Published in final edited form as:

IEEE Int Workshop Genomic Signal Process Stat. 2012 December ; 2012: 113–117. doi:10.1109/
GENSIPS.2012.6507742.

Sparse Bayesian Graphical Models for RPPA Time Course Data

Riten Mitra,

ICES, University of Texas at Austin, rmitra1@math.utexas.edu

Peter Mueller,

Department of Mathematics, University of Texas at Austin, pmueller@math.utexas.edu

Yuan Ji,

CCRI, NorthShore University HealthSystem, yji@northshore.org

Gordon Mills, and

Department of Systems Biology, UT MD Anderson Cancer Centre, gmills@mdanderson.org

Yiling Lu

Department of Systems Biology, UT MD Anderson Cancer Centre, yilinglu@mdanderson.org

Abstract

Advances in functional proteomic technologies have significantly enriched our knowledge of protein functions and their interactions in bio-molecular pathways. We discuss inference for RPPA (reverse phase protein array) data that measure the expression of the protein markers over time. We exploit the dynamical nature of the experiment to build a directed network of protein interactions. For this, we employ a Bayesian graphical model with an informative prior that favors sparsity. Conditional on the network, we model dependence at the level of latent binary indicators rather than the raw expression measurements. One of the key features of the proposed approach is a hierarchical model that allows for the dependence structure to be shared across different experiments, in the case of the motivating application across different drugs and doses. This is critical to facilitate meaningful inference with the limited available sample sizes. The second key feature is a sparsity inducing prior on the dependence structure.

We show an application of the method to data measuring abundance of phosphorylated proteins in a human ovarian cell line.

I. Introduction

We propose a sparse Bayesian graphical model to analyze data from protein expression time course experiments. The model generates a network that describes a dynamic dependence structure of a set of proteins over time. The proposed model distinguishes itself from earlier dependence models in three important aspects. First, instead of directly modeling the raw data, we propose graphical prior models based on latent binary signals that represents activation status of the proteins. This step filters out the high noise that is typically seen in high throughput protein or gene expression data. Second, we specify the prior model to

favor sparse graphs. This is achieved by imposing a suitable prior distribution on the probability of edge inclusion. Third, and perhaps most important, we set up a hierarchical model that allows us to share the same dependence structure across several related experiments. Without this assumption meaningful inference on the dependence structure would be impossible with the limited available sample sizes. In this paper we review the main features of the approach. A more detailed description of the statistical methodology and the data analysis appears in [1]. Recent approaches to analyzing time course data include Bayesian clustering, network modeling and Boolean propagation networks ([2], [3], [4])

II. Data

Reverse phase protein arrays (RPPA) [5] are high-throughput functional proteomic experiments that quantify expression for targeted proteins. Each slide contains multiple samples and is hybridized against an antibody that binds to one specific protein. Separate slides are used for each target protein. In this paper we analyze data from a pathway-inhibition RPPA experiment on ovarian cancer cell lines. The experiment aimed to detect the time-course behavior of $m = 67$ disease markers in the PI3K pathway in response to three different drugs at four different dose levels. For each combination of drug and dose level, we had three replicate measurements for each protein. The three replicate measurements were averaged to give a single measurement for each dose, drug, protein and time point. The index, $i = 1, \dots, m$ represents a protein, the index $\ell = 1, \dots, L$ represents a drug, the index $k = 1, \dots, K$ represents a dose level and the index $t = 1, \dots, 8$, represents a time point. That is, the measurement $y_{ik\ell t}$ reports the expression of protein i for drug ℓ at dose k and time t .

A key challenge in inference for this data are the small sample sizes for each drug and dose combination that make separate inference for each drug or drug/dose combination impractical.

