INFERENCE AND LEARNING IN COMPUTATIONAL SYSTEMS BIOLOGY

CLAUDIA RANGEL

NATIONAL INSTITUTE OF GENOMIC MEDICINE

DAVID WILD, WARWICK UNIV. U.K.
JOHN ANGUS, CLAREMONT GRADUATE UNIV., CA
FRANCESCO FALCIANI, UNIVERSITY OF BIRMINGHAM, U.K
ZOUBIN GHAHRAMANI, UNIV. OF CAMBRIDGE, U.K.

CIMAT May 5, 2010

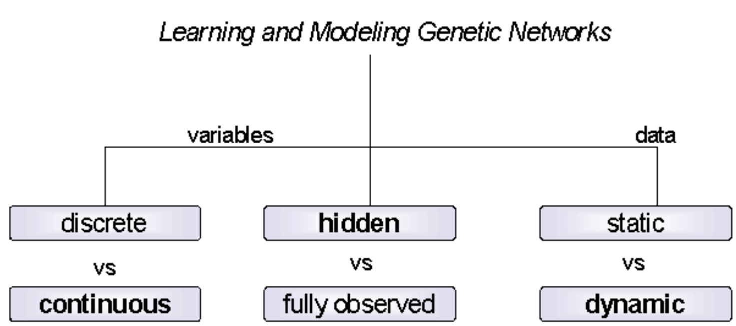
Gene Expression

- Our body is a machine regulated by proteins which are in turn regulated by genes
- Genes are found in the nucleous of every cell in our body
- Understanding how genes are regulated meaning being turned on and off is key in understanding diseases, pathologies, even how our brain operates
- Differential gene expression

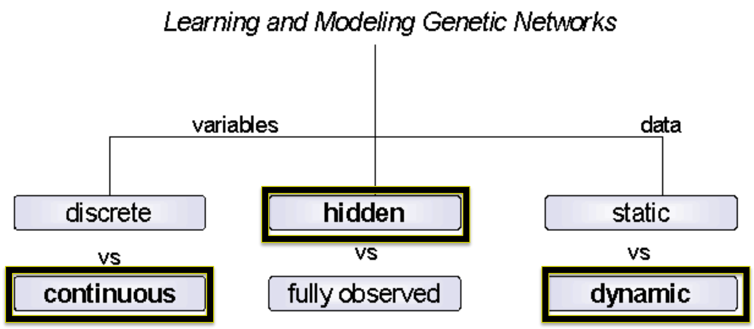
Motivation

- Questions people ask when they do microarray experiments?
- High percentage of publications that involve microarray experiments are designed to answer the first 2 questions)
  - 1) Are genes differentially expressed? Ctrol vs Treatment
  - 2) Do they cluster together? Do they have common functions?
  - 3) What can we understand about the underlying genome protein regulatory networks

Possible scenarios

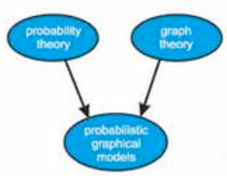


LDS / SSM



Probabilistic Graphical Models

- Graphical representations of probability distributions
  - new insights into existing models
  - motivation for new models
  - graph based algorithms for calculation and computation

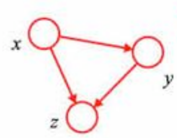


Directed Graphs: Decomposition

- Consider an arbitrary joint distribution

$$p(x, y, z)$$

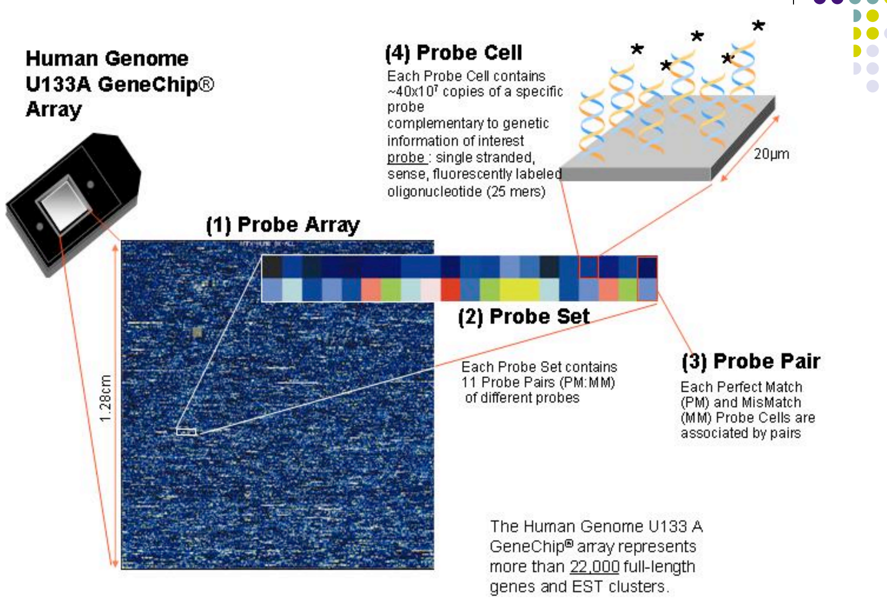
- By successive application of the product rule



$$p(x,y,z) = p(x)\,p(y,z|x)$$
$$= p(x)\,p(y|x)\,p(z|x,y)$$

Data

- Data is generated with a high throughput technology called microarrays
- These are capable of measure thousands of genes simultaneously
- The technology is expensive - about 700 dlls per chip. Having the budget for generating a reasonable sample size is difficult
- The technology is noisy



http://www.weizmann.ac.il/

Data structure:
Time series 10 x 44 x 58 {0,2,4,6,8,18,24,48,72}

g_1 g_2 … g_{58}

t_1
t_2
:
t_{10} expression levels } replicate 1

---

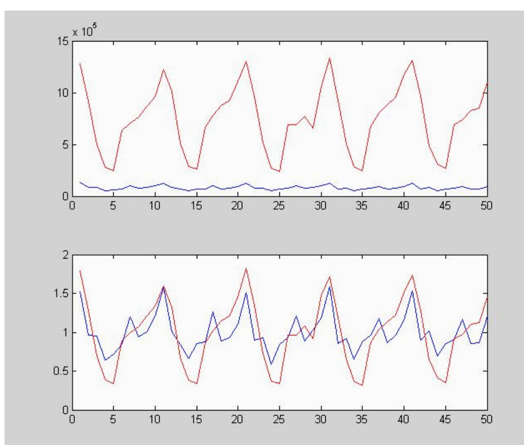
t_1
t_2
:
t_{10} expression levels } replicate 2

