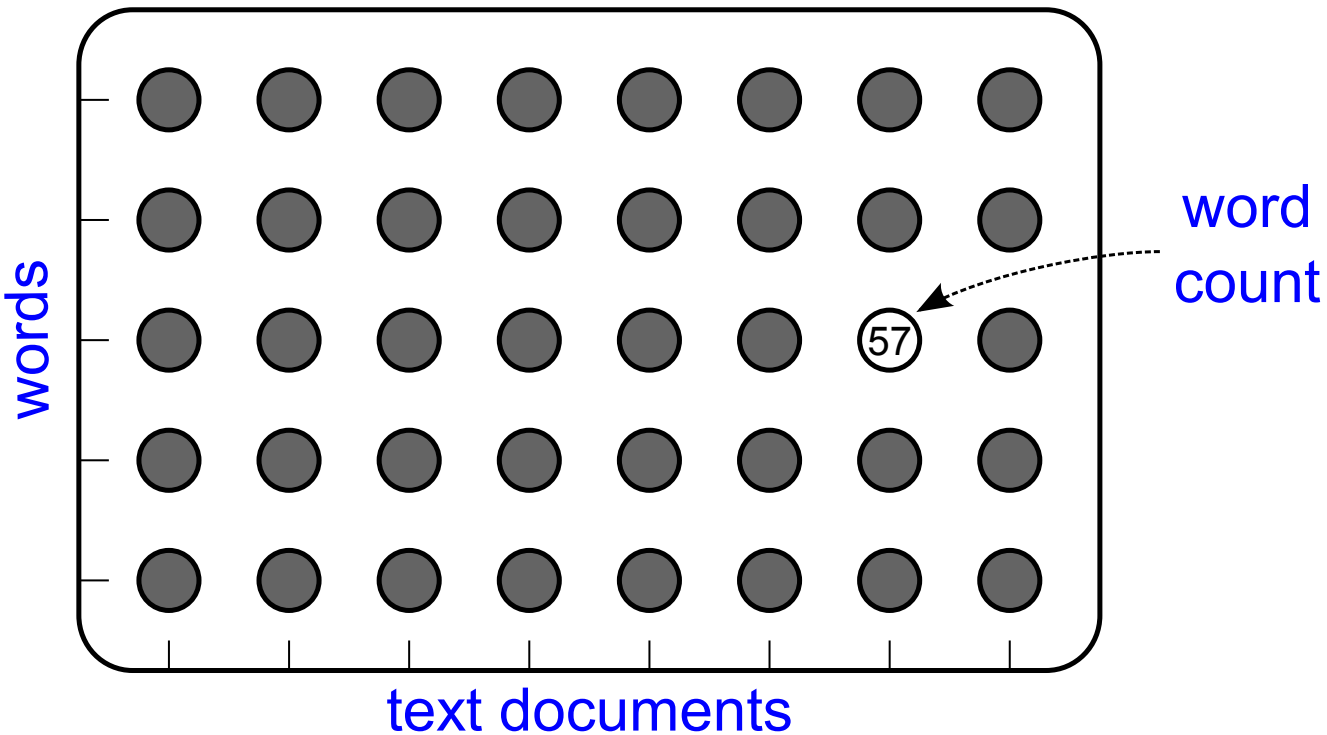
Matrix factorisation in data processing

Data often available in matrix form.



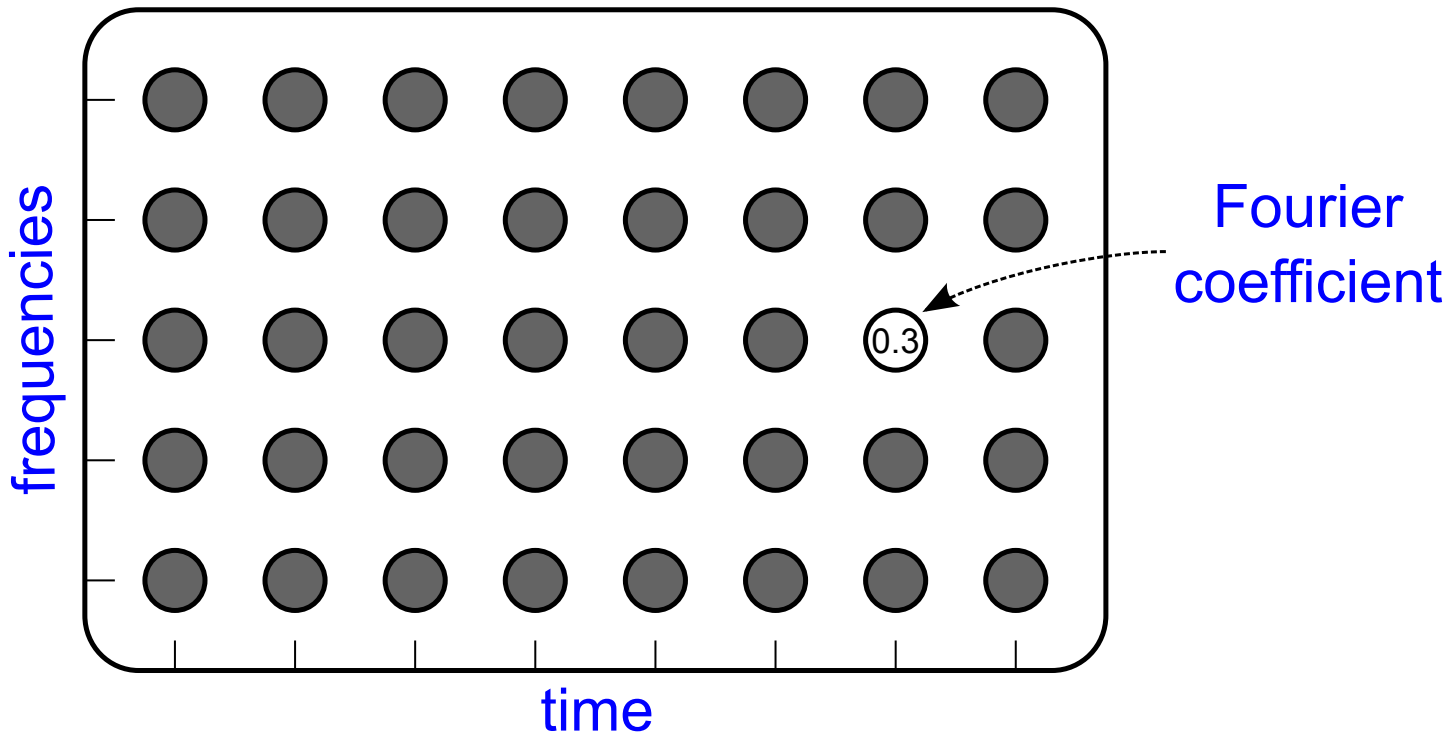
Matrix factorisation in data processing

Data often available in matrix form.



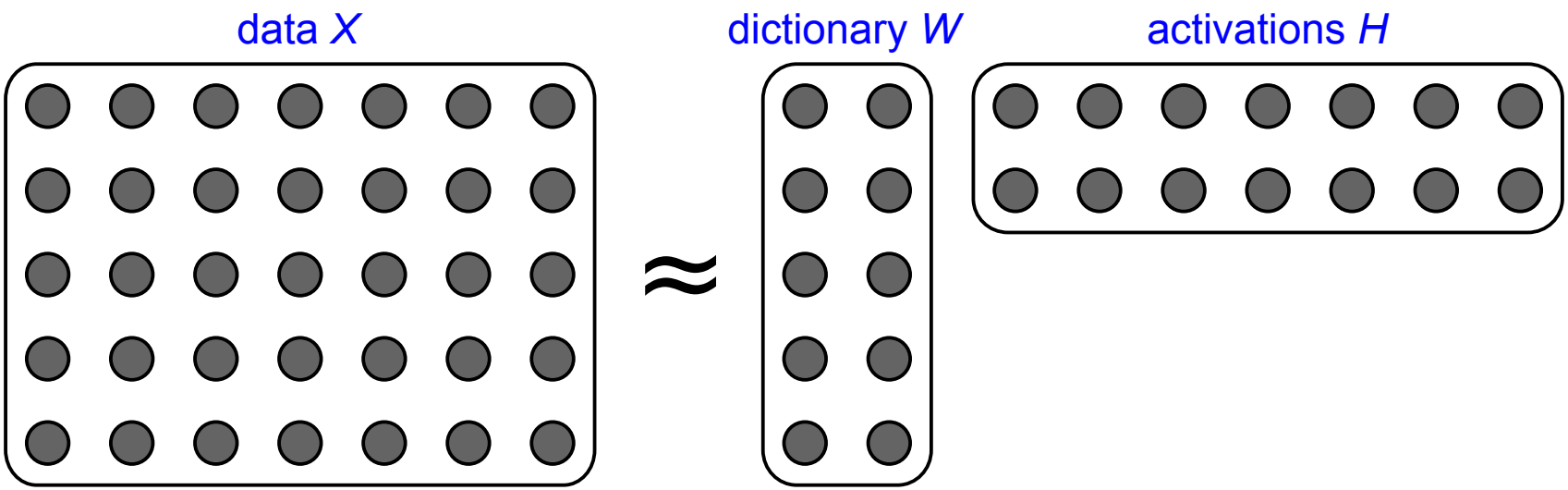
Matrix factorisation in data processing

Data often available in matrix form.



