Supplementary Materials: Modeling Heterogeneous Statistical Patterns In High-dimensional Data By Adversarial Distributions: An Unsupervised Generative Framework

Han Zhang[1], Wenhao Zheng[3], Charley Chen[1], Kevin Gao[1], Yao Hu[3], Ling Huang[2] and Wei Xu[1]
[1]Tsinghua University, Beijing, China
[2]AHI Fintech, China, [3]Youku Cognitive and Intelligent Lab, Alibaba Group

1 Proof of Theorem 1

As the ELBO is concave w.r.t. all variables, it can be easily shown that the solution in Theorem 1 is optimal. To show the uniqueness, we prove that each adversarial component $p_k(x_{nm}|d_n, \theta)$ will fit the corresponding pattern $p^{\star}_{k'}(x_{nm}|d_n)$. According to the definition

$$D_{KL}\left(p^{\star}_{k'}(x_{nm}|d_n) \| p_k(x_{nm}|d_n, \theta)\right) = \mathbb{E}_{x \sim p^{\star}_{k'}}[\log p_k(x_{nm}|d_n, \theta)] + H,$$

where H is the entropy of $p^{\star}_{k'}(x_{nm}|d_n)$, the KL divergence assumption in Theorem 1 indicates for all possible d_n and $\forall j \neq k$,

$$\mathbb{E}_{x \sim p^{\star}_{k'}}[p_k(x_{nm}|d_n, \theta)] > \mathbb{E}_{x \sim p^{\star}_{k'}}[p_j(x_{nm}|d_n, \theta)]. \tag{S1}$$

Since $\tilde{\mu}_{nmk} > 0$, multiplying Eq. (S1) with $q(d_n)\tilde{\mu}_{nmk}$ and summing over m, k and d_n gives

$$\begin{aligned}
&\lim_{N\to\infty} \frac{1}{N} ELBO(p_k \to p^{\star}_{k'}) \\
=&\lim_{N\to\infty} \frac{1}{N} \sum_{n,d_n,m,k} q(d_n)\tilde{\mu}_{nmk} \log p_k(x_{nm}|d_n, \theta) + const \\
\geq&\lim_{N\to\infty} \frac{1}{N} \sum_{n,d_n,m,k} q(d_n)\tilde{\mu}_{nmk} \log p_j(x_{nm}|d_n, \theta) + const \\
=&\lim_{N\to\infty} \frac{1}{N} ELBO(p_j \to p^{\star}_{k'})
\end{aligned}$$

which indicates that the optimal solution use $p_k(x_{nm}|d_n, \theta)$ to approximate the corresponding pattern $p^{\star}_{k'}(x_{nm}|d_n)$. Using EM algorithm gives the estimation of other parameters in Theorem 1, which completes the proof.

2 Derivation of EM Updates

In E step we approximate the likelihood by finding the best variational distribution, which is given by

$$p(d_n, \mathbf{f}_n | \mathbf{x}_n, \hat{\boldsymbol{\theta}}) = \frac{\prod_{g=1}^{G} \left\{ \hat{\pi}_g \prod_{m=1}^{M} \gamma_{ngm}^{f_{nm}} \cdot \bar{\gamma}_{ngm}^{1-f_{nm}} \right\}^{d_{ng}}}{\sum_{g'=1}^{G} \phi_{ng'}}.$$

The information of this distribution is summarized in Eq. (6). In M step we optimize the log-likelihood with respect to the model parameters:

$$\begin{aligned} \max_{\boldsymbol{\theta}} \quad Q(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) \quad &= \sum_{n=1}^{N} \sum_{d_n, \mathbf{f}_n} p(d_n, \mathbf{f}_n | \mathbf{x}_n, \hat{\boldsymbol{\theta}}) \log p(\mathbf{x}_n, \mathbf{f}_n, d_n | \boldsymbol{\theta}) \\ &- \sum_{g=1}^{G} \lambda_g^{(1)} \log \pi_g - \sum_{gmi} \lambda_{gmi}^{(2)} \left(\log \alpha_{gmi} - \log \beta_{gmi} \right), \end{aligned}$$

