Maximum Marginal Likelihood Estimation For Nonnegative Dictionary Learning

Onur Dikmen and **Cédric Févotte**

CNRS LTCI; Télécom ParisTech

dikmen@telecom-paristech.fr
perso.telecom-paristech.fr/∼dikmen

26.5.2011

Outline

Problem & Motivation

- NMF: Approximate $\mathbf{V} \approx \mathbf{WH}$ by minimising $D(\mathbf{V}|\mathbf{WH})$ (Kullback-Leibler (KL), Itakura-Saito (IS), Euclidean)

$$D_{KL}(\mathbf{A}|\mathbf{B}) = \sum_{f=1}^{F}\sum_{n=1}^{N}\left(a_{fn}\log\frac{a_{fn}}{b_{fn}} - a_{fn} + b_{fn}\right)$$

- $\mathbf{W}$: dictionary ($F \times K$), $\mathbf{H}$: expansion coefficients ($K \times N$), nonnegative

- Efficient majorization-minimization algorithms (optimisation of an auxiliary function)
- Expectation Maximisation (EM) algorithm (Poisson observation model)

- Optimality of $\mathbf{W}$ is in question ($FK + KN$ parameters)

Problem & Motivation

- Our aim: to learn $\mathbf{W}$ from the marginal likelihood

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

- $p(\mathbf{V}|\mathbf{W})$: marginal likelihood of $\mathbf{W}$, not evidence $p(\mathbf{V})$
- Not directly possible. Variational approximation will be pursued.

Model

- Minimising $D_{KL}(\mathbf{V}|\mathbf{WH})$ = ML on the Poisson observation model

$$v_{fn} \sim \mathcal{PO}(v_{fn}|\sum_k w_{fk}h_{kn})$$

- With the introduction of latent variables $\mathbf{C}$

$$v_{fn} = \sum_{k=1}^{K} c_{k,fn}, \quad c_{k,fn} \sim \mathcal{PO}(c_{k,fn}|w_{fk}h_{kn})$$

- We assign a prior distribution on $\mathbf{H}$

$$h_{kn} \sim \mathcal{G}(h_{kn}|\alpha_k, \beta_k)$$

- $\mathcal{G}$ is conjugate prior for $\mathcal{PO}$ observation model
- $\mathbf{W}$ is a deterministic variable

Estimators & Algorithms (MJLE)

- Maximum joint likelihood estimation (MJLE)

$$C_{JL}(\mathbf{V}|\mathbf{W},\mathbf{H}) \stackrel{\text{def}}{=} \log p(\mathbf{V},\mathbf{H}|\mathbf{W}) = \log p(\mathbf{V}|\mathbf{W},\mathbf{H}) + \log p(\mathbf{H})$$

- Optimisation w.r.t $\mathbf{H}$ can be performed by majorization-minimization

$$h_{kn} \leftarrow \frac{h_{kn}\sum_f w_{fk}v_{fn}/[\mathbf{WH}]_{fn} + (\alpha_k - 1)}{\sum_f w_{fk} + \beta_k}$$

- Nonnegativity is ensured when $\alpha_k \geq 1$
- Akin to standard NMF

$$w_{fk} \leftarrow w_{fk}\frac{\sum_n h_{kn}v_{fn}/[\mathbf{WH}]_{fn}}{\sum_n h_{kn}}$$

Estimators & Algorithms (MMLE)

- Maximum marginal likelihood estimation (MMLE)

$$C_{ML}(\mathbf{V}|\mathbf{W}) \overset{\text{def}}{=} \log p(\mathbf{V}|\mathbf{W}) = \log \int_{\mathbf{H}} p(\mathbf{V}|\mathbf{W},\mathbf{H})p(\mathbf{H})\, d\mathbf{H}$$

- We cannot obtain $\log p(\mathbf{V}|\mathbf{W})$ analytically.
- EM algorithm

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

- Again, no analytical solution
- Approximate the posterior $p(\mathbf{C},\mathbf{H}|\mathbf{V},\tilde{\mathbf{W}})$ using variational Bayes or MCMC