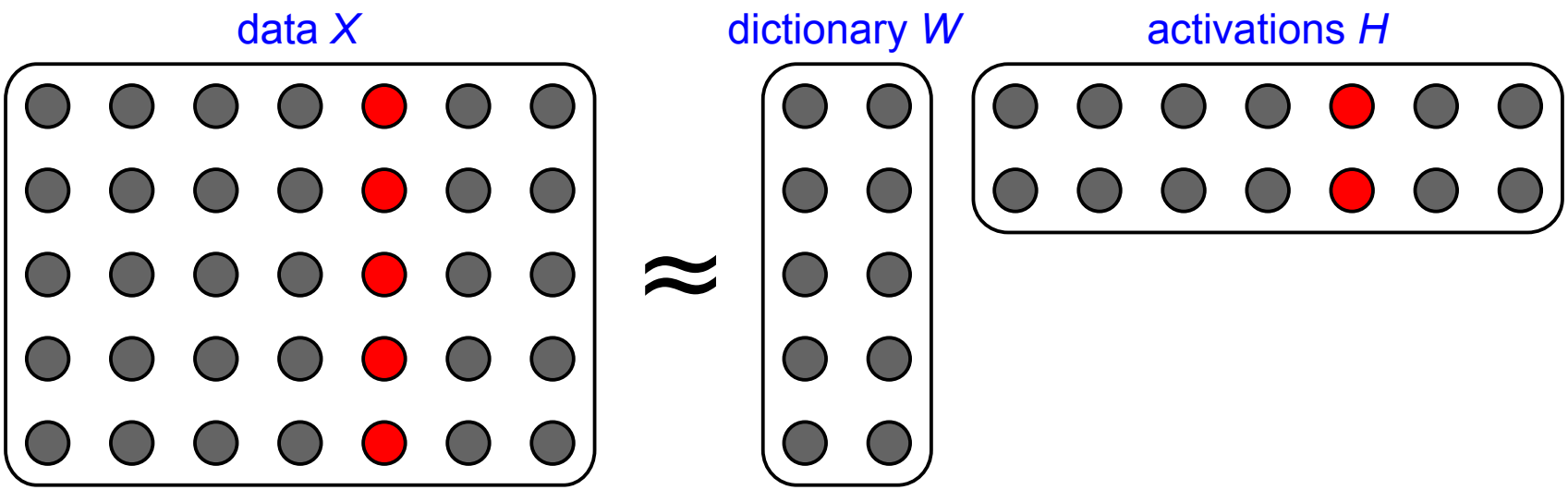
Matrix factorisation in data processing

$\approx$ **dictionary learning**
low-rank approximation
factor analysis
latent semantic analysis



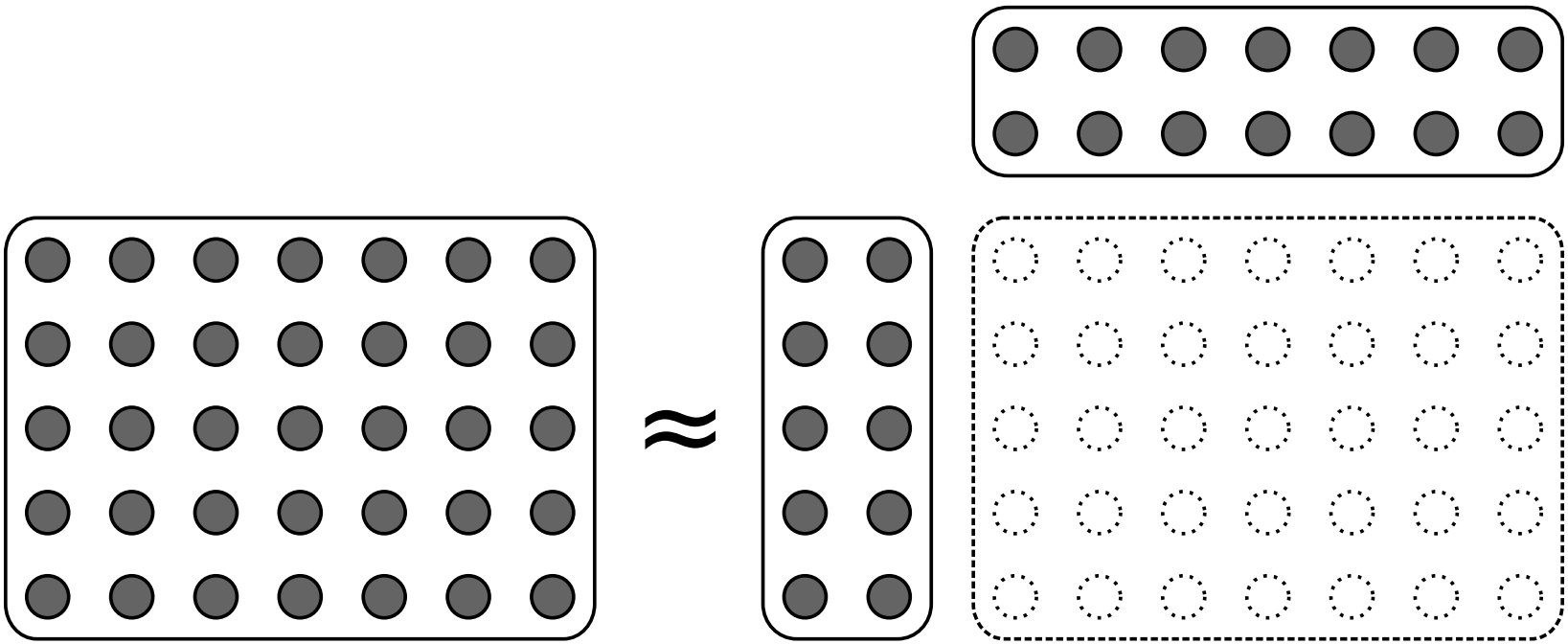
Matrix factorisation in data processing

$\approx$ **dictionary learning**
low-rank approximation
factor analysis
latent semantic analysis



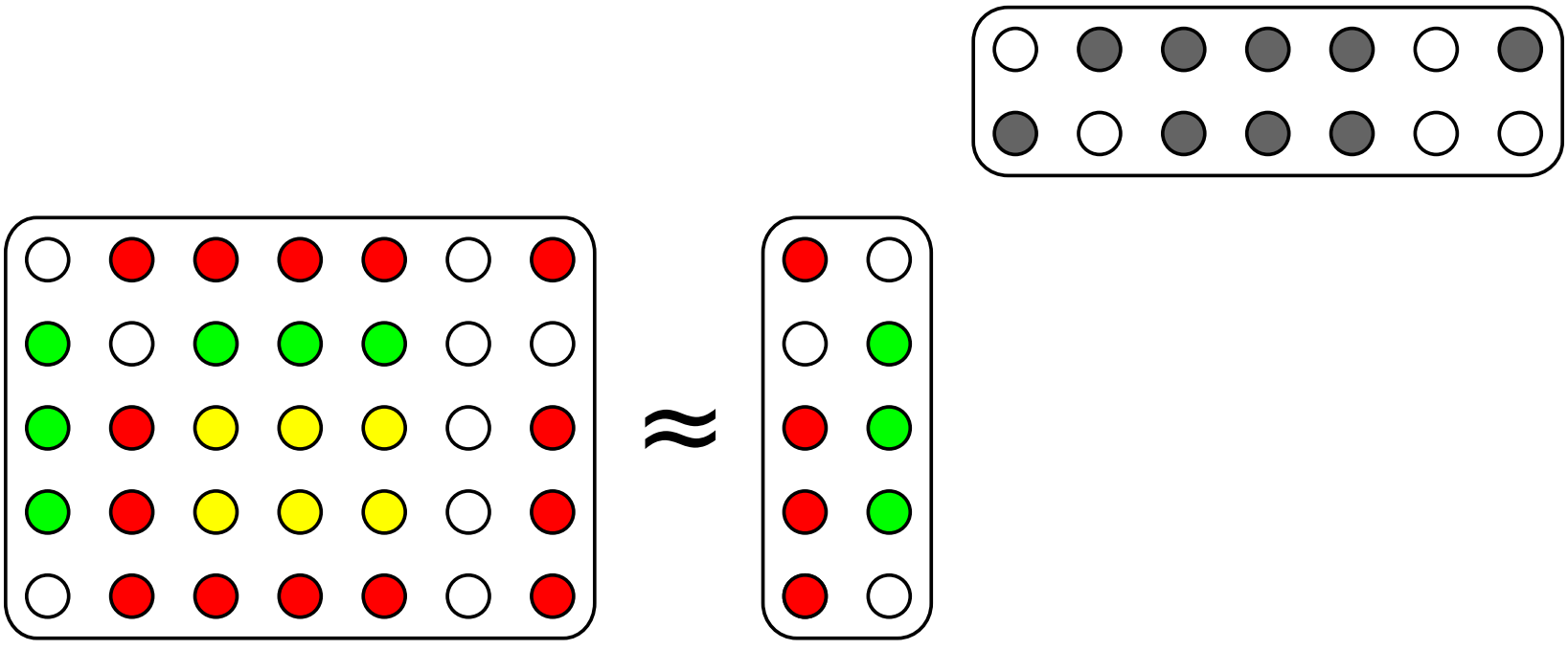
Matrix factorisation in data processing

for dimensionality reduction (coding, low-dimensional embedding)



