Inferring gene regulatory networks with hidden variables using state space models

Andrea Rau, F. Jaffrézic, J.-L. Foulley, R. W. Doerge

February 9, 2012
Réunion du réseau méthodologique “Inférence de réseaux” (INRA / MIA)
AgroParisTech

Outline

1. Introduction
   - State space models
2. EBDBN Method
   - Model selection
   - Hidden state estimation
   - Parameter estimation
3. Results
   - Simulations
   - T-cell data analysis
4. Discussion

Inferring gene regulatory networks

Gene regulatory networks:

Set of genes that interact with one another (directly or indirectly) through other genes, transcription factors, protein products

⇒ **Goal**: Reverse-engineer the structure of a gene regulatory network from (continuous) time-course gene expression data

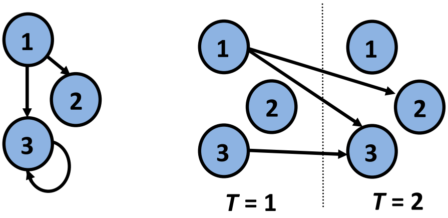
- Statistical challenges: noisy data, short time series, few biological replicates, many potential interactions

Dynamic Bayesian Networks (DBN)

Bayesian network

- Graphic structure, $M = (V, E)$, family of conditional distributions, F, and their parameters q
- Topology describes relationships between nodes in terms of conditional dependencies, must be a directed acyclic graph (DAG)

$\Rightarrow$ Unfold over time to make a **Dynamic Bayesian Network**



Vector Autoregressive (VAR) Process

Let $\mathbf{y}_t$ be the P-dimensional expression observations at time t. We may model the observations using a VAR process :

$$\mathbf{y}_t = D\mathbf{y}_{t-1} + \mathbf{v}_t, t \geq 2$$

with D being a sparse $(P \times P)$ coefficient matrix and $\mathbf{v}_t \sim \mathcal{N}(\mathbf{0}, \Sigma)$ for diagonal covariance matrix Σ.

- Non-zero elements of D define interactions ($d_{ij} \neq 0 \Rightarrow$ gene j regulates gene i)
- Assumptions: time-homogeneous interactions, direct interactions among genes from one time point to the next

Vector Autoregressive (VAR) Process

Let $\mathbf{y}_t$ be the P-dimensional expression observations at time t. We may model the observations using a VAR process :

$$\mathbf{y}_t = D\mathbf{y}_{t-1} + \mathbf{v}_t, t \geq 2$$

with D being a sparse $(P \times P)$ coefficient matrix and $\mathbf{v}_t \sim \mathcal{N}(\mathbf{0}, \Sigma)$ for diagonal covariance matrix Σ.

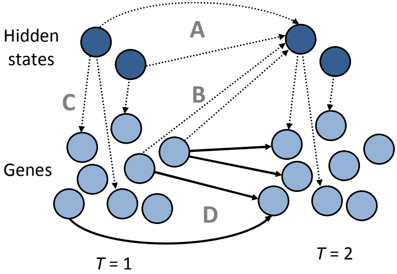
- Non-zero elements of D define interactions ($d_{ij} \neq 0 \Rightarrow$ gene j regulates gene i)
- Assumptions: time-homogeneous interactions, direct interactions among genes from one time point to the next
- **Improvements**: D_t (e.g., ARTIVA), include hidden states in the model (EBDBN)

State-space model with feedback loops

$$\mathbf{x}_t = A\mathbf{x}_{t-1} + B\mathbf{y}_{t-1} + \mathbf{w}_t$$
$$\mathbf{y}_t = C\mathbf{x}_t + D\mathbf{y}_{t-1} + \mathbf{v}_t$$

$$\mathbf{w}_t \sim \mathcal{N}(\mathbf{0}, I), \mathbf{v}_t \sim \mathcal{N}(\mathbf{0}, V = \text{diag}(\mathbf{v}^{-1}))$$

- $\mathbf{x}_1, ..., \mathbf{x}_T$ are the K-dimensional hidden states $\Rightarrow$ K is fixed
- D and $CB + D$ are "sub-identifiable" matrices (Rangel et al. 2004)