Estimators & Algorithms (MMLE with VBEM)

- Approximate $p(\mathbf{C}, \mathbf{H}|\mathbf{V}, \tilde{\mathbf{W}})$ with a variational distribution (independent Gamma and Multinomial distributions)

$$q(\mathbf{C}, \mathbf{H}) = \prod_{f=1}^{F} \prod_{n=1}^{N} q(\mathbf{c}_{fn}) \prod_{k=1}^{K} \prod_{n=1}^{N} q(h_{kn})$$

by minimising

$$\mathrm{KL}(q(\mathbf{C}, \mathbf{H})||p(\mathbf{C}, \mathbf{H}|\mathbf{V}, \mathbf{W})) = \log p(\mathbf{V}|\mathbf{W}) + \mathrm{KL}(q(\mathbf{C}, \mathbf{H})||p(\mathbf{V}, \mathbf{C}, \mathbf{H}|\mathbf{W}))$$

- Fixed point equations

$$\log q(h_{kn}) =^{+} \langle \log p(\mathbf{V}, \mathbf{C}, \mathbf{H}|\mathbf{W}) \rangle_{q(\mathbf{H}_{-(kn)})q(\mathbf{C})}$$
$$\log q(\mathbf{c}_{fn}) =^{+} \langle \log p(\mathbf{V}, \mathbf{C}, \mathbf{H}|\mathbf{W}) \rangle_{q(\mathbf{H})q(\mathbf{C}_{-(fn)})}$$

Estimators & Algorithms (MMLE with VBEM)

- Approximate E-step

$$\hat{Q}(\mathbf{W}|\tilde{\mathbf{W}}) \equiv \int_{\mathbf{H}} \log p(\mathbf{V}, \mathbf{C}, \mathbf{H}|\mathbf{W}) q(\mathbf{C}, \mathbf{H})\, d\mathbf{C}\, d\mathbf{H}$$

- $\hat{Q}(\mathbf{W}|\tilde{\mathbf{W}})$ is analytically available because $q(\mathbf{C}, \mathbf{H})$ can be factorized
- Multiplicative update rules for $\mathbf{W}$:

$$w_{fk} \leftarrow w_{fk} \frac{\sum_n \exp(\langle \log h_{kn} \rangle) v_{fn} / [\mathbf{W} \exp(\langle \log \mathbf{H} \rangle)]_{fn}}{\sum_n \langle h_{kn} \rangle}$$

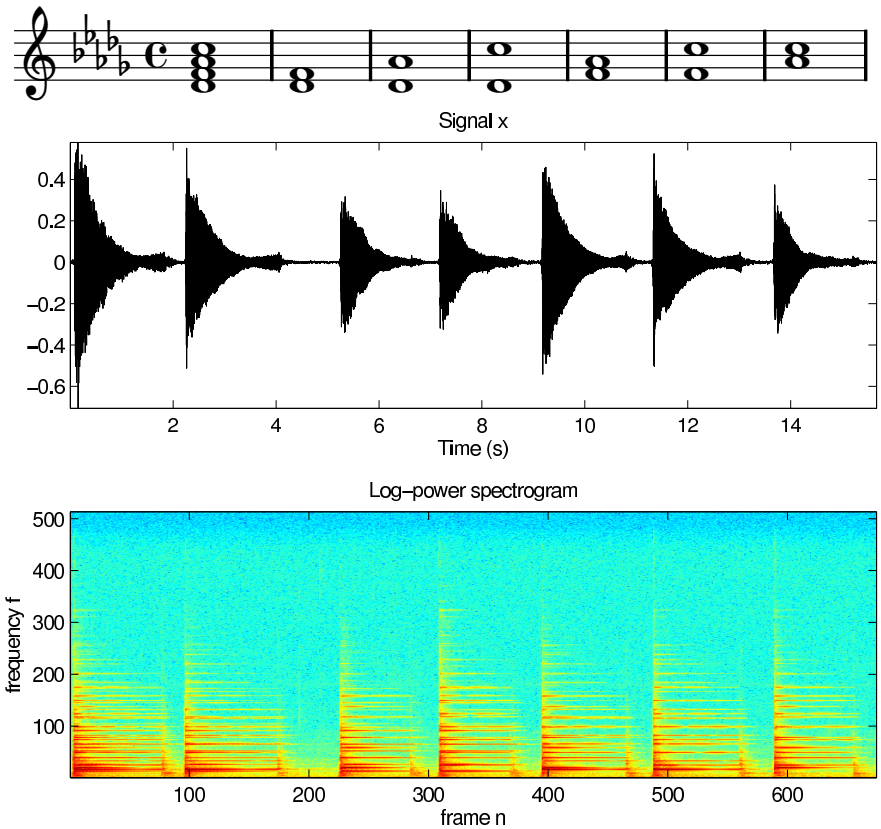
- $-\mathrm{KL}(q||p(\mathbf{V}, \mathbf{C}, \mathbf{H}|\mathbf{W}))$ is a lower bound for $\log p(\mathbf{V}|\mathbf{W})$