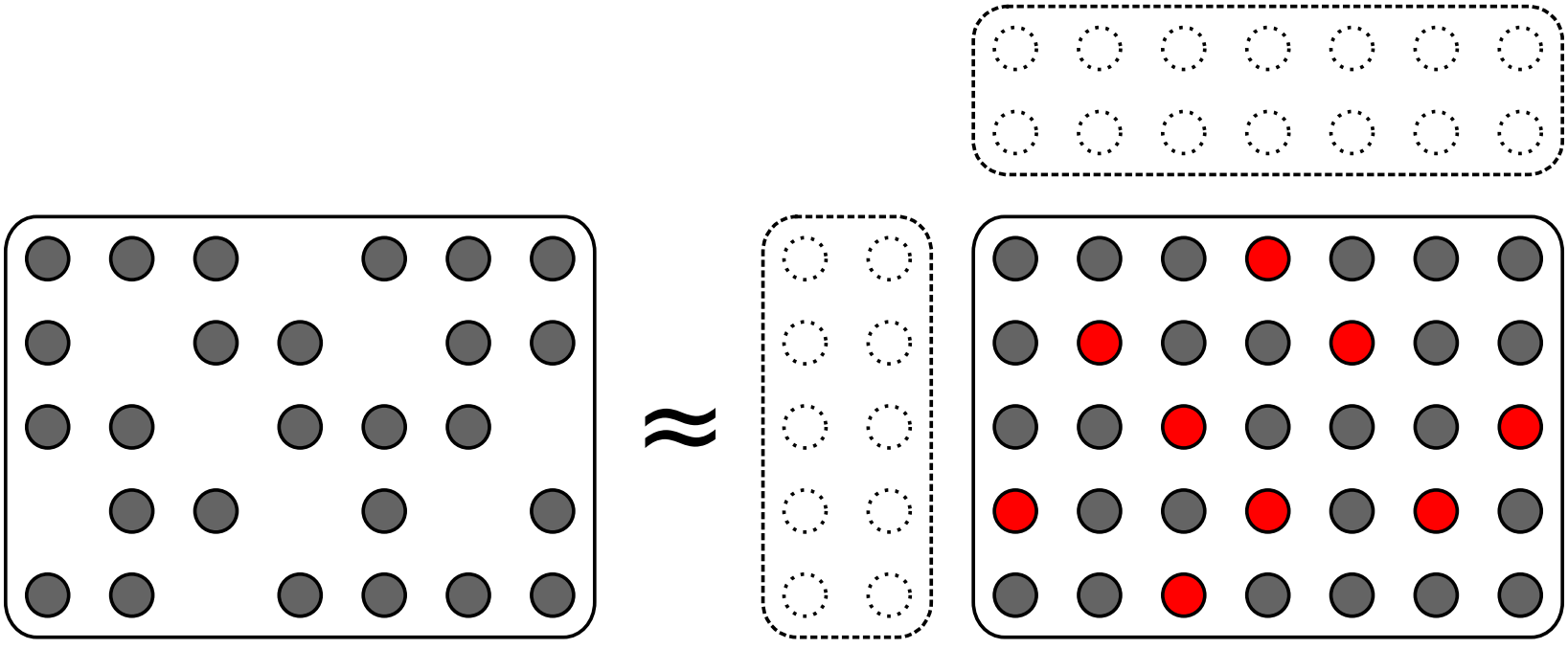
Matrix factorisation in data processing

for unmixing (source separation, latent topic discovery)

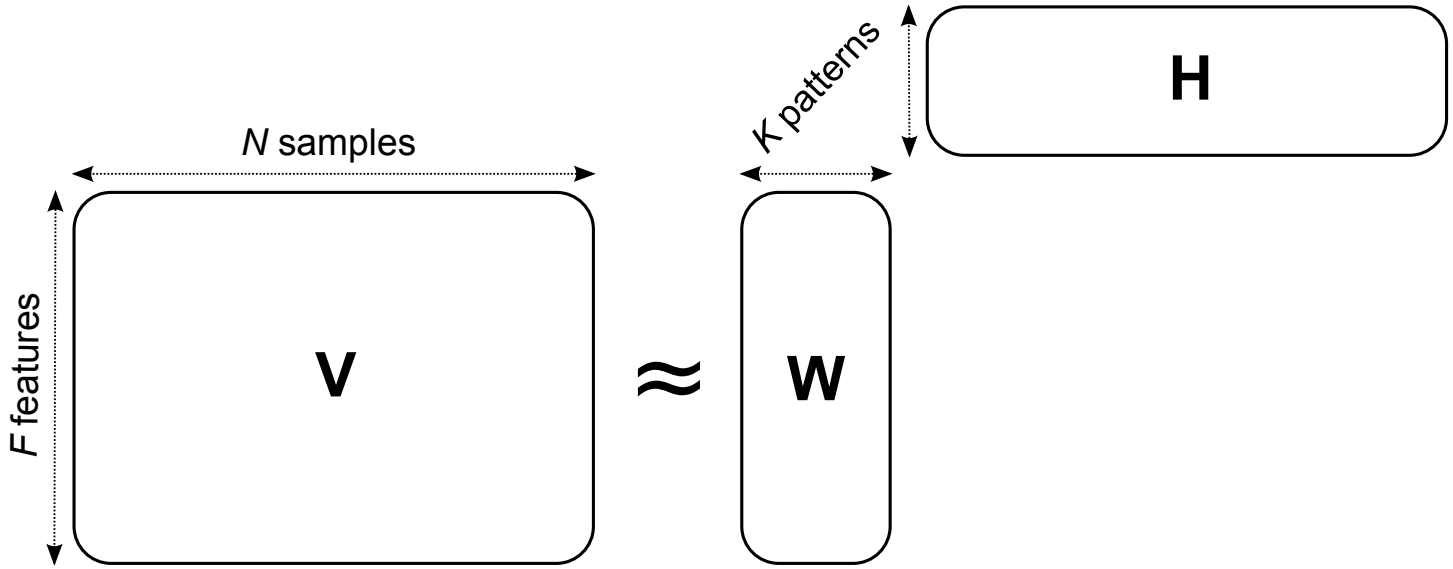


Matrix factorisation in data processing

for interpolation (collaborative filtering, image inpainting)