Hierarchical Bayesian state-space model

Hierarchical Bayesian structure motivated by Beal et al. (2005):

$$\mathbf{x}_t = A\mathbf{x}_{t-1} + B\mathbf{y}_{t-1} + \mathbf{w}_t$$
$$\mathbf{y}_t = C\mathbf{x}_t + D\mathbf{y}_{t-1} + \mathbf{v}_t$$
$$\mathbf{w_t} \sim \mathcal{N}(\mathbf{0}, I), \mathbf{v_t} \sim \mathcal{N}(\mathbf{0}, V = \text{diag}(\mathbf{v}^{-1}))$$

$j = 1, \ldots, K$ hidden states, $i = 1, \ldots, P$ genes

Hierarchical Bayesian state-space model

Hierarchical Bayesian structure motivated by Beal et al. (2005):

$$\mathbf{x}_t = A\mathbf{x}_{t-1} + B\mathbf{y}_{t-1} + \mathbf{w}_t$$
$$\mathbf{y}_t = C\mathbf{x}_t + D\mathbf{y}_{t-1} + \mathbf{v}_t$$
$$\mathbf{w_t} \sim \mathcal{N}(\mathbf{0}, I), \mathbf{v_t} \sim \mathcal{N}(\mathbf{0}, V = \mathrm{diag}(\mathbf{v}^{-1}))$$

$j = 1, \ldots, K$ hidden states, $i = 1, \ldots, P$ genes

Prior distributions:

$$\mathbf{x}_0 \sim \mathcal{N}_k(\boldsymbol{\mu}_0, \Sigma_0)$$
$$A_{rows} \sim \mathcal{N}_k(\mathbf{0}, \mathrm{diag}(\boldsymbol{\alpha})^{-1})$$
$$B_{rows} \sim \mathcal{N}_p(\mathbf{0}, \mathrm{diag}(\boldsymbol{\beta})^{-1})$$
$$C_{rows} \sim \mathcal{N}_k(\mathbf{0}, v_i^{-1}\mathrm{diag}(\boldsymbol{\gamma})^{-1})$$
$$D_{rows} \sim \mathcal{N}_p(\mathbf{0}, v_i^{-1}\mathrm{diag}(\boldsymbol{\delta})^{-1})$$

Variational Bayes State Space Model (Beal et al. 2005)

- Approximate marginal likelihood $p(\mathbf{y}|m)$ for model m with the *a posteriori* variational probability:

$$\begin{aligned} \ln p(\mathbf{y}|m) &\geq \int q_{\mathbf{x}}(\mathbf{x}) q_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \ln \frac{p(\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}|m)}{q_{\mathbf{x}}(\mathbf{x}) q_{\boldsymbol{\theta}}(\boldsymbol{\theta})} d\mathbf{x} d\boldsymbol{\theta} \\ &= \mathcal{F}_m(q_{\mathbf{x}}(\mathbf{x}) q_{\boldsymbol{\theta}}(\boldsymbol{\theta}), \mathbf{y}) \end{aligned}$$

- Variational Bayes EM algorithm for hidden state and parameter estimation, model selection performed by choosing K which maximizes $\mathcal{F}_m(\cdot)$
- Implemented in Matlab (but rather slow to run)

Motivation

$\Rightarrow$ Propose a method based on the SSM of Beal et al. (2005) that is computationally efficient.

Outline

1. Introduction
   - State space models

2. EBDBN Method
   - Model selection
   - Hidden state estimation
   - Parameter estimation

3. Results
   - Simulations
   - T-cell data analysis

4. Discussion

1. Choice of hidden state dimension (K)

Time series method for model selection (still an open research question):

- Construct a block-Hankel matrix of autocovariances of time-series gene expression observations:

$$H = \begin{pmatrix} \hat{\Gamma}_1 & \hat{\Gamma}_2 & \cdots & \hat{\Gamma}_m \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\Gamma}_m & \hat{\Gamma}_{m+1} & \cdots & \hat{\Gamma}_{2m-1} \end{pmatrix}$$

  where $\hat{\Gamma}_i = \frac{1}{T}\sum_{t=1}^{T-i} \mathbf{y}_t \mathbf{y}'_{t+i}$, m is the maximum pertinent biological time-lag between genes and their regulators ($m \leq 3$).

