[4] Olivier Chapelle, Vladimir Vapnik, Olivier Bousquet, and Sayan Mukherjee. Choosing multiple parameters for support vector machines. *Machine Learning*, 46:131–159, 2002.

[5] Ronald A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7 Part II:179–188, 1936.

[6] A. E. Gelfand and A. F. M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85:398–409, 1990.

[7] Amir Globerson and Sam Roweis. Metric learning by collapsing classes. In *Advances in Neural Information Processing Systems 18*, 2006.

[8] Jacob Goldberger, Sam Roweis, Geoff Hinton, and Ruslan Salakhutdinov. Neighbourhood components analysis. In *Advances in Neural Information Processing Systems 17*, 2005.

[9] Yuhong Guo and Suicheng Gu. Multi-label classification using conditional dependency networks. In *Proceedings of the 22nd International Joint Conference on Artifical Intelligence*, 2011.

[10] Shuiwang Ji, Lei Tang, Shipeng Yu, and Jieping Ye. A shared-subspace learning framework for multi-label classification. *ACM Transactions on Knowledge Discovery from Data*, 4(2):8:1–8:29, 2010.

[11] Shuiwang Ji and Jieping Ye. Linear dimensionality reduction for multi-label classification. In *Proceedings of the 21st International Joint Conference on Artifical Intelligence*, 2009.

[12] Neil D. Lawrence and Michael I. Jordan. Semi-supervised learning via Gaussian processes. In *Advances in Neural Information Processing Systems 17*, 2005.

[13] Kai Mao, Feng Liang, and Sayan Mukherjee. Supervised dimension reduction using Bayesian mixture modeling. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, 2010.

[14] Radford M. Neal. *Bayesian Learning for Neural Networks*. Lecture Notes in Statistics 118. Springer, 1996.

[15] Cheong Hee Park and Moonhwi Lee. On applying linear discriminant analysis for multi-labeled problems. *Pattern Recognition Letters*, 29:878–887, 2008.

[16] Karl Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2:559–572, 1901.

[17] Francisco Pereira and Geoffrey Gordon. The support vector decomposition machine. In *Proceedings of the 23rd International Conference on Machine Learning*, 2006.

[18] James Petterson and Tiberio Caetano. Reverse multi-label learning. In *Advances in Neural Information Processing Systems 23*, 2010.

[19] Buyue Qian and Ian Davidson. Semi-supervised dimension reduction for multi-label classification. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence*, 2010.

[20] Piyush Rai and Hal Daumé III. Multi-label prediction via sparse infinite CCA. In *Advances in Neural Information Processing Systems 22*, 2009.

[21] Irina Rish, Genady Grabarnik, Guillermo Cecchi, Francisco Pereira, and Geoffrey J. Gordon. Closed-form supervised dimensionality reduction with generalized linear models. In *Proceedings of the 25th International Conference on Machine Learning*, 2008.

[22] Sajama and Alon Orlitsky. Supervised dimensionality reduction using mixture models. In *Proceedings of the 22nd International Conference on Machine Learning*, 2005.

[23] Liang Sun, Shuiwang Ji, and Jieping Ye. Hypergraph spectral learning for multi-label classification. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2008.

[24] Michael K.-S. Tso. Reduced-rank regression and canonical analysis. *Journal of the Royal Statistical Society: Series B (Methodological)*, 43:183–189, 1981.

[25] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. Mining multi-label data. In Oded Maimon and Lior Rokach, editors, *Data Mining and Knowledge Discovery Handbook*, pages 667–685. Springer, 2009.

[26] Hua Wang, Chris Ding, and Heng Huang. Multi-label linear discriminant analysis. In *Proceedings of the 11th European Conference on Computer Vision*, 2010.

[27] Kilian Q. Weinberger and Lawrence K. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10:207–244, 2009.

[28] Kai Yu, Shipeng Yu, and Volker Tresp. Multi-label informed latent semantic indexing. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2005.

[29] Shipeng Yu, Kai Yu, Volker Tresp, Hans-Peter Kriegel, and Mingrui Wu. Supervised probabilistic principal component analysis. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006.

[30] Min-Ling Zhang. LIFT: Multi-label learning with label-specific features. In *Proceedings of the 22nd International Joint Conference on Artifical Intelligence*, 2011.

[31] Min-Ling Zhang and Zhi-Hua Zhou. ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognition*, 40:2038–2048, 2007.

[32] Wei Zhang, Xiangyang Xue, Jianping Fan, Xiaojing Huang, Bin Wu, and Mingjie Liu. Multi-kernel multi-label learning with max-margin concept network. In *Proceedings of the 22nd International Joint Conference on Artifical Intelligence*, 2011.