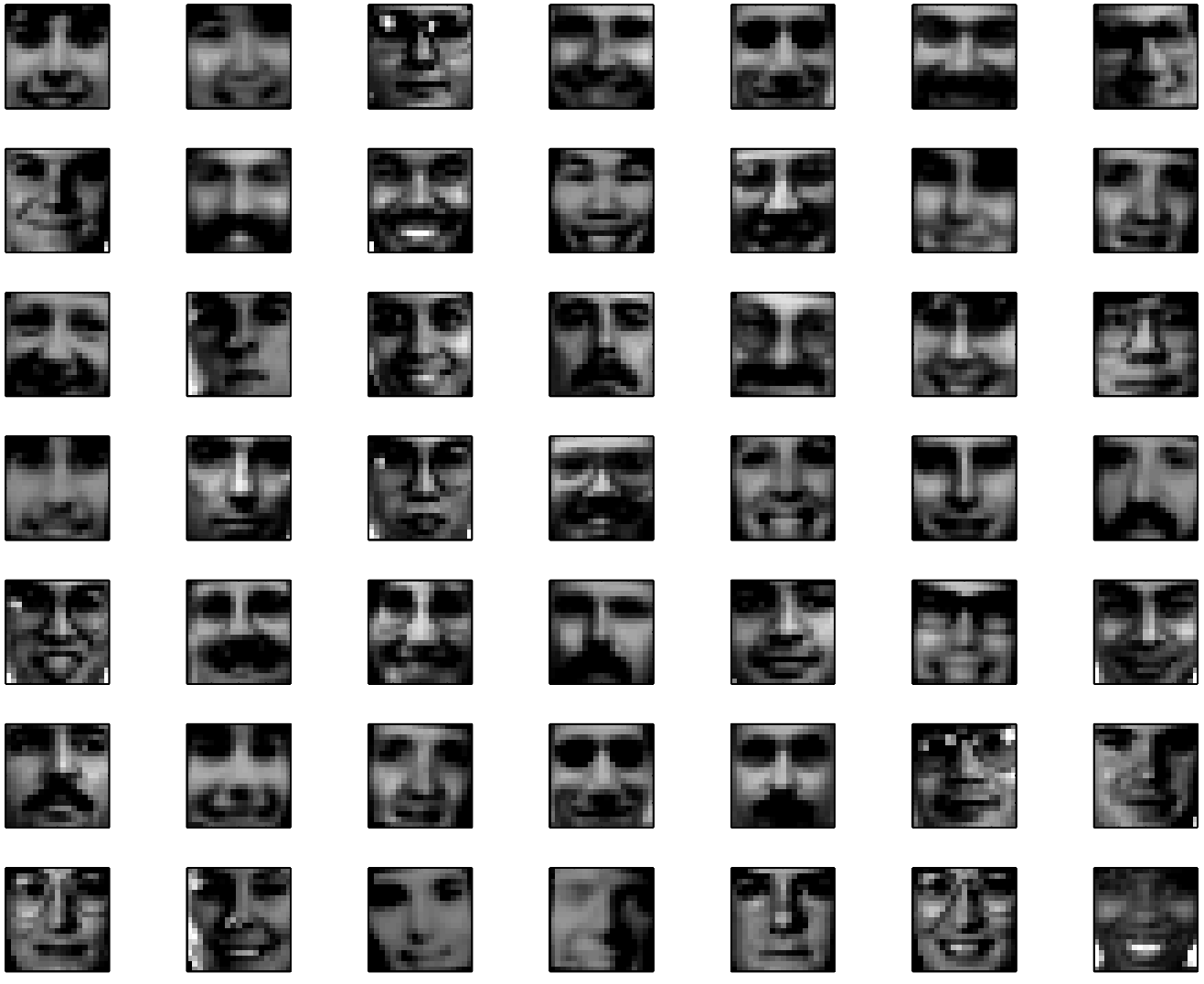
Nonnegative matrix factorisation



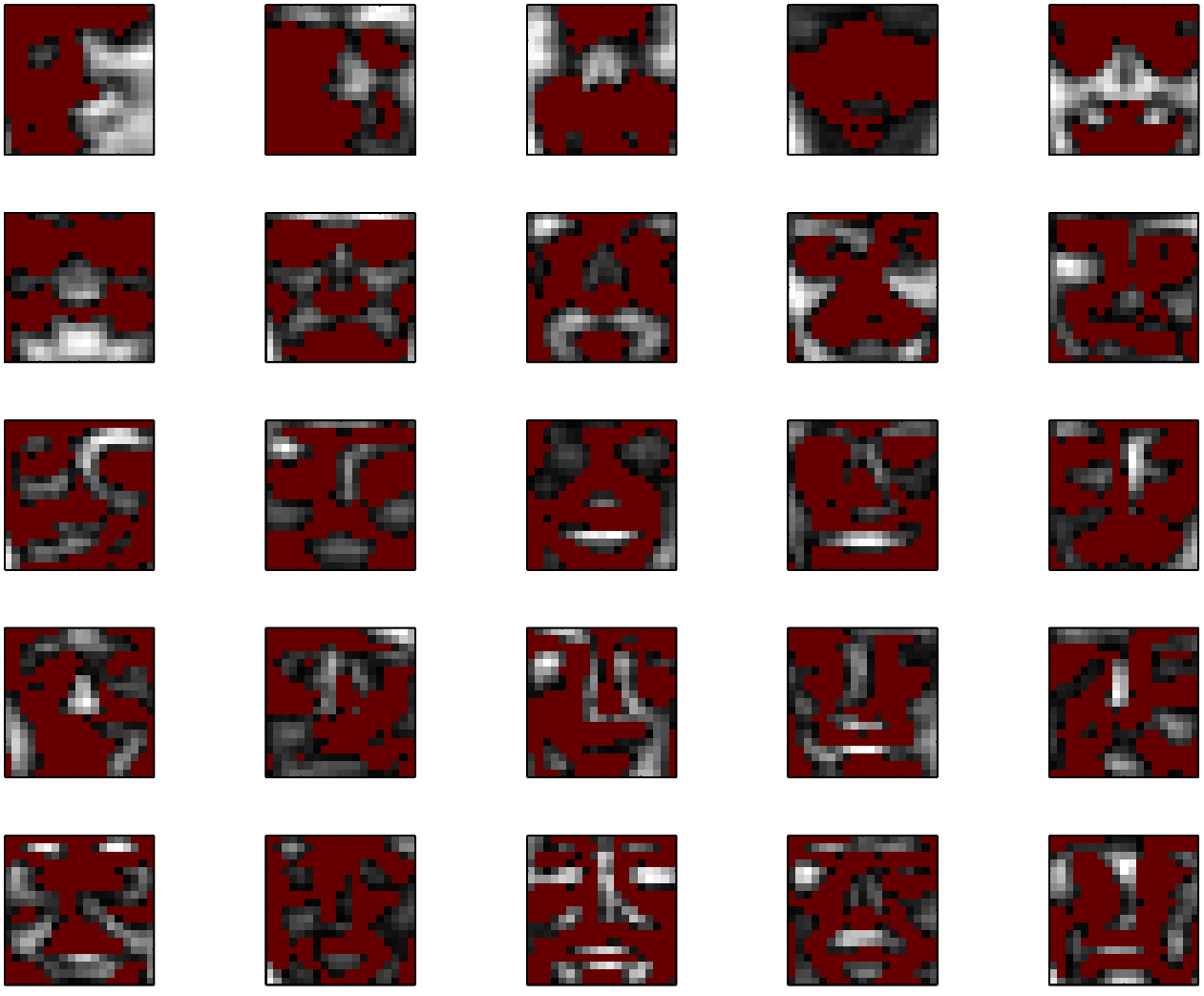
- data $\mathbf{V}$ and factors $\mathbf{W}$, $\mathbf{H}$ have nonnegative entries.
- nonnegativity of $\mathbf{W}$ ensures interpretability of the dictionary, because patterns $\mathbf{w}_k$ and samples $\mathbf{v}_n$ belong to the same space.
- nonnegativity of $\mathbf{H}$ tends to produce part-based representations, because subtractive combinations are forbidden.

Early work by Paatero and Tapper (1994), landmark *Nature* paper by Lee and Seung (1999)

49 images among 2429 from MIT's CBCL face dataset

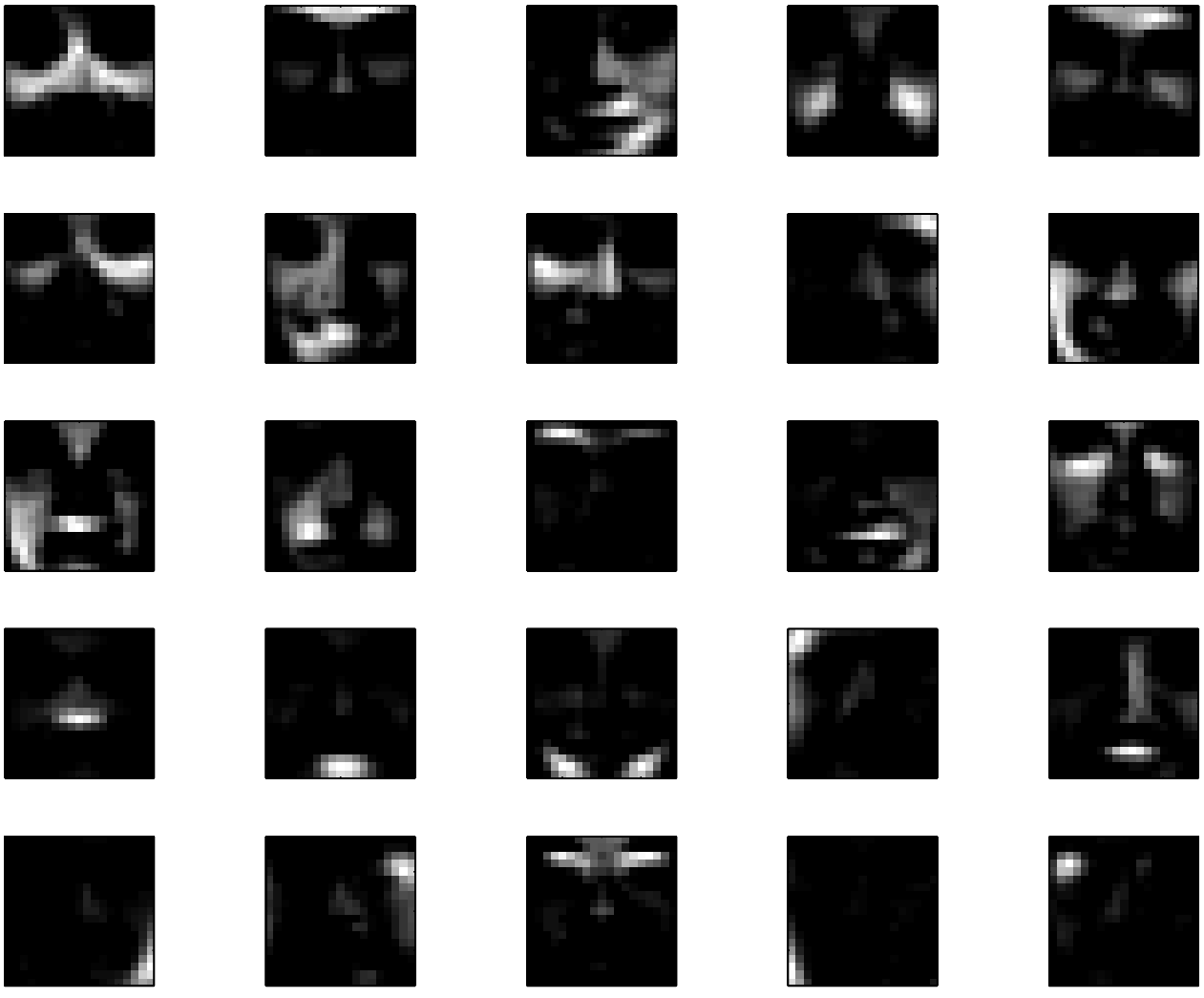


PCA dictionary with $K = 25$



red pixels indicate negative values

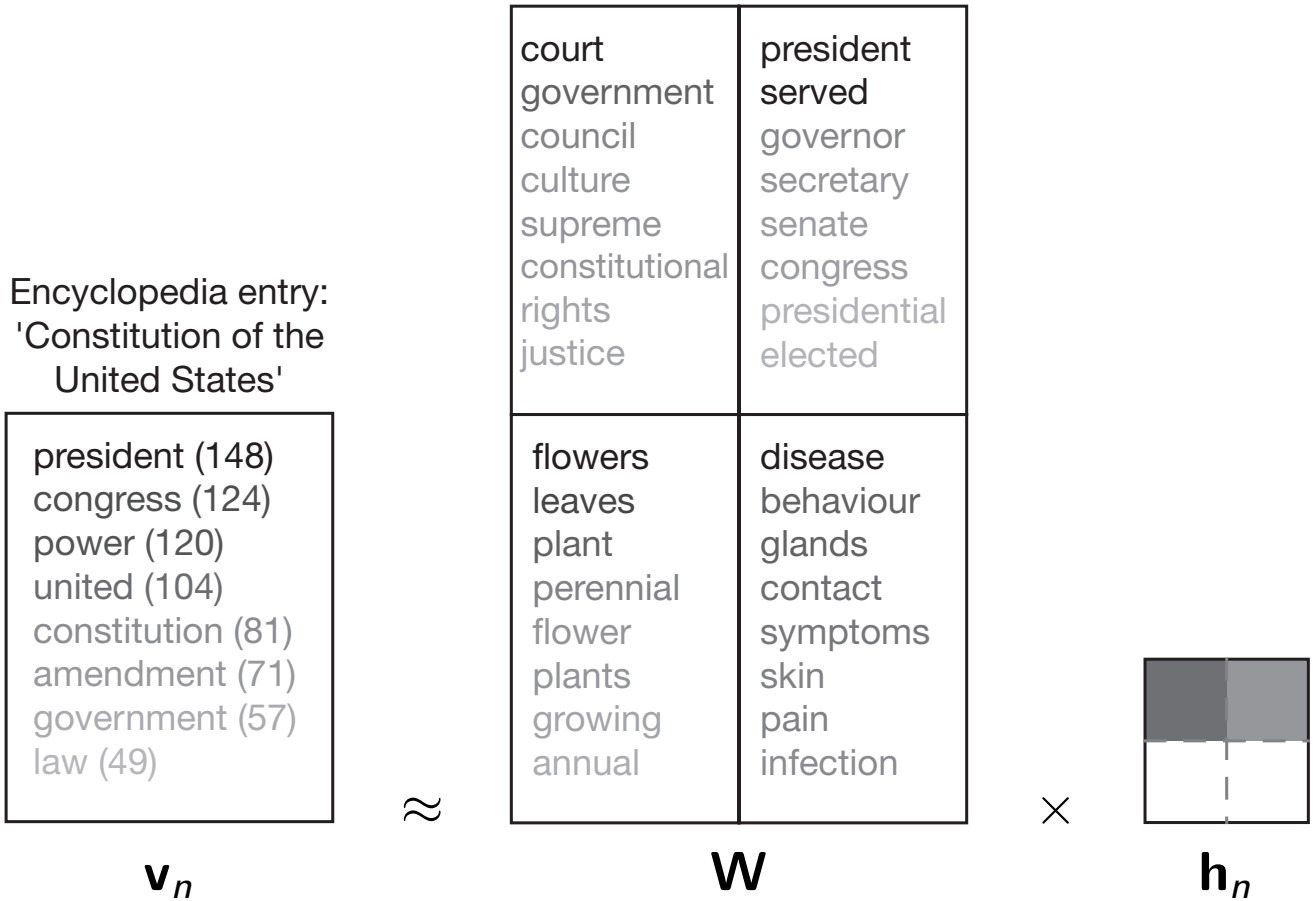
NMF dictionary with $K = 25$



experiment reproduced from (Lee and Seung, 1999)

