Technical University of Denmark

02460

Advanced Machine Learning

Appendix for derivations

S153758 Peter Ebert Christensen
S144234 Martin Hatting Petersen
S154317 Niklas Refsgaard
S163502 Oldouz Majidi

May 23, 2018

DTU Compute
Department of Applied Mathematics and Computer Science

Contents

1 Appendix A

Derivations for the paper: Bayesian inference for Nonnegative Factorisation Models

We need to define the following in order to do the derivations from the mentioned paper

$$X \in \mathbb{R}^{WK}, \quad \mathbf{T} \in \mathbb{R}^{WI}, \quad V \in \mathbb{R}^{IK},$$

$$x_{\nu\tau} \approx [TV]_{\nu\tau} = \Sigma_i \tau_{\nu,i} \nu_{i,\tau} \tag{1}$$

$$x_{\nu\tau} = \Sigma_i s_{\nu,i,\tau} \tag{2}$$

$$s_{\nu,i,\tau} \sim PO(s_{\nu,i,\tau} | \tau_{\nu,i} \nu_{i,\tau}) \tag{3}$$

Where $PO(s|\lambda) = \exp(s \cdot \log\lambda - \lambda - \ln\Gamma(s+1))$ (and it is known that $\Gamma(s+1) = s!$)is the Poisson distribution, and the variables s is a latent source. To get equation 7 from the paper (Bayesian inference for nonnegative matrix factorisation models), following the equation below, where we first use the sum rule and then the product rule, realise that the logarithm of the probability of X given the factorised matrices T and V is the product of all possible sums of the elements two matrices can take to get X and then use the superposition property of Poisson variables, that is $s_i \sim PO(s_i|\lambda)$ and that $x = s_1 + s_2 + ... + s_n$, then the marginal probability is given by $p(x) = PO(x|\Sigma_i \lambda_i)$ and then finally use equation 2 and 1 above;

$$\ln(p(X|T,V)) \overset{\text{sum rule}}{=} \ln(\Sigma_S p(X,S|T,V)) \overset{\text{product rule}}{=} \tag{4}$$

$$\ln(\Sigma_S p(X|S) p(S|T,V) \overset{\text{superposition property}}{=} \tag{5}$$

$$\ln(\prod_{\nu,\tau} PO(x_{\nu\tau}|\Sigma_i \tau_{\nu,i} \nu_{i,\tau})) \overset{\text{(1) \& Poisson dist}}{=} \tag{6}$$

$$\Sigma_\nu \Sigma_\tau (x_{\nu\tau} \ln([TV]_{\nu\tau} - [TV]_{\nu\tau} - \ln\Gamma(x_{\nu,\tau}+1)) \tag{7}$$

Now we consider deriving a maximum likelihood based EM algorithm for updating a instrumental distribution q for the lower bound of the likelihood $\mathcal{L}$ and then compute the new estimates for the factorised matrices T and V in the M step. The reason for this is that direct optimisation of $P(X|T,V)$ is difficult unlike the complete data log likelihood $P(X,S|T,V)$ The log likelihood of our observed data X is:

$$\ell_X(T,V) \overset{(4)\&(5)}{=} \ln(p(X|T,V)) \overset{\text{bishop 9.70}}{=} \mathcal{L}(q,T,V) + \text{KL}(q||p) \geq \tag{8}$$

$$\mathcal{L}(q,T,V) \overset{\text{bishop 9.71}}{=} \Sigma_S q(S) \ln(p(X,S|T,V) q(S)^{-1}) = \tag{9}$$

$$\Sigma_S q(S) \ln(p(X,S|T,V)) - \Sigma_S q(S) \ln(q(S)) \overset{\text{bishop 10.1}}{=} \tag{10}$$

$$< \ln(p(S,X|T,V)) >_{q(S)} + H[q] = \mathcal{B}_{\text{EM}}[q] \tag{11}$$

We see that this corresponds to equation (11) in the paper, plus the entropy $H[q]$.Note here that $\mathcal{L}$ is a functional of the distribution q and a function of the matrices T and V, which is our lower bound to the log-likelihood. This lower bound can be used in an EM algorithm, that will indeed maximise the log likelihood, defined in (8). Thus in the E step we maximise $\mathcal{L}$ with respect to q, while holding T and V fixed (that is using the previous estimate):

