Sparse Overcomplete Latent Variable Decomposition of Counts Data: Supplemental Material

Madhusudana Shashanka
Mars, Incorporated
Hackettstown, NJ
shashanka@cns.bu.edu

Bhiksha Raj
Mitsubishi Electric Research Labs
Cambridge, MA
bhiksha@merl.com

Paris Smaragdis
Adobe Systems
Newton, MA
paris@adobe.com

A Parameter estimation for Latent Variable Model

In this appendix, we derive update equations for the latent variable model as described in Section 2.

The model is given by equation (1) as reproduced below:

$$P_n(f) = \sum_z P(f|z)P_n(z).$$

The goal is to estimate the parameters of the model such that they best explain the collection of all observations V_{nf}, where V_{nf} represents the counts of f in the n^{th} data set in the collection. The parameters to be estimated are $P(f|z)$ and $P_n(z)$. z is the hidden variable and f is the feature observed at any particular draw from the distribution $P_n(f)$. The subscript n signifies that the generative distribution $P_n(f)$ and mixture weights $P_n(z)$ are specific to the n^{th} data set. We use a maximum likelihood formulation of the problem. The log-likelihood of all the observed data sets is given by

$$\mathcal{P} = \sum_n \sum_f V_{fn} \log P_n(f). \tag{9}$$

The maximum-likelihood method estimates parameters such that the log-likelihood is maximized.

The standard procedure for maximum likelihood estimation in latent variable models is the Expectation Maximization (EM) algorithm. EM alternates two steps: (1) an expectation (E) step where the *a posteriori* probabilities of the latent variables are computed based on the current estimates of the parameters, and (2) a maximization (M) step, where parameters are updated such that the expected complete data log-likelihood is maximized.

For the E-step, we obtain the *a posteriori* probability for the latent variable as

$$P_n(z|f) = \frac{P_n(z)P(f|z)}{\sum_z P_n(z)P(f|z)}. \tag{10}$$

In the M-step, we maximize the expected complete data log-likelihood. Let Λ represent the set of parameters of the model, i.e. $\Lambda = \{P(f|z), P_n(z)\}$. The expected log-likelihood can be written as

$$\mathcal{L} = E_{\bar{z}|\bar{f};\Lambda} \log P(\bar{f}, \bar{z}), \tag{11}$$

where $\bar{f}$ and $\bar{z}$ represent the set of all observations of f and z in the draws that generated all data sets. The complete data likelihood can be written as

$$P(\bar{f}, \bar{z}) \propto \prod_{j,n} P_n(f_j, z_j) = \prod_{j,n} P_n(z_j)P(f_j|z_j), \tag{12}$$

where f_j and z_j are the observed values of variables f and z in the j-th draw. Hence, we can write the function $\mathcal{L}$ as (ignoring the constant terms)

$$\begin{aligned}
\mathcal{L} &= E_{\bar{z}|\bar{f};\Lambda} \sum_{j,n} \log P_n(f_j, z_j) \\
&= \sum_{j,n} E_{z_j|f_j;\Lambda} \log P_n(f_j, z_j) \\
&= \sum_{j,n} E_{z_j|f_j;\Lambda} \log P_n(z_j) + \sum_{j,n} E_{z_j|f_j;\Lambda} \log P(f_j|z_j) \\
&= \sum_{j,n} \sum_{z} P(z|f_j) \log P_n(z) + \sum_{j,n} \sum_{z} P(z|f_j) \log P(f_j|z). \qquad (13)
\end{aligned}$$

In the above equation, we can change the summation over draws j to a summation over features f by accounting for how many times f was observed, i.e. the f-th entry in the observed data set[1]. The expected log-likelihood can now be written as

$$\mathcal{L} = \sum_{n} \sum_{f} \gamma V_{fn} \sum_{z} P_n(z|f) \log P_n(z) + \sum_{n} \sum_{f} \gamma V_{fn} \sum_{z} P_n(z|f) \log P(f|z). \qquad (14)$$

In order to take care of the normalization constraints, the above equation must be augmented by appropriate Lagrange multipliers τ_n and ρ_z,

$$Q = \mathcal{L} + \sum_{n} \tau_n \Big(1 - \sum_{z} P_n(z)\Big) + \sum_{z} \rho_z \Big(1 - \sum_{f} P(f|z)\Big) \qquad (15)$$

Maximization of Q with respect to $P_n(z)$ and $P(f|z)$ leads to the following sets of equations

$$\sum_{f} \gamma V_{fn} P_n(z|f) + \tau_n P_n(z) = 0, \qquad (16)$$

$$\sum_{n} \gamma V_{fn} P_n(z|f) + \rho_z P(f|z) = 0. \qquad (17)$$

After eliminating the Lagrange multipliers, we obtain the M-step re-estimation equations

$$P(f|z) = \frac{\sum_n V_{fn} P_n(z|f)}{\sum_f \sum_n V_{fn} P_n(z|f)}, \qquad P_n(z) = \frac{\sum_f V_{fn} P_n(z|f)}{\sum_z \sum_f V_{fn} P_n(z|f)}. \qquad (18)$$