Relation to Other Works

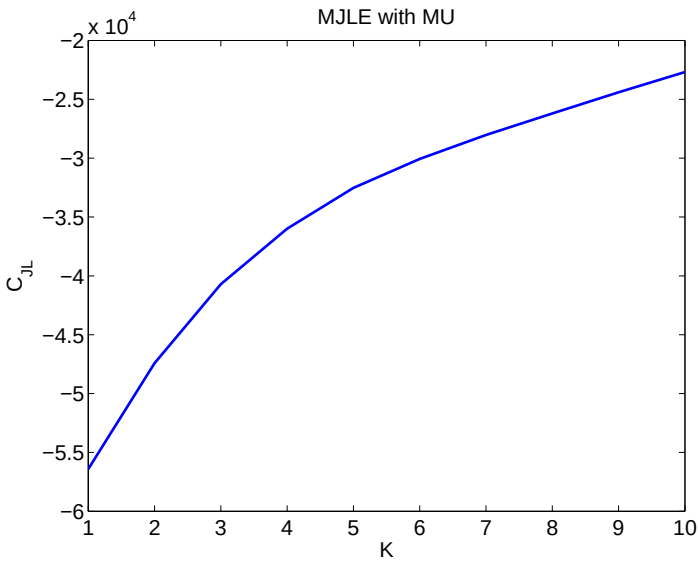
- Our model coincides with the Gamma-Poisson model in Canny'04. MJLE is derived there also.
- In Buntine'06, $\mathbf{W}$ has Dirichlet priors and inferred using variational Bayes.
- In Cemgil'09, $\mathbf{W}$ has Gamma priors. Model selection is done using $p(\mathbf{V})$, after full Bayesian treatment.

A short piano excerpt

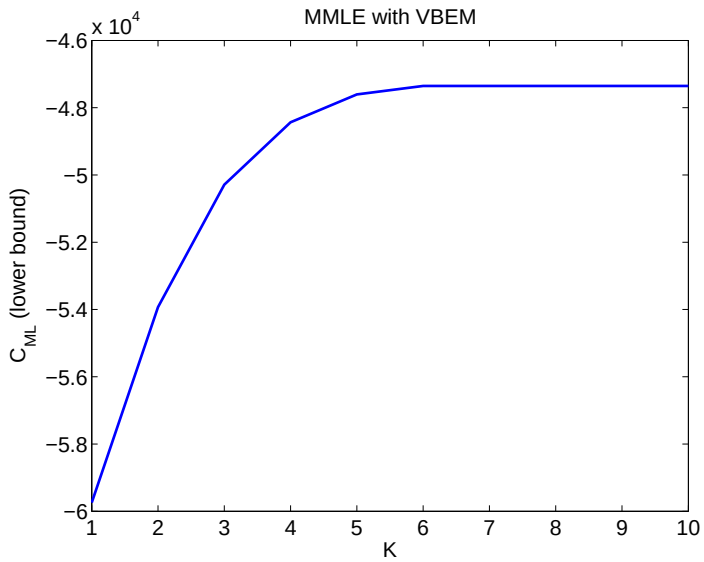
- 15 sec., 22.05 kHz., 16 bits
- STFT, overlapping windows of size 1024
- $F = 513$, $N = 676$
- Combinations of 4 notes are played



