Regulatory Inference from Gene Expression

CMSC858P Spring 2012
Hector Corrada Bravo
Genetic networks

- Microarrays provide a snapshot of gene expression in a cell. Genes are not expressed independently, they regulate each others activity.

  Goal: Reconstruct the gene regulation network!

- We want a probabilistic method that can handle noisy data.

- Causality, not correlation! Is the effect of a mutated gene on a target direct, or mediated by other genes? What is the nature of the interaction between genes (e.g. does gene A inhibit gene B)?
• Let $y$ be a vector-valued random variable

• Suppose some conditional independence properties hold for some variables in $y$.
  – Example: variable $y_2$ and $y_3$ are independent given remaining variables in $y$.

• We can encode this conditional independence properties in a graph
Graphical Model

• Hammersley-Clifford theorem: all probability distributions that satisfy conditional independence properties in a graph can be written as

\[ P(y) = \frac{1}{Z} \exp \left\{ \sum_{c \in C} f_c(y_c) \right\}, \]

- cliques in graph
- potential function
- variables in clique
Graphical Models

• The probability distribution is determined by choice of potential functions

• Example:

1. \( f_c(y_i) = -\frac{\tau_{ii}y_i^2}{2} \)

2. \( f_c(\{y_i, y_j\}) = -\frac{\tau_{ij}y_i y_j}{2} \)

3. \( f_c(y_c) = 0 \) for \(|y_c| \geq 3\).
Gaussian Graphical Models

• Define matrix as

1. $\Sigma_{ij}^{-1} = \tau_{ij}$ if there is an edge between $y_i$ and $y_j$

2. $\Sigma_{ij}^{-1} = 0$ otherwise

$$\Sigma^{-1} = \begin{pmatrix} \tau_{11}^2 & \tau_{12} & 0 & \tau_{14} \\ \tau_{12} & \tau_{22}^2 & \tau_{23} & 0 \\ 0 & \tau_{23} & \tau_{33}^2 & \tau_{34} \\ \tau_{14} & 0 & \tau_{34} & \tau_{44}^2 \end{pmatrix}$$
Network Discovery

• Gene expression vector $y$ is assumed to be distributed as

$$y \sim N(0, \Sigma)$$

• Likelihood given data ($n$ arrays) is then

$$l(\Sigma) = (2\pi \det(\Sigma))^{-n/2} \exp \left\{ -\frac{1}{2} \sum_{i=1}^{n} y_i^T \Sigma^{-1} y_i \right\}$$
Genome-wide discovery of transcriptional modules from DNA sequence and gene expression

E. Segal*, R. Yelensky and D. Koller

Computer Science Department of Stanford University, Stanford, CA 94305-9010, USA

Received on January 6, 2003; accepted on February 20, 2003
Module Networks: Segal 2003

Pre-processing
- Regulator selection

Expression data
- Data selection

Clustering
- Gene partition
- Gene reassignment to modules

Functional modules

Module network procedure

Motif search
- Annotation analysis
- Graphic presentation
- Hypotheses & validation

Post-processing
- Modules
- Motifs
- Annotations

Conditions
Module Networks: Segal 2003
Module Networks: Segal 2003
Segal and Widom, 2009
Network Discovery

• Recent efforts in discovering networks directly from expression data

• Main idea:
  – treat entire vector of gene expression for a given sample as multivariate normal
  – A sparse (inverse) covariance matrix induces a network
    • This is related to gaussian graphical models

• Banerjee, et al. [ICML 2006, JMLR 2008], Friedman [Biostatistics 2007]
Gaussian Graphical Model

• We can write $P$ as

$$P(y) = \frac{1}{Z} \exp \left\{ -\frac{y^T \Sigma^{-1} y}{2} \right\}$$

• By a nice property of exponential families, this makes $y$ be multivariate normal
  – implies $Z = \sqrt{2\pi \det(\Sigma)}$
ML Estimation

• Maximum Likelihood Estimate is given by inverse of solution to

\[
\max_{X \succ 0} \log \det X - (SX)
\]

• \(S\) is the sample covariance

\[
S = \sum_{i=1}^{n} y_i y_i^T
\]
Conditional Independence

- Conditional independence is given by sparse inverse covariance

\[
\Sigma^{-1} = \begin{pmatrix}
\tau_{11} & \tau_{12} & 0 & \tau_{14} \\
\tau_{12} & \tau_{22} & \tau_{23} & 0 \\
0 & \tau_{23} & \tau_{33} & \tau_{34} \\
\tau_{14} & 0 & \tau_{34} & \tau_{44}
\end{pmatrix}
\]
Sparse-inducing penalty

• Use a penalized likelihood method to induce sparsity

\[
\max_{X \succ 0} \log \det X - (SX) - \lambda \|X\|_1
\]

• where

\[
\|X\|_1 = \sum_{ij} |X_{ij}|
\]
Block-coordinate ascent

• Solve by maximizing one column of matrix at a time

\[
W = \begin{pmatrix} W_{11} & w_{12} \\ w_{12}^T & w_{22} \end{pmatrix}, \quad S = \begin{pmatrix} S_{11} & s_{12} \\ s_{12}^T & s_{22} \end{pmatrix}
\]

• Turns out this is equivalent to solving

\[
\min_\beta \frac{1}{2} \| W_{11}^{-\frac{1}{2}} \beta - z \|^2 + \lambda \| \beta \|_1
\]

where

\[
z = W_{11}^{-\frac{1}{2}} s_{12}
\]

• Solution is then

\[
w_{12} = W_{11} \beta
\]
Block-coordinate ascent

• So, we have $l_1$-regularized regression (lasso) problem at each step

$$
\min_{\beta} \frac{1}{2} \| W_{11}^{\frac{1}{2}} \beta - z \|^2 + \lambda \| \beta \|_1
$$

— Each of these can be solved via coordinate descent as before

— Recall: soft-thresholding at each step
We conclude that the method of analyzing the data obtained from the senators' records is applicable to a wide range of samples, including those from various sources such as LDL receptor and estrogenic receptors. The analysis of these data reveals a set of genes associated with iron homeostasis and cellular membrane fusion. These genes include those involved in the regulation of lipid metabolism and receptor function, as well as some unannotated genes that may be of interest for further study.

**Figure 1:** An example of the data obtained from the senators' records. The first-order variables used in the analysis are highlighted in blue, while the second-order variables are highlighted in red. The analysis is performed using the Gaussian method, and the results are used to predict the likelihood of the senators' behavior in various situations.
Summary

• Using gaussian graphical model representation
  — multivariate normal probability over a sparse graph
  — we can take resulting graph as gene network

• Use sparsity-inducing regularization (l1-norm)

• Block-coordinate ascent method leads to l1-regularized regression at each step
  — Can use efficient coordinate descent (soft-thresholding) to solve regression