B Maximum Likelihood and KL minimization

Maximimum likelihood method estimates parameters such that the log-likelihood $\mathcal{P}$, given by equation (9), is maximized. We can rewrite this as

$$\mathcal{P} = \sum_{n} \big(\sum_{f} V_{fn}\big) \sum_{f} \frac{V_{fn}}{\sum_{f'} V_{f'n}} \log P_n(f) \qquad (19)$$

$V_{fn} / \sum_{f'} V_{f'n}$ represents the normalized histogram for the n^{th} data set. Representing this term by $\bar{V}_{fn}$,

$$\begin{aligned}
\mathcal{P} &= \sum_{n} \big(\sum_{f} V_{fn}\big) \sum_{f} \bar{V}_{fn} \log\left(\frac{P_n(f)}{\bar{V}_{fn}}\right) + \sum_{n} \big(\sum_{f} V_{fn}\big) \sum_{f} \bar{V}_{fn} \log(\bar{V}_{fn}) \qquad (20) \\
&= \sum_{n} \big(\sum_{f} V_{fn}\big) \sum_{f} \bar{V}_{fn} \log\left(\frac{P_n(f)}{\bar{V}_{nf}}\right) + C \qquad (21) \\
&= -\sum_{n} \big(\sum_{f} V_{fn}\big) KL(\bar{V}_{fn}, P_n(f)) + C \qquad (22)
\end{aligned}$$

[1] Since observed dataset is modeled as a histogram, entries should be integers. To account for this, we weight the data by an unkown scaling factor γ

where C is a constant term that is not dependent on $P_n(f)$.

$$\text{argmax}_{V_{fn}} \mathcal{P} = \text{argmin}_{V_{fn}} \sum_n \big(\sum_f V_{fn} \big) KL(\bar{V}_{fn}, P_n(f)) \tag{23}$$

Maximizing $\mathcal{P}$ with respect to $P_n(f)$ is therefore equivalent to minimizing the sum of the KL distances between the normalized histograms $\bar{V}_{fn}$ and $P_n(f)$ for each data set, scaled by the total number of draws in that data set.

C Parameter estimation: sparse latent variable model

The model is given by the equation

$$P_n(f) = \sum_z P(f|z) P_n(z).$$

The set of parameters to be estimated are $P(f|z)$ and $P_n(z)$ i.e. $\Lambda = \{P(f|z), P_n(z)\}$. We impose an *a priori* probability on the parameters given by

$$P(\Lambda) \propto \prod_z e^{\bar{\alpha} \sum_f P(f|z) \log P(f|z)} \prod_n e^{\bar{\beta} \sum_z P_n(z) \log P_n(z)},$$

where $\bar{\alpha}$ and $\bar{\beta}$ are parameters indicating the extent of sparsity desired on $P(f|z)$ and $P_n(z)$ respectively. The log-prior (logarithm of the above *a priori* probability) can be written as

$$\bar{\alpha} \sum_z \sum_f P(f|z) \log P(f|z) + \bar{\beta} \sum_n \sum_z P_n(z) \log P_n(z), \tag{24}$$

We use *maximum a posteriori* estimation and use the EM algorithm.

For the E-step, we compute the *a posteriori* probability of the latent variable as before:

$$P_n(z|f) = \frac{P_n(z) P(f|z)}{\sum_z P_n(z) P(f|z)}. \tag{25}$$

In the M-step, instead of maximizing the log-likelihood, we maximize the log-posterior (the logarithm of the *a posteriori* probability of the model parameters). The log-posterior to be maximized is given by

$$\mathcal{L} = E_{\bar{z}|\bar{f};\Lambda} \log P(\bar{f}, \bar{z}) + \log P(\Lambda) \tag{26}$$

where $\bar{f}$ and $\bar{z}$ represent the set of all observations of f and z in the draws that generated all data sets. The first term of equation (26), corresponding to the log-likelihood, can be derived as shown in the previous appendix and is given by equation (14). The second term corresponding to the log-prior is given by equation (24). Hence, we can write the function $\mathcal{L}$ as (ignoring the constant terms)

$$\begin{aligned} \mathcal{L} \;\; = \;\; & \sum_n \sum_f \gamma V_{fn} \sum_z P_n(z|f) \log P_n(z) + \sum_n \sum_f \gamma V_{fn} \sum_z P_n(z|f) \log P(f|z) \\ & + \bar{\alpha} \sum_z \sum_f P(f|z) \log P(f|z) + \bar{\beta} \sum_n \sum_z P_n(z) \log P_n(z). \end{aligned} \tag{27}$$

Here, γ is a parameter that weights the data while $\bar{\alpha}$ and $\bar{\beta}$ are parameters weighting the prior.

