Automatic Relevance Determination in Nonnegative Matrix Factorization with the β-Divergence

VINCENT Y. F. TAN[†], CÉDRIC FÉVOTTE[§]

[†]Department of ECE, University of Wisconsin-Madison (email: vtan@wisc.edu)

[§]CNRS LTCI, TELECOM ParisTech (email: fevotte@telecom-paristech.fr)

Nonnegative Matrix Factorization

- Given a nonnegative data matrix $\mathbf{V} \in \mathbb{R}_+^{F\times N}$, find a decomposition
$$\mathbf{V} \approx \hat{\mathbf{V}} \triangleq \mathbf{W}\mathbf{H}$$
where $\mathbf{W} \in \mathbb{R}_+^{F\times K}$ (basis) and $\mathbf{H} \in \mathbb{R}_+^{K\times N}$ (coefficients)
- Common dimension K is chosen such that $FK + KN \ll FN$
- Overall number of parameters to describe the data is reduced
- Alternating minimization usually performed
$$\min_{\mathbf{H}\geq 0} D(\mathbf{V}|\mathbf{W}\mathbf{H}), \qquad \min_{\mathbf{W}\geq 0} D(\mathbf{V}|\mathbf{W}\mathbf{H})$$
- The measure of fit $D(\mathbf{V}|\mathbf{W}\mathbf{H})$ is separable, i.e.,
$$D(\mathbf{V}|\hat{\mathbf{V}}) = \sum_{f=1}^{F}\sum_{n=1}^{N} d(v_{fn}|\hat{v}_{fn})$$
- We take the scalar cost function $d(v_{fn}|\hat{v}_{fn})$ to be the so-called β-divergence. Special cases include:

$\beta = 0$: Itakura-Saito divergence

$\beta = 1$: (Generalized) Kullback-Leibler divergence

$\beta = 2$: (Squared) Euclidean distance

Main Contributions

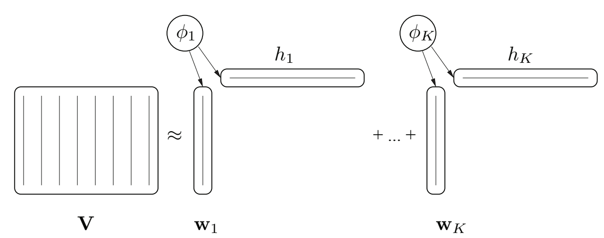
- In practical applications, model order K is hard to choose
- If K is too large $\Rightarrow$ overfitting ☹
- If K is too small $\Rightarrow$ data does not fit well to the model ☹
- Bayesian NMF model based on Automatic Relevance Determination to estimate K and get a better decomposition ☺
- ARD has been previously employed in Bayesian PCA (Bishop 1999) and sparse Bayesian learning (Tipping 2001).

Majorization-Minimization (MM) for β-NMF

- Algorithms are based on the MM framework
- Let the cost function to be minimized be $C(\mathbf{H})$
- Build an auxiliary function $G(\mathbf{H}|\tilde{\mathbf{H}})$ such that
$$G(\mathbf{H}|\tilde{\mathbf{H}}) \geq C(\mathbf{H}), \qquad G(\tilde{\mathbf{H}}|\tilde{\mathbf{H}}) = C(\tilde{\mathbf{H}})$$
- Then, optimizing $G(\cdot|\mathbf{H}^{(i)})$ yields
$$C(\mathbf{H}^{(i+1)}) \leq G(\mathbf{H}^{(i+1)}|\mathbf{H}^{(i)}) \leq G(\mathbf{H}^{(i)}|\mathbf{H}^{(i)}) = C(\mathbf{H}^{(i)})$$
- Hence, the MM updates consists in performing
$$\mathbf{H}^{(i+1)} = \arg\min_{\mathbf{H}\geq 0} G(\mathbf{H}|\mathbf{H}^{(i)}).$$
- For β-NMF, $G(\mathbf{H}|\tilde{\mathbf{H}})$ can be found (Févotte and Idier 2011) and this amounts to the simple multiplicative update rule
$$h_{kn} = \tilde{h}_{kn}\left(\frac{p_{kn}}{q_{kn}}\right)^{\gamma(\beta)}$$
where p_{kn}, q_{kn} are simple functions of the previous iterate and the data and $\gamma(\beta)$ is simple exponent in β.