[33] Yin Zhang and Zhi-Hua Zhou. Multilabel dimensionality reduction via dependence maximization. *ACM Transactions on Knowledge Discovery from Data*, 4(3):14:1–14:21, 2010.

A Variational Lower Bound for Multilabel Learning

The variational lower bound of our multilabel learning model can be written as

$$\mathcal{L} = \mathrm{E}_{q(\boldsymbol{\Theta},\boldsymbol{\Xi})}[\log p(\mathbf{Y},\boldsymbol{\Theta},\boldsymbol{\Xi}|\mathbf{X})] - \mathrm{E}_{q(\boldsymbol{\Theta},\boldsymbol{\Xi})}[\log q(\boldsymbol{\Theta},\boldsymbol{\Xi})]$$

where the joint likelihood is defined as

$$p(\mathbf{Y},\boldsymbol{\Theta},\boldsymbol{\Xi}|\mathbf{X}) = p(\boldsymbol{\Phi})p(\mathbf{Q}|\boldsymbol{\Phi})p(\mathbf{Z}|\mathbf{Q},\mathbf{X})p(\boldsymbol{\lambda})p(b|\boldsymbol{\lambda}) \\ p(\boldsymbol{\Psi})p(\mathbf{W}|\boldsymbol{\Psi})p(\mathbf{T}|\boldsymbol{b},\mathbf{W},\mathbf{Z})p(\mathbf{Y}|\mathbf{T}).$$

Using these definitions, the variational lower bound becomes

$$\begin{aligned}\mathcal{L} = {} & \mathrm{E}_{q(\boldsymbol{\Phi})}[\log p(\boldsymbol{\Phi})] + \mathrm{E}_{q(\boldsymbol{\Phi})q(\mathbf{Q})}[\log p(\mathbf{Q}|\boldsymbol{\Phi})] \\ & + \mathrm{E}_{q(\mathbf{Q})q(\mathbf{Z})}[\log p(\mathbf{Z}|\mathbf{Q},\mathbf{X})] + \mathrm{E}_{q(\boldsymbol{\lambda})}[\log p(\boldsymbol{\lambda})] \\ & + \mathrm{E}_{q(\boldsymbol{\lambda})q(\boldsymbol{b},\mathbf{W})}[\log p(\boldsymbol{b}|\boldsymbol{\lambda})] + \mathrm{E}_{q(\boldsymbol{\Psi})}[\log p(\boldsymbol{\Psi})] \\ & + \mathrm{E}_{q(\boldsymbol{\Psi})q(\boldsymbol{b},\mathbf{W})}[\log p(\mathbf{W}|\boldsymbol{\Psi})] \\ & + \mathrm{E}_{q(\mathbf{Z})q(\boldsymbol{b},\mathbf{W})q(\mathbf{T})}[\log p(\mathbf{T}|\boldsymbol{b},\mathbf{W},\mathbf{Z})] \\ & + \mathrm{E}_{q(\mathbf{T})}[\log p(\boldsymbol{y}|\mathbf{T})] - \mathrm{E}_{q(\boldsymbol{\Phi})}[\log q(\boldsymbol{\Phi})] \\ & - \mathrm{E}_{q(\mathbf{Q})}[\log q(\mathbf{Q})] - \mathrm{E}_{q(\mathbf{Z})}[\log q(\mathbf{Z})] \\ & - \mathrm{E}_{q(\boldsymbol{\lambda})}[\log q(\boldsymbol{\lambda})] - \mathrm{E}_{q(\boldsymbol{\Psi})}[\log q(\boldsymbol{\Psi})] \\ & - \mathrm{E}_{q(\boldsymbol{b},\mathbf{W})}[\log q(\boldsymbol{b},\mathbf{W})] - \mathrm{E}_{q(\mathbf{T})}[\log q(\mathbf{T})]\end{aligned}$$

where the exponential form expectations of the distributions in the joint likelihood can be calculated as

$$\mathrm{E}_{q(\boldsymbol{\Phi})}[\log p(\boldsymbol{\Phi})] = \sum_{f=1}^{D}\sum_{s=1}^{R}\Bigg((\alpha_\phi - 1)\widetilde{\log \phi_s^f} - \frac{\widetilde{\phi_s^f}}{\beta_\phi} \\ - \log\Gamma(\alpha_\phi) - \alpha_\phi \log \beta_\phi\Bigg)$$

