Temporal Graphical Models for Cross-Species Gene Regulatory Network Discovery

Yan Liu [*†] Alexandru Niculescu-mizil [†] Aurelie Lozano [†] Yong Lu [‡§]

April 5, 2010

Abstract

Many genes and biological processes function in similar ways across different species. Cross-species gene expression analysis, as a powerful tool to characterize the dynamical properties of the cell, has found a number of applications, such as identifying a conserved core set of cell cycle genes. However, to the best of our knowledge, there is limited effort on developing appropriate techniques to capture the causality relations between genes from time-series microarray data across species. In this paper, we present hidden Markov random field regression with L_1 penalty to uncover the regulatory network structure for different species. The algorithm provides a framework for sharing information across species via hidden component graphs and is able to incorporate domain knowledge across species easily. We demonstrate our method on two synthetic dataset and to discover causal graphs using innate immune response data.

1 Introduction

The activity of genes in a living cell is coordinated by a regulatory network that regulates gene expression conditioned on environmental stimuli. With genome-wide expression profiles, it is possible to reverse-engineer gene regulatory networks [12], which is essential for understanding how the cell functions. However, it remains a challenging task due to inherent and observational noise in expression data, the need to identify for each gene a small number of regulators among thousands of genes, and a limited number of samples in each experiment.

Combining expression data from different species has been shown to help discovery of true association between genes [3]. This is because many genes across species perform similar functions or share the same regulatory relations, and one can exploit information on related genes in different species. Similarly, expression data from different environmental conditions or from multiple cell types can be used to improve prediction of gene functions [31], because many genes may share similar activity and regulatory patterns across conditions and cell types.

In addition to improving prediction quality, cross-species expression analysis can identify conserved/common regulatory relations, which are more likely to play essential roles, as well as species-specific regulatory relations [37]. In the case of different environmental conditions and cell types, a combined analysis can identify common regulatory patterns as well as those specific to one cell type and/or one condition. With the exponential accumulation of microarray datasets, the benefits of cross-species analysis of expression data become increasingly apparent [29].

Computationally, inferring regulatory network by cross-species analysis can be viewed as a multitask learning problem. A multitask learning method performs several related learning tasks simultaneously, borrowing information across tasks, instead of learning each task independently. In our application, each task refers to learning a regulatory network from time-series microarray data generated from one cell type, in a single species, and under one environmental condition. To the best of our knowledge, there is no systematic

[*]Corresponding author: phone (914)245-1224; email {liuya@us.ibm.com}
[†]IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, USA
[‡]Harvard Medical School, Harvard University, Boston, MA 02115
[§]Corresponding author: phone (617)432-3552; email {Yong_Lu@hms.harvard.edu}

approach to jointly discover regulatory networks for several species by leveraging information across multiple species, cell types, and environmental conditions.

A number of methods have been proposed for learning regulatory networks in a single species [19]. However, these methods do not take into account temporal patterns in time-series gene expression data. Other methods have been proposed to exploit information in temporal expression patterns. [27] propose to apply temporal causal modeling methods to causality inference on expression data, which provides useful insights on the regulatory relationships between genes. The method in [27] combines Granger causality [20], an operational definition of causality well known in econometrics, with L1 regression algorithms, to perform causality inference involving many variables. Similar methods have received considerable attention in data mining problems [9, 34].

In this paper, we propose a novel probabilistic graphical model approach for learning regulatory networks from expression data in multiple species, cell types, and environmental conditions. It is based on the temporal causal models [9, 34, 27], but unlike [27], which can handle only one task (i.e. one species, cell type or environmental condition), our proposed method performs regulatory network discovery in a multitask learning manner, simultaneously learning from multiple datasets. We assume the regulatory networks are composed of a mixture of hidden component networks, which may be shared across species, cell types, and environmental conditions. Depending on the combination of species, cell type, and condition, the selection of component networks can be different, and the similarity can be guided by parameters, e.g. the evolutionary distance between species.

We infer the hidden component networks as well as the mixture assignments by defining the joint probability of observations (i.e. the microarray data from all species, cell types, and environmental conditions) and the hidden states (i.e. the mixture assignment) over graph G via hidden Markov random field (hMRF) with L1 penalty. One major advantage of our model is that domain knowledge on the evolution distance of species can be naturally incorporated in the graph G in order to provide constraint on mixture selections during inference.

In a related work, [6] proposes to use differential equations to infer regulatory networks by combining evolutionary cost and gene expression data across species. Unlike their work, which does not model time lag effect, our method explicitly takes into account the information from multiple previous time points when inferring causality, which is able to better capture the biological system. There are other related work that address alignment of biological networks across species [37]. Network alignment methods take networks of the same type from several species as input, and the goal is to identify functionally conserved subnetworks. In contrast, our method takes gene expression time series from multiple species, and simultaneously infers the causal relationship between genes as well as similar subnetworks across species. Another related work is local alignment of network motifs [2], but it aims to address different goals, i.e. given the input of a single network in one species and a list of motifs, finding significant motifs present in the network. In [30, 31], a genes dynamic property is summarized by computing an expression score from the time series, while our method uses all time points to infer causality, without first collapsing them into a single score.

The rest of the paper is organized as follows: we first review the temporal graphical modeling based on Granger causality in Section 2; then we motivate the challenges in cross-species analysis and describe the details of our proposed algorithms. We show experiment results on two synthetic datasets and one application data on cross-species in Section 3. Finally, we summarize the paper and conclude with future work.

2 Methodology

Learning the graph structures of regulatory networks from microarray data have found great success. Recent progress on structure learning establishes L_1-based algorithms as one of the most promising techniques for this task, especially for applications with inherent sparse graph structures [33, 42, 43, 18]. Additionally, L_1-based algorithms have been adapted to Granger casuality to discover the temporal "causal" networks between genes from time-series microarray data, which reveals important dependency information between current observations and histories [27]. Such approaches are based on the notion of Granger causality [20] between time-series. In what follows, we first review the Granger graphical models, then introduce the approach of hidden Markov random field regression for cross-species gene regulatory network discovery.