The Model for ARD in β-NMF

- Main idea is to tie column $\mathbf{w}_k$ and row h_k through their prior, a common (variance-like) relevance weight ϕ_k



- For ℓ_2-ARD, each element of the matrices $\mathbf{W}$ and $\mathbf{H}$ is assigned a Half-Normal Prior
$$p(h_{kn}|\phi_k) = \sqrt{\frac{2}{\pi\phi_k}}\exp\left(-\frac{h_{kn}^2}{2\phi_k}\right)$$
- For ℓ_1-ARD, each element of the matrices $\mathbf{W}$ and $\mathbf{H}$ is assigned an Exponential Prior
$$p(h_{kn}|\phi_k) = \frac{1}{\phi_k}\exp\left(-\frac{h_{kn}}{\phi_k}\right)$$
- Relevance weights ϕ_k assigned inverse-Gamma priors:
$$p(\phi_k; a, b) = \frac{b^a}{\Gamma(a)}\exp\left(-\frac{b}{\phi_k}\right)$$
- Different β's underlie different statistical noise models:

IS-NMF :	$\beta = 0$	$v_{fn} \sim \mathcal{G}(v_{kn}\|\alpha, \hat{v}_{fn}/\alpha)$
KL-NMF :	$\beta = 1$	$v_{fn} \sim \mathcal{P}(v_{kn}\|\hat{v}_{fn})$
EUC-NMF :	$\beta = 2$	$v_{fn} \sim \mathcal{N}(v_{kn}\|\hat{v}_{fn}, \sigma^2)$

- The likelihood is given by
$$-\log p(\mathbf{V}|\mathbf{W},\mathbf{H}) = \rho\, D_\beta(\mathbf{V}|\mathbf{W}\mathbf{H}) + \mathrm{cst}$$
where ρ is some regularization constant reflecting our belief in the noise power, e.g., $\rho = 1/\sigma^2$ for $\beta = 2$.

The Overall Cost Function (Posterior)

- Combining the likelihood and the priors gives the cost function (posterior): $C(\mathbf{W},\mathbf{H},\boldsymbol{\lambda}) = -\log p(\mathbf{W},\mathbf{H},\boldsymbol{\lambda}|\mathbf{V}) =$
$$\rho\, D_\beta(\mathbf{V}|\mathbf{W}\mathbf{H}) + \sum_{k=1}^{K}\frac{1}{\phi_k}\left(f(\mathbf{w}_k) + f(h_k) + b\right) + c\log\phi_k + \mathrm{cst}.$$
where $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_2^2, c = (F+N)/2 + a + 1$ for ℓ_2-ARD and $f(\mathbf{x}) = \|\mathbf{x}\|_1, c = F + N + a + 1$ for ℓ_1-ARD
- MAP optimization $\Rightarrow$ some relevance weights converge to a small constant and corresponding components are pruned
- Optimizing only over $\boldsymbol{\lambda}$ leads to the cost function $C(\mathbf{W},\mathbf{H}) =$
$$\rho\, D_\beta(\mathbf{V}|\mathbf{W}\mathbf{H}) + c\sum_{k=1}^{N}\log(f(\mathbf{w}_k) + f(h_k) + b) + \mathrm{cst}$$
- This cost has connections to reweighted ℓ_1 minimization [Candès et al. 2008] and group LASSO [Yuan and Lin 2006]

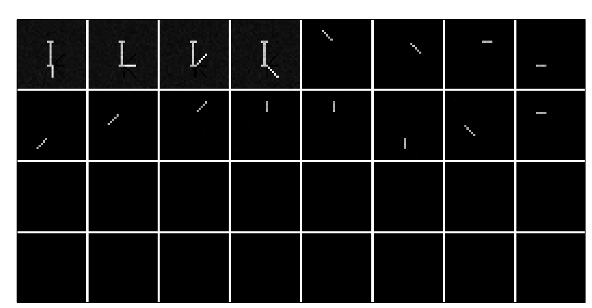
The Inference Algorithms

- Build auxiliary functions to optimize $C(\mathbf{W},\mathbf{H},\boldsymbol{\lambda})$ over $\mathbf{H}$
- In the end, the multiplicative updates are
$$\ell_2 - \mathrm{ARD}: \quad h_{kn} = \tilde{h}_{kn}\left(\frac{p_{kn}}{q_{kn} + \tilde{h}_{kn}/(\rho\,\phi_k)}\right)^{\xi(\beta)}$$
$$\ell_1 - \mathrm{ARD}: \quad h_{kn} = \tilde{h}_{kn}\left(\frac{p_{kn}}{q_{kn} + 1/(\rho\,\phi_k)}\right)^{\gamma(\beta)}$$
- Updates of ϕ_k proceed as follows
$$\phi_k = \frac{f(\mathbf{w}_k) + f(h_k) + b}{c}$$
- Because we use MM-based algorithms, the cost function $C(\mathbf{W},\mathbf{H},\boldsymbol{\lambda})$ decreases monotonically
- Choose the hyperparameter b using the method of moments:
$$\ell_2 - \mathrm{ARD}: \quad \hat{b} = \frac{\pi(a-1)\hat{\mu}_{\mathbf{V}}}{2K}$$
$$\ell_1 - \mathrm{ARD}: \quad \hat{b} = \left(\frac{(a-1)(a-2)\hat{\mu}_{\mathbf{V}}}{K}\right)^{1/2}$$