$$q(S)^{(n)} = \arg\max_{q(S)} \mathcal{B}_{\text{EM}}[q] = p(S|X, T^{(n-1)}, V^{(n-1)}) \tag{12}$$

subsequently in the M step we fix q and the lower bound $\mathcal{L}$ is maximised with respect to T and V:

$$(T^{(n)}, V^{(n)}) = \arg\max_{T,V} \mathcal{B}_{\text{EM}}[q] = \arg\max_{T,V} < \ln(p(S, X|, T, V)) > \tag{13}$$

It will now be shown how we find the E step, that is;

$$\arg\max_{q(S)} \mathcal{B}_{\text{EM}}[q] = p(S|X, T, V)$$

This we recognise as the posterior of the latent sources;

$$p(S|X, T, V) = p(S, X|T, V)(p(X|T, V))^{-1} \tag{14}$$

From the latent sources in equation (3) and using equation (7) we can express the following,

$$\ln p(S, X|T, V) = \Sigma_\nu \Sigma_\tau (\Sigma_i (s_{\nu,i,\tau} \ln(t_{\nu,i} v_{i,\tau}) - t_{\nu,i} v_{i,\tau} \tag{15}$$

$$-\ln\Gamma(t_{\nu,i} v_{i,\tau} + 1) + \ln(\delta(x_{\nu\tau} - \Sigma_i s_{\nu,i,\tau}))) \tag{16}$$

In which δ is the Kronecker delta function.Using the above equation we can find E step in the posterior to be

$$\begin{aligned}
\ln(p(S|X, T, V)) \overset{\text{eq\_14}}{=} \ln(p(X, S|T, V)) - \ln(p(S|T, V)) \overset{\text{eq\_7}}{=} \\
\Sigma_\nu \Sigma_\tau (\Sigma_i (s_{\nu,i,\tau} \ln(t_{\nu,i} v_{i,\tau}) - t_{\nu,i} v_{i,\tau} - \ln\Gamma(t_{\nu,i} v_{i,\tau} + 1) \\
+\ln(\delta(x_{\nu\tau} - \Sigma_i s_{\nu,i,\tau})) - x_{\nu,\tau} \ln[t, v] + [TV]_{\nu\tau} + \ln\Gamma(x_{\nu,\tau} + 1)) \overset{\text{log\_rule}}{=} \\
\Sigma_\nu \Sigma_\tau (\Sigma_i (s_{\nu,i,\tau} \ln(p_{\nu,i,\tau}) - \ln(\Gamma(s_{\nu,i,\tau} + 1)) \\
+\ln(\Gamma(x_{\nu,\tau} + 1)) + \ln(\delta(x_{\nu\tau} - \Sigma_i s_{\nu,i,\tau})))
\end{aligned}$$

Here we have also used

$$- x_{\nu,\tau} \ln[t, v] = -\Sigma_i s_{\nu,i,\tau} \ln(\Sigma_i t_{\nu,i} v_{i,\tau}) \tag{17}$$

as well as defined the cell probabilities

$$p_{\nu,i,\tau} = \frac{t_{\nu,i} v_{i,\tau}}{\Sigma_j t_{\nu,j} v_{j,\tau}} \tag{18}$$

It should be noted that the above result for the posterior can also be written in a form of a multinomial distribution, with mean $< s_{\nu,i,\tau} >= x p_{\nu,i,\tau}$.
Now we can proceed with the M step, we see from the posterior of the latent sources that when maximising $T^{(n)}$,$V^{(n)}$, then X and S does not play any role and are therefore seen as constants. Thus we have $H[q]$ disappears from equation (11) and we end p with equation (13), in which $q(S)$ is found in equation (12).