2.1 Graphical Granger Modeling

"Granger Causality" [20] was introduced by the Nobel prize winning economist, Clive Granger, and has proven useful as an *operational* notion of causality in time series analysis in the area of econometrics. It is based on the intuition that a cause should necessarily precede its effect, and in particular that if a time series variable causally affects another, then the past values of the former should be helpful in predicting the future values of the latter. More specifically, let $\{x_{1,t}\}_{t=1}^{T}$ denote the time series variables for x_1 and $\{x_{2,t}\}_{t=1}^{T}$ the same for x_2. A time series x_1 is said to "Granger cause" another time series x_2, if given the following two regressions:

$$x_{2,t} \approx \sum_{j=1}^{d} a_j \cdot x_{2,t-j} + \sum_{j=1}^{d} b_j \cdot x_{1,t-j} \ (1), \qquad x_{2,t} \approx \sum_{j=1}^{d} a_j \cdot x_{2,t-j} \ (2)$$

where d is the maximum "lag" allowed in past observations, (2) is more accurate than (1) with a statistically significant advantage, such as F-test [1].

The notion of Granger causality was defined only for a pair of time series. Recently, several graphical modeling approaches have been developed to determine the causal relationships between *multiple* time series variables [1, 28, 27]. These approaches are based on L_1 regularized regression algorithms, such as Lasso, as a more convenient, efficient, and effective alternative to the application of exhaustive pairwise Granger tests among all the time series. Taking three time series x_1, x_2, x_3 as an example, for all i, these approaches regress $x_{i,t}$ in terms of the previous d values of all the time series, applying an L_1 penalty on the coefficients:

$$\hat{\beta} = \arg\min_{\beta} \sum_{t} (x_{1,t} - \sum_{i=1}^{3} \sum_{j=1}^{d} \beta_{i,j} x_{i,t-j})^2 + \lambda(\|\beta\|_1)$$

L_1 regularization is well known for variable selection, i.e. variables that are not significantly improving the accuracy of the model will have their values set to 0. This can be readily used to determine causality in the Granger sense: if any of the coefficients corresponding to a past value of x_j is non-zero, it means that it helps significantly to improve the accuracy of modeling the current value of x_i, and thus x_j is a cause of x_i in a Granger sense. We can represent the causal relationships between variables via a feature graph (see Figure 1(a) for an example).

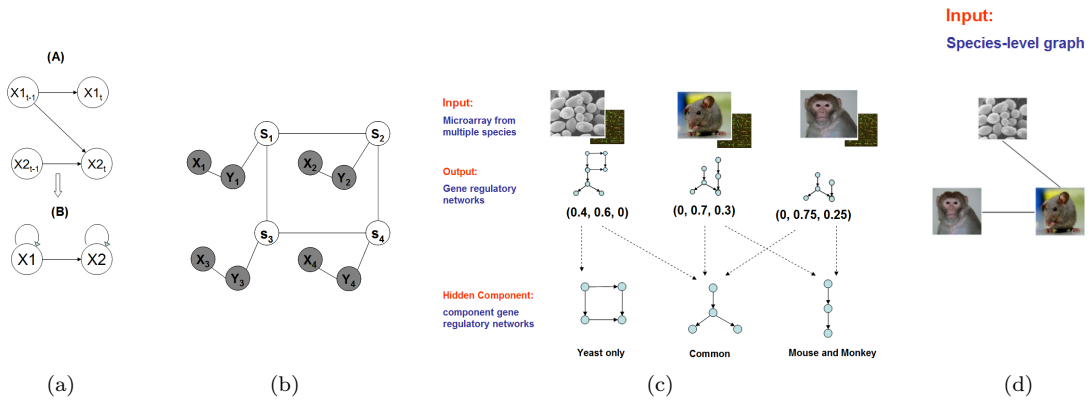
In this paper, we use the term "feature" to mean a time series (e.g. x) and use temporal variables or lagged variables to refer to the individual variables (e.g. x_t). In the context of microarray time series, a feature denotes the time series of expression levels of a gene, while a temporal variable or a lagged variable refers to the expression level of a gene at a given time point. Notice that temporal causal modeling is not limited to one single method, but a family of models that capture the temporal causal relations between time-series data. For example, recent advances in regularization theory have led to a series of extensions of the original lasso algorithm, such as elastic net and group lasso, which have been adapted to temporal graphical modeling and demonstrated effectiveness on different applications [28, 27].

2.2 Cross-species Regulatory Network Discovery

In Section 1, we identify our task of cross-species microarray analysis as an application of multi-task learning. One of the dominant approaches in multitask learning is to model the tasks as generated from a linear combination of a set of base components (classifiers or networks). In other words, the relatedness of multiple tasks can be explained by the fact that they share a certain number of hidden components [7]. Borrowing the idea, one simple approach to cross-species structure learning is to assume that the networks are from a mixture of hidden base networks. Mapping back to the Granger graphical models, we can think of the observations at time t, x_t, as generated from a mixture of regressions of previous observations x_{t-1}. Depending on the species, cell type and environment condition, the assignment of mixtures is different.

Notice that one major difference between our application and most previous multitask learning settings is: we are also given rich prior knowledge on the relations between these tasks (for example, the evolutionary

[1] Notice that the Granger Causality is not meant to be equivalent to true causality, but is merely intended to provide useful information regarding causation.



(a) (b) (c) (d)

Figure 1: (a) Demonstration to convert a temporal graph to feature graph; (b) Graphical model representation of hidden Markov random field regression; (c,d) Demonstration of hMRF-regression for cross-species gene regulatory network discovery. Input: microarray data from multiple species (c) and domain knowledge on the similarity between species (d). Output: the regulatory networks of each species. The algorithm assumes the regulatory networks of each species are generated from a linear mixture of common composite graphs.

distance between species). This information can be abstracted as a graph G, in which node corresponds to a cell type in a species under certain condition and there is an edge between nodes if they share the same cell type, species, or condition. In this way, we are able to better infer the hidden component networks and select component for each species.

2.2.1 Preliminary: Data processing and graph construction

In order to conduct the cross-species microarray analysis, we need to determined the subset of genes that are orthologous across species. For each species, we first select the genes that appear in all the experiments; then we obtain the orthologs between species A and B, and select the common regulatory genes that either themselves or their orthologs can be found in the microarray data.

The prior knowledge can be represented by a high-level graph. In general, a node represents one species (or a cell type under some condition); there is an edge between the nodes for genes in the same species but in different experiments because we expect many of them would exhibit similar regulatory relations; there is also an edge between the same cell type/condition across different species if the distance between the two species is smaller than a threshold, since some of the genes may share similar regulated functions as their orthologs. One example of the species-level graph for three species is as shown in Figure 1.