25,520 data points

⋮

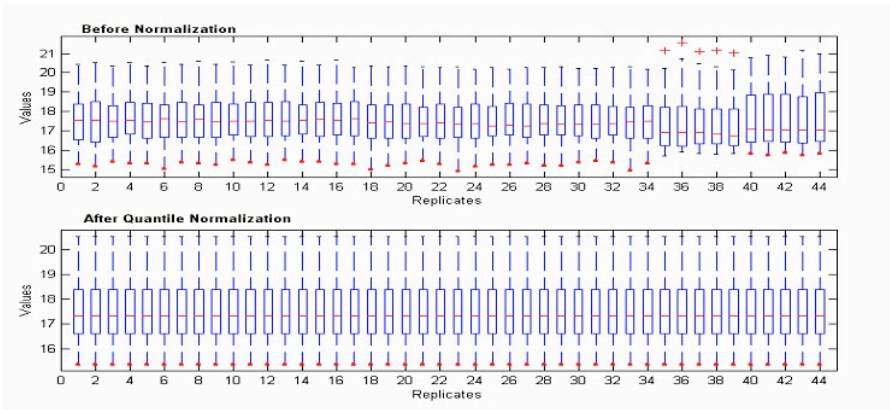
replicate 44

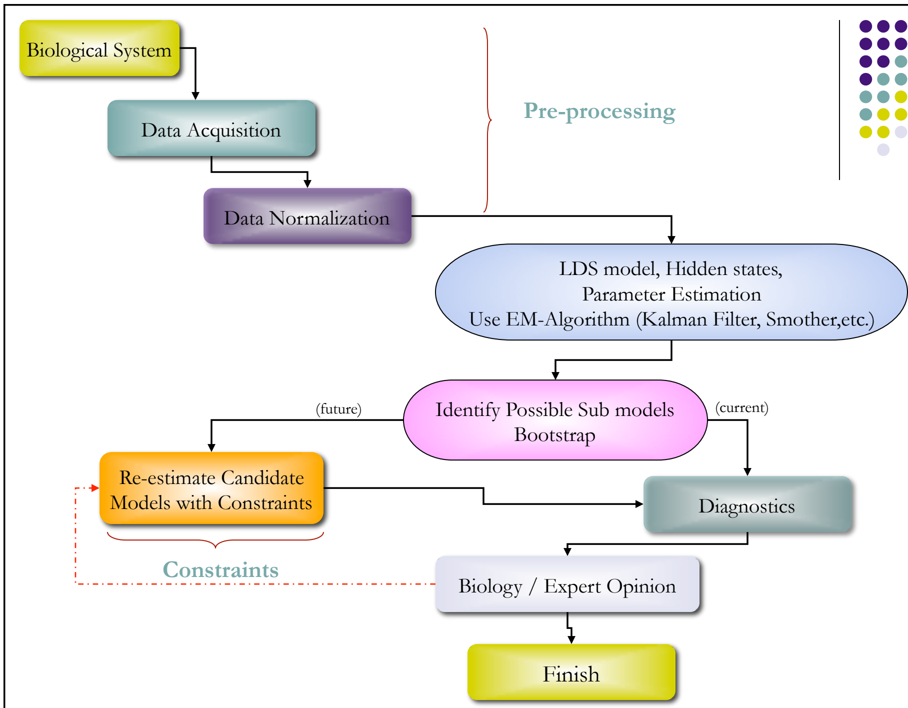
Data Normalization



Data Normalization

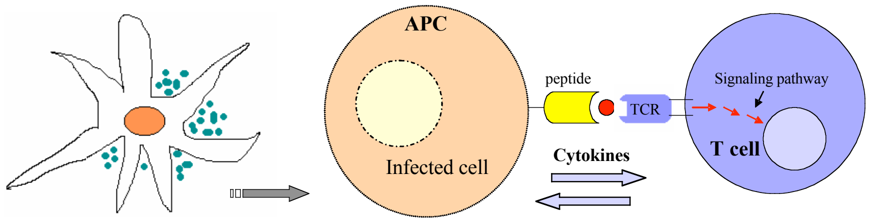
- *Motivation:* Common distribution of intensities across replicates.
- *Algorithm: Quantile Normalization* [Bolstad et al.] (Based on the Q-Q plots)





T cell Activation

The central event in the generation of an immune response is the activation of T cells.



T cell recognizes complex of viral peptide and kills infected ce ll.
T cell activation is initiated by the interaction between the T cell receptor (TCR) and the antigen peptide presented on the surface of an antigen -presenting cell. This event triggers a cascade of events that couple the stimulatory signal received form TCR to gene transcription events in the nucleus.

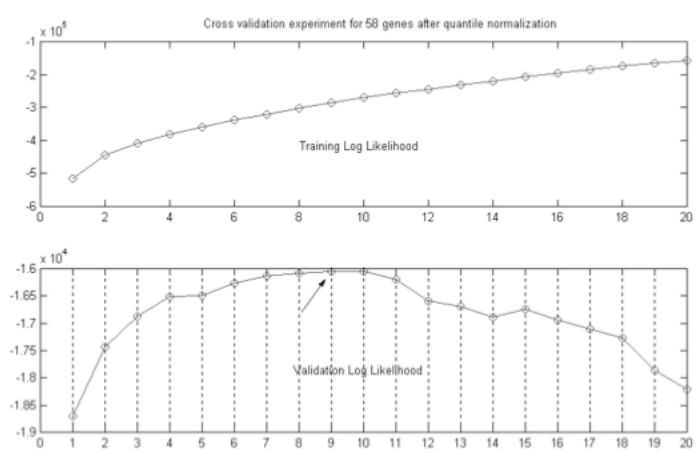
Why Linear Dynamical Systems (LDS)?

- Linear Dynamical Systems or Linear State-Space models provide a methodology for treating problems in time series analysis.
- Multivariate case is easily handled by simple extensions of univariate theory
- LDS assume the existence of a hidden state variable which evolves with Markovian dynamics.
  - Hidden variables can model
    - The effects of genes that have not been included on the microarray
    - Levels of regulatory proteins
    - The effects of mRNA degradation
  - Continuous variables
- Approach is based on the structural analysis of the problem.