Likelihoods vs. Component Number

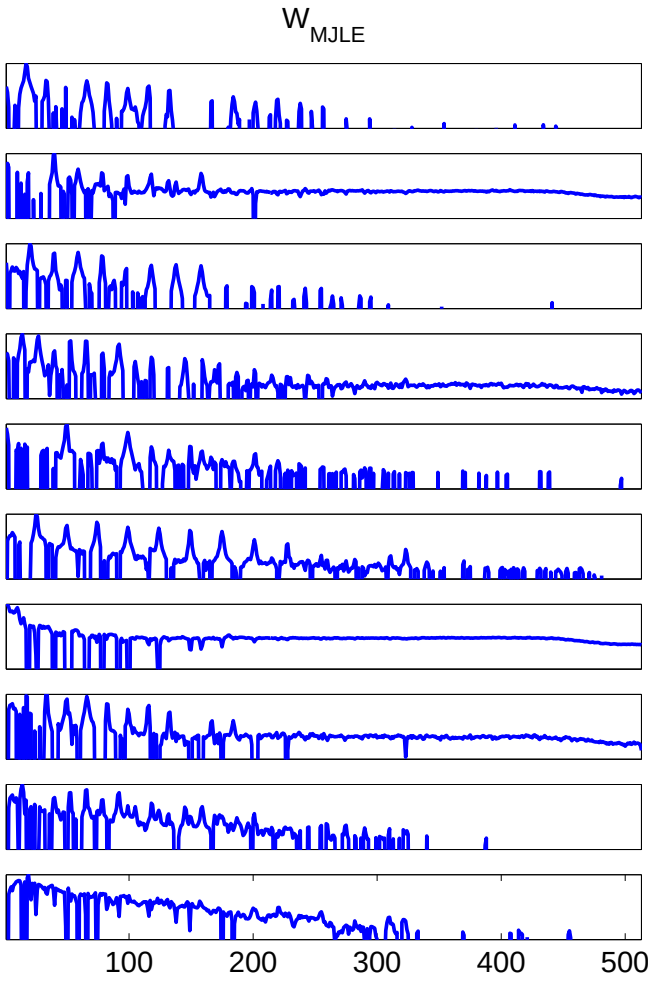


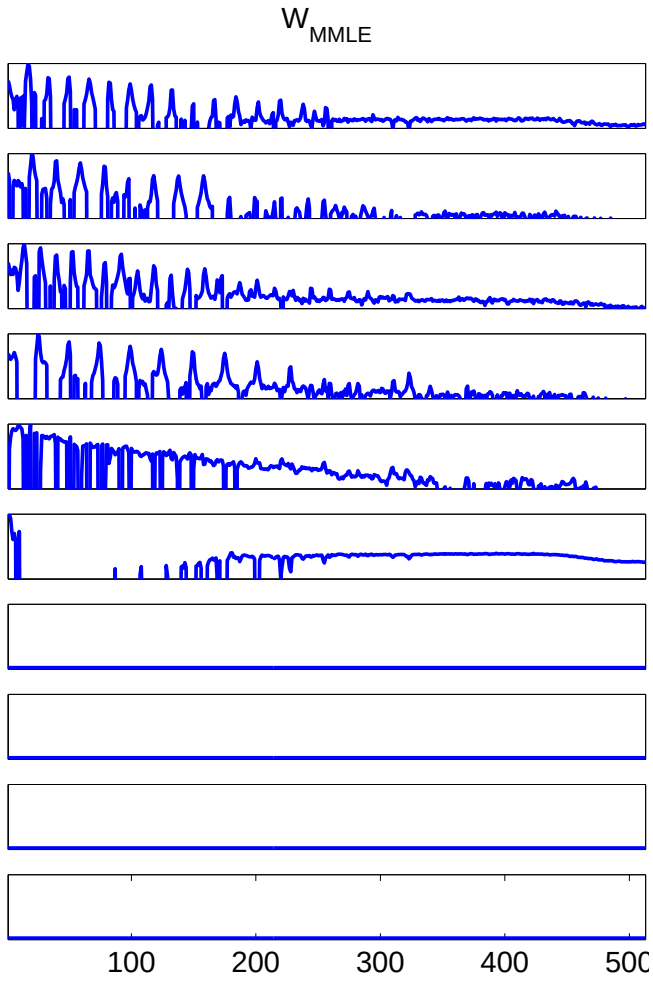
(a) Joint likelihood



(b) Marginal likelihood

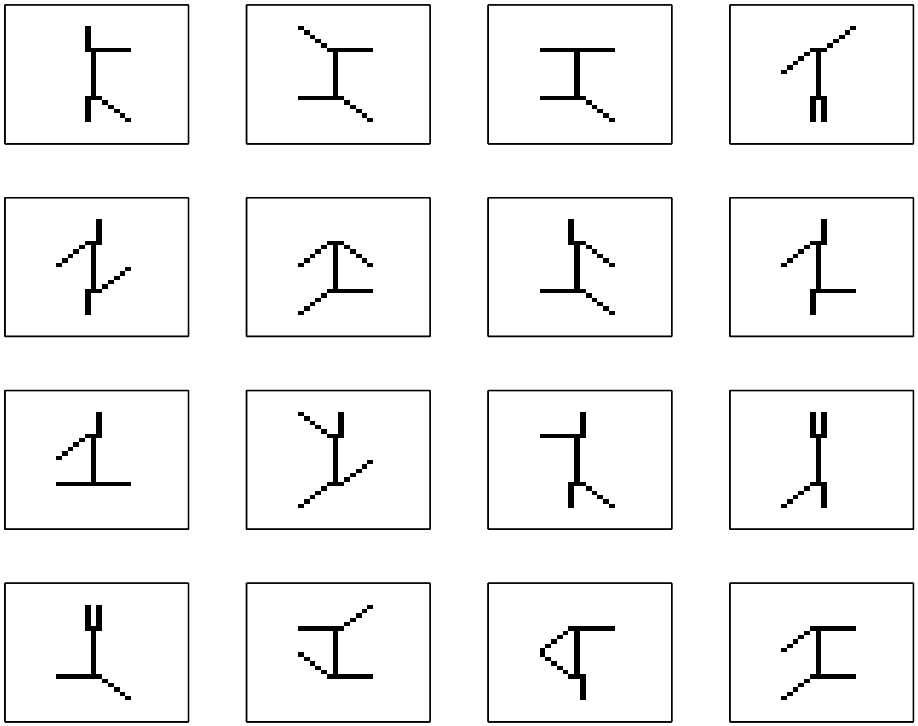
Estimated Dictionaries



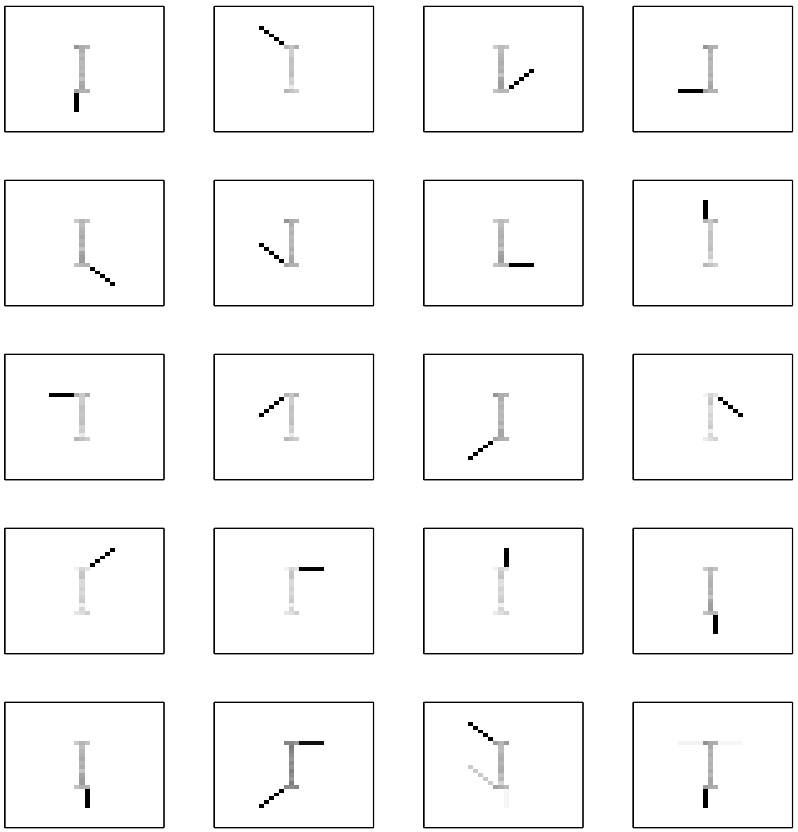


Swimmer Dataset

