A COMPARISON OF BAYESIAN INFERENCE TECHNIQUES FOR SPARSE FACTOR MODELS

RESEARCH PROPOSAL FOR AMSI VACATION RESEARCH SCHOLARSHIPS 2019/20

Yong See Foo

Factor analysis is a dimension reduction technique which linearly maps high dimensional data onto a lower dimensional subspace. Given p observations $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_p]$ each with k variables, factor analysis attempts to describe the data using l latent variables (where $l \ll k$) by finding a loading matrix $\mathbf{W} \in \mathbb{R}^{k \times l}$ and factors $\mathbf{X} \in \mathbb{R}^{l \times p}$ such that

$$\mathbb{E}[\mathbf{Y}] = \mathbf{W}\mathbf{X}.$$

For example, in computational biology, dimension reduction techniques are widely used for studies of gene regulatory networks. In this context, $\mathbf{Y}$ represents gene expression data measured on k genes from p samples. Biological theory expects that the gene expression levels are regulated by a number of unobserved transcription factors far smaller than the number of genes. These may be discovered through the latent variables in factor analysis. Hartemink [2005] demonstrates that the gene regulatory networks are sparsely connected, where each transcription factor is responsible for regulating only a small number of gene expression levels. Hence, to obtain latent variables that are more likely to represent transcription factors, *sparse factor models* are used such that many entries of $\mathbf{W}$ are zero. This can be achieved within a Bayesian framework by using a sparsity-inducing prior for $\mathbf{W}$. The work of Sharp [2011] focuses on using zero-norm priors that have a non-zero prior probability of an entry of $\mathbf{W}$ being precisely zero. In general, sparse factor models are capable of providing clearer model interpretations.

As a direct calculation of the posterior is computationally intractable, approximate Bayesian inference techniques are used instead, e.g. Markov chain Monte Carlo methods (MCMC) and variational inference. MCMC simulates the posterior by constructing a Markov chain whose equilibrium distribution is the posterior; whereas variational inference selects a member from a family of densities such that the dissimilarity between said member and the posterior is minimised. These methods have their respective advantages and disadvantages: MCMC is capable of achieving high accuracy yet is computationally intensive as it suffers from slow mixing due to correlated samples; whereas variational inference reaches convergence more quickly yet is less accurate due to the restriction of a predefined family of densities (Sharp [2011]). The R package `slfm` (Duarte and Mayrink [2019]) implements a MCMC approach to sparse factor models, however there is no publicly available R implementation of variational inference for sparse factor models.

The aims and tentative timeline of the project are as follows:

- Review and derive MCMC and variational inference techniques for sparse factor models - *2.5 weeks*
- Implement variational inference for sparse factor models - *1.5 weeks*
- Compare the performance of MCMC and variational inference techniques by using RNA-seq data for inference of gene regulatory networks - *1 week*
- Create an R package implementing variational inference for sparse factor models - *1 week*

References

J. D. N. Duarte and V. D. Mayrink. slfm: An R package to evaluate coherent patterns in microarray data via factor analysis. *Journal of Statistical Software*, 90(9):1–22, 2019. doi: 10.18637/jss.v090.i09.

A. J. Hartemink. Reverse engineering gene regulatory networks. *Nature Biotechnology*, 23(5):554–555, 05 2005. doi: 10.1038/nbt0505-554.

K. Sharp. *Effective Bayesian Inference for Sparse Factor Analysis Models*. PhD thesis, University of Manchester, 2011.