NMF for latent semantic analysis

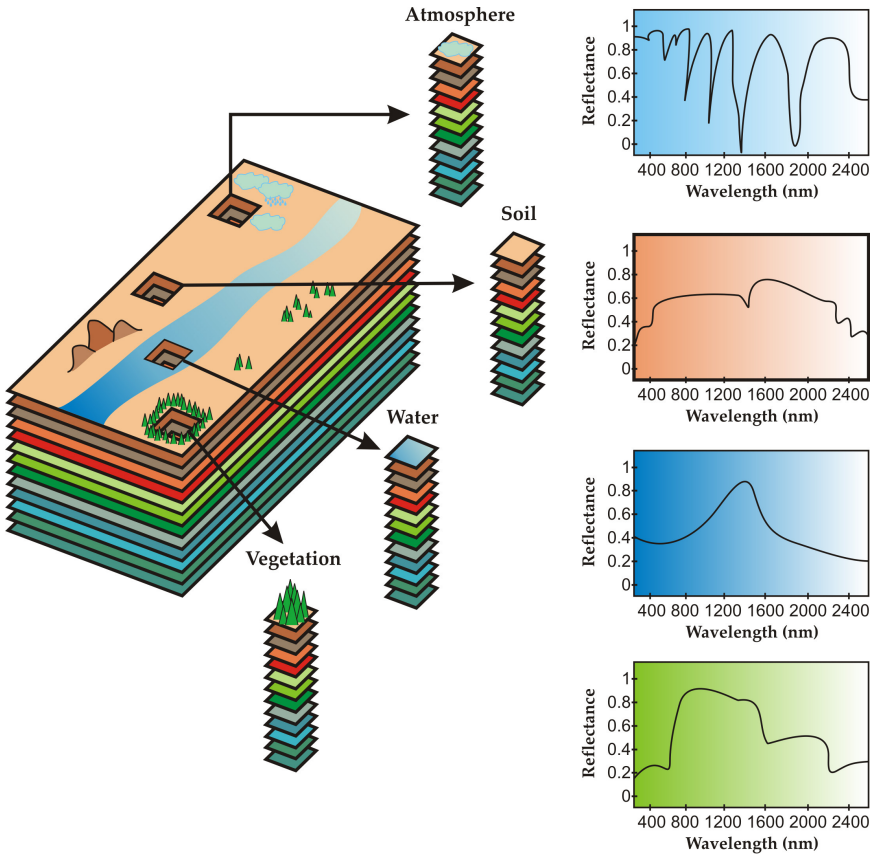
(Lee and Seung, 1999; Hofmann, 1999)



reproduced from (Lee and Seung, 1999)

NMF for hyperspectral unmixing

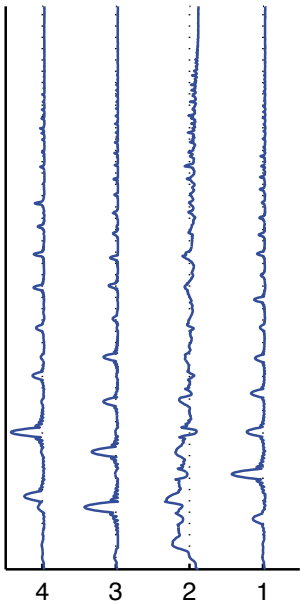
(Berry, Browne, Langville, Pauca, and Plemmons, 2007)

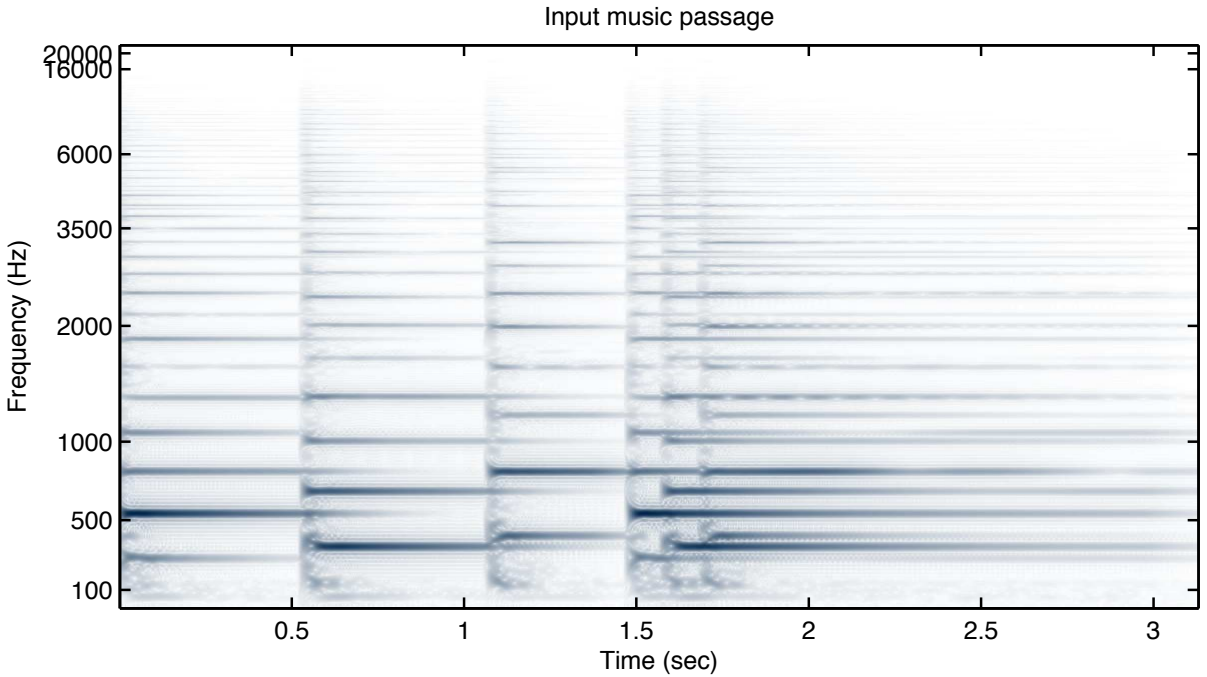


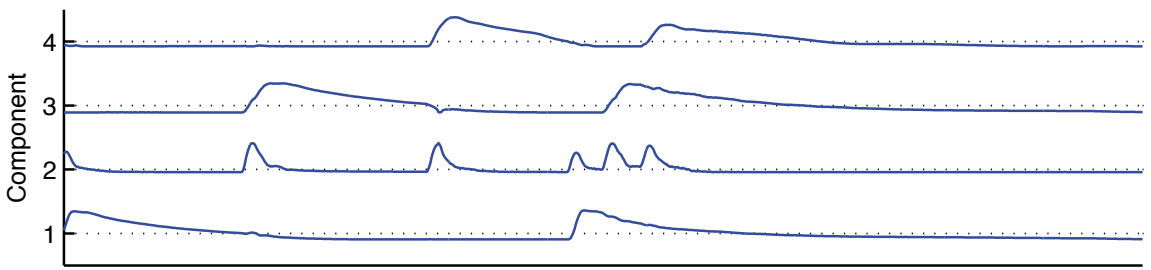
reproduced from (Bioucas-Dias et al., 2012)

NMF for audio spectral unmixing

(Smaragdis and Brown, 2003)







reproduced from (Smaragdis, 2013)

Outline

Traditional NMF

Minimise a measure of fit between $\mathbf{V}$ and $\mathbf{WH}$, subject to nonnegativity :

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

where $d(x|y)$ is a scalar cost function, e.g.,

- squared Euclidean distance (Paatero and Tapper, 1994; Lee and Seung, 2001)
- generalised KL divergence (Lee and Seung, 1999; Finesso and Spreij, 2006)
- Itakura-Saito divergence (Févotte, Bertin, and Durrieu, 2009)
- α-divergence (Cichocki et al., 2008)
- β-divergence (Cichocki et al., 2006; Févotte and Idier, 2011)
- Bregman divergences (Dhillon and Sra, 2005)
- and more in (Yang and Oja, 2011)

Regularisation terms often added to $D(\mathbf{V}|\mathbf{WH})$ for sparsity, smoothness, dynamics, etc.