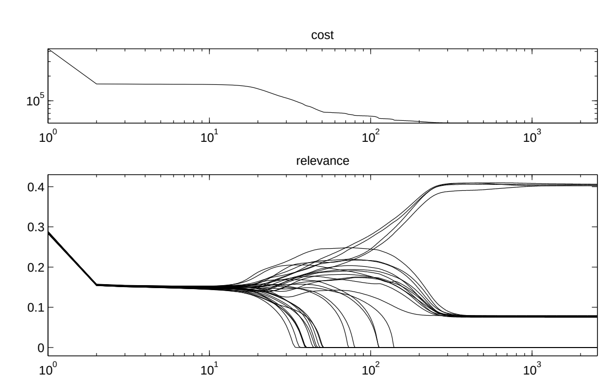
Noisy Swimmer Dataset

- Synthetic dataset of $N = 256$ images size $F = 32 \times 32 = 1024$
- Each image represents a swimmer composed of an invariant torso and four limbs. Each limb in four different positions.

- Dictionary learned using one run of ℓ_1-ARD with $a = 100$

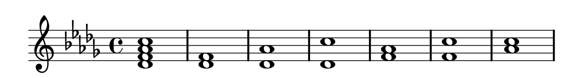


- Monotonic decrease in cost and evolution of relevances

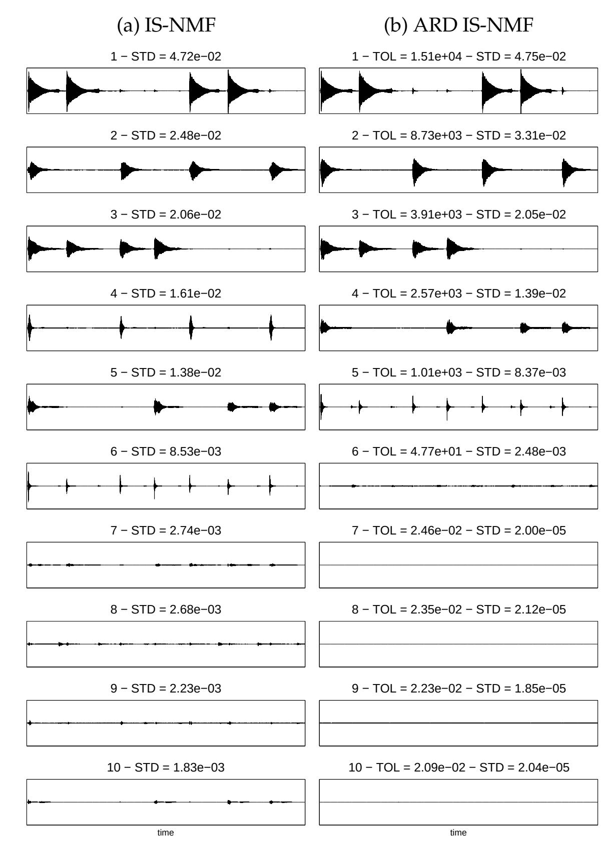


Music Decomposition Example

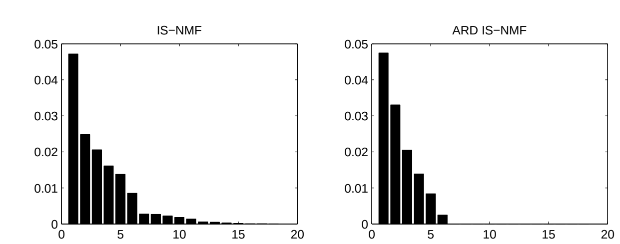
- Use the Itakura-Saito (IS) divergence ($\beta = 0$) which is suited for audio applications (Févotte et al. 2009)
- Underlies a generative statistical model of superimposed Gaussian components in the squared STFT domain
- Sequence is composed of 4 piano notes, played all at once in the first measure and then played by pairs in all possible combinations in the subsequent measures.



- Given an approximate factorization $\mathbf{W}\mathbf{H}$ of the spectrogram $v_{fn} = |x_{fn}|^2$, the STFT estimate $\hat{c}_{k,fn}$ of component k is
$$\hat{c}_{k,fn} = \frac{w_{fk}h_{kn}}{\sum_j w_{fj}h_{jn}}x_{fn}$$



- Histogram shows that 6 components retained by ℓ_1-ARD



- First four components extract the individual notes and the next two components extract the sound of hammer hitting the strings and the sound produced by the sustain pedal