Hidden States

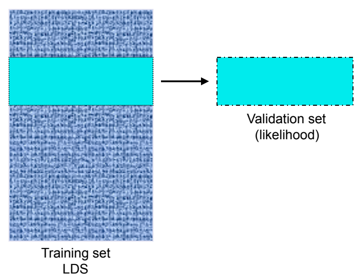
- Learning probabilistic models using hidden variables means that we should account for unobserved variables interacting with the observables
- A hidden variable can induce network structures or substructures improving the accuracy of the network
- By adding one or more hidden variables in the structure can result in a higher score
- Having too many hidden variables makes the model more complex affecting the accuracy of the parameters

How do we determine the number of hidden states?

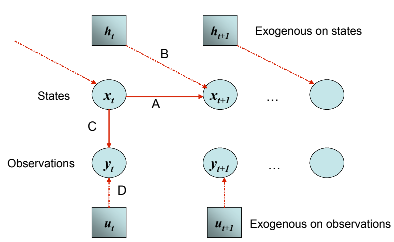


Bootstrap Cross Validation

- 44-way cross validation experiment to find the optimal number of hidden states
- In general in a R-fold cross-validation experiment, the data set is randomly divided into R mutually exclusive subsets of equal size. Data is trained R times, each time leaving out one of the subsets from training, but using only the omitted subset to compute the likelihood.



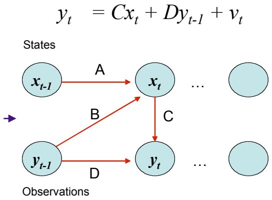
Definition of Model Structure



$$x_{t+1} = Ax_t + Bh_t + w_t$$
$$y_t \;\; = Cx_t + Du_t + v_t$$

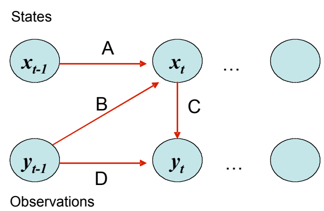
Gene expression data

$$x_{t+1} = Ax_t + By_t + w_t$$
$$y_t \;\; = Cx_t + Dy_{t-1} + v_t$$



Assumptions: $\{w_t\} \sim WN(0, Q)$
$\{v_t\} \sim WN(0, R)$, $\{w_t\} \perp \{v_t\}$

Model Parameters



$$x_{t+1} = Ax_t + By_t + w_t$$

$$y_t = Cx_t + Dy_{t-1} + v_t$$

$$\{w_t\} \sim WN(0, Q)$$

$$\{v_t\} \sim WN(0, R), \quad \{w_t\} \perp \{v_t\}$$

A: *K x K* transition matrix (K is the number of hidden states)
B: *K x 58* input to state matrix
C: *58 x K* influence of hidden states on gene expression at each time point
D: *58 x 58* gene to gene expression level influence at a consecutive time points

Notes:

1. We are interested in the CB+D matrix but that does not involve additional parameter estimation.
2. K=9

General Structural Properties

There are basically three important properties that must be verified

- First of all, we want to know if the system is asymptotically stable. This property is known as **Stability**. For the genetic model it is required that the matrix A has spectral radius less than one. In other words we will require that the eigenvalues of the matrix *A* be less than one in magnitude

- The other two properties are **Controllability and Observability**. These properties address information about the dimension of the state-space vector. Given a state-space model for the data {Y } we want to find the smallest possible dimension of the state vector x_t

Identifiability

- Consider an observable random vector (or matrix) Y defined on some probability space (Ω, F, P) having probability distribution $P_\theta \in \{P_{\theta_0} : \theta_0 \in \Theta\}$ where the parameter space Θ is an open subset of a n dimensional Euclidean space.

  We say that this probabilistic model is **identifiable** if the family $P_\theta \in \{P_{\theta_0} : \theta_0 \in \Theta\}$ has the property that $P_{\theta 1}(B) = P_{\theta 2}(B)$ for all Borel sets B if and only if $\theta_1 = \theta_2$ *both in* Θ. It is conventional in this parametric setting to say that in this case, the parameter θ is identifiable.

  It is easy to see why this property is important, for without it, it would be possible for different values of the parameter θ to give rise to identically distributed observables, making the statistical problem of estimating θ ill-posed.

Importance

- The identifiability problem has been studied extensively for the linear dynamic system model of th form

$$x_{t+1} = Ax_t + Bu_t + w_t$$
$$y_t = Cx_t + Du_t + v_t$$

- Taking the unknown parameter $\boldsymbol{\theta}$ to be the composite of A,B,C,D,Q,R, it is known that without any restrictions on the parameter, this model is not identifiable. In fact, it is easily seen that by a coordinate transformation of the state variable x_t,

$$\tilde{x}_t = Tx_t$$

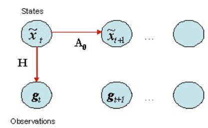
$$\tilde{x}_{t+1} = TAT^{-1}\tilde{x}_t + TBu_t + Tw_t$$

$$y_t = CT^{-1}\tilde{x}_t + Du_t + v_t$$

Another way of Representing the Gene Expression Model

The gene expression model can be expressed in a simpler state-space form

$$\tilde{x}_{t+1} = A_0\tilde{x}_t + \tilde{w}_t$$
$$g_t = H\tilde{x}_t$$



where,

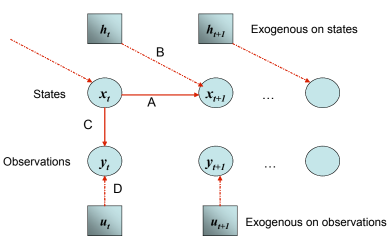
$$\tilde{x}_t = \begin{pmatrix} x_t \\ y_t \end{pmatrix} \;;\; A_0 = \begin{pmatrix} A & B \\ CA & CB + D \end{pmatrix} \;;\; H = \begin{bmatrix} 0 & I \end{bmatrix} \;;\; \tilde{w}_t = \begin{pmatrix} w_t \\ Cw_t + v_{t+1} \end{pmatrix}$$

and the white noise term in the state equation now has variance

$$\tilde{Q} = \begin{pmatrix} Q & QC' \\ CQ & CQC' + R \end{pmatrix}$$

Simpler form allows us to address stability, controllability, observability and identifiability in terms of known results.