Probabilistic viewpoint

- Let $\mathbf{V} \sim p(\mathbf{V}|\mathbf{WH})$ such that
  (1) $p(\mathbf{V}|\mathbf{WH}) = \prod_{fn} p(v_{fn}|[\mathbf{WH}]_{fn})$
  (2) $\mathrm{E}[\mathbf{V}|\mathbf{WH}] = \mathbf{WH}$
- then the following correspondences apply with

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

data support	distribution/noise	divergence	examples
real-valued	additive Gaussian	squared Euclidean	many
integer	multinomial	weighted KL	word counts
integer	Poisson	generalised KL	photon counts
nonnegative	multiplicative Gamma	Itakura-Saito	spectral data
generally non-negative	Tweedie	β-divergence	generalises above models

Probabilistic viewpoint

- Let $\mathbf{V} \sim p(\mathbf{V}|\mathbf{WH})$ such that
  (1) $p(\mathbf{V}|\mathbf{WH}) = \prod_{fn} p(v_{fn}|[\mathbf{WH}]_{fn})$
  (2) $\mathrm{E}[\mathbf{V}|\mathbf{WH}] = \mathbf{WH}$
- then the following correspondences apply with

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

data support	distribution/noise	divergence	examples
real-valued	additive Gaussian	squared Euclidean	many
integer	multinomial	weighted KL	word counts
integer	Poisson	generalised KL	photon counts
nonnegative	multiplicative Gamma	Itakura-Saito	spectral data
generally non-negative	Tweedie	β-divergence	generalises above models

- NMF sometimes cast as maximum likelihood estimation of $\mathbf{W}$ and $\mathbf{H}$.
- ill-posed estimation, because the number of parameters grows with data (one $\mathbf{h}_n$ for every $\mathbf{v}_n$)

Maximum *marginal* likelihood estimation

(Dikmen and Févotte, 2011, 2012)

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

Maximum *marginal* likelihood estimation

(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

(Dikmen and Févotte, 2011, 2012)

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

(Dikmen and Févotte, 2011, 2012)

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

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

+ better-posed than traditional NMF (fixed number of parameters)
+ better-behaved with respect to scales
+ self-regularisation of the rank observed in practice
☹ hard-to-obtain estimator that mingles optimisation and integration steps

Maximum *marginal* likelihood estimation

(Dikmen and Févotte, 2011, 2012)

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

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

+ better-posed than traditional NMF (fixed number of parameters)
+ better-behaved with respect to scales
+ self-regularisation of the rank observed in practice
☹ hard-to-obtain estimator that mingles optimisation and integration steps

Background

- inspired by ICA & sparse coding, latent Dirichlet allocation (LDA), statistical estimation with nuisance parameters
- fully-Bayesian methods treat both $\mathbf{W}$ and $\mathbf{H}$ as random parameters and aim at $p(\mathbf{W}, \mathbf{H}|\mathbf{V})$

Maximum *marginal* likelihood estimation

(Dikmen and Févotte, 2011, 2012)

Expectation-Maximisation : complete data $\mathbf{V}$ with $\mathbf{H}$ and optimise

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

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

Maximum *marginal* likelihood estimation

(Dikmen and Févotte, 2011, 2012)

Expectation-Maximisation : complete data $\mathbf{V}$ with $\mathbf{H}$ and optimise

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

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

(Dikmen and Févotte, 2011, 2012)

Expectation-Maximisation : complete data $\mathbf{V}$ with $\mathbf{H}$ and optimise

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

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

Setup

Models

	Gamma-Poisson	Gamma-Exponential
$p(\mathbf{V}\|\mathbf{WH})$	$\prod_{fn} \mathrm{Pois}(v_{fn}\|[\mathbf{WH}]_{fn})$ (Poisson)	$\prod_{fn} \mathrm{Exp}(v_{fn}\|[\mathbf{WH}]_{fn})$ (multiplicative exponential noise)
$p(\mathbf{H}\|\boldsymbol{\beta})$	$\prod_{kn} \mathrm{G}(h_{kn}\|\alpha_k, \beta_k)$	

Estimators

- Maximum marginal likelihood estimation (MMLE)

$$C_{\mathrm{ML}}(\mathbf{W}, \boldsymbol{\beta}) = -\log p(\mathbf{V}|\mathbf{W}, \boldsymbol{\beta})$$

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

- Maximum joint likelihood estimation (MJLE)

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

  Equivalent to penalised NMF.
  Optimisation with state-of-the-art majorisation-minimisation.

Scales

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

MMLE is scale-invariant

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

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

Scales

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

MMLE is scale-invariant

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

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

MJLE is not scale-invariant

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

$\Rightarrow$ degenerate solutions $\|\mathbf{W}\| \to \infty$, $\|\mathbf{H}\| \to 0$, $\|\boldsymbol{\beta}\| \to 0$.

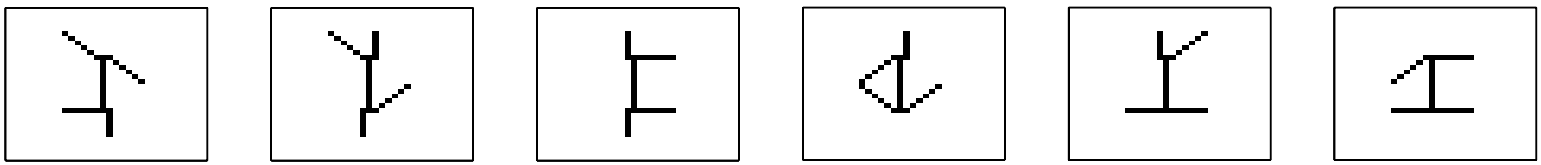
if $\alpha_k > 1$, we may set $\beta_k = 1$ and let $\mathbf{W}$, $\mathbf{H}$ free.
if $\alpha_k \leq 1$ the norm of $\mathbf{W}$ needs to be controlled.

swimmer data

Gamma-Exponential model

Swimmer dataset corrupted by multiplicative exponential noise.

(a) pure samples

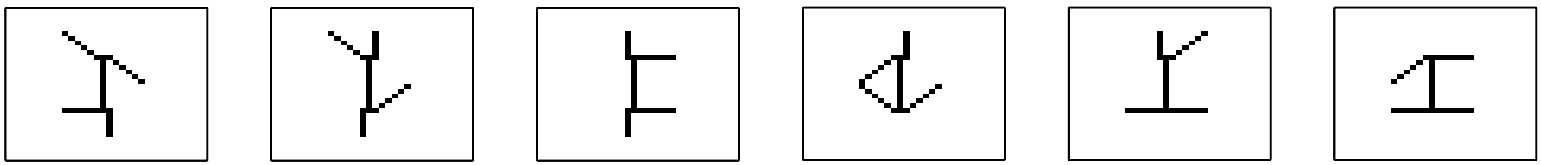


swimmer data

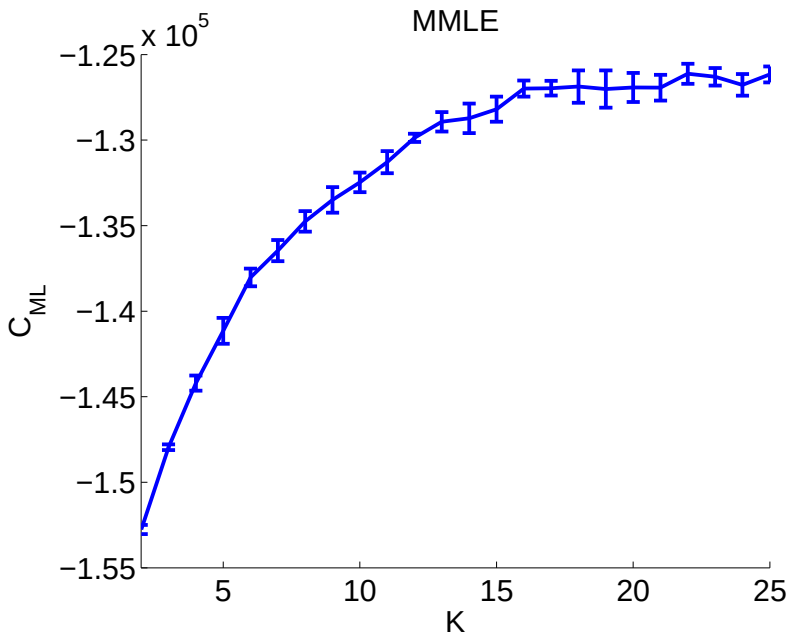
Gamma-Exponential model

Swimmer dataset corrupted by multiplicative exponential noise.