Equation (13) can also be expressed as follows, due to equation (15) and (16), as well as due to the linearity of the expectation operator;

$$< \ln(p(S,X|,T,V)) >_{q(S)} = \Sigma_\nu \Sigma_\tau (\Sigma_i (< s_{\nu,i,\tau} > \ln(t_{\nu,i} v_{i,\tau}) - t_{\nu,i} v_{i,\tau} \\ - < \ln\Gamma(t_{\nu,i} v_{i,\tau} + 1) > + < \ln(\delta(x_{\nu\tau} - \Sigma_i s_{\nu,i,\tau})) >)$$

However as the optimisation is with respect to T and V we only have this much simpler function to evaluate;

$$Q(T,V) = \Sigma_\nu \Sigma_\tau (\Sigma_i (< s_{\nu,i,\tau} >^{(n)} \ln(t_{\nu,i} v_{i,\tau}) - t_{\nu,i} v_{i,\tau}) \tag{19}$$

Where $< s_{\nu,i,\tau} >$ can be found by the mentioned mean of the multinomial distribution as well as the cell probabilities. Maximising Q and substituting $< s_{\nu,i,\tau} >$ in equation 19 yields the following point equations;

$$\frac{\partial Q}{\partial t_{\nu,i}} = -\Sigma_\tau v_{i,\tau}^{(n)} + (t_{\nu,i})^{-1} \Sigma_\tau < s_{\nu,i,\tau} >^{(n)} \tag{20}$$

$$t_{\nu,i}^{(n+1)} = (\Sigma_\tau v_{i,\tau}^{(n)})^{-1} \Sigma_\tau < s_{\nu,i,\tau} >^{(n)} \tag{21}$$

$$\frac{\partial Q}{\partial v_{i,\tau}} = -\Sigma_\nu t_{\nu,i}^{(n)} + (v_{i,\tau})^{-1} \Sigma_\nu < s_{\nu,i,\tau} >^{(n)} \tag{22}$$

$$v_{i,\tau}^{(n+1)} = (\Sigma_\nu v_{\nu,i}^{(n)})^{-1} \Sigma_\nu < s_{\nu,i,\tau} >^{(n)} \tag{23}$$

Equation (21) and (23) is found by setting the derivative in equation (20) equal to 0 and solve for $t_{\nu,i}^{(n)}$ and similarly for the derivative in equation (22), in order to get $v_{i,\tau}^{(n)}$.

In the following derivations we are going focus on full bayesian inference, where we are working with the lower bound of the marginal log likelihood (the evidence) $P(X|\Theta) = \int dTdV \sum_S p(X|S)p(S|T,V)p(TV|\Theta)$. Similar to before we can get the lower bound by letting us inspire from equation (8) and (9) on the conditional log likelihood $p(X|\Theta)$ to get

$$\mathcal{L}_X(\Theta) \geq \langle \ln p(X,S,V,T|\theta) \rangle_q + H[q] = B_{VB}[q], \quad q = q(S,T,V) \tag{24}$$

Where H is our entropy and q is an instrumental distribution which is used to approximate p. Here we can make use of the mean field approximation, that is q can be factorised into 3 terms $q(S,T,V) = q(S)q(T)q(V)$ in order to make our approximating distribution less complex, however we are not guaranteed that we will find the exact marginal likelihood $\mathcal{L}_X(\Theta)$. This is however not too troublesome as the strategy of variational inference is to optimise a lower bound. Going back to $B_V B[q]$ and looking at the lower bound and the entropy, we make use of things;

- Impose a structure on the elements of T and V such as a gamma distribution $\mathcal{G}(x, a, b) = \exp((a-1)\ln x - xb^{-}1 - \ln\Gamma(a) - a\ln b)$; $t_{\nu,i} \sim \mathcal{G}(t_{\nu,i}; a^t_{\nu,i}, \frac{b^t_{\nu,i}}{a^t_{\nu,i}})$ and $v_{i,\tau} \sim \mathcal{G}(v_{i,\tau}; a^v_{i,\tau}, \frac{b^v_{i,\tau}}{a^t_{i,\tau}})$. Here the parameters a and b are the variables in the factor Θ. We use this assumption due to computational convenience, as the Gamma distribution is the conjugate prior to the Poisson intensity.