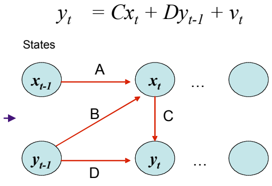
Definition of Model Structure



$$x_{t+1} = Ax_t + Bh_t + w_t$$
$$y_t \;= Cx_t + Du_t + v_t$$

Gene expression data

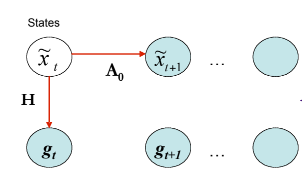
$$x_{t+1} = Ax_t + By_t + w_t$$
$$y_t \;= Cx_t + Dy_{t-1} + v_t$$



$\{w_t\} \sim WN(0,Q)$
$\{v_t\} \sim WN(0,R)$
$\{w_t\} \perp \{v_t\}$

Identify Model Properties

$$\tilde{x}_{t+1} = A_0\tilde{x}_t + \tilde{w}_t$$
$$g_t = H\tilde{x}_t$$



- **Controllability** is associated with the inputs. The state space model is controllable if the state vector can be "controlled" to evolve from a given, arbitrary initial state x_0 to a given, arbitrary final state x_t at a future time by a judicious choice of the inputs $\{w_t\}$. For the genetic model we have (by iterating the state equation)

$$x_t = A^t x_0 + \sum_{j=1}^{t} A^{j-1} B h_{t-j} + \sum_{j=1}^{t} A^{j-1} w_{t-1}$$

So we can write

$$x_t = A^t x_0 + \underbrace{[B, AB, A^2B, ..., A^{t-1}B]}_{V} \underbrace{\begin{bmatrix} h_{t-1} \\ h_{t-2} \\ \vdots \\ h_0 \end{bmatrix}}_{U^*} + \underbrace{[I, A, A^2, ..., A^{t-1}]}_{\mathcal{C}} \underbrace{\begin{bmatrix} w_{t-1} \\ w_{t-2} \\ \vdots \\ w_0 \end{bmatrix}}_{W^*}$$

which can be expressed as

$$x_t - A^t x_0 - VU^* = \mathcal{C}W^*$$

If $\mathcal{C}$ is full rank, then we can solve

$$W^* = \mathcal{C}'(\mathcal{C}\mathcal{C}')^{-1}(x_t - A^t x_0 - VU^*)$$

and we have *controllability* if $[I, A, A^2, ..., A^{t-1}]$ is of full rank for some $t \geq 1$.

On the other hand **observability** is associated with the outputs. The state space model is observable if, when the noise vectors are all taken to be 0 vectors, the initial state vector can be reconstructed from a sequence of output observations y_t. When there is no noise, we have

$$y_t = CA^t x_0 + C\sum_{j=1}^{t} A^{j-1} B h_{t-j} + D u_t$$

Letting

$$Y_t = y_t - C\sum_{j=1}^{t} A^{j-1} B h_{t-j} - D u_t$$

we can write

$$\begin{bmatrix} Y_0 \\ Y_1 \\ \vdots \\ Y_{K-1} \end{bmatrix} = \underbrace{\begin{bmatrix} C \\ CA \\ \vdots \\ CA^{K-1} \end{bmatrix}}_{\mathcal{O}} x_0$$

So if $\mathcal{O}$ is full rank, we can solve for x_0:

$$x_0 = (\mathcal{O}'\mathcal{O})^{-1}\mathcal{O}' \begin{bmatrix} Y_0 \\ \vdots \\ Y_{K-1} \end{bmatrix}$$

Therefore,

- In the genetic model we have that

$$\theta = \left[\begin{pmatrix} A & B \\ C & D \end{pmatrix}, Q, R \right]$$

- But for each transformation we have

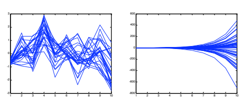
$$\left\{ \theta_T : \theta_T = \left[\begin{pmatrix} TAT^{-1} & TB \\ CT^{-1} & D \end{pmatrix}, TQT', R \right], \det(T) \neq 0 \right\}$$

- it is clear that D remains "identifiable," in some sense, as it is invariant to the choice of T. By inspection, other invariants can be seen to include
  $CB + D$, CB, and CA^kB, $k = 1, 2, \ldots$

Model Properties - Genetic Model

- Stability (parameters) the state variable does not "explode" exponentially -
  The Model will be **stable** *iff* the matrix

$$A_0 = \begin{pmatrix} A & B \\ CA & CB + D \end{pmatrix}$$



  has spectral radius less than one,
- Controllability (inputs) ability to move the state from any given initial value to a predetermined final value by manipulation of the noise - The model will be **controllable** *iff* the matrix

$$[I, A_0, A_0^2, \ldots, A_0^{K-1}] \qquad K = \dim(\tilde{x}_t)$$

  is full rank,
- Observability (outputs) ability to determine the initial state from a sequence of noiseless observations – The model will be **observable** *iff* the matrix

$$\begin{bmatrix} H & HA_0 & HA_0^2 & \cdots & HA_0^{K-1} \end{bmatrix} \qquad K = \dim(\tilde{x}_t)$$

  is full rank.

Methodology

- Expectation–Maximization (EM) algorithm
  - The motivation for using EM algorithm is that it iteratively computes the MLE for incomplete data sets.
- Filtering
  - Filtering is aimed at updating our knowledge of the system as each observation y_t comes in
- Smoothing
  - Smoothing enables us to base our estimates of quantities of interest on the entire sample $y_1,\ldots,y_T$.
- Bootstrapping
  - Bootstrap methods can be used for estimating confidence bounds for network outputs

EM Algorithm

$$x_{t+1} = Ax_t + By_t + w_t \qquad w_t \sim N(0,Q)$$

$$y_t = Cx_t + Dy_{t-1} + v_t \qquad v_t \sim N(0,R)$$

E-step

- Use $x_0, P_0, A, B, C, D, Q, R$

Kalman Filter / Smoother

$\hat{x}_t, P_t$

- Compute the *expected* log likelihood given the data

M-step

- Use $\hat{x}_t, P_t$

To Re-estimate

$x_0, P_0, \hat{A}, \hat{B}, \hat{C}, \hat{D}, \hat{Q}, \hat{R}$

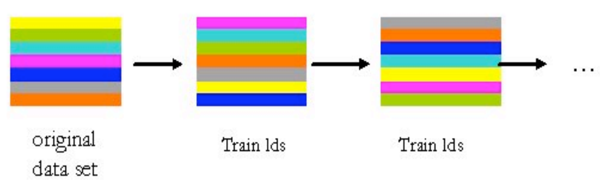
By *maximizing* the log likelihood

Kalman Filtering & Smoothing

- The likelihood can be calculated by a routine application of the Kalman filter, considered the optimal linear estimator.
- The **Kalman filter** estimates the current value of our variables incorporating all information available.
  - Knowledge of the system
  - The statistical description of any uncertainty of the dynamics of the model
  - Noises and measurement errors
  - Initial conditions
- The **Smoother** solves the problem of estimating the state at time *t* given the parameters and the observations.