2.2.2 Hidden Markov Random Field Regression

A hidden Markov random field (hMRF) [24] is a generalization of hidden Markov model (HMM). It has an underlying Markov random field, i.e. an undirected graphical model with some graph structure, instead of a simple chain structure as in HMM. hMRF has been successfully applied to many applications with correlated data, such as image segmentation , genetics , and disease mapping. Therefore we use the framework to model the species-level constraint.

In order to integrate the species-level constraint with regulatory network discovery using Granger temporal models, we extend the hMRF to handle regression. The basic assumption is that the time-series are generated from a stochastic process, where the current observation of node i (i.e. species i) $x_t^{(i)}$ is conditionally dependent on the histories $x_{t-1}^{(i)}, \ldots, x_{t-L}^{(i)}$, as well as an underlying process of hidden states $s^{(i)}$. The semantic of the hidden states are the assignment of hidden component networks. More specifically, given the time-series observations of node i, $x^{(i)} = [x_1^{(i)}, \ldots, x_N^{(i)}]^T$, we can define a pairwise hMRF as a product

of node potentials and edge potentials:

$$P(x^{(1)},\ldots,x^{(M)},s^{(1)},\ldots,s^{(M)}|\beta,\Sigma,w)=\frac{1}{Z}\prod_{i=1}^{M}\Phi(x^{(i)},s^{(i)}|\beta,\Sigma)\prod_{(i,j)\in\text{edge}}\Phi(s^{(i)},s^{(j)}|w) \quad (1)$$

where the node potential is a product of multivariate Gaussian distributions, i.e.

$$\Phi(x^{(i)},s^{(i)}|\beta,\Sigma)=\prod_{t=1}^{N_{(i)}}\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}\times\exp(-\frac{1}{2}(x_t^{(i)}-o_t^{(i)}\beta_{s^{(i)}})^T\Sigma_{s^{(i)}}^{-1}(x_t^{(i)}-o_t^{(i)}\beta_{s^{(i)}})$$

$o_t^{(i)}$ is a concatenated matrix of previous observations $[x_{t-1}^{(i)},\ldots,x_{t-L}^{(i)}]^T$, and p is the dimension of $x_t^{(i)}$; The edge potential $\Phi(s^{(i)},s^{(j)}|w)$ is defined as $\Phi(s^{(i)},s^{(j)}|w)=\exp(\sum_k w_k\delta_k(s^{(i)},s^{(j)}))$, where δ is the indicator function, i.e. $\delta_{k=\{s,s'\}}(s^{(i)},s^{(j)})=1$ if $s^{(i)}=s$ and $s^{(j)}=s'$, and 0 otherwise; $w_{s^{(i)},s^{(j)}}$ is the parameter to evaluate the similarity between state s and s', similar to the transition probability in HMM; Z is the normalization constant. By our definition of node potentials, the value of Z will only be affected by the edge potentials, i.e. $Z=\sum_{s^{(1)},\ldots,s^{(M)}}\exp(\sum_{(i,j)\in\text{edge}}w_{s^{(i)},s^{(j)}}\delta(s^{(i)},s^{(j)}))$.
Figure 1 shows the graphical model representation of hMRF. As we can see, the model aims to infer the hidden regulatory networks (captured by the regression coefficients β_s) via mixture of regression [5] (i.e. the node potential) and the mixture selection is constrained by species-level graph (i.e. the edge potential).

There are three sets of parameters in the model, namely β, Σ and w. Since values of the state variables $s^{(1)},\ldots,s^{(M)}$ is not known, EM algorithm is applied to estimate the parameters [4]: Specifically, we calculate the expected value of the log likelihood function Q as follows:

$$Q=\sum_{s^{(1)},\ldots,s^{(M)}}P(\{s^{(i)}\}|\{x^{(i)}\},\widetilde{\beta},\widetilde{\Sigma},\widetilde{w})\log P(\{x^{(i)}\},\{s^{(i)}\}|\beta,\Sigma,w)$$

For the M-step, we estimate the values of the parameters β, Σ and w that maximize Q. Taking the derivatives of Q with respect to β_s and Σ_s respectively, we have

$$\begin{aligned}\hat{\beta}_s &= (V(x,s)^TV(x,s))^{-1}V(x,s)U(x,s)\\ \hat{\Sigma}_s &= (U(x,s)-\hat{\beta}_s^TV(x,s))^T(U(x,s)-\hat{\beta}_s^TV(x,s))/\sum_{n=1}^{N_{(i)}}\widetilde{P}^{(i)}(s)\end{aligned}$$

where

$$\widetilde{P}^{(i)}(s)=P(s^{(i)}=s|x^{(1)},\ldots,x^{(M)},\widetilde{\beta},\widetilde{\Sigma},\widetilde{w}),U(x,s)=[\sqrt{(\widetilde{P}^{(L)}(s))}x^{(L)}\ldots\sqrt{(\widetilde{P}^{(M)}(s))}x^{(M)}]^T \quad (2)$$

$$V(x,s)=\begin{pmatrix}\sqrt{(\widetilde{P}^{(1)}(s))}x^{(1)} & \ldots & \sqrt{(\widetilde{P}^{(M-L)}(s))}x^{(M-L+2)}\\ \sqrt{(\widetilde{P}^{(2)}(s))}x^{(2)} & \ldots & \sqrt{(\widetilde{P}^{(M-L+3)}(s))}x^{(M-L+3)}\\ \ldots & & \\ \sqrt{(\widetilde{P}^{(L-1)}(s))}x^{(L-1)} & \ldots & \sqrt{(\widetilde{P}^{(M)}(s))}x^{(M)}\end{pmatrix}^T$$

We can see that the solution can be achieved as a normal linear regression by reweighing the observed variables and response variables with weights $\widetilde{P}^{(i)}(s)$. For the parameter w, there is no closed form solution. We use iterative searching algorithms with the first derivative of Q with respect to $w_{s,s'}$ as follows:

$$\frac{\partial Q}{w_{s,s'}}=\sum_{(i,j)\in\text{edge}}(P(s^{(i)}=s,s^{(j)}=s'|x,\widetilde{\beta},\widetilde{\Sigma},\widetilde{w})\delta(s,s'))-\sum_{(i,j)\in\text{edge}}E[\delta(s^{(i)},s^{(j)})|x,\beta,\Sigma,w]$$

Gradient descent algorithm is applied to compute the solution of w.