- Using the product rule of probability we can write $p(X, S, T, V|\Theta) = p(X|S)p(S|T, V)p(T|\Theta^T)p(V|\Theta^V)$, as Θ consists of 2 set of variables of the priors on T and V respectively.

- The entropy can be ignored under the optimisation of the lower bound, as we are making an approximation in the first place. In case one wants to calculate the bound $B_V B[q]$ we can no longer ignore it.

- Using theorem 11.9 In Probabilistic Graphical Models, Principles and Techniques by Daphne Koller 2009, then the distribution q can attain a local optimum by the following iterative equation;

  $$q^{(n+1)}_\alpha \propto \exp(\langle \ln p(X, S, T, V|\Theta)\rangle_{q\neg\alpha^{(n)}}) \tag{25}$$

  Here α denotes the set consisting of the following S, T, V.

- Due to the factoring assumptions we have that S,T and V are disjoint and we can optimise over one of the factors S,T or V at a time by fixing X, Θ and the 2 other factors in α. Which would give us for example;

  $$q(T)^{(n+1)} \propto \exp(\langle \ln p(X, S, T, V|\Theta)\rangle_{q(S)^{(n)}, q(V)^{(n)}}) \tag{26}$$

We should note that in order for us to use the second bullet point, we calculate the term $p(X|S)p(S|T, V)$ by referring to equation (5) and the two other terms $p(T|\Theta^T)p(V|\Theta^V)$ by making direct use of the Gamma distribution as mentioned in the first bullet point. Doing so yields the full joint distribution;

$$\begin{aligned}\ln p(S, X, T, V|\Theta) &= \Sigma_\nu\Sigma_i\Sigma_\tau(-t_{\nu,i}v_{i,\tau} + s_{\nu,i,\tau}\ln(t_{\nu,i}v_{i,\tau}) - \Gamma(s_{\nu,i,\tau}+1)) \\ &+\Sigma_\nu\Sigma_\tau \ln\delta(x_{\nu,\tau} - s_{\nu,i,\tau}) + \Sigma_\nu\Sigma_i(a^t_{\nu,i} - 1)\ln(t_{\nu,i}) - \frac{a^t_{v,i}}{b^t_{v,i}}t_{v,i} - \ln\Gamma(a^t_{\nu,i}) \\ -a^t_{\nu,i}\ln\frac{b^t_{nu,i}}{a^t_{\nu,i}} &+ \Sigma_i\Sigma_\tau(a^v_{i,\tau} - 1)\ln(v_{i,\tau}) - \frac{a^v_{i,\tau}}{b^v_{i,\tau}}v_{i,\tau} - \ln\Gamma(a^v_{i,\tau}) - a^v_{i,\tau}\ln\frac{b^v_{i,\tau}}{a^v_{i,\tau}}\end{aligned}$$

With the above equation we can then insert this into the update equation (25) to get the following;

$$\begin{aligned}q(s_{\nu,1:I,\tau}) &\propto \exp(\Sigma_i(s_{\nu,i,\tau}(\langle \ln t_{\nu,i}\rangle + \langle \ln v_{i,\tau}\rangle) - \ln\Gamma(s_{\nu,i,\tau}+1)))\delta(x_{\nu,\tau} - \Sigma_i s_{\nu,i,\tau}) \\ &\propto \mathcal{M}(s_{\nu,1,\tau}, ..., s_{\nu,i,\tau}, ..., s_{\nu,I,\tau}; x_{\nu,\tau}; p_{\nu,1,\tau}, ..., p_{\nu,i,\tau}, ...p_{\nu,I,\tau})\end{aligned}$$

Where we have defined the following and made use of the multinomial distribution;

$$\mathcal{M}(s,x,p) = \delta(x - \Sigma_i s_i)\exp(\ln\Gamma(x+1) + \sum_i^I (s_i \ln p_i - \ln\Gamma(s_i+1)))$$
$$p_{\nu,i,\tau} = \frac{\exp(\langle \ln t_{\nu,i}\rangle + \langle \ln v_{i,\tau}\rangle)}{\Sigma_i \exp(\langle \ln t_{\nu,i}\rangle + \langle \ln v_{i,\tau}\rangle)}$$
$$\langle s_{\nu,i,\tau} = x_{\nu,\tau} p_{\nu,i,\tau}\rangle$$