1. Choice of hidden state dimension (K)

Time series method for model selection (still an open research question):

- Construct a block-Hankel matrix of autocovariances of time-series gene expression observations:

$$H = \begin{pmatrix} \hat{\Gamma}_1 & \hat{\Gamma}_2 & \cdots & \hat{\Gamma}_m \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\Gamma}_m & \hat{\Gamma}_{m+1} & \cdots & \hat{\Gamma}_{2m-1} \end{pmatrix}$$

  where $\hat{\Gamma}_i = \frac{1}{T} \sum_{t=1}^{T-i} \mathbf{y}_t \mathbf{y}'_{t+i}$, m is the maximum pertinent biological time-lag between genes and their regulators ($m \leq 3$).

- If signal-noise ratio is large, singular value decomposition will yield K "large" singular values $\Rightarrow$ choose K to be smallest number of singular values needed to explain 90% of total variance

▸ Details

2. Hidden state estimation: Kalman filtering and smoothing

When A, B, C, D, and V are known, the Kalman filter/smoother may be used to recursively estimate the hidden variables:

Kalman filter (prediction and update)

$$\hat{\mathbf{x}}_t^- = A\hat{\mathbf{x}}_{t-1} + B\mathbf{y}_{t-1}$$
$$\hat{\mathbf{x}}_t = \hat{\mathbf{x}}_t^- + \mathbb{K}(\mathbf{y}_t - C\hat{\mathbf{x}}_t^- - D\mathbf{y}_{t-1})$$

Kalman smoother (smooth estimates using all data)

$$\hat{\mathbf{x}}_t^T = \hat{\mathbf{x}}_t + \mathbb{J}(\hat{\mathbf{x}}_{t-1}^T - A\hat{\mathbf{x}}_t - B\hat{\mathbf{y}}_{t-1})$$

- $\mathbb{K}$ and $\mathbb{J}$ are the Kalman gain and smoothing matrices defined in Kalman (1960)

3. Parameter estimation of $\{A, B, C, D, V\}$

- Parameter set: $\boldsymbol{\theta} = \{A, B, C, D, V\}$
- Hyperparameter set: $\boldsymbol{\psi} = \{\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\delta}, \boldsymbol{\mu}_0, \Sigma_0\}$
- Joint likelihood:

$$\begin{aligned} p(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta} | \boldsymbol{\psi}) = & \ p(A|\boldsymbol{\alpha}) p(B|\boldsymbol{\beta}) p(V) p(C|V, \boldsymbol{\gamma}) p(D|V, \boldsymbol{\delta}) \times \\ & \times p(\mathbf{x}_0 | \boldsymbol{\mu}_0, \Sigma_0) \times \\ & \prod_{t=1} p(\mathbf{x}_t | \mathbf{x}_{t-1}, \mathbf{y}_{t-1}, A, B) p(\mathbf{y}_t | \mathbf{x}_t, \mathbf{y}_{t-1}, C, D, V) \end{aligned}$$

- ⇒ Use EM algorithm for hyperparameter estimation, fixing the current values of $\mathbf{x}$

Two-step implementation of the EM algorithm in practice

Fix initial values $\psi^{(0)}$, $\mathbf{v}^{(0)}$, $\mathbf{x}^{(0)}$.

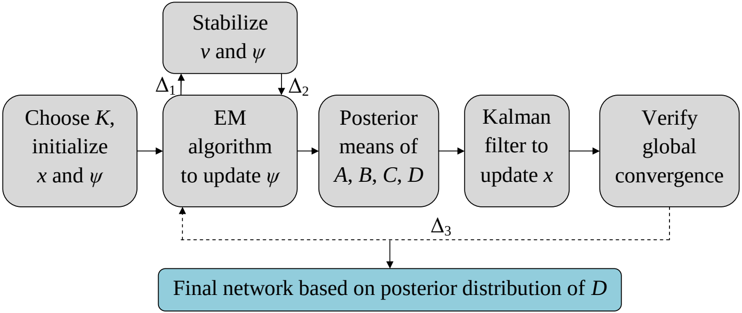
At iteration i:

1. **EM algorithm I**, with $\mathbf{v}^{(i)}$ and $\mathbf{x}^{(i)}$ fixed, to estimate $\tilde{\psi}$
   - Calculate $\tilde{\mathbf{v}}^{(i+1)}$, the innovation variances:

     $$\tilde{v}_m^{(i+1)} = \sum_{t=1}^{T} (y_{tm} - \hat{C}\mathbf{x}_t^{(i-1)} - \hat{D}\mathbf{y}_{t-1})^2/(T-1),$$

     where $\hat{C}$ and $\hat{D}$ are the a posteriori means of C and D given $\tilde{\psi}$ and $\mathbf{x}^{(i)}$
   - Convergence criterion Δ_1
2. **EM algorithm II**, with $\tilde{\mathbf{v}}^{(i+1)}$ and $\mathbf{x}^{(i)}$ fixed, to estimate $\hat{\psi}^{(i+1)}$
   - Convergence criterion Δ_2

Empirical Bayes - Dynamic Bayesian Network (EBDBN) algorithm (Rau et al. 2010)


Stabilize v and ψ
Δ_1
Δ_2
Choose K, initialize x and ψ
EM algorithm to update ψ
Posterior means of A, B, C, D
Kalman filter to update x
Verify global convergence
Δ_3
Final network based on posterior distribution of D

⇒ Standard Z-statistics may be computed for each edge

Outline

1. Introduction
   - State space models
2. EBDBN Method
   - Model selection
   - Hidden state estimation
   - Parameter estimation
3. Results
   - Simulations
   - T-cell data analysis
4. Discussion

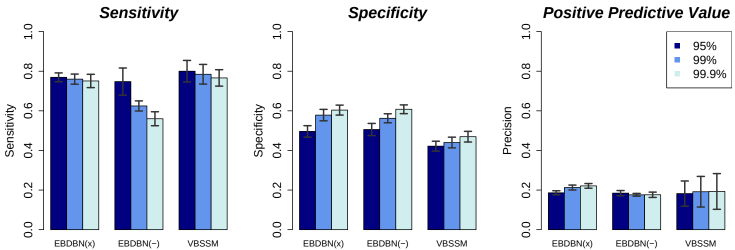
Data-based simulations (Zak et al. 2003)

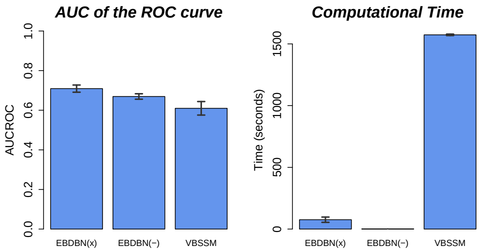
- $P = 10$ genes (+ 45 other observed quantities) with expression level derived from "realistic" interactions with regulatory motifs taken from biological literature
- Simulations in Matlab by integration of ordinary differential equations
- $T = 500$ time points
  - Sub-sampling of time ($T = \{5, 12, 35, 50, 75, 120\}$), replicates generated by adding Gaussian noise

Comparison criteria

- Area Under the Curve of the ROC curve, sensitivity, specificity, positive predictive value, computational time

Simulation results





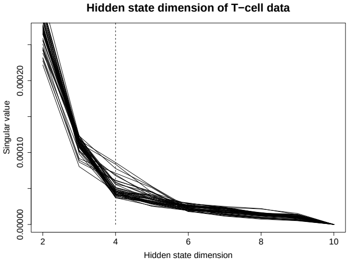
R = 32 replicates
T = 50 time points
(25 datasets)

T-cell Activation Data (Rangel et al. 2004)

T-cell data

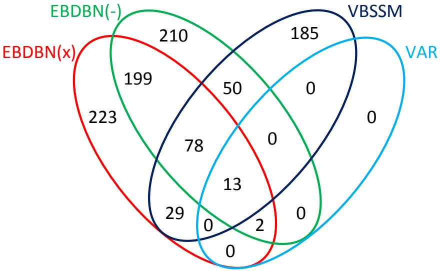
- Study of the response on the expression of T-cells in humans after an ionomicine treatment; genes pre-selected for modulation following activation, reproducibility over replicates
- Pre-treatment: log-transformation and quantile normalization
- $P = 58$ genes, $T = 10$ time points, $R = 44$ replicates