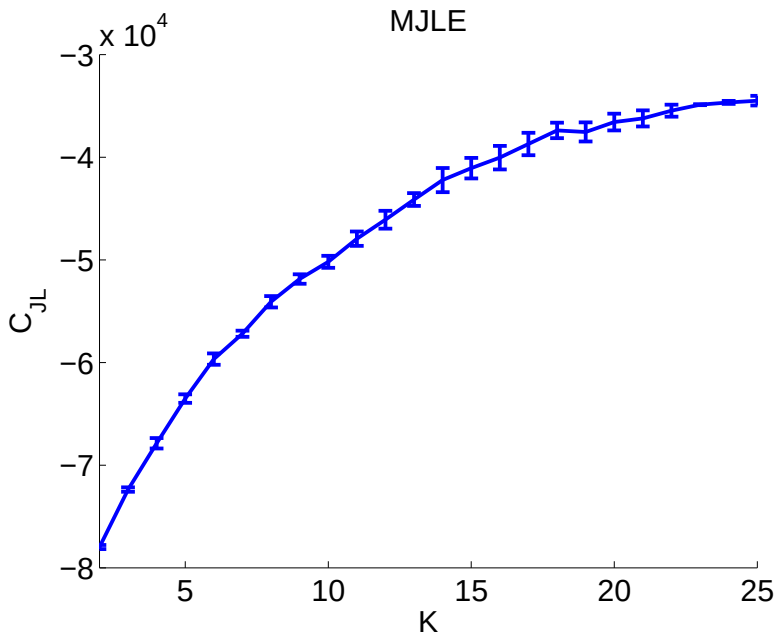
(a) pure samples



(b) marginal log-likelihoods

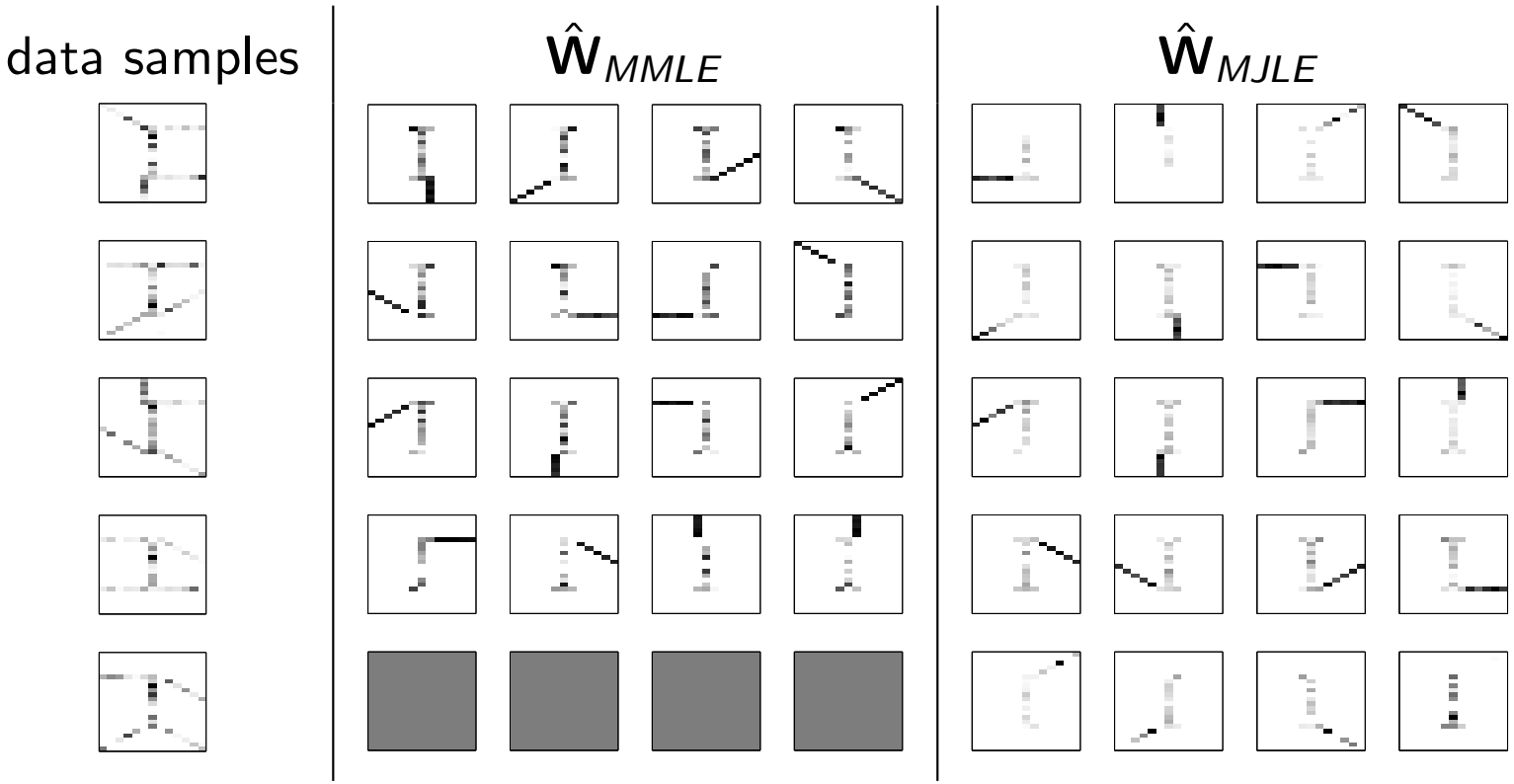


(c) joint log-likelihoods



swimmer data

Gamma-Exponential model



MMLE returns four null columns in $\hat{\mathbf{W}}$ (self-regularisation of rank).
MJLE overfits.

musiXmatch data

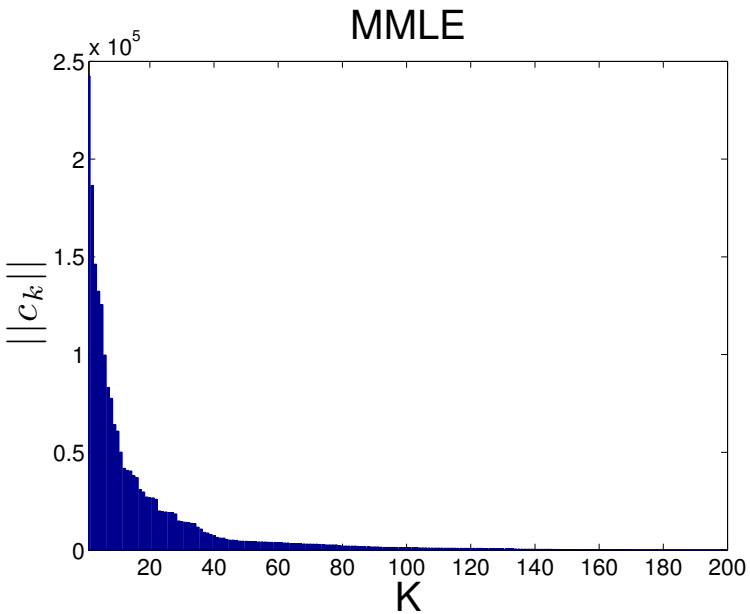
Gamma-Poisson model

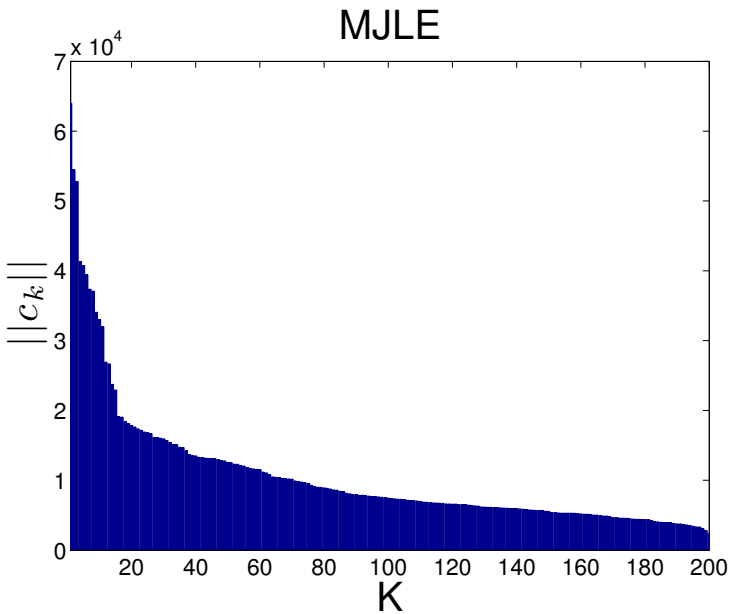
- lyrics of $10,000$ songs.
- bag-of-words representation of each song using the $5,000$ most frequent (stemmed) words.
- semantic analysis with MMLE and MJLE.
- $K = 200$.
- # occurrences of word f from topic k in song n is reconstructed by :

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

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

Norms $\|\hat{\mathbf{C}}_k\|$ of the components from the two estimators.





MMLE cancels out about 50 of the components (self-regularisation of rank).

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

(a) hip-hop

($k = 2$) 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

(b) metal

($k = 8$) 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

(c) girls

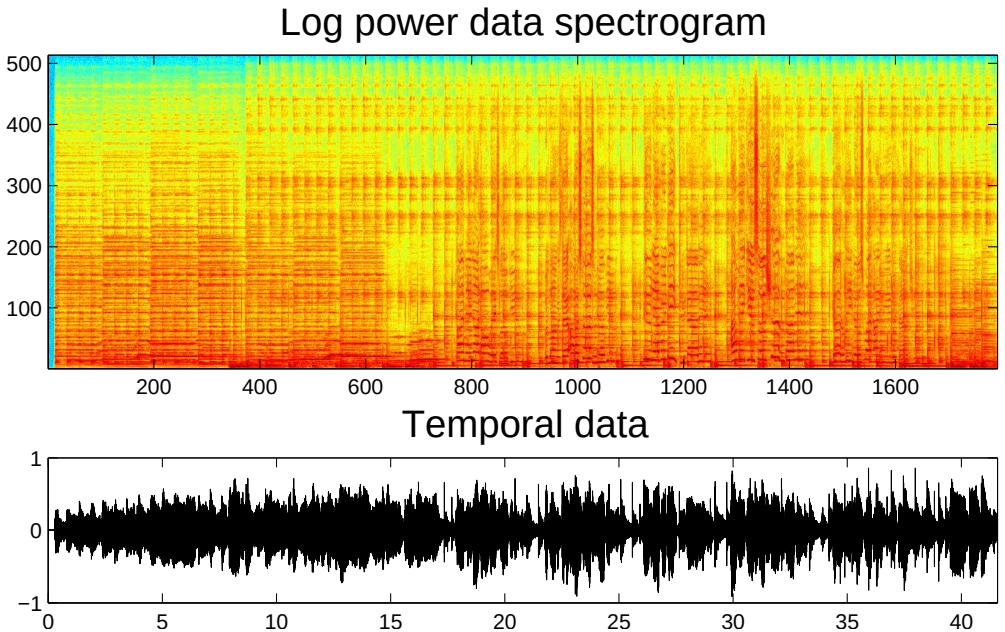
($k = 26$) she her girl beauti woman & queen sex sexi cloth herself doll shes pink gypsi bodi midnight callin dress hair	
Headhunter - Sex & Drugs & Rock'N Roll	& sex drug rock roll n is good veri inde and not my are all need dead bodi brain i
Holy Barbarians - She	she of kind girl my is the a littl 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