For E-step, we use loopy belief propagation to compute the marginal of individual node $\widetilde{P}^{(i)}(s)$ and edges $\widetilde{P}(s^{(i)},s^{(j)})$ [35].

2.3 Extending hMRF regression with L_1 penalty

Next we examine how to extend hMRF regression to incorporate L_1 penalty. Following the idea in [26], we add a Laplacian prior for β as follows:

$$P(\beta|\lambda)=(\lambda/2)^N\exp(-\lambda\|\beta\|_1).$$

Algorithm I: hMRF Regression for temporal graph structure learning

1. **Input:** For each gene i, we are given time series data $x^{(i)} = \{x_1^{(i)} \dots x_{N_{(i)}}^{(i)}\}$ where $x_t^{(i)}$ is a p-dimensional vector;
   Parameters: (1) time lag L; (2) number of hidden states K;
   Function input: regression function f
2. Run hMRF regression and output coefficients β_s for each state s, the mixing of hidden states $\widetilde{P}^{(i)}(s)$
3. For each gene i, iterate the following steps:

3.1 Initialize the adjacency matrix for the p features, i.e. $G = \langle V, E\rangle$ where V is the set of p features.

3.2 For each feature $x_u \in V$ place an edge $x_u \rightarrow x_v$ into E, if and only if at least one of the corresponding coefficients for x_u in $\sum_s \widetilde{P}^{(i)}(s)\beta_s$ is above threshold θ.

As a result, the terms relevant to β in auxiliary function Q include

$$Q_\lambda = -\sum_{i=1}^{M}\sum_{t=1}^{N^{(i)}}\sum_{s^{(i)}} P(s^{(i)}|x^{(1)},\dots,x^{(M)},\widetilde{\beta},\widetilde{\Sigma},\widetilde{w}) \times (x_t^{(i)} - x_{t-1..t-L}^{(i)}\beta_{s^{(i)}})^T \Sigma_{s^{(i)}}^{-1}(x_t^{(i)} - x_{t-1..t-L}^{(i)}\beta_{s^{(i)}}) - \lambda\|\beta\|_1$$

Recent reexamination of gradient-based optimization algorithms, such as the coordinate descent, has shown that they are very effective to solve lasso-type regressions [17]. We compute the first derivative of Q_λ with respect to β_s as follows:

$$Q'_{\beta_s} = \begin{cases} -\sum_{i=1}^{M}\sum_{t=1}^{N^{(i)}} W_i(s)(x_t^{(i)} - x_{t-1..t-L}^{(i)}\beta_{s^{(i)}})^T - \lambda & \text{if } \beta_s >= 0 \\ -\sum_{i=1}^{M}\sum_{t=1}^{N^{(i)}} W_i(s)(x_t^{(i)} - x_{t-1..t-L}^{(i)}\beta_{s^{(i)}})^T + \lambda & \text{if } \beta_s < 0 \end{cases}$$

where $W_i(s) = P(s^{(i)}|x^{(1)},\dots,x^{(M)},\widetilde{\beta},\widetilde{\Sigma},\widetilde{w})(\Sigma_s^{-1} + \Sigma_s^{-1^T})$. Then we apply coordinate descent algorithms to get the solution to β. Other regression algorithms with L_1 penalty, such as elastic net and group lasso, can also be extended similarly. We skip the full discussions.

2.4 hMRF Regression for Learning Dynamic Temporal Graphs

To apply hMRF regression to learn dynamic temporal graphs, we need to determine how to combine component graphs associated with each state into one. In this paper, we use a heuristic weighted average approach: i.e. for each gene i, reweighing the base graph of state s (represented by coefficient β_s) with its mixing proportion $\widetilde{P}^{(i)}(s)$. Then we decide that there is an edge between two nodes if and only if the corresponding coefficients in the weighted average matrix $\sum_s \widetilde{P}^{(i)}(s)\beta_s$ are above some threshold. In our experiment, the threshold is set to 0.05. Algorithm I shows the details.

3 Experimental Results

To examine the effectiveness of the proposed algorithm algorithm, we conduct experiments on two simulation dataset and one application data of cross-species innate immune response analysis.

3.1 Simulation data

The two simulation datasets are both generated from a 2-state MRF, whose graph structure is a 10×10 grid (notice that this corresponds to species-level graph in the application of cross-species regulatory network discovery), and the coefficients are defined as follows: $w(i,i) = 1$ and $w(i,i') = 0.5$ for $i \neq i'$. The observations of each node (i.e. each task) are generated from Gaussian distributions using examples of the AR models used in [43, 18] (notice that this corresponds to the gene-regulatory networks of individual species in the application of cross-species regulatory network discovery). More specifically:
Simulation Data I: assume that state 1 corresponds to a sparse scenario, i.e. the inverse of the covariance matrix is an AR(1) model as follows: $(\sum^{-1})_{ii} = 1$, $(\sum^{-1})_{i,i-1} = (\sum^{-1})_{i-1,i} = 0.5$, and state 2 corresponds to dense scenario, $(\sum^{-1})_{ii} = 2$, $(\sum^{-1})_{ii'} = 1$. The goal of conducting experiments on this dataset is to verify whether our algorithm is able to recover the sparse component graph from data mixed with dense

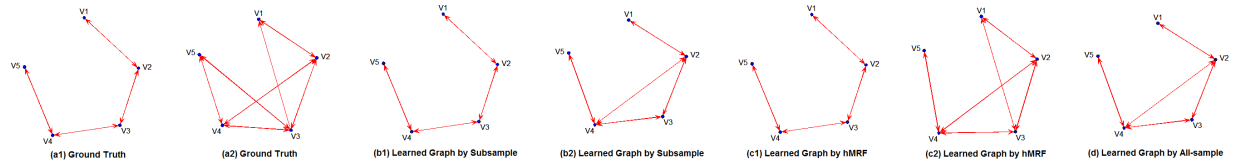


Figure 2: Example of learned graphs by different methods: (1a, 1b) true component graphs of state 1 and state 2; (2a, 2b) learned graphs by baseline method SUB; (3a, 3b) learned graphs by hMRF; (4) learned graph by baseline method ALL

Table 1: Comparison Results of Structure Learning on Simulation Data (sample size per node = 500)

Algorithm	Simulation I (F_1)		Simulation I (F_1)	
	State 1	State 2	State 1	State 2
hMRF	**0.9251**	**0.7577**	**1.000**	**0.9085**
ALL	0.8191	0.6388	0.8160	0.7664
SUB	0.7429	0.7273	1.000	0.7273

component graph.