Similarly we can express the other factors of q (Where Ψ is the digamma function);

$$q(t_{\nu,i}) \propto \exp((a^t_{\nu,i} + \Sigma_\tau \langle s_{\nu,i,\tau}\rangle - 1)\ln(t_{\nu,i}) - (\frac{a^t_{\nu,i}}{b^t_{\nu,i}} + \Sigma_\tau \langle v_{i,\tau}\rangle) t_{\nu,i})$$
$$\propto \mathcal{G}(t_{\nu,i}; \alpha^t_{\nu,i}, \beta^t_{\nu,i}), \quad \alpha^t_{\nu,i} = a^t_{\nu,i} + \Sigma_\tau \langle s_{\nu,i\tau}\rangle, \quad \beta^t_{\nu,i} = (\frac{a^t_{\nu,i}}{b^t_{\nu,i}} + \Sigma_\tau \langle v_{i,\tau}\rangle)^{-1}$$
$$\exp(\langle \ln t_{\nu,i}\rangle) = \exp(\Psi(\alpha^t_{\nu,i}))\beta^t_{\nu,i}, \quad \langle t_{\nu,i}\rangle = \alpha^t_{\nu,i}\beta^t_{\nu,i}$$

and

$$q(v_{i,\tau}) \propto \exp((a^v_{i,\tau} + \Sigma_\nu \langle s_{\nu,i,\tau}\rangle - 1)\ln(v_{i,\tau}) - (\frac{a^v_{i,\tau}}{b^v_{i,\tau}} + \Sigma_\nu \langle t_{\nu,i}\rangle) v_{i,\tau})$$
$$\propto \mathcal{G}(v_{i,\tau}; \alpha^v_{i,\tau}, \beta^v_{i,\tau}), \quad \alpha^v_{\nu,i} = a^v_{\nu,i} + \Sigma_\tau \langle s_{\nu,i\tau}\rangle, \quad \beta^v_{i,\tau} = (\frac{a^v_{i,\tau}}{b^v_{i,\tau}} + \Sigma_\nu \langle t_{\nu,i}\rangle)^{-1}$$
$$\exp(\langle \ln v_{i,\tau}\rangle) = \exp(\Psi(\alpha^v_{i,\tau}))\beta^v_{i,\tau}, \quad \langle v_{i,\tau}\rangle = \alpha^v_{i,\tau}\beta^v_{i,\tau}$$

which can be efficiently implemented.

NMF GPP

In our second paper the notation and methods is a little different, but the main problem remains the same. Namely that we have a data matrix $X = DH + N \in R^{K\times L}$, that is factorised as the product of two element-wise non-negative matrices, $D \in R_+^{K\times M}$, $H \in R_+^{M\times L}$

Ex.1:least squares NMF
In our first example we are looking at a maximum likelihood NMF approach in which the noise from the factor N is i.i.d Guassian distributed with variance σ^2. Then the maximum likelihood estimate will be the gaussian PDF as below;

$$P^{LS}_{X|D,H}(X|D,H) = (\sqrt{2\pi}\sigma_N)^{-KL} exp(-(2\sigma_N^2)^{-1}||X - DH||_F^2) \tag{27}$$

I order to find optimal values for D and H we are going to continue our calculations with the negative log likelihood, which also serves as a cost function:

$$ln(P^{LS}_{X|D,H}) = ln(\sqrt{2\pi}\sigma_N)^{-KL} + ln(-2\sigma_N^2)^{-1} + ln||X - DH||_F^2 \tag{28}$$
$$\tag{29}$$

and this is equal to:

$$ln\frac{||X-DH||_F^2}{(\sqrt{2\pi}\sigma_N)^{KL}(-2\sigma_N^2)} \propto \frac{1}{2\sigma_N^2}||X-DH||_F^2 = \ell^{LS}_{X|D,H} \tag{30}$$

