Supplementary material to "Automatic Relevance Determination in Nonnegative Matrix Factorization with the β-Divergence"

Vincent Y. F. Tan, Cédric Févotte

April 22, 2012

Equation numbers refer to those in the main paper. All equations here are appended with an A-.

1 Connection to group LASSO and reweighted ℓ_1-minimization

In this section, we comment on how the $\boldsymbol{\lambda}$-optimized cost function $C(\mathbf{W},\mathbf{H})$ in (21) is related to group LASSO [1] and reweighted ℓ_1-minimization [2]. By using Lagrange multiplers, it can be shown that the optimization of $C(\mathbf{W},\mathbf{H})$ over $\mathbf{H}$ is equivalent to the following non-convex minimization problem over $\mathbf{H}$:

$$\begin{aligned} \underset{\mathbf{H}}{\text{minimize}} \quad & \sum_k \log(f(\mathbf{w}_k)+f(\underline{h}_k)+b)) \\ \text{subject to} \quad & \mathbf{H}\geq 0, \quad D_\beta(\mathbf{V}|\mathbf{WH})\leq \delta \end{aligned} \tag{A-1}$$

The equivalence can be formalized as follows: For a particular c in (21), there is a corresponding $\delta>0$ in the optimization in (A-1). We focus on ℓ_1-ARD where $f(\mathbf{x})=\|\mathbf{x}\|_1$. Then the objective is concave in $\mathbf{H}$. One natural way to solve (A-1) iteratively is to use an MM procedure by upper bounding the objective function with its tangent (first-order Taylor expansion) at the current iterate $\mathbf{H}$. This yields the following convex program

$$\begin{aligned} \underset{\mathbf{H}}{\text{minimize}} \quad & \sum_{k,n} \frac{h_{kn}}{\sum_f w_{fk}+\sum_{n'}\tilde{h}_{kn'}+b} \\ \text{subject to} \quad & \mathbf{H}\geq 0, \quad D_\beta(\mathbf{V}|\mathbf{WH})\leq \delta \end{aligned} \tag{A-2}$$

The notation $\tilde{h}_{kn}$ denotes the parameter h_{kn} at the previous iteration. Note that if $\beta=2$, $D_\beta(\mathbf{V}|\mathbf{WH})$ is (one-half) the square of the Frobenius norm of $\mathbf{V}-\mathbf{WH}$ and so the optimization problem in (A-2) is in fact a quadratically-constrained ℓ_1-minimization problem, which can be solved efficiently. Furthermore, defining the *weights*

$$\theta_k \triangleq \frac{1}{\sum_f w_{fk}+\sum_{n'}\tilde{h}_{kn'}+b}, \tag{A-3}$$

we see that the objective function in (A-2) is $\sum_k \theta_k \sum_n h_{kn}$. This linear combination in the objective reinforces that $\mathbf{w}_k$ and $\underline{h}_k$ are intimately tied together; if $\|\mathbf{w}_k\|_1$ is small, then weight assigned to the entire k^{th} row of $\mathbf{H}$, namely θ_k, would be large. This penalizes the k^{th} row of $\mathbf{H}$ heavily, thus forcing its ℓ_1 norm to go toward zero at the next iteration. Note that this is also the intuition and interpretation behind the success of the reweighted ℓ_1-minimization by Candès et al. [2]. However, for our Bayesian model, only K distinct weights (the θ_k's) are required for the KN elements in $\mathbf{H}$. Each row of $\mathbf{H}$ (a group) is assigned a *single* weight θ_k and all the elements in a particular row of $\mathbf{H}$ are penalized *equally*, while different rows are penalized differently. As in group Lasso [1], this has the effect of sparsifying the matrix $\mathbf{H}$ row-wise (and the matrix $\mathbf{W}$ column-wise).

