Learning Graphical Model Structure with Sparse Bayesian Factor Models and Process Priors

Ricardo Henao, Ole Winther
rhenao@binf.ku.dk, owi@imm.dtu.dk

February 18, 2009

Abstract

In this paper we present an algorithm to learn factor models and directed acyclic graphs (DAG) within the same framework. It is based upon starting with inference of an identifiable sparse Bayesian factor model. A stochastic search over variable and latent factor orderings gives a candidate set of variable permutation compatible with a lower triangular loading matrix representation. These candidate orderings are then used as starting point for inference in a sparse DAG model. In previous work, we considered a heavy-tailed independent identically prior for the factors. We demonstrated in simulations that we could recover the correct ordering from purely observational data (no interventions) for both artificial settings and to some degree recover the text book protein-signaling network in [1]. In this work we consider E. coli gene expression profile data recorded across time [2], in particular, samples from 100 genes were taken at 5, 15, 30 and 60 min, and every hour until 6 hours after transition from glucose to acetate. Our model is extended to handle the smoothness of the data by using a Gaussian process prior for the factors and tested jointly with its original version to highlight the importance of using process priors for the factors. We also compare with a number of different approaches for gene networks inference based on time-series expression profiles [3].

Now we give some details about the model. Assuming that the observed variables can be ordered in such way they can be represented as a DAG and that the value of each variable is a linear combination of values already taken by previous variables plus a driving signal, we can write a data vector $\mathbf{x}$ with d variables as $\mathbf{x} = \mathbf{PAPx} + \mathbf{z}$, where $\mathbf{A}$ is a strictly lower triangular weight matrix, $\mathbf{P}$ is a permutation matrix encoding the correct order of the variables and $\mathbf{z}$ is the driving signal. If $\mathbf{A}$ is square we can rewrite the problem as $\mathbf{x} = \mathbf{Bz} = \mathbf{P}(\mathbf{I}-\mathbf{A})^{-1}\mathbf{Pz}$ and we end up with a linear factor model with two restrictions, (i) $\mathbf{B}$ must be permutable to a triangular form and (ii) $\mathbf{z}$ must be non-Gaussian independent variables or a process accounting for time correlations depending on the data. In order to estimate the factor model we specify a Bayesian model where $\mathbf{B}$ has a slap and spike mixture prior to allow for sparsity [4], $\mathbf{z}$ has a Laplace distribution or a Gaussian process if the observed data is a time-series and the inference process is carried out by Gibbs sampling. The factor model is invariant to $\mathbf{P}$ and $\mathbf{P}_c$ but we can make a stochastic search for $\mathbf{P}$ and $\mathbf{P}_c$ within the Gibbs sampling by accepting new permutations matrices according log likelihood ratios (Metropolis-Hastings) for $\mathbf{PBP}_c$ *masked to be lower triangular*. This produces a list of candidate orderings that can be used in the DAG estimation step by specifying a similar Bayesian model on $\mathbf{x} = \widehat{\mathbf{P}}\mathbf{A}\widehat{\mathbf{P}}\mathbf{x} + \mathbf{z}$, were $\widehat{\mathbf{P}}$ is a candidate ordering and $\mathbf{A}$ is strictly lower triangular, again with slap and spike priors. The final outcome of the algorithm is an ensemble of factor and DAG models. Model selection among these are performed using log likelihoods.

[1] K. Sachs, O. Perez, D. Pe'er, D.A. Lauffenburger, and G.P. Nolan, Causal protein-signaling networks derived from multiparameter single-cell data, Science 308:523–529 (2005).
[2] K.C. Kao, Y-L. Yang, R. Boscolo, C. Sabatti, V. Roychowdhury and J.C. Liao, Transcriptome-based determination of multiple transcription regulator activities in Escherichia coli by using network component analysis, PNAS 13:641–646 (2004).
[3] M. Bansal, V. Belcastro, A. Ambesi-Impiombato and D. di Bernardo, How to infer gene networks from expression proles, Molecular Systems Biology 3:78 (2007).
[4] M. West, Bayesian factor regression models in the "large p, small n" paradigm, in J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, and M. West, editors, Bayesian Statistics 7, pages 723–732, Oxford University Press (2003).