III. Network Model

A. Dependence graph G

The inference goal is to describe the dependence of a set of protein marker expressions at time t on the same expressions at time $t - 1$. We model the dependence through a graph $G = \{V, E\}$ where V is the set of vertices and E is the set of edges. The vertices $i = 1, \dots, m$ of the graph represent the proteins and presence of a directed edge $(i, j) \in E$ implies that the expression of protein i at time $t - 1$ influences the expression of protein j at time t . The graph G can be displayed as a directed network with nodes $i = 1, \dots, m$ that includes edges in both directions as well as self-cycles. A self-cycle (i, i) implies that activation of this protein at time t depends on activation of the same protein at time $t - 1$.

B. Prior $p(G)$

Two modeling choices related to G are critical to delivering the desired inference on the dependence structure of proteins in the pathway. The first assumption is shared dependence structure G across all drugs and dose levels. In other words, G is not indexed by ℓ and k .

Without this assumption, and keeping all other model elements unchanged, posterior inference on G is too diffuse to be of practical use.

The second assumption is an informative sparsity-inducing prior on G .

With small sample sizes, vague priors would lead to a practical identifiability problems in the posterior probability model. Let $G_{ij} = I\{(i, j) \in E\}$ denote an indicator for the presence of an edge connecting nodes i and j in graph G . We define $p(G)$ as an exchangeable prior on G_{ij} with a hyper-parameter π .

$$p(G_{ij}=1)=\pi \text{ and } \pi \sim \text{Be}(a, b). \quad (\text{III.1})$$

The hyper-parameter π represents the probability of edge inclusion. The model shrinks the posterior probabilities $p(G_{ij} = 1 | y)$ to the overall inclusion probability π and allows for learning about π .

C. Sampling model

To complete the model, we state the joint distribution of y conditional on G . Gaussian Graphical models (GGMs) [6], [7], [8] are popular approaches that model y as a multivariate normal distribution. However, typically in many bio-molecular pathways, the dependence acts only between the latent activation states of the proteins. That is, the dependence of protein expression is a result of co-activation of the corresponding proteins, controlled by a common mechanism such as co-expression of mRNAs. Motivated by this, we introduce an additional layer of latent binary variables $e_{ik\ell}$ between the graph and the continuous measurements.

The $e_{ik\ell}$ are interpreted as indicators for protein activation.

We specify the sampling model for the observed expressions $y_{ik\ell}$ conditioned on $e_{ik\ell}$ as a mixture of normals.

Let $\theta = (\mu_{ci}, \sigma_{ci}, i = 1, \dots, m, c = 1, 2)$ denote a set of parameters that index the sampling model. We assume

$$p(y_{ik\ell} | e_{ik\ell}, \theta) = \begin{cases} N(\mu_{1i}, \sigma_{1i}^2) & e_{ik\ell}=0, \\ N(\mu_{2i}, \sigma_{2i}^2) & e_{ik\ell}=1. \end{cases} \quad (\text{III.2})$$

D. Dynamic prior on protein activations $e_{ik\ell}$

We complete the model construction with an autoregressive model for $e_{ik\ell}$ conditional on $e_{i,k,l,t-1}$.

Let $e_{\bullet,t} = (e_{ik\ell}, i = 1, \dots, m, k = 1, \dots, K, \ell = 1, \dots, L)$ denote the set of indicators at time t .

$$p(e_{ik\ell}=1 | e_{\bullet,t-1}, \beta, G) = \frac{\exp(\beta_i + \sum_{i \rightarrow j} \beta_{ij} e_{ik\ell} e_{jk,l,t-1})}{1 + \exp(\beta_i + \sum_{i \rightarrow j} \beta_{ij} e_{ik\ell} e_{jk,l,t-1})},$$

and $\text{logit}\{p(e_{ikl} = 1 | \beta)\} = \beta_i$ at $t = 1$. The coefficients β_{ij} are zero for directed edges (i, j) not present in G . We highlight this by including G in the conditioning set for $p(e_{ikl} | \dots)$. We assume independent normal priors $p(\beta_{ij} | G) \propto N(0, 50)$ and $\beta_i \propto N(0, 50)$. We specify the priors for the parameters in θ as $\mu_{ik} \sim N(0, \tau = 100)$, $k = 1, 2$. For efficient MCMC performance, we fix σ_{ik} to $.5s_i$ where s_i is the empirical standard deviation of the expressions of protein i . We fixed the hyper-parameters of the beta prior in (III.1) as $a = 2$ and $b = 10$ respectively.