2 Detailed derivation of ℓ_1-ARD for β-NMF

In this algorithm, we assume that $\mathbf{W}$ and $\mathbf{H}$ have Exponential priors as in (13) and thus, the regularizer can be expressed as

$$R_1(\mathbf{H}) \triangleq \sum_k \frac{1}{\lambda_k} f(\underline{h}_k) = \sum_{kn} \frac{1}{\lambda_k} h_{kn}. \tag{A-4}$$

We will derive the updates for two separate cases: $\beta < 1$ and $\beta \geq 1$. For $\beta < 1$, we do not need to modify the auxiliary function $F(\mathbf{H}|\tilde{\mathbf{H}}) \triangleq G(\mathbf{H}|\tilde{\mathbf{H}}) + R_1(\mathbf{H})$ to get a simple update rule. Indeed, for this case, it can be seen from Table 1 that

$$F(\mathbf{H}|\tilde{\mathbf{H}}) = \sum_{kn} h_{kn} \left(\frac{1}{\phi} q_{kn} + \frac{1}{\lambda_k} \right) - \frac{p_{kn} \tilde{h}_{kn}}{\phi(\beta - 1)} \left(\frac{h_{kn}}{\tilde{h}_{kn}} \right)^{\beta - 1}. \tag{A-5}$$

Differentiating $F(\mathbf{H}|\tilde{\mathbf{H}})$ w.r.t. h_{kn} and setting the result to zero yields the update rule

$$h_{kn} = \tilde{h}_{kn} \left(\frac{p_{kn}}{q_{kn} + \phi/\lambda_k} \right)^{1/(2-\beta)}. \tag{A-6}$$

Note that the exponent $1/(2-\beta)$ corresponds to the $\beta < 1$ case in the definition of $\gamma(\beta)$ in (11). For $\beta \geq 1$, we need to leverage on Lemma 1. Setting $\nu = h_{kn}/\tilde{h}_{kn}$ in Lemma 1, we can conclude that for $\beta \geq 1$,

$$\frac{h_{kn}}{\tilde{h}_{kn}} - 1 \leq \frac{1}{\beta} \left[\left(\frac{h_{kn}}{\tilde{h}_{kn}} \right)^{\beta} - 1 \right]. \tag{A-7}$$

In other words,

$$\frac{1}{\lambda_k} h_{kn} \leq \frac{1}{\beta \lambda_k} \tilde{h}_{kn} \left(\frac{h_{kn}}{\tilde{h}_{kn}} \right)^{\beta} + \text{cst}. \tag{A-8}$$

As in the $\beta \geq 2$ case in ℓ_2-ARD, we replace the regularizer $R_1(\mathbf{H})$ in (A-4) with the upper bound in (A-8). Zeroing the gradient of the resulting auxiliary function yields

$$h_{kn} = \tilde{h}_{kn} \left(\frac{p_{kn}}{q_{kn} + \phi/\lambda_k} \right)^{\gamma(\beta)}. \tag{A-9}$$

We omit the details. We emphasize that the update in (A-9) holds for all β with the exponent $\gamma(\beta)$ defined in (11). Hence, the only difference vis-à-vis the MM update rule for β-NMF in (10) is that there is an additional $1/\lambda_k$ term in the denominator in (32).

3 Alternative updates for ℓ_2-ARD

In this appendix, we provide an alternative set of updates for ℓ_2-ARD for the special and important cases $\beta = 0, 1, 2$ without recourse to further upper bounding $F(\mathbf{H}|\tilde{\mathbf{H}}) = \phi^{-1} G(\mathbf{H}|\tilde{\mathbf{H}}) + R_2(\mathbf{H})$ using another auxiliary function $J(\mathbf{H}|\tilde{\mathbf{H}})$. Note that these are not the updates we employ in the actual ℓ_2-ARD.

$\beta = 0$: In this case, using the definition of $G(\mathbf{H}|\tilde{\mathbf{H}})$ in Table 1 and $R_2(\mathbf{H})$ in (22) yields

$$F(\mathbf{H}|\tilde{\mathbf{H}}) = \sum_{kn} \frac{1}{\phi} q_{kn} h_{kn} + \frac{1}{\phi} p_{kn} \tilde{h}_{kn} \left(\frac{\tilde{h}_{kn}}{h_{kn}} \right) + \frac{1}{2\lambda_k} h_{kn}^2. \tag{A-10}$$

Note that $\tilde{h}_{kn}$ is the estimate of h_{kn} at the *previous* iteration and p_{kn} and q_{kn} are defined in (9). Finding the gradient w.r.t. h_{kn} and setting it to zero gives

$$\frac{1}{\phi} q_{kn} - \frac{1}{\phi} p_{kn} \left(\frac{\tilde{h}_{kn}}{h_{kn}} \right)^2 + \frac{1}{\lambda_k} h_{kn} = 0. \tag{A-11}$$

Hence, we solve the following cubic equation in h_{kn}:

$$\frac{\phi}{\lambda_k} h_{kn}^3 + q_{kn} h_{kn}^2 - p_{kn} \tilde{h}_{kn}^2 = 0. \tag{A-12}$$

We can check graphically from the signs of the coefficients of the cubic that there always exists a nonnegative solution to (A-12). In fact, there is either one real positive root or three real roots comprising two negative roots and one positive root. The positive root is to be chosen.