(d) French

($k = 13$) 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

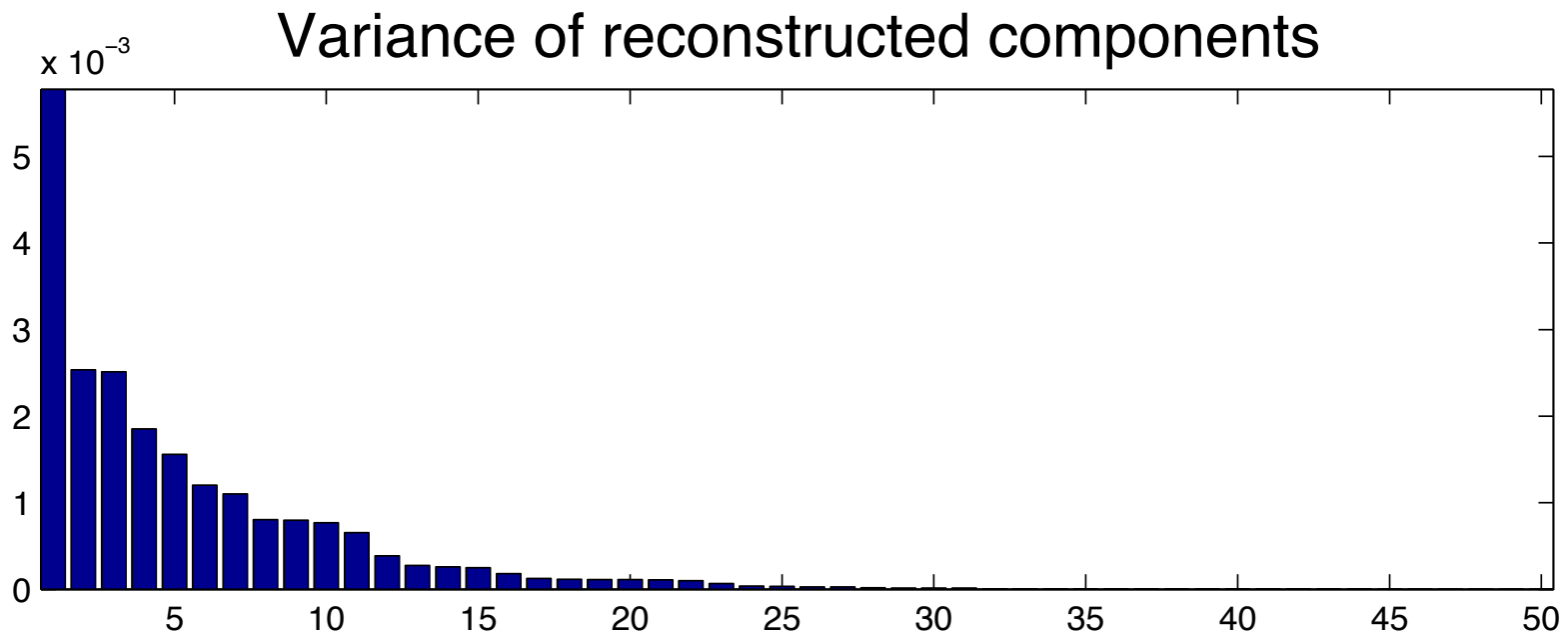
Gamma-Exponential model



- 40 seconds of *God Only Knows* by the Beach Boys.
- MMLE decomposition of the spectrogram $v_{fn} = |x_{fn}|^2$ with $K = 50$ components.
- Gamma-Exponential model shown to be a valid generative model of the spectrogram in (Févotte et al., 2009).
- component reconstruction $\hat{c}_{k,fn} = \frac{\hat{w}_{fk}\hat{h}_{kn}}{[\mathbf{WH}]_{fn}} x_{fn}$.

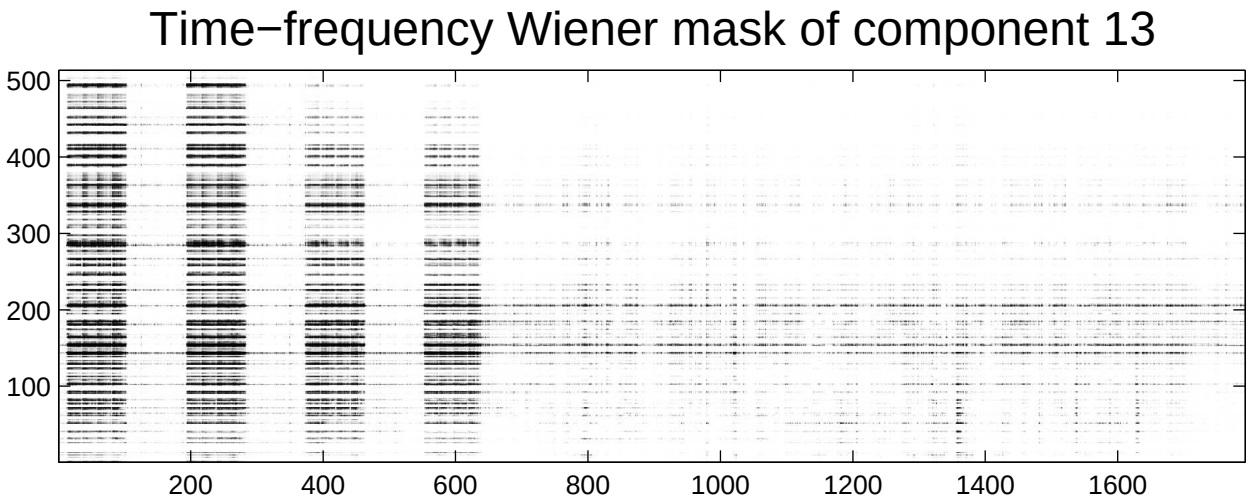
Spectral data

Gamma-Exponential model



Spectral data

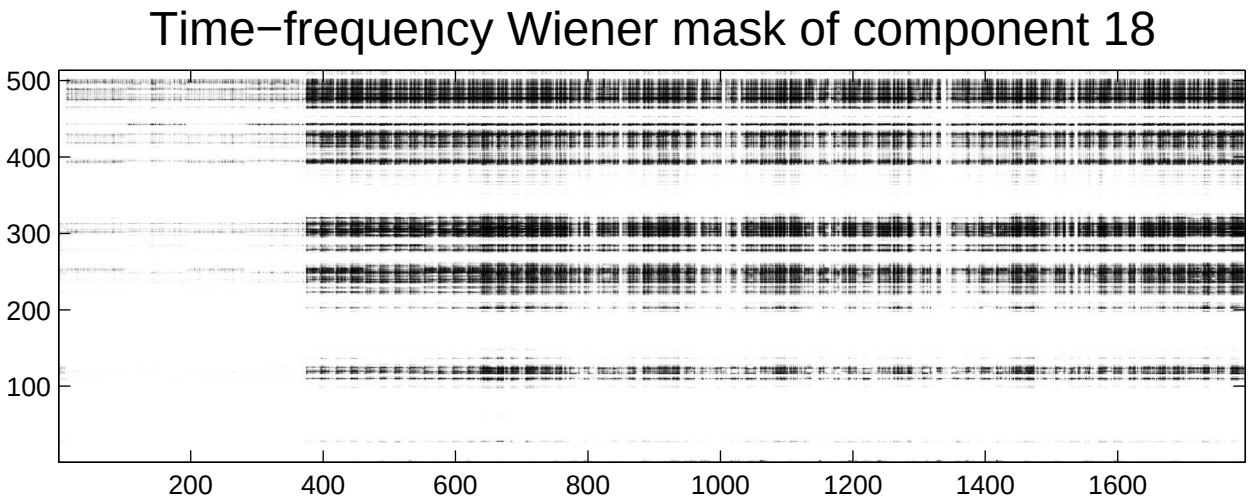
Gamma-Exponential model



audio

Spectral data

Gamma-Exponential model



audio

Conclusions

- MMLE leads to a better-posed estimator than MAP/MJLE
  - statistically well-posed (finite number of parameters)
  - scale-invariant
- MMLE found empirically to self-regularise rank (for the two models considered)
  - surprising and very appealing result
  - Laplace approximation of the marginal likelihood provides a start to explain this phenomenon, see (Dikmen and Févotte, 2012)
  - similar findings in Bayesian Matrix Factorisation in (Nakajima and Sugiyama, 2011; Nakajima et al., 2013), "model-induced regularisation"

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.

J. M. Bioucas-Dias, A. Plaza, N. Dobigeon, M. Parente, Q. Du, P. Gader, and J. Chanussot. 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, 2012.

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

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 *Proc. 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.

A. Cichocki, R. Zdunek, and S. Amari. Csiszar's divergences for non-negative matrix factorization : Family of new algorithms. In *Proc. International Conference on Independent Component Analysis and Blind Signal Separation (ICA)*, pages 32–39, Charleston SC, USA, Mar. 2006.

A. Cichocki, H. Lee, Y.-D. Kim, and S. Choi. Non-negative matrix factorization with α-divergence. *Pattern Recognition Letters*, 29(9) :1433–1440, July 2008.

I. S. Dhillon and S. Sra. Generalized nonnegative matrix approximations with Bregman divergences. In *Advances in Neural Information Processing Systems (NIPS)*, 2005.

References II

O. Dikmen and C. Févotte. Nonnegative dictionary learning in the exponential noise model for adaptive music signal representation. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2267–2275, Granada, Spain, Dec. 2011. URL `https://www.irit.fr/~Cedric.Fevotte/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 `https://www.irit.fr/~Cedric.Fevotte/publications/journals/ieee_sp_mmle.pdf`.

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 `https://www.irit.fr/~Cedric.Fevotte/publications/journals/neco11.pdf`.

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 `https://www.irit.fr/~Cedric.Fevotte/publications/journals/neco09_is-nmf.pdf`.

L. Finesso and P. Spreij. Nonnegative matrix factorization and I-divergence alternating minimization. *Linear Algebra and its Applications*, 416 :270–287, 2006.

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 (Statistical Methodology)*, 49(2) :127–162, 1987.

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

References III

D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural and Information Processing Systems 13*, pages 556–562, 2001.

S. Nakajima and M. Sugiyama. Theoretical analysis on bayesian matrix factorization. *Journal of Machine Learning Research*, 12 :2579–2644, 2011.

S. Nakajima, M. Sugiyama, S. D. Babacan, and R. Tomioka. Global analytic solution of fully-observed variational bayesian matrix factorization. *The Journal of Machine Learning Research*, 14(1) :1–37, 2013.

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.

P. Smaragdis. About this non-negative business. WASPAA keynote slides, 2013. URL `http://web.engr.illinois.edu/~paris/pubs/smaragdis-waspaa2013keynote.pdf`.

P. Smaragdis and J. C. Brown. Non-negative matrix factorization for polyphonic music transcription. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 2003.

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 `https://www.irit.fr/~Cedric.Fevotte/publications/journals/pami13_ardnmf.pdf`.

M. Tweedie. An index which distinguishes between some important exponential families. In *Proc. Indian Statistical Institute Golden Jubilee International Conference*, 1984.

References IV

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.

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.

Additional slides

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 :

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

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 :

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

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).