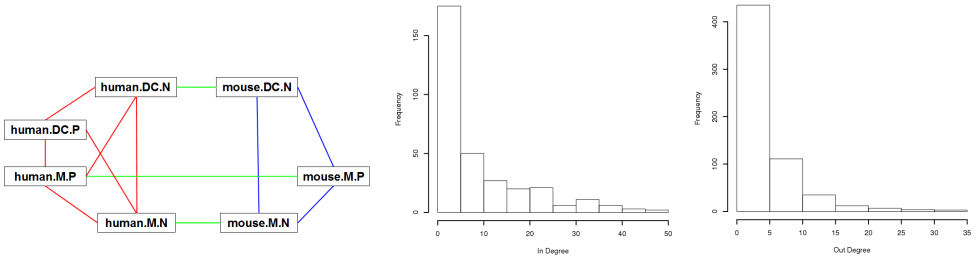
Simulation Data II: contain data generated from Gaussian distributions of inverse covariance with similar graph structures: state 1 corresponds to the same distribution as state 1 in Simulation Data I, and state 2 corresponds $(\sum^{-1})_{ii} = 1$, $(\sum^{-1})_{i,i-1} = (\sum^{-1})_{i-1,i} = 0.5$, $(\sum^{-1})_{i,i-2} = (\sum^{-1})_{i-2,i} = 0.25$ (see Figure 2 (1a) and (2a) for graph structure). Our goal is to examine whether the algorithm can recover the true graphs when the underlying two component graphs are similar, which better mimics our application data on cross-species gene regulatory networks. Therefore we focus our discussion on the results of this dataset.

In the experiment, we sample the values of underlying hidden states for all the nodes using Gibbs sampling; then for each node, we generate N samples from the underlying distributions determined by the value of hidden states. We vary the number of samples N ranging from 10, 20, 50, 100, 200, 300, 400, 500 to 1000. The penalty terms of Lasso are selected by cross-validation. We evaluate the performance of structure learning methods using the F1-measure, i.e. viewing the causal modeling problem as that of predicting the inclusion of the edges in the true graph, or the corresponding adjacency matrix. Recall that, given precision P and recall R, the F1-measure is defined as $F1 = 2PR/(P + R)$, and hence strikes a balance in the trade-off between the two measures (see example [38] for use of these metrics in evaluation of structural learning methods). In addition, we also evaluate the performance of state assignment of each node using the F1-measure.

We compare the performance of hMRF with two other baselines: one is aggregating all the data from different tasks and learn one graph (referred to as "ALL"), and the other is to learning a graph using data of individual task only (referred to as "SUB"). We repeated each sample size N 30 times and report the average on in Figure 2 and Table 1. Figure 2 show an example of learned graph by different methods when the sample size N is 200. As we can see, hMRF produces graphs closest to the ground truth. Table 1 shows that hMRF achieves better performance than competing methods on both Simulation Data I and II.

3.2 Applications to Cross-Species Gene Regulatory Network Discovery

Most multicellular organisms rely on their immune system to defend against the infection from a multitude of pathogens. There have been many studies using microarrays to compare immune gene expression programs under different conditions [23, 8, 22]. To understand the roles and possible interplays between different types of immune cells, it is important to identify both regulatory relations common to different immune cells, as well as those specific to a certain cell type. While each of these subsets of experiments (macrophages vs. dendritic, human vs. mouse etc.) can be analyzed separately and then compared to each other, the learned biological networks become much less reliable due to the noise in gene expression data. It is therefore desirable to combine microarray gene expression datasets from different studies to overcome noise in the datasets and jointly infer regulatory networks involved in immune response.



(a) Species-level Graph (b) In-degree Distribution (c) Out-degree Distribution

Figure 3: (a) Species-level Graph. Red/blue edges: dependency due to same species (i.e. human/mouse); green edges: dependency due to same experiments; (conveniently generated from domain knowledge) (b) Distribution of in-degree counts (c) Distribution of out-degree counts

We applied our algorithm to learn the causal networks between genes for immune response system. Specifically, we collected time-series microarray datasets on innate immune response of human and mouse from the supporting websites of [8, 11, 14, 21, 22, 23, 25, 32, 41]. The gene expression experiments were done on macrophages (M) and dendritic cells (DC) in humans and mice, under the infection of two types of bacteria, Gram-positive (P) and Gram-negative (N). The only exception is mouse dendritic cells, where we only found data on Gram-negative bacteria. The 39 microarrayexperiments are grouped into seven datasets, and referred to as "human.DC.N", "human.DC.P", "human.M.N", "human.M.P", "mouse.DC.N", "mouse.M.N" and "mouse.M.P" respectively (see [30] for full details of the data).

In order to exploit information shared across species/cell types, we process the data as follows: for those experiments on the same species, we only select the genes that appear in all the experiments. This results in 3869 genes for mouse microarray data and 1651 genes for human microarray data; next we obtain the human and mouse orthologs from Mouse Genome Database [15], and select the common candidate regulatory genes where either themselves or their orthologs can be found in our dataset. This results in a set of 789 common genes across species. We construct the species-level graph as follows: there is an edge between two experiments on the same species if they share the same cell type or the same infection type; there is also an edge between the same cell type and infection type across different species because we expect that some of the genes may share the similar regulated functions as their orthologs. This results in the species-level graph as Figure 3(a).

We varied the number of hidden component graphs from 2 to 7 and set to 4 by Bayesian information criterion (BIC) score. We ran experiments for a maximum lags of 2. There is an edge in the component graph if and only if the absolute value of its corresponding coefficients are larger than 0.05. In the end, we have around 2000 edges in each component graph. Generally, the degree of the nodes in the component graph roughly follows the power law (Figure 3(b, c)).

3.2.1 Component-Independent Regulations

In order to better examine the results, we divided the genes in a component graph into three classes based on their connectivity: genes with only out-going edges, genes with only incoming edges, and genes with both types of edges. As we show later, genes in each class have demonstrated different characteristics.