We also need to know the gradient of the negative log likelihood for obtaining the maximum likelihood estimate of D and H:

$$\begin{aligned} \bigtriangledown_H \ell_{X|D,H} &= \bigtriangledown_H(\frac{1}{2\sigma_N^2}(X-DH)^T(X-DH)) \qquad (31)\\ &= \frac{1}{2\sigma_N^2}((2(-D)^T(X-DH)) \qquad (32)\\ &= \sigma_N^{-2}D^T(DH-X) \qquad (33) \end{aligned}$$

2.3:
From our assumptions of the gaussian process prior we use a gaussian process vector h that when transformed with a link function f_h we map into the elements of H. Here the underlying vector h is a multivariate zer mean gaussian distribution with covariance matrix Σ_h and linked to H via the link function f_h (Which can be any strictly increasing function that maps from $\mathbb{R}^+$ into $\mathbb{R}$, as well as having an inverse function f_h^{-1} and that the derivative of both functions exits); $d \in R^{M\times K}$, $h \in R^{L\times M}$

$$\begin{aligned} P_h(h) &\sim N(h|0,\Sigma_h), \qquad (34)\\ h &= f_h(vec(H)) \qquad (35) \end{aligned}$$

With these assumptions we can calculate the prior over H by using the Jacobian of transformation h;

$$\frac{\partial h}{\partial H} = |\frac{\partial f_h(H)}{\partial H}| \tag{36}$$

So when we are for example integrating or similar under a region R the region then becomes changed to Q under a transformation f_h and so we multiply $|J(H)|$ onto the transformed integral. In this case we have the following;

$$\begin{aligned} P_H(H) = P_h(H)|J(H)| = N(H|0,\Sigma_h).\prod_i |f_h'(H)| \propto \qquad (37)\\ exp(-\frac{1}{2}H^T\Sigma_h^{-1}H)\prod_i |f_h'(H)| \qquad (38) \end{aligned}$$

2.4:
In order for us to use unconstrained optimisation methods we iIntroduce δ and η related to D and H by

$$H = g_h(\eta) = vec^{-1}(f_h^{-1}(C_h^T\eta)) \tag{39}$$

where C_h is matrix square root of Σ_h(covariance matrix)to make $\eta \sim N(\eta|0,I)$ that η is standard iid Gauss ($\sigma^2 = 1$) thus we can calculate the prior distribution of the transformed variable η:

$$\begin{aligned} P_\eta(\eta) = P_H(g_h(\eta))|J(g_h(\eta))| = exp(\frac{-1}{2}(C_n^T\eta)^T\Sigma_h^{-1}(C_n^T\eta).C \qquad (40)\\ = \frac{1}{(2\pi)^{LM}}exp(\frac{-1}{2}\eta_\eta^T) \qquad (41) \end{aligned}$$

Taking the log of this factor we arrive at the following log negative prior (and the result for $\ell_\delta(\delta)$ is equivalent to $\ell_\eta(\eta)$);

$$\ell_\eta(\eta) = -ln(P_\eta(\eta)) = -(\frac{-NLM}{2} - \frac{1}{2}\Sigma_n^N \eta_\eta^T \eta_\eta \propto \tag{42}$$

$$\frac{1}{2}\eta_\eta^T, \ell_\delta(\delta) \propto \frac{1}{2}\delta^T \delta \tag{43}$$

and from before we had that the log likelihood cost function is proportional to the following.

$$\ell_{D,H|X}(D,H) \propto \ell_{X|D,H}(D,H) + \ell_{D,H}(D,H) \tag{44}$$

$$\tag{45}$$

Here we can use equation (27) together with the negative log prior from above to get a computable the cost function;

$$\ell_{\delta,\eta|X}(\delta,\eta) = \ell_{X|D,H}(g_d(\delta), g_h(\eta)) + \tag{46}$$

$$\ell_\delta(\delta) + \ell_\eta(\eta)) = \ell_{X|D,H} + \frac{1}{2}\delta^T \delta + \frac{1}{2}\eta^T \eta \tag{47}$$