$\beta = 1$: In this case, we have

$$F(\mathbf{H}|\tilde{\mathbf{H}}) = \sum_{kn} \frac{1}{\phi} q_{kn} h_{kn} - \frac{1}{\phi} p_{kn} \tilde{h}_{kn} \log\left(\frac{h_{kn}}{\tilde{h}_{kn}}\right) + \frac{1}{2\lambda_k} h_{kn}^2. \tag{A-13}$$

Hence, doing exactly the same as in the above, we see that minimizer of $F(\mathbf{H}|\tilde{\mathbf{H}})$ is given by the positive solution to the following quadratic equation:

$$\frac{\phi}{\lambda_k} h_{kn}^2 + q_{kn} h_{kn} - p_{kn} \tilde{h}_{kn} = 0. \tag{A-14}$$

The discriminant $q_{kn}^2 + 4p_{kn}\phi/\lambda_k > 0$ so there are two real roots. It can be verified using the quadratic formula that one of these two roots is positive and the other is negative. Naturally, the positive one is chosen as the minimizer of $F(\mathbf{H}|\tilde{\mathbf{H}})$.

$\beta = 2$: This case is the simplest since the regularizer $R_2(\mathbf{H})$ "fits" nicely with the auxiliary function $G(\mathbf{H}|\tilde{\mathbf{H}})$ (both consist of terms linear in h_{kn}^2). We have

$$F(\mathbf{H}|\tilde{\mathbf{H}}) = \sum_{kn} \frac{1}{2\phi} q_{kn} \tilde{h}_{kn} \left(\frac{h_{kn}}{\tilde{h}_{kn}}\right)^2 - \frac{1}{\phi} p_{kn} h_{kn} + \frac{1}{2\lambda_k} h_{kn}^2. \tag{A-15}$$

Thus, the minimizer of $F(\mathbf{H}|\tilde{\mathbf{H}})$ is given by the formula:

$$h_{kn} = p_{kn} \left(\frac{q_{kn}}{\tilde{h}_{kn}} + \frac{\phi}{\lambda_k}\right)^{-1}. \tag{A-16}$$

4 Estimating both a and b using the method of moments

Assume that

$$v_{fn} \sim p(v_{fn}|\hat{v}_{fn}) \tag{A-17}$$

where the density p denotes the noise model governed by the scale parameter β. We have that

$$\mathbb{E}[v_{fn}|\hat{v}_{fn}] = \hat{v}_{fn} \tag{A-18}$$

$$\mathrm{var}[v_{fn}|\hat{v}_{fn}] = \phi \hat{v}_{fn}^{2-\beta} \tag{A-19}$$

The relation in (A-18) means that

$$\mathbb{E}[v_{fn}] = \mathbb{E}[\hat{v}_{fn}] \tag{A-20}$$

by the law of iterated expectations. Consider the second moment:

$$\mathbb{E}[v_{fn}^2|\hat{v}_{fn}] = \mathrm{var}[v_{fn}|\hat{v}_{fn}] + \mathbb{E}[v_{fn}|\hat{v}_{fn}]^2 \tag{A-21}$$

$$= \phi \hat{v}_{fn}^{2-\beta} + \hat{v}_{fn}^2 \tag{A-22}$$

By taking expectations of (A-22) with respect to $\hat{v}_{fn}$, we have

$$\mathbb{E}(\mathbb{E}[v_{fn}^2|\hat{v}_{fn}]) = \mathbb{E}[\phi \hat{v}_{fn}^{2-\beta} + \hat{v}_{fn}^2] \tag{A-23}$$

Now using the law of iterated expectations and linearity of expectation, the left-hand-side of (A-23) is exactly $\mathbb{E}[v_{fn}^2]$, hence we have

$$\mathbb{E}[v_{fn}^2] = \phi\mathbb{E}[\hat{v}_{fn}^{2-\beta}] + \mathbb{E}[\hat{v}_{fn}^2] \tag{A-24}$$