We identified the top ten well-connected genes that have only out-going edges and common to all component graphs (Table 2). The list contains a number of chemokines and receptors, which are consistent with the hypothesis that genes in this class serve to sense environment and inter-cellular communication. E.g. IL1R2 is a receptor for pro-inflammatory interleukin 1 (IL-1) and related to cell migration [40]. NFkB is a transcription factor that can be activated by intra-/extra-cellular stimuli including cytokines and bacterial products. CXCL10 is a chemokine which can trigger many effects including stimulation of immun cells.

We identified the top ten well-connected genes with only incoming edges (Table 3). These genes are involved in various celluar processes. E.g. CCT5 is a member of TCP1 ring complex that folds various proteins including actin and tubulin. CYP2E1 is an enzyme that catalyzes many reactions involved in drug

Table 2: Top 10 Well-connected Genes with only Out-going Edges

Out-degree	Symbol	Description
43	FTH1	ferritin, heavy polypeptide 1
25	RPL37	ribosomal protein L37
20	IL1R2	interleukin 1 receptor, type II
18	NFKB1	nuclear factor of kappa light polypeptide gene enhancer in B-cells 1
18	CXCL10	chemokine (C-X-C motif) ligand 10
17	CYTIP	cytohesin 1 interacting protein
14	DUSP2	dual specificity phosphatase 2
12	PTGS2	prostaglandin-endoperoxide synthase 2
12	MMP12	matrix metallopeptidase 12
12	LSP1	lymphocyte-specific protein 1

Table 3: Top 10 Well-connected Genes with only Incoming Edges

In-degree	Symbol	Description
28	CCT5	chaperonin TCP1 subunit 5
28	PCNA	proliferating cell nuclear antigen
21	CYP2E1	cytochrome P450, family 2, subfamily E, polypeptide 1
20	NEDD4	neural precursor developmentally down-regulated 4
16	ZFHX3	zinc finger homeobox 3
15	EXT2	exostoses (multiple) 2
13	CLK3	CDC-like kinase 3
12	NMT1	N-myristoyltransferase 1
10	CBX5	chromobox homolog 5
10	CD1D	T-cell surface glycoprotein CD1d

Table 4: Example of Component-specific Hubs

Component1	HLA-DRA, ID1, CTSB, ELK1, CDKN2A	Component2	TSC22D3, ACVR2A, EPHA5, NFE2, PCTK3
Component3	PIK3R1, TK2, IL1R1	Component4	ASNS, MAP4K1, KCNH2, INPPL1, COL9A2

metabolism. CD1 mediates the presentation of primarily lipid and glycolipid antigens of self or microbial origin to T cells.

Next, we look at densely connected subgraphs in the learned component graphs. To further enforce sparcity, we apply a more stringent threshold (0.2) on the absolute value of the edge weights. Here we show one example of the subgraphs (Figure 4). The genes with only out-going edges in this subgraph include a number of genes located on the membrane, e.g. CD14 [13], and HLA class II histocompatibility antigen (HLA-DRA), which are expressed in antigen presenting cells and plays a central role in the immune system [39]. The middle layer of the sub-graph includes GTP binding protein (GTPBP1), and CDKN1A, a cell cycle regulator, and VIM, which is involved in attachment, migration, and cell signaling [36]. The bottom level of the graph includes genes mediate signal transductions (CD83), important chemokines (CCL5), and interferon-induced GTPase (GBP1).

3.2.2 Component-specific Regulations

Next we compare the component graphs and identify characteristics specific to each graph. First, we compare the hub genes in each component graph, which are defined as genes with at least 5 outgoing edges and no incoming edges. We identify a total of 40 component specific hub genes. For component graph 1, the list includes genes involved in cell cycle control (E2F1, CDKN2A) and wound repair (MMP3). In addition, ITGA7 is involved in cell-cell interaction, and MAP3K8 can induce the production of NFkB. For component graph 2, the list includes TSC22D3, which plays a key role in the anti-inflammatory process. Hub genes in component 3 include Il1R1, interleukin 1 receptor, and PIK3R1, which are involved in metabolism of insulin. For component 4, hub genes include IRF9, a regulatory factor of interferons (proteins released by cells in response to pathogens), and MYH9, which has a function in the maintenance of cell shape. Some of the component-specific hub genes are listed in Table 4.

To characterize the genes with high incoming edges in each component graph, we examine the genes with at least 20 incoming edges and confirmed enriched GO categories [16]. For example, some of the top enriched categories include "Regulation of Glucose Transport" (component 1; corrected pval=0.002), "Leukocyte Homeostasis" (component 2 and 3; corrected $pval < 0.001$), "Locomotory Behavior" (component 2 and 3; corrected $pval < 0.006$), and "Double-strand break repair" (component 4; corrected pval=0.034).

3.2.3 Comparison with Other Approaches

We also compare the learned networks generated by hMRF with the networks by two other baselines: one is aggregating all the data from different tasks and learn one graph ("ALL"), and the other is to learning a

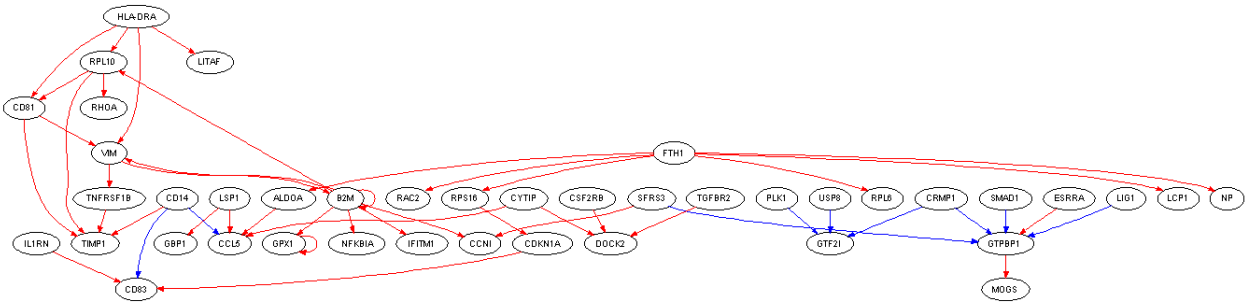


Figure 4: An example of densely connected subgraphs in the learned component graphs. The regulation relation can be either positive (red edges) or negative (blue edges).