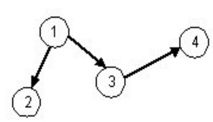
Bootstrapping

Develop Bootstrapping algorithm (on replicates) for estimation of confidence intervals on $\hat{C}\hat{B}+\hat{D}$

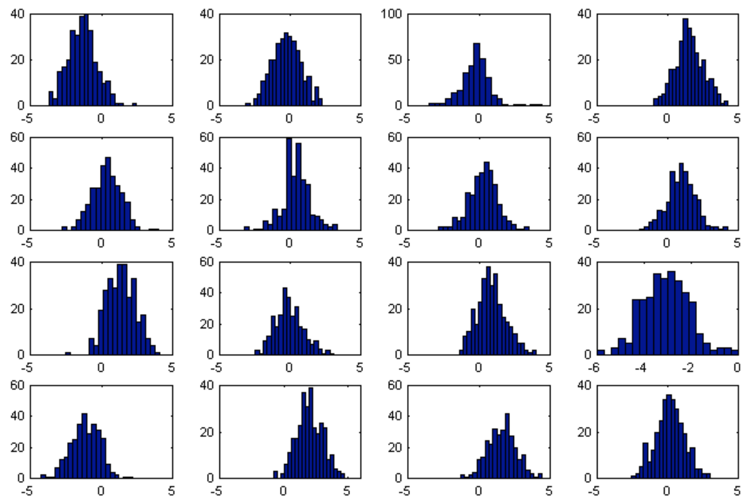


$$x_0, P_0, \hat{A}^*, \hat{B}^*, \hat{C}^*, \hat{D}^*, \hat{Q}^*, \hat{R}^*$$

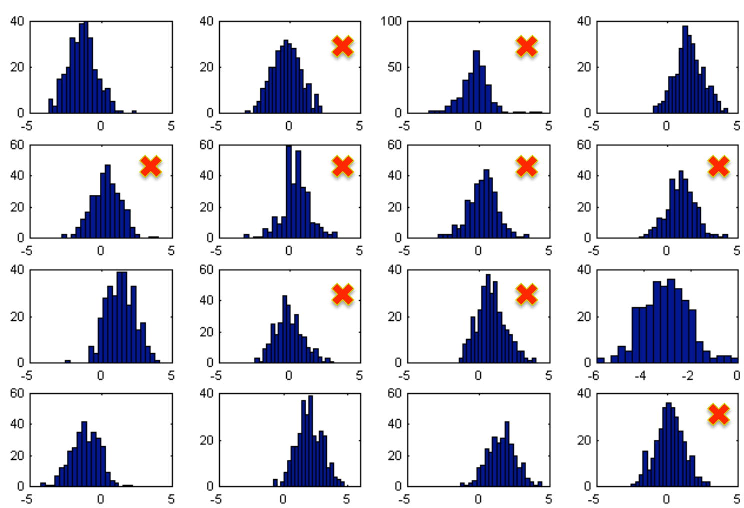
$$\overline{C^*B^*+D^*} \Rightarrow \begin{pmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \Rightarrow$$



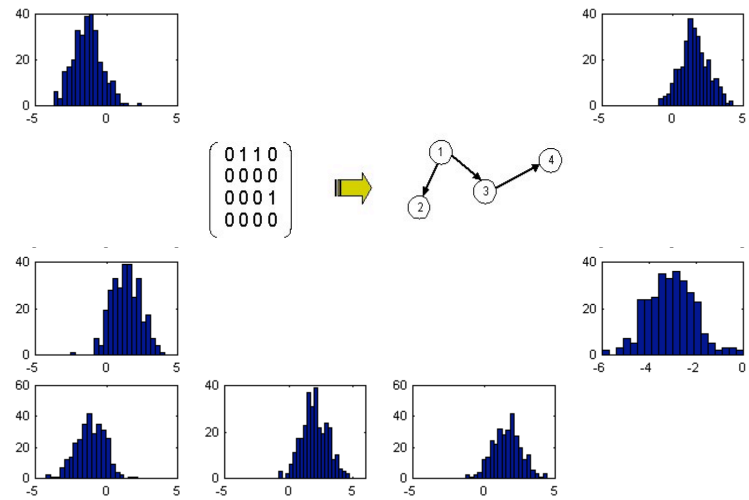
Usamos resultados del Bootstrapping



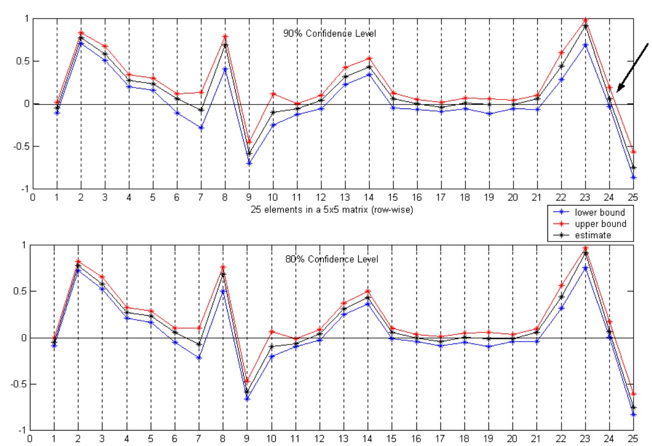
Usamos resultados del Bootstrapping



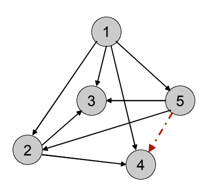
Usamos resultados del Bootstrapping



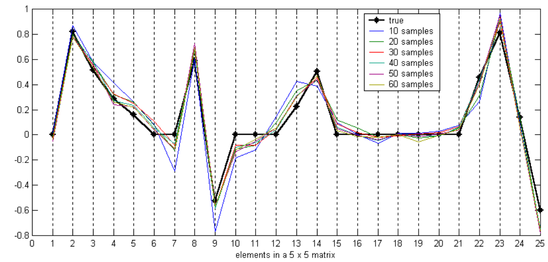
Results Simulated Data : 40 samples, 10 time points, 5 genes