$$\mathrm{E}_{q(\boldsymbol{\Phi})q(\mathbf{Q})}[\log p(\mathbf{Q}|\boldsymbol{\Phi})] = \sum_{s=1}^{R}\Bigg(-\frac{1}{2}\operatorname{tr}(\operatorname{diag}(\widetilde{\boldsymbol{\phi}_s})\widetilde{\boldsymbol{q}_s\boldsymbol{q}_s^\top}) \\ - \frac{1}{2}D\log 2\pi + \frac{1}{2}\log|\operatorname{diag}(\widetilde{\boldsymbol{\phi}_s})|\Bigg)$$

$$\mathrm{E}_{q(\mathbf{Q})q(\mathbf{Z})}[\log p(\mathbf{Z}|\mathbf{Q},\mathbf{X})] = \sum_{i=1}^{N}\Bigg(-\frac{1}{2}\widetilde{\boldsymbol{z}_i^\top\boldsymbol{z}_i} + \boldsymbol{x}_i^\top\widetilde{\mathbf{Q}}\widetilde{\boldsymbol{z}_i} \\ - \frac{1}{2}\operatorname{tr}(\widetilde{\mathbf{Q}\mathbf{Q}^\top}\boldsymbol{x}_i\boldsymbol{x}_i^\top) - \frac{1}{2}R\log 2\pi\Bigg)$$

$$\mathrm{E}_{q(\boldsymbol{\lambda})}[\log p(\boldsymbol{\lambda})] = \sum_{o=1}^{L}\Bigg((\alpha_\lambda - 1)\widetilde{\log\lambda_o} - \frac{\widetilde{\lambda_o}}{\beta_\lambda} - \log\Gamma(\alpha_\lambda) \\ - \alpha_\lambda\log\beta_\lambda\Bigg)$$

$$\mathrm{E}_{q(\boldsymbol{\lambda})q(\boldsymbol{b},\mathbf{W})}[\log p(\boldsymbol{b}|\boldsymbol{\lambda})] = \sum_{o=1}^{L}\Bigg(-\frac{1}{2}\widetilde{\lambda_o}\widetilde{b_o^2} - \frac{1}{2}\log 2\pi \\ + \frac{1}{2}\log\widetilde{\lambda_o}\Bigg)$$

$$\mathrm{E}_{q(\boldsymbol{\Psi})}[\log p(\boldsymbol{\Psi})] = \sum_{s=1}^{R}\sum_{o=1}^{L}\Bigg((\alpha_\psi - 1)\widetilde{\log\psi_o^s} - \frac{\widetilde{\psi_o^s}}{\beta_\psi} \\ - \log\Gamma(\alpha_\psi) - \alpha_\psi\log\beta_\psi\Bigg)$$

$$\mathrm{E}_{q(\boldsymbol{\Psi})q(\boldsymbol{b},\mathbf{W})}[\log p(\mathbf{W}|\boldsymbol{\Psi})] = \\ \sum_{o=1}^{L}\Bigg(-\frac{1}{2}\operatorname{tr}(\operatorname{diag}(\widetilde{\boldsymbol{\psi}_o})\widetilde{\boldsymbol{w}_o\boldsymbol{w}_o^\top}) - \frac{1}{2}R\log 2\pi \\ + \frac{1}{2}\log|\operatorname{diag}(\widetilde{\boldsymbol{\psi}_o})|\Bigg)$$

$$\mathrm{E}_{q(\mathbf{Z})q(\boldsymbol{b},\mathbf{W})q(\mathbf{T})}[\log p(\mathbf{T}|\boldsymbol{b},\mathbf{W},\mathbf{Z})] = \sum_{o=1}^{L}\sum_{i=1}^{N}\Bigg(-\frac{1}{2}\widetilde{(t_i^o)^2} \\ + (\widetilde{\boldsymbol{w}_o^\top}\widetilde{\boldsymbol{z}_i} + \widetilde{b_o})\widetilde{t_i^o} - \frac{1}{2}\Big(\operatorname{tr}(\widetilde{\boldsymbol{w}_o\boldsymbol{w}_o^\top}\widetilde{\boldsymbol{z}_i\boldsymbol{z}_i^\top}) + 2\widetilde{b_o\boldsymbol{w}_o^\top}\widetilde{\boldsymbol{z}_i} + \widetilde{b_o^2}\Big) \\ - \frac{1}{2}\log 2\pi\Bigg)$$

$$\mathrm{E}_{q(\mathbf{T})}[\log p(\boldsymbol{y}|\mathbf{T})] = 0$$

and the negative entropies of the approximate posteriors in the ensemble are given as