- Choose $K = 4$ via block-Hankel matrix
- Cutoff of 99.9% (Z-scores) used for edge selection in EBDBN



Results for T-cell activation data, by method

Method	# Activation	# Inhibition	Total Edges (%)
EBDBN(x)	435	109	544 (16.2)
EBDBN(-)[1]	338	214	552 (16.4)
VBSSM	233	122	355 (10.6)
VAR[2]	9	6	15 (0.4)



[1] EBDBN method with no hidden states

[2] Vector Auto-Regressive (VAR) model of Opgen-Rhein and Strimmer (2007)

Outline

1. Introduction
   - State space models
2. EBDBN Method
   - Model selection
   - Hidden state estimation
   - Parameter estimation
3. Results
   - Simulations
   - T-cell data analysis
4. Discussion

Discussion

EBDBN method

- Straightforward, EM-like estimation procedure using a state-space model for continuous time-course gene expression data,
- Improved computational speed (implemented in R package `ebdbNet`)

- All methods (EBDBN, VBSSM, VAR, ...) require a minimum number of replicates (≈ 10) and time points (≈ 10) to be effective
- Need for a set of realistic, time-course benchmark datasets

- **Open questions**: What can reliably be inferred from the available data (sub-networks, specific interactions, specific motifs)? How to include other sources of information (e.g., ChIP-chip)? How to define a consensus network? What about NGS data?

Thank you!

RWD research group (Purdue)
PSGen research group (INRA)

My Troung
Doug Crabill
NSF Plant Genome (DBI 0733857)

Beal, M. J., F. Falciani, Z. Ghahramani, C. Rangel, and D. L. Wild (2005). *Bioinformatics*.

Kalman, R. E. (1960). *Transactions of the ASME - Journal of Basic Engineering*.

Opgen-Rhein, R. and K. Strimmer (2007). *BMC Bioinformatics*.

Rangel, C., J. Angus, Z. Ghahramani, M. Lioumi, E. Southeran, A. Gaiba, D. L. Wild, and F. Falciani (2004). *Bioinformatics*.

Rau, A., F. Jaffrézic, J.-L. Foulley, and R. W. Doerge (2010). *Statistical Applications in Genetics and Molecular Biology*.

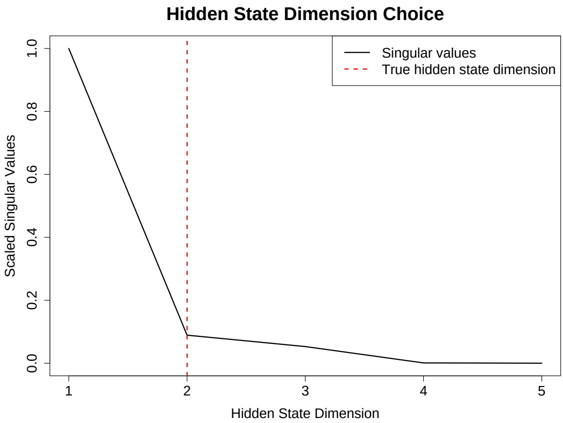
Zak, D. E., G. E. Gonye, J. S. Schwaber, and F. J. Doyle (2003). *Genome Research*.

Appendix: Model selection

- AIC and BIC tend to perform poorly due to the large number of observations and model parameters

- In absence of error, the rank of H equals the number of hidden states K needed to characterize the time series (obviously not true for noisy gene expression data)
- After finding the singular value decomposition of H, there will be K singular values of "large" amplitude, provided the signal-to-noise ratio (SNR) is also large (SNR $\gg 1$).
    - Note that for T time points in data, only the first $T-1$ singular values will be non-zero
- Similar to choosing the number of components in a Principal Components Analysis: choose smallest number of singular values needed to explain 90% of the total variance

Appendix: Model selection example



- Simulated data: $P = 10$ genes, $K = 2$ hidden states, $T = 10$ time points, sample all elements of $\{A, B, C, D\}$ from $\mathcal{U}(-1, 1)$

▸ Back