In order to take care of the normalization constraints, the above equation must be augmented by appropriate Lagrange multipliers τ_n and ρ_z,

$$Q = \mathcal{L} + \sum_n \tau_n \Big(1 - \sum_z P_n(z) \Big) + \sum_z \rho_z \Big(1 - \sum_f P(f|z) \Big) \tag{28}$$

Maximization of Q with respect to $P_n(z)$ and $P(f|z)$ leads to the following sets of equations

$$\frac{\sum_n V_{fn} P_n(z|f)}{P(f|z)} + \alpha + \alpha \log P(f|z) + \rho_z = 0, \tag{29}$$

$$\frac{\sum_f V_{fn} P_n(z|f)}{P_n(z)} + \beta + \beta \log P_n(z) + \tau_n = 0, \tag{30}$$

where $\alpha = \bar{\alpha}/\gamma$ and $\beta = \bar{\beta}/\gamma$. We have replaced two parameters weighting the data and prior separately (γ and $\bar{\alpha}$ for equation (29), γ and $\bar{\beta}$ for equation (30)) by a single parameter that weights the prior with respect to the data (α and β in equations (29) and (30) respectively).

Now, consider solving for $P_n(z)$. Equation (30) can be written as

$$\frac{\omega_z}{P_n(z)} + \beta + \beta \log P_n(z) + \tau_n = 0, \tag{31}$$

where ω_z represents $\sum_f V_{fn} P_n(z|f)$. The above set of simultaneous transcendental equations for $P_n(z)$ can be solved using the Lambert's $\mathcal{W}$ function ([3]) as proposed by [1].

Lambert's $\mathcal{W}$ function is an inverse mapping satisfying

$$\mathcal{W}(y)e^{\mathcal{W}(y)} = y \qquad \Longrightarrow \qquad \log \mathcal{W}(y) + \mathcal{W}(y) = \log y$$

As shown in [1], we can set $y = e^x$ and work backwards towards equation (31) as follows,

$$\begin{aligned} 0 &= -\mathcal{W}(e^x) - \log \mathcal{W}(e^x) + x \\ &= \frac{-1}{1/\mathcal{W}(e^x)} - \log \mathcal{W}(e^x) + x + \log q - \log q \\ &= \frac{-q}{q/\mathcal{W}(e^x)} + \log q/\mathcal{W}(e^x) + x - \log q \end{aligned}$$

Setting $x = 1 + \tau_n/\beta + \log q$ and $q = -\omega_z/P_n(z)$, the above equation simplifies to equation (31):

$$\begin{aligned} 0 &= \frac{\omega_z/\beta}{-(\omega_z/\beta)/\mathcal{W}(-\omega_z e^{1+\tau_n/\beta}/\beta)} + \log \frac{-\omega_z/\beta}{\mathcal{W}(-\omega_z e^{1+\tau_n/\beta}/\beta)} \\ &\quad +1 + \frac{\tau_n}{\beta} \\ &= \frac{\omega_z/\beta}{P_n(z)} + \log P_n(z) + 1 + \frac{\tau_n}{\beta} \end{aligned}$$

which implies that

$$\hat{P}_n(z) = \frac{-\omega_z/\beta}{\mathcal{W}(-\omega_z e^{1+\tau_n/\beta}/\beta)}, \tag{32}$$

where equations (31) and (32) form a set of fixed-point iterations for τ_n, and thus the M-step for finding $P_n(z)$. [1] points out that these equations typically converge in 2-5 iterations. [2] provides details about how to compute the lambert's $\mathcal{W}$ function.

We can similarly solve for $P(f|z)$ by solving the set of transcendental equations given by equation (29) using Lambert's $\mathcal{W}$ function. It can be shown that it can be estimated as

$$\hat{P}(f|z) = \frac{-\xi/\alpha}{\mathcal{W}(-\xi e^{1+\rho_z/\alpha}/\alpha)}, \tag{33}$$

where we have let ξ represent $\sum_n V_{fn} P_n(z|f)$. Equations (29) and (33) form a set of fixed-point iterations and correspond to the M-step updates for $P(f|z)$.

References

[1] ME Brand. Pattern discovery via entropy minimization. In *Uncertainty 99: AISTATS 99*, 1999.

[2] ME Brand. Structure learning in conditional probability models via an entropic prior and parameter extinction. *Neural Computation*, 1999.

[3] RM Corless, GH Gonnet, DEG Hare, DJ Jeffrey, and DE Knuth. On the lambert $\mathcal{W}$ function. *Advances in Computational mathematics*, 1996.