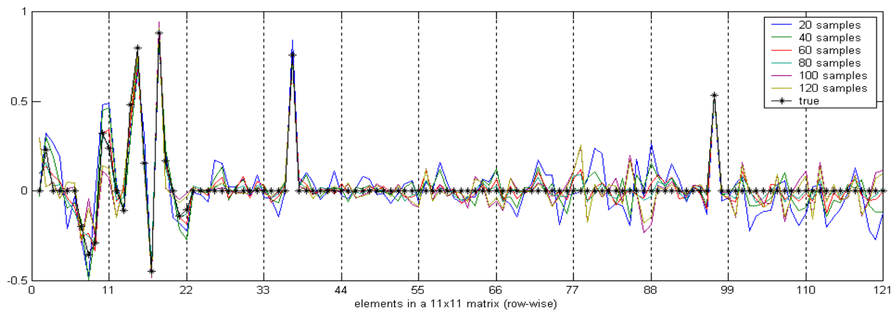


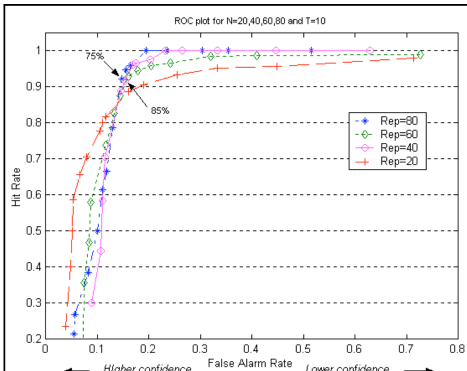
0	1	1	1	1
0	0	1	1	0
0	0	1	0	0
0	0	0	0	0
0	1	1	1	1



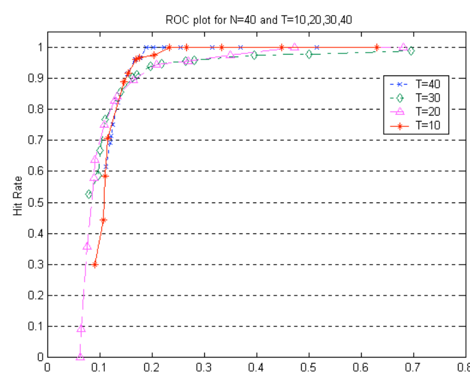
Results Simulated Data: 5 and 11 nodes

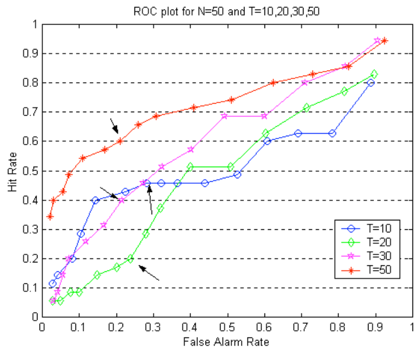


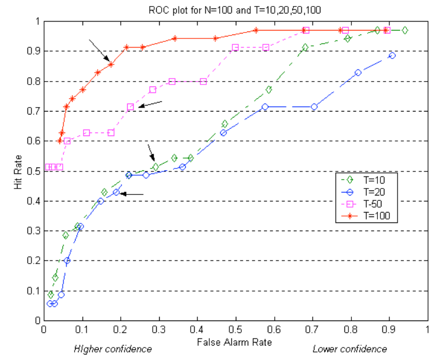


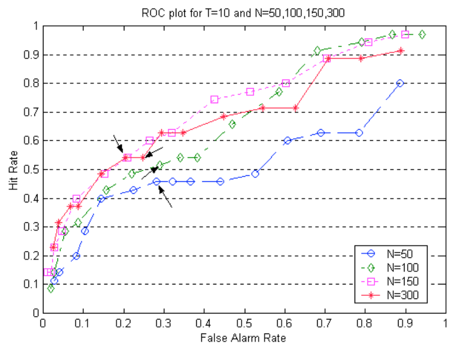


11 genes (nodes)









39 Nodes

Artificial time series are not stationary for a few time points sample -> bias

Diagnostics on Fitted Model

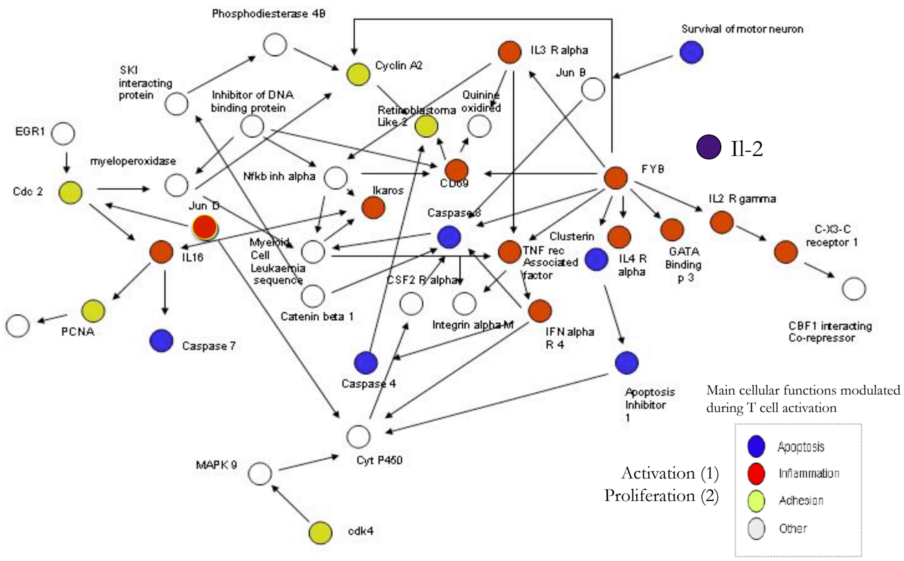
- Common Methods
  - Examination of standardized innovations for lack of correlation / pattern $\hat{v}_t = y_t - E(y_t \mid y_1, \ldots, y_{t-1})$
- Check that estimates of A, B, C, D are in the observable, controllable, stable region of the parameter space:

$$\rho(A_0) < 1;$$

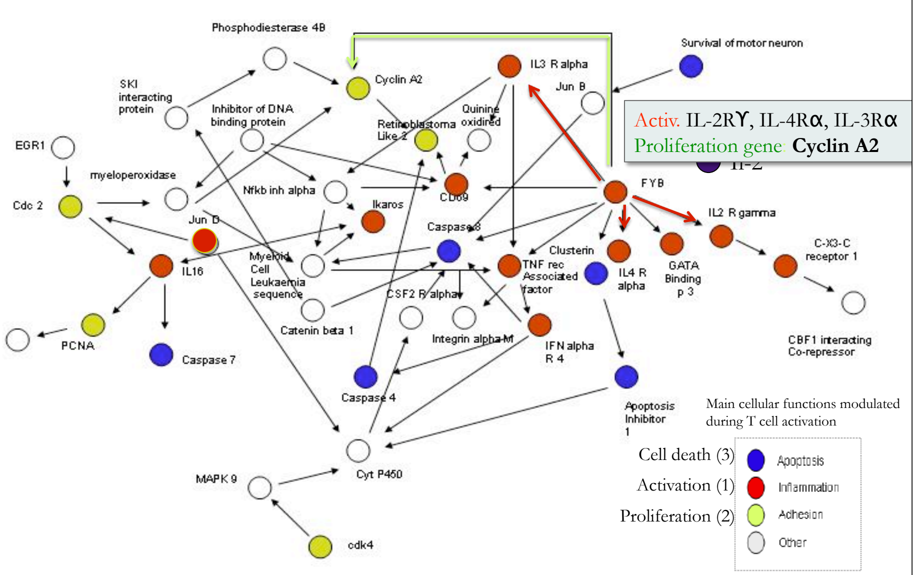
$$\begin{bmatrix} I & A_0 & A_0^2 & . & . & . & A_0^{K-1} \end{bmatrix} \text{ full rank}$$

$$\begin{bmatrix} H' & A_0'H' & A_0^2{}'H' & . & . & . & A_0^{K-1}{}'H' \end{bmatrix} \text{ full rank}$$

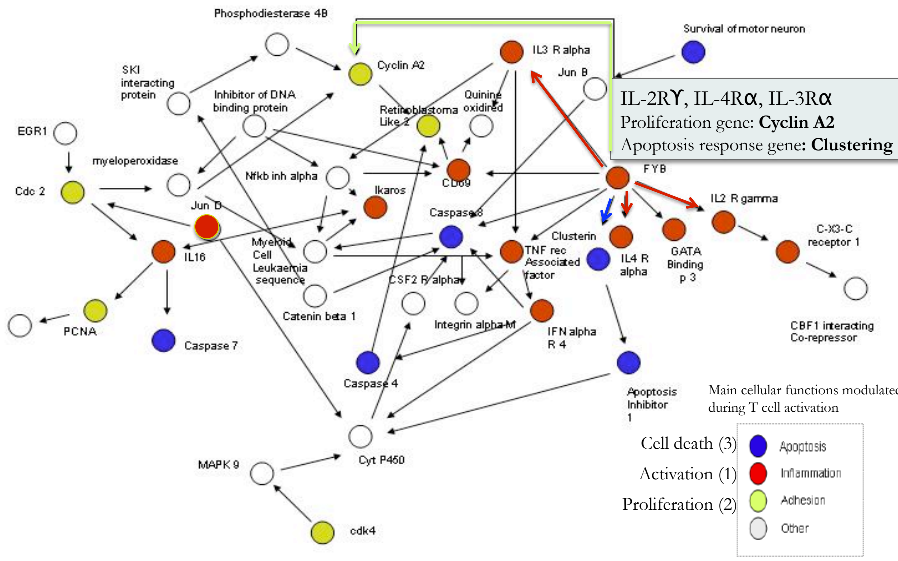
$$\text{where H=}\begin{bmatrix} 0 & I \end{bmatrix}; A_0 = \begin{bmatrix} A & B \\ CA & CB+D \end{bmatrix}$$