By the law of large numbers, we have that the empirical moments are close to the population moments:

$$\hat{\mu}_{\mathbf{V}} := \frac{1}{FN}\sum_{f,n} v_{fn} \approx \mathbb{E}[v_{fn}] \tag{A-25}$$

$$\hat{\xi}_{\mathbf{V}} := \frac{1}{FN}\sum_{f,n} v_{fn}^2 \approx \mathbb{E}[v_{fn}^2] \tag{A-26}$$

So we can use the empirical moments as proxies for $\mathbb{E}[v_{fn}]$ and $\mathbb{E}[v_{fn}^2]$. It remains to compute $\mathbb{E}[\hat{v}_{fn}^{2-\beta}]$ for $\beta = 0, 1, 2$. Clearly, for $\beta = 2$, $\mathbb{E}[\hat{v}_{fn}^{2-\beta}] = \mathbb{E}[1] = 1$. For $\beta = 1$, $\mathbb{E}[\hat{v}_{fn}^{2-\beta}]$ we have

$$\mathbb{E}[\hat{v}_{fn}] = \sum_k \mathbb{E}\left[w_{fk}h_{kn}\right] = K\,\mathbb{E}[w_{fk}h_{kn}] \tag{A-27}$$

for every $k = 1, \ldots, K$ since the random variables $g_k \triangleq w_{fk}h_{kn}$ are identically distributed. For $\beta = 0$, $\mathbb{E}[\hat{v}_{fn}^{2-\beta}]$ reduces to the power $\mathbb{E}[\hat{v}_{fn}^2]$. We have

$$\mathbb{E}[\hat{v}_{fn}^2] = \mathrm{var}[\hat{v}_{fn}] + (\mathbb{E}[\hat{v}_{fn}])^2 \tag{A-28}$$

By the i.i.d. assumption of the random variables $g_k = w_{fk}h_{kn}$ for $k = 1, \ldots, K$, we have

$$\mathrm{var}[\hat{v}_{fn}] = K\ \mathrm{var}[w_{fk}h_{kn}] = K\ \left(\mathbb{E}[w_{fk}^2h_{kn}^2] - (\mathbb{E}[w_{fk}h_{kn}])^2\right) \tag{A-29}$$

where (A-29) holds for every $k = 1, \ldots, K$. Thus, using Eq. (A-27),

$$\mathbb{E}[\hat{v}_{fn}^2] = K\left[\mathbb{E}[w_{fk}^2h_{kn}^2] + (K-1)(\mathbb{E}[w_{fk}h_{kn}])^2\right]. \tag{A-30}$$

Now, using conditional independence of w_{fk} and h_{kn} given λ_k and the tower property, we have

$$\mathbb{E}[w_{fk}h_{kn}] = \mathbb{E}[\mathbb{E}[w_{fk}h_{kn}|\lambda_k]] = \mathbb{E}\left[\mathbb{E}\left[w_{fk}|\lambda_k\right]\mathbb{E}\left[h_{kn}|\lambda_k\right]\right], \tag{A-31}$$

and

$$\mathbb{E}[w_{fk}^2h_{kn}^2] = \mathbb{E}[\mathbb{E}[w_{fk}^2h_{kn}^2|\lambda_k]] = \mathbb{E}[\mathbb{E}[w_{fk}^2|\lambda_k]\mathbb{E}[h_{kn}^2|\lambda_k]], \tag{A-32}$$

which remained to be assessed under our specific prior assumptions.

Half Normal model

$$\mathbb{E}\left[w_{fk}|\lambda_k\right] = \mathbb{E}\left[h_{kn}|\lambda_k\right] = \sqrt{\frac{2\lambda_k}{\pi}} \tag{A-33}$$

$$\mathbb{E}[w_{fk}^2|\lambda_k] = \mathbb{E}[h_{kn}^2|\lambda_k] = \lambda_k \tag{A-34}$$

so

$$\mathbb{E}[w_{fk}h_{kn}] = \mathbb{E}\left[\frac{2\lambda_k}{\pi}\right] = \frac{2b}{\pi(a-1)} \tag{A-35}$$

$$\mathbb{E}[w_{fk}^2h_{kn}^2] = \mathbb{E}[\lambda_k^2] = \frac{b^2}{(a-1)(a-2)}. \tag{A-36}$$