Table 5: Top Five Component-Specific Enriched Biological Processes

Component 1	Component 2	Component 3	Component 4
regulation of glucose import	response to organic substance	response to organic substance	double-strand break repair
glucose import	leukocyte homeostasis	leukocyte homeostasisstimulus	response to abiotic stimulus
regulation of glucose transport	homeostasis of number of cells	cellular response to stimulus	cellular response to stimulus
response to organic substance	cellular response to stimulus	response to chemical stimulus	response to heat
response to peptide hormone stimulus	positive regulation of cellular process	positive regulation of cellular process	positive regulation of catabolic process

graph using data of individual task only ("SUB"). Compared with SUB, our method has major advantages since some of the datasets, for example Human.DC.P and Mouse.M.N, have very limited number of time-series observations (1-2), and no reasonable graph can be generated by SUB. For fair comparison (in favor of SUB method), we choose the dataset with the largest number of time-series observations, i.e. Human.M.N, to compare the results of different methods. One general observation is that the networks by ALL (31,218 edges) and SUB (14,346 edges) are much denser than that by hMRF (7458 edges) while the three graphs share 4,071 edges in common. Sparse graphs do not necessarily suggest better performance, but around 54.6% commonality suggests that hMRF is able to provide a network with much higher precisions. Figure 6 lists an example of 10 genes with the highest number of out-degrees in the learned networks. From the results, we can see that hMRF not only shares some top-ranked genes with the other two algorithms, such as CXCL10, but also uniquely identifies important immune genes, such as IL1R2, HLA-DRA, and CD14, as well as B2M (Beta-2-microglobulin), which is a serum protein found in association with the major histocompatibility complex (MHC) class I heavy chain on the surface of nearly all nucleated cells; MSN (Moesin), which is localized to filopodia and other membranous protrusions that are important for cell-cell recognition and functions as cross-linkers between plasma membranes and actin-based cytoskeletons.

3.2.4 Bootstrap Evaluation

In addition, we also evaluate the performance of our method by applying the Bootstrap procedure, which is a technique widely used in statistics for evaluating statistical accuracy (see, [10] for a review). More precisely, given the original lagged data matrix, we randomly draw B datasets by sampling with replacement the rows of the original data matrix, so that each dataset has the same number of rows as the original lagged data matrix. We then apply our method to each of the B bootstrap datasets. Comparing the original network (i.e. the network obtained by using the original dataset) with the bootstrap networks (i.e. those obtained using the bootstrap datasets) allows us to get a measure of confidence in the causal relationships identified in the original network. In particular, for each causal relationship identified in the original network, we can get confidence in that relationship by counting the number of times it appears in the bootstrap networks. As shown in Table 7, the causal relationships identified by our method in the original network appear on

Table 6: Top 10 Genes by Out-degress in the Learned Networks by Different Methods

hMRF		ALL		SUB	
EntrezID	Edge #	EntrezID	Edges #	EntrezID	Edges #
FTH1	182	PTGS2	170	ACVR2A	224
IL1R2	110	ACVR2A	157	VPS45	179
B2M	104	CXCL10	154	PTGS2	175
VIM	75	DUSP2	145	NFE2	172
CXCL10	74	PPIB	140	FTH1	168
RPL37	71	FMO1	136	FOS	167
LSP1	70	PECAM1	135	PECAM1	162
HLA-DRA	68	NR4A1	132	FPR1	160
MSN	66	MCM4	131	CDC6	157
CD14	60	IL7R	128	LSP1	140

Table 7: Percentage of Overlap between Bootstrap Networks and Original Networks

Species Type	% of Overlap
human.DC.N	0.7572
human.DC.P	0.7569
human.M.N	0.7541
human.M.P	0.7575
mouse.DC.N	0.7713
mouse.M.N	0.7510
mouse.M.P	0.7527

the average 75.2% of the time in the bootstrap networks, which demonstrates that hMRF produces stable networks.

4 Conclusion

In this paper we examine the problem of discovering regulatory networks from multi-species time-series microarray data by leveraging the common regulation information across species. We develop hidden Markov random field regression with L_1 penalty to extend temporal Granger modeling for multitask learning. We show that our method is able to uncover causal relations on two synthetic datasets, as well as conserved regulatory network common to two types of cells in humans and mice and shared between response to different types of bacteria. For future work, we are interested in more systematic evaluation of the experiment results. We also plan to apply our model for other types of cross-species regulatory network discovery, such as antifungal drug resistance.

References

[1] A. Arnold, Y. Liu, and N. Abe. Temporal causal modeling with graphical granger methods. In *Proceedings of International Conference on Knowledge Discovery and Data Mining (SIGKDD-07)*, 2007.

[2] J. Berg and M. Lssig. Cross-species analysis of biological networks by bayesian alignment. *PNAS*, 103(29):1096710972, 2004.

[3] S. Bergmann, J. Ihmels, and N. Barkai. Similarities and differences in genome-wide expression data of six organisms. *PLoS Biol*, 2(1):e9, 2003.

[4] J. Bilmes. A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. Technical report, 1998.

[5] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2007.

[6] G. Bourque and D. Sankoff. Improving gene network inference by comparing expression time-series across species, developmental stages or tissues. *J Bioinform Comput Biol.*, 2(4):765–83, 2004.

[7] R. Caruana. Multitask learning. In *Ph.D. Thesis, School of Computer Science, Carnegie Mellon University*. 1997.

[8] D. Chaussabel, R. Semnani, M. McDowell, D. Sacks, A. Sher, and T. Nutman. Unique gene expression profiles of human macrophages and dendritic cells to phylogenetically distinct parasites. *Blood*, 202:672681, 2003.

[9] R. Dahlhaus and M. Eichler. Causality and graphical models in time series analysis. *Highly structured stochastic systems*, 2003.

[10] A. C. Davison and D. V. Hinkley. *Bootstrap Methods and Their Application (Cambridge Series in Statistical and Probabilistic Mathematics , No 1)*. Cambridge University Press, 1997.

[11] C. Detweiler, D. Cunanan, and S. Falkow. Host microarray analysis reveals a role for the salmonella response regulator phop in human macrophage cell death. *Proc. Natl. Acad. Sci.*, 98:5850–855, 2001.