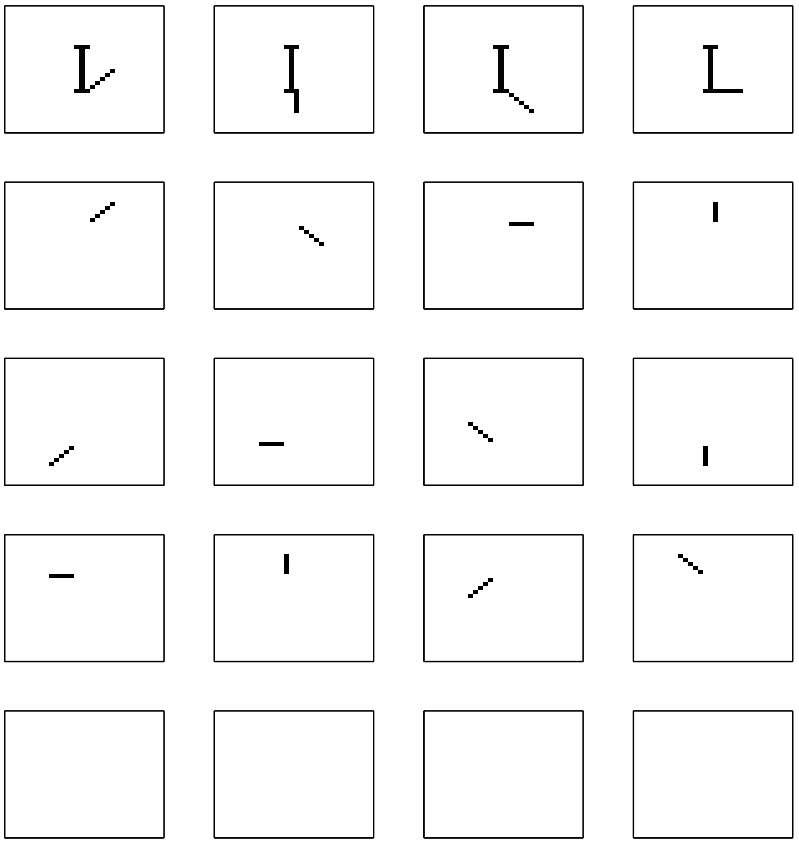
- 4 joints / 4 angles
- 256 figures, 32x32 pixels each



Estimated Dictionaries



(a) $\mathbf{W}_{\mathrm{MJLE}}$



(b) $\mathbf{W}_{\mathrm{MMLE}}$

Conclusions

- Two approaches: Joint likelihood versus marginal likelihood
- MMLE with VBEM has comparable complexity to MJLE
- When K_{opt} is used, they perform similarly

- MMLE has an intrinsic way of selecting the model order by automatically cancelling "irrelevant" columns in $\mathbf{W}$
- More efficient than computing and comparing $p(\mathbf{V})$ for many values of K in a full Bayesian setting