$$\mathrm{E}_{q(\boldsymbol{\Phi})}[\log q(\boldsymbol{\Phi})] = \sum_{f=1}^{D}\sum_{s=1}^{R}(-\alpha(\phi_s^f) - \log\beta(\phi_s^f) \\ - \log\Gamma(\alpha(\phi_s^f)) - (1-\alpha(\phi_s^f))\psi(\alpha(\phi_s^f)))$$

$$\mathrm{E}_{q(\mathbf{Q})}[\log q(\mathbf{Q})] = \\ \sum_{s=1}^{R}\Bigg(-\frac{1}{2}D(\log 2\pi + 1) - \frac{1}{2}\log|\Sigma(\boldsymbol{q}_s)|\Bigg)$$

$$\mathrm{E}_{q(\mathbf{Z})}[\log q(\mathbf{Z})] = \sum_{i=1}^{N}\Bigg(-\frac{1}{2}R(\log 2\pi + 1) - \frac{1}{2}\log|\Sigma(\boldsymbol{z}_i)|\Bigg)$$

$$\mathrm{E}_{q(\boldsymbol{\lambda})}[\log q(\boldsymbol{\lambda})] = \sum_{o=1}^{L}(-\alpha(\lambda_o) - \log\beta(\lambda_o) - \log\Gamma(\alpha(\lambda_o)) - (1-\alpha(\lambda_o))\psi(\alpha(\lambda_o)))$$

$$\mathrm{E}_{q(\boldsymbol{\Psi})}[\log q(\boldsymbol{\Psi})] = \sum_{s=1}^{R}\sum_{o=1}^{L}(-\alpha(\psi_o^s) - \log\beta(\psi_o^s) - \log\Gamma(\alpha(\psi_o^s)) - (1-\alpha(\psi_o^s))\psi(\alpha(\psi_o^s)))$$

$$\mathrm{E}_{q(\boldsymbol{b},\mathbf{W})}[\log q(\boldsymbol{b},\mathbf{W})] = \sum_{o=1}^{L}\left(-\frac{1}{2}(R+1)(\log 2\pi + 1) - \frac{1}{2}\log|\Sigma(b_o, \boldsymbol{w}_o)|\right)$$

$$\mathrm{E}_{q(\mathbf{T})}[\log q(\mathbf{T})] = \sum_{o=1}^{L}\sum_{i=1}^{N}\left(-\frac{1}{2}(\log 2\pi + \Sigma(t_i^o)) - \log\mathcal{Z}_i^o\right)$$

where $\Gamma(\cdot)$ denotes the gamma function and $\psi(\cdot)$ denotes the digamma function. The only nonstandard distribution we need to operate on is the truncated normal distribution used for the auxiliary variables. From our model definition, the truncation points for each auxiliary variable are defined as

$$(l_i^o, u_i^o) = \begin{cases}(-\infty, 0) & \text{if } y_i^o = -1\\ (0, +\infty) & \text{otherwise}\end{cases}$$

where l_i^o and u_i^o denote the lower and upper truncation points, respectively. The normalization coefficient, the expectation, and the variance of the auxiliary variables can be calculated as

$$\begin{aligned}\mathcal{Z}_i^o &= \Phi(\beta_i^o) - \Phi(\alpha_i^o)\\ \widetilde{t_i^o} &= \widetilde{\boldsymbol{w}_o^\top}\widetilde{\boldsymbol{z}}_i + \widetilde{b_o} + \frac{\phi(\alpha_i^o) - \phi(\beta_i^o)}{\mathcal{Z}_i^o}\\ \widetilde{(t_i^o)^2} - \widetilde{t_i^o}^2 &= 1 + \frac{\alpha_i^o\phi(\alpha_i^o) - \beta_i^o\phi(\beta_i^o)}{\mathcal{Z}_i^o} - \frac{(\phi(\alpha_i^o) - \phi(\beta_i^o))^2}{(\mathcal{Z}_i^o)^2}\end{aligned}$$

where $\phi(\cdot)$ is the standardized normal probability density function and $\{\alpha_i^o, \beta_i^o\}$ are defined as

$$\begin{aligned}\alpha_i^o &= l_i^o - \widetilde{\boldsymbol{w}_o^\top}\widetilde{\boldsymbol{z}}_i - \widetilde{b_o}\\ \beta_i^o &= u_i^o - \widetilde{\boldsymbol{w}_o^\top}\widetilde{\boldsymbol{z}}_i - \widetilde{b_o}.\end{aligned}$$