Approximating the posterior distribution of mixture weights with application to transcript expression estimation

Panagiotis Papastamoulis[1] **and Magnus Rattray**[1]

[1] *Faculty of Life Sciences, University of Manchester*

Abstract

This study focuses on approximating the posterior distribution of mixture weights ($\boldsymbol{\theta}$) given some data ($\boldsymbol{x}$) using Variational Bayes (VB) methods [1]. Standard VB implementation [4] for this problem approximates the joint posterior distribution $p(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{x})$ of parameters and latent variables ($\boldsymbol{z}$). It is demonstrated via simulation that this approach leads to variance underestimation. For this reason a new variational scheme is developed by integrating out the latent variables and targeting the marginal posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{x})$. The new approximation belongs to the richer family of Generalized Dirichlet distributions [8], while a stochastic approximation algorithm [6] performs the optimization in the corresponding spaces arising from two different parameterizations. Moreover, it is proven that the new solution leads to a better marginal log-likelihood bound compared to the former.

The method is applied to transcript expression estimation using high throughput sequencing of RNA (RNA-seq) technology. Mixture models are a natural way to deal with such problems, and Gibbs sampling has already been applied [3]. The application of Variational methods to these datasets is novel and leads to encouraging results. Finally, the variational solution is exploited in order to improve Markov Chain Monte Carlo (MCMC) sampling with the Delayed Rejection algorithm [7].

Keywords: Kullback-Leibler divergence, marginal likelihood bound, BitSeq, RNA-seq

AMS subject classifications: 62F15, 81T80, 92B15

Bibliography

[1] Bishop, C. (2006). *Pattern Recognition and Machine Learning*, Springer, New York.

[2] Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*, Springer, New York.

[3] Glaus, P., Honkela, A. and Rattray, M. (2012). Identifying differentially expressed transcripts from RNA-seq data with biological variation. *Bioinformatics* 28, 1721–1728.

[4] Hensman, J., Rattray, M. and Lawrence, N.D. (2012). Fast Variational Inference in the Conjugate Exponential Family. *Advances in Neural Information Processing Systems*, arXiv:1206.5162v2.

[5] Pan, Q., Shai, O., Lee, L.J., Frey, B.J. et al. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics* 40:1413–1415.

[6] Spall, J.C. (1992). Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control* 37, 332–341.

[7] Tierney, L. and Mira, A. (1999). Some adaptive Monte Carlo methods for Bayesian inference. *Statistics in Medicine* 18, 2507–2515.

[8] Wong, T.T. (1998). Generalized Dirichlet distribution in Bayesian analysis. *Applied Mathematics and Computation* 97, 165–181.