[12] P. D'haeseleer, S. Liang, and R. Somogyi. Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics*, 16(8):707, 2000.

[13] S. DL, T. S, T. DG, N.-W. A, and S. B. Monocyte antigen cd14 is a phospholipid anchored membrane protein. *Blood*, 73(1):284–9, 1989.

[14] D. Draper, H. Bethea, and Y. He. Toll-like receptor 2-dependent and-independent activation of macrophages by group B streptococci. *Immunology letters*, 102(2):202–214, 2006.

[15] J. Eppig, C. Bult, J. Kadin, J. Richardson, and J. Blake. The Mouse Genome Database (MGD): from genes to mice–a community resource for mouse biology. *Nucleic acids research*, 33(Database Issue):D471, 2005.

[16] J. Ernst and Z. Bar-Joseph. Stem: a tool for the analysis of short time series gene expression data. *BMC Bioinformatics.*, 7(191), 2006.

[17] J. Friedman, T. Hastie, and R. Tibshirani. Regularized paths for generalized linear models via coordinate descent. 2008.

[18] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostat*, 9(3):432–441, 2008.

[19] N. Friedman. Inferring cellular networks using probabilistic graphical models. *Science*, 303(5659):799, 2004.

[20] C. Granger. Testing for causlity: A personal viewpoint. *Journal of Economic Dynamics and Control*, 2:329–352, 1980.

[21] F. Granucci, C. Vizzardelli, N. Pavelka, S. Feau, M. Persico, E. Virzi, M. Rescigno, G. Moro, and P. Ricciardi-Castagnoli. Inducible IL-2 production by dendritic cells revealed by global gene expression analysis. *Nature immunology*, 2(9):882–888, 2001.

[22] R. Hoffmann, K. van Erp, K. Trulzsch, and J. Heesemann. Transcriptional responses of murine macrophages to infection with Yersinia enterocolitica. *Cellular Microbiology*, 6(4):377–390, 2004.

[23] Q. Huang, D. Liu, P. Majewski, L. Schulte, J. Korn, R. Young, E. Lander, and N. Hacohen. The plasticity of dendritic cell responses to pathogens and their components. *Science*, 294(5543):870–875, 2001.

[24] H. Kunsch, S. Geman, and A. Kehagias. Hidden markov random fields. *Ann. Appl. Probab.*, 5(3):577–602, 1995.

[25] R. Lang, D. Patel, J. Morris, R. Rutschman, and P. Murray. Shaping gene expression in activated and resting primary macrophages by IL-10. *Journal of Immunology*, 169(5):2253–2263, 2002.

[26] S.-I. Lee, H. Lee, P. Abbeel, and A. Y. Ng. Efficient l1 regularized logistic regression. In *AAAI*, 2006.

[27] A. Lozano, N. Abe, Y. Liu, and S. Rosset. Grouped graphical granger modeling for gene expression regulatory networks discovery. In *Proceedings of International Conference on Intelligent Systems for Molecular Biology (ISMB-09)*, 2009.

[28] A. Lozano, N. Abe, Y. Liu, and S. Rosset. Grouped graphical granger modeling methods for temporal causal modeling. In *Proceedings of International Conference on Knowledge Discovery and Data Mining (SIGKDD-09)*, 2009.

[29] Y. Lu, P. Huggins, and Z. Bar-Joseph. Cross species analysis of microarray expression data. *Bioinformatics*, 25(12):1476, 2009.

[30] Y. Lu, S. Mahony, P. Benos, R. Rosenfeld, I. Simon, L. Breeden, and Z. Bar-Joseph. Combined analysis reveals a core set of cycling genes. *Genome Biology*, 8(7):R146, 2007.

[31] Y. Lu, R. Rosenfeld, G. J. Nau, and Z. Bar-Joseph. Cross species expression analysis of innate immune response. In *RECOMB 2'09: Proceedings of the 13th Annual International Conference on Research in Computational Molecular Biology*, pages 90–107, Berlin, Heidelberg, 2009. Springer-Verlag.

[32] R. McCaffrey, P. Fawcett, M. O'Riordan, K. Lee, E. Havell, P. Brown, and D. Portnoy. A specific gene expression program triggered by Gram-positive bacteria in the cytosol. *Proceedings of the National Academy of Sciences*, 101(31):11386–11391, 2004.

[33] N. Meinshausen and P. Buhlmann. High dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34(6):1436–1462, 2006.

[34] N. Mukhopadhyay and S. Chatterjee. Causality and pathway search in microarray time series experiment. *Bioinformatics*, 23(4), 2007.

[35] K. P. Murphy, Y. Weiss, and M. I. Jordan. Loopy belief propagation for approximate inference: An empirical study. In *Proceedings of Uncertainty in AI*, pages 467–475, 1999.

[36] D. Phua, P. Humbert, and W. Hunziker. Vimentin regulates scribble activity by protecting it from proteasomal degradation. *Mol Biol Cell.*, 20(12):2841–55, 2009.

[37] R. Sharan and T. Ideker. Modeling cellular machinery through biological network comparison. *Nature Biotechnology*, 24:427433, 2006.

[38] R. Silva, R. Scheine, C. Glymour, and P. Spirtes. Learning the structure of linear latent variable models. *J. Mach. Learn. Res.*, 7:191–246, 2006.

[39] P. Stumptner-Cuvelette, S. Morchoisne, M. Dugast, S. L. Gall, G. Raposo, O. Schwartz, and P. Benaroch. Hiv-1 nef impairs mhc class ii antigen presentation and surface expression. *PNAS*, 98(21):12144–9, 2001.

[40] C. SY, S. PF, and L. TC. Ectopic expression of interleukin-1 receptor type ii enhances cell migration through activation of the pre-interleukin 1alpha pathway. *Cytokine*, 45(1):32–8, 2009.

[41] K. van Erp, K. Dach, I. Koch, J. Heesemann, and R. Hoffmann. Role of strain differences on host resistance and the transcriptional response of macrophages to infection with Yersinia enterocolitica. *Physiological Genomics*, 25(1):75, 2006.

[42] M. J. Wainwright, P. Ravikumar, and J. D. Lafferty. High-dimensional graphical model selection using ℓ_1-regularized logistic regression. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1465–1472. 2007.

[43] M. Yuan and Y. Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.