E. Posterior simulation

The full conditional posterior distributions for e , β , and the sampling model parameters are straightforward allowing us to implement posterior Markov chain Monte Carlo (MCMC) simulation with Gibbs sampling transition probabilities. Updating G requires transdimensional MCMC, which is implemented by a reversible jump MCMC [9] transition probability with birth and death steps. Posterior estimates are computed using MCMC posterior simulation with a total of 10000 iterations and an initial burn-in of 4500 iterations. We then estimate the graph G by including all edges (i, j) with posterior inclusion probability

$p(G_{ij} = 1 | \mathbf{y}) > .5$. Figure 1 shows the estimated posterior network.

An important feature of our approach is that we can report posterior inference on any subgraphs of interest and quantify uncertainties over these graphs through posterior distribution. In Figure 2 we show the posterior distribution over edges connected to the protein TSC2 which has the highest number of edges in the posterior estimate graph in Figure 1.

IV. Conclusion

Network modeling has become an important tool to study interactions in proteomics, and hence necessary for developing novel drugs and therapies. We presented an approach to infer signaling pathways from time course experiments that is based on a hierarchical Bayesian graphical model. Inference allows investigators to discover interactions between proteins at consecutive time points. The novelty in our approach arises from borrowing strength across multiple related experiments, using a sparse prior on graphs to accommodate large networks and introducing an additional layer of latent variables for protein activation.

References

1. Riten, M.; Mueller, P.; Ji, Y. A hierarchical sparse graphical model: Bayesian inference for multiple related RPPA studies. ICES, University of Texas at Austin, Tech. Rep. 2012.
2. Inoue L, Neira M, Nelson C, Gleave M, Etzioni R. Cluster-based network model for time-course gene expression data. *Biostatistics*. 2007; 8(3):507–525. [PubMed: 16980695]
3. Telesca D, Inoue L, Neira M, Etzioni R, Gleave M, Nelson C. Differential expression and network inferences through functional data modeling. *Biometrics*. 2009; 65(3):793–804. [PubMed: 19053995]
4. Bender C, Henjes F, Fröhlich H, Wiemann S, Korf U, Beißbarth T. Dynamic deterministic effects propagation networks: learning signalling pathways from longitudinal protein array data. *Bioinformatics*. 2010; 26(18):i596–i602. [PubMed: 20823327]

5. Tibes R, Qiu Y, Lu Y, Hennessy B, Andreeff M, Mills G, Kornblau S. Reverse phase protein array: validation of a novel proteomic technology and utility for analysis of primary leukemia specimens and hematopoietic stem cells. *Molecular cancer therapeutics*. 2006; 5(10):2512. [PubMed: 17041095]
6. Jones B, Carvalho C, Dobra A, Hans C, Carter C, West M. Experiments in Stochastic Computation for High-Dimensional Graphical Models. *Statistical Science*. 2004; 20:388–400.
7. Daniels M, Pourahmadi M. Bayesian analysis of covariance matrices and dynamic models for longitudinal data. *Biometrika*. 2002; 89:553–566.
8. Dobra A, Hans C, Jones B, Nevins JR, West M. Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*. 2004; 90:196–212.
9. Green R. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*. 1995; 82

Author Manuscript

(a) 57

(b) 28

(c) 5

(d) 4

Fig. 2.

The four most frequent configurations (a through d) of a subgraph consisting of 5 edges. The posterior probabilities (in percent) are given below each subgraph.