Gene Regulatory Network

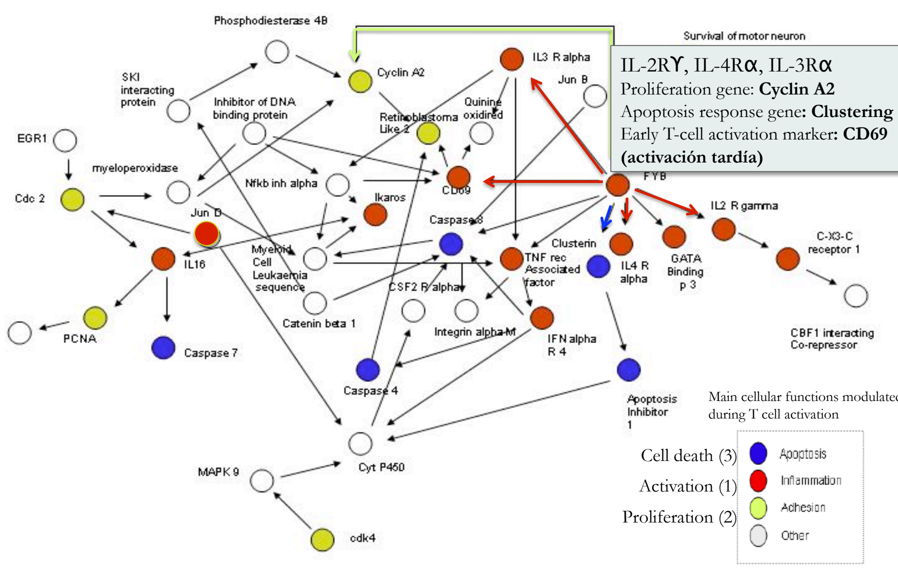


Gene Regulatory Network



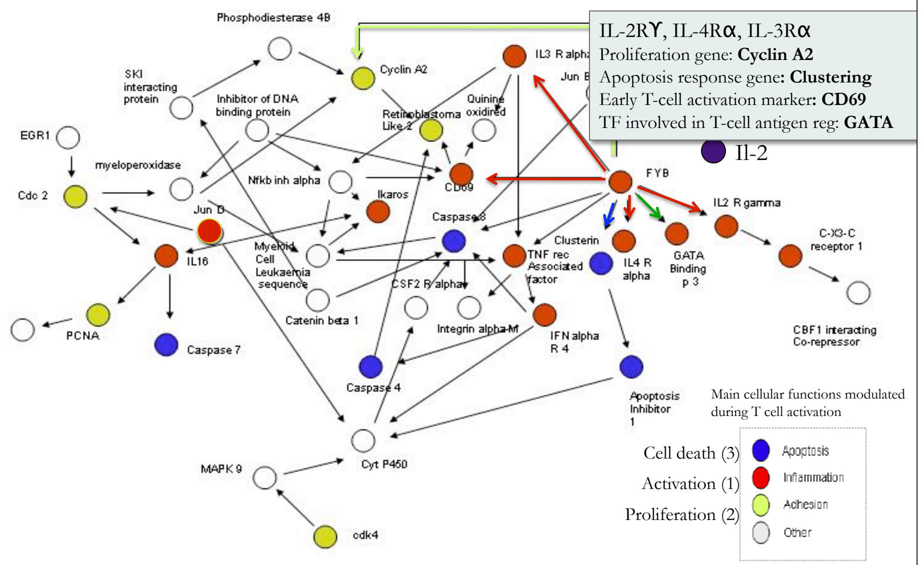
IL-2Rϒ, IL-4Rα, IL-3Rα
Proliferation gene: **Cyclin A2**
Apoptosis response gene**: Clustering**

Gene Regulatory Network

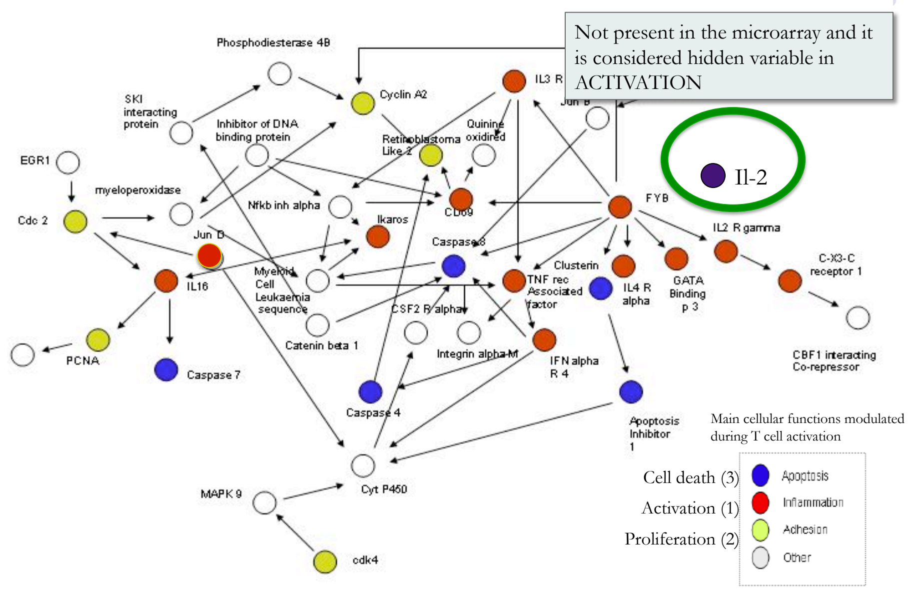


IL-2Rϒ, IL-4Rα, IL-3Rα
Proliferation gene: **Cyclin A2**
Apoptosis response gene**: Clustering**
Early T-cell activation marker**: CD69**
(activación tardía)

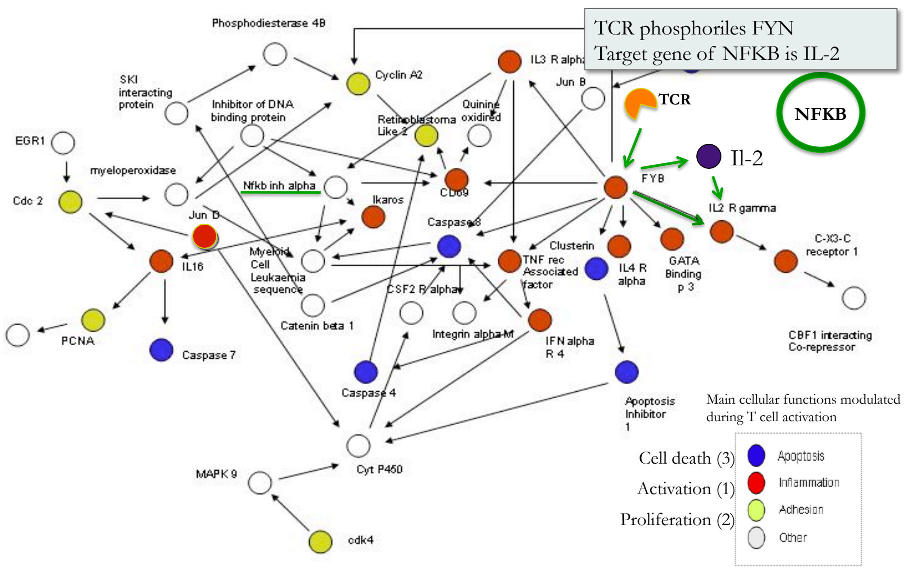
Gene Regulatory Network



Gene Regulatory Network

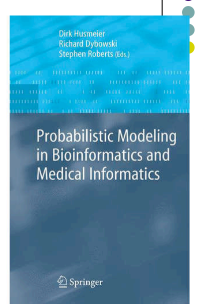


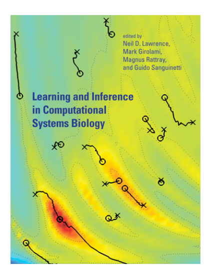
Gene Regulatory Network



Follow-up Research

- VBSSM
  - Variational Bayesian State-Space Model
- Synthetic Data
  - Genome Research Dirk Husmeier
- Constraints
  - Learning and Inference in Computational Biology MIT press - 2010





Incorporating Biological Knowledge, Knocking Out: Implementing Constraints

By thinking of each element in D as the connection strength with which gene *i* influences gene *j* over time, allows the matrix D to be constrained to have zero values where there is no connection between two genes.

- Two types of constraints on D of the form
  - $DF=G$ $_{(*)}$
  - $F\text{vec}(D) = G$ $_{(**)}$
- Constrained model address the estimation of fewer number of parameters
- Implemented by Lagrange multipliers by doing constrained maximization in the M step.

* *Shumway and Stoffer 1982*

** *New result*