where the parameters lie in probability simplexes. For $\boldsymbol{\mu}_g$ we set

$$\begin{aligned} \frac{\partial Q}{\partial \mu_{gm}} &= \sum_{n=1}^{N} \sum_{d_n, \mathbf{f}_n} p(d_n, \mathbf{f}_n | \mathbf{x}_n, \hat{\boldsymbol{\theta}}) \frac{\partial \log p(\mathbf{f}_n | d_n, \boldsymbol{\mu})}{\partial \mu_{gm}} \\ &= \sum_{n=1}^{N} \frac{\hat{\pi}_g \prod_{m' \neq m} (\gamma_{ngm'} + \bar{\gamma}_{ngm'})}{\sum_{g'=1}^{G} \phi_{ng'}} \left(\frac{\gamma_{ngm}}{\mu_{gm}} - \frac{\bar{\gamma}_{ngm}}{1 - \mu_{gm}} \right) \end{aligned}$$

to zero and we obtain the update equation in Eq. (7). For β_{gmi} we need to introduce Lagrange multiplier ω to eliminate the probability simplex constraint:

$$\begin{aligned} \frac{\partial Q}{\partial \beta_{gmi}} &= \sum_{n=1}^{N} \sum_{d_n, \mathbf{f}_n} p(d_n, \mathbf{f}_n | \mathbf{x}_n, \hat{\boldsymbol{\theta}}) \frac{\partial \log p(x_{nm} | f_{nm}, d_n, \boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial \beta_{gmi}} + \lambda_{gmi}^{(2)} - \omega \\ &= \sum_{n=1}^{N} \sum_{n=1}^{N} \frac{\hat{\pi}_g \prod_{m' \neq m} (\gamma_{ngm'} + \bar{\gamma}_{ngm'})}{\sum_{g'=1}^{G} \phi_{ng'}} \cdot \frac{x_{nmi}}{\beta_{gmi}} + \lambda_{gmi}^{(2)} - \omega. \end{aligned}$$

Setting $\partial Q / \partial \beta_{gmi}$ to zero and using $\sum_i \beta_{gmi} = 1, \sum_i x_{nmi} = 1$ we obtain the result in Eq. (7).

Optimization w.r.t. α_{gmi} and π_g is difficult since the objective function is no longer convex. [1] proposed a technique to optimize such concave problems with convex regularizers in a probability simplex. Given a variable $\boldsymbol{\theta}$ in a $(K+1)$-simplex we can introduce a pseudo-Dirichlet prior to promote sparsity

$$Q(\boldsymbol{\theta}; \lambda, \epsilon) = \sum_{k=1}^{K} c_k \log \theta_k - \sum_{k=1}^{K} \lambda \log(\theta_k + \epsilon) + \mathit{constant},$$

where c_k is the observation counts for k-th possible value and $\lambda > 0$. [1] proved that $Q(\boldsymbol{\theta}; \lambda, \epsilon)$ has one global maximum, and we can use the fixed point iteration to calculate this maximum. Suppose $\sum_k c_k = c$, the update rule is

$$\theta_k = \frac{c_k + \theta_k \cdot \lambda \cdot \sum_{k=1}^{K} \theta_k / (\theta_k + \epsilon)}{c + \lambda / (\theta_k + \epsilon)}.$$

Using the fact that the sum of observation counts for $\boldsymbol{\pi}$ is N and $\sum_{n=1}^{N} \widetilde{\gamma}_{ngm}\widetilde{\phi}_{ng}$ for $\boldsymbol{\alpha}_{gm}$ and letting $\epsilon \to 0$ we obtain the update rules in Eq. (8).

References

[1] Martin O. Larsson and Johan Ugander. A concave regularization technique for sparse mixture models. In *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain.*, pages 1890–1898, 2011.