As a result, for the Half Normal model, the first moment is

$$\mathbb{E}[\hat{v}_{fn}] = \frac{2Kb}{\pi(a-1)} =: \psi_{\text{HN}}(a,b,K), \tag{A-37}$$

and the second moment is

$$\mathbb{E}[\hat{v}_{fn}^2] = K\left[\frac{b^2}{(a-1)(a-2)} + (K-1)\frac{4b^2}{\pi^2(a-1)^2}\right] \tag{A-38}$$

$$= \frac{Kb^2}{(a-1)(a-2)}\left[1 + \frac{4(K-1)(a-2)}{\pi^2(a-1)}\right] =: \gamma_{\text{HN}}(a,b,K). \tag{A-39}$$

Hence, using (A-20) and (A-24), we can estimate a and b as the solutions to the simultaneous equations

$$\hat{\mu}_{\mathbf{V}} = \psi_{\text{HN}}(\hat{a},\hat{b},K), \tag{A-40}$$

$$\hat{\xi}_{\mathbf{V}} = \begin{cases} (\phi+1)\,\gamma_{\text{HN}}(\hat{a},\hat{b},K) & \beta = 0 \\ \phi\psi_{\text{HN}}(\hat{a},\hat{b},K) + \gamma_{\text{HN}}(\hat{a},\hat{b},K) & \beta = 1 \\ \phi + \gamma_{\text{HN}}(\hat{a},\hat{b},K) & \beta = 2 \end{cases}. \tag{A-41}$$

Note that if a is fixed as in our experiments, we only need to solve for b in (A-40).

Exponential model

$$\mathbb{E}\left[w_{fk}|\lambda_k\right] = \mathbb{E}\left[h_{kn}|\lambda_k\right] = \lambda_k \tag{A-42}$$

$$\mathbb{E}[w_{fk}^2|\lambda_k] = \mathbb{E}[h_{kn}^2|\lambda_k] = 2\lambda_k^2 \tag{A-43}$$

so

$$\mathbb{E}[w_{fk}h_{kn}] = \mathbb{E}[\lambda_k^2] = \frac{b^2}{(a-1)(a-2)} \tag{A-44}$$

$$\mathbb{E}[w_{fk}^2 h_{kn}^2] = \mathbb{E}[4\lambda_k^4] = \frac{4b^4}{(a-1)(a-2)(a-3)(a-4)}. \tag{A-45}$$

As a result, for the Exponential model, the first moment is

$$\mathbb{E}[\hat{v}_{fn}] = \frac{Kb^2}{(a-1)(a-2)} =: \psi_{\text{Exp}}(a,b,K), \tag{A-46}$$

and the second moment is

$$\mathbb{E}[\hat{v}_{fn}^2] = K\left[\frac{4b^4}{(a-1)(a-2)(a-3)(a-4)} + (K-1)\frac{b^4}{(a-1)^2(a-2)^2}\right] \tag{A-47}$$

$$= \frac{Kb^4}{(a-1)(a-2)(a-3)(a-4)}\left[4 + \frac{(K-1)(a-3)(a-4)}{(a-1)(a-2)}\right] =: \gamma_{\text{Exp}}(a,b,K). \tag{A-48}$$

Hence, using (A-20) and (A-24), we can estimate a and b as the solutions to the simultaneous equations

$$\hat{\mu}_{\mathbf{V}} = \psi_{\text{Exp}}(\hat{a},\hat{b},K), \tag{A-49}$$

$$\hat{\xi}_{\mathbf{V}} = \begin{cases} (\phi+1)\,\gamma_{\text{Exp}}(\hat{a},\hat{b},K) & \beta = 0 \\ \phi\psi_{\text{Exp}}(\hat{a},\hat{b},K) + \gamma_{\text{Exp}}(\hat{a},\hat{b},K) & \beta = 1 \\ \phi + \gamma_{\text{Exp}}(\hat{a},\hat{b},K) & \beta = 2 \end{cases}. \tag{A-50}$$

Note that if a is fixed as in our experiments, we only need to solve for b in (A-49).

References

[1] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society, Series B*, vol. 68, no. 1, pp. 49–67, 2007.

[2] E. J. Candès, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *Journal of Fourier Analysis and Applications*, vol. 14, pp. 877–905, Dec 2008.

[3] D. R. Hunter and K. Lange, "A tutorial on MM algorithms," *The American Statistician*, vol. 58, pp. 30 – 37, 2004.