L-46: Classical and Bayesian approaches to reconstructing genetic regulatory networks

Matthew J. Beal[1], Claudia Rangel[2], Francesco Falciani[3], Zoubin Ghahramani[4], David Wild[5]

[1] Computer Science and Engineering Department, SUNY at Buffalo, NY, USA mbeal@cse.buffalo.edu
[2] Department of Computational and Molecular Biology, University of Southern California, USA rangelc@usc.edu
[3] School of Biosciences, University of Birmingham, UK f.falciani@bham.ac.uk
[4] Gatsby Computational Neuroscience Unit, University College London, UK zoubin@gatsby.ucl.ac.uk
[5] Keck Graduate Institute of Applied Life Sciences, CA, USA david_wild@kgi.edu

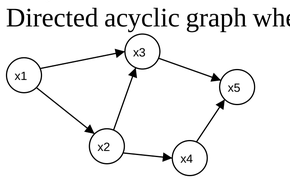
1. Objectives

Can we "reverse engineer" the regulatory networks involved in T-cell activation using highly replicated gene expression profiling time series data and graphical models?

2. Methods

Graphical Models

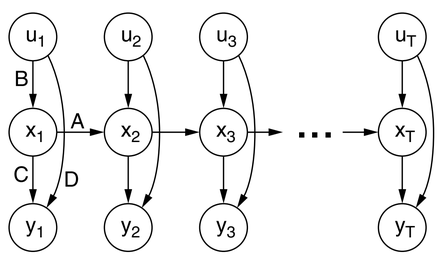
(Bayesian Networks, Belief Nets and Probabilistic Independence Nets.)
Directed acyclic graph where each node corresponds to a random variable.



$$P(\mathbf{x}) = P(\mathbf{x}_1)P(\mathbf{x}_2|\mathbf{x}_1)P(\mathbf{x}_3|\mathbf{x}_1,\mathbf{x}_2) P(\mathbf{x}_4|\mathbf{x}_2)P(\mathbf{x}_5|\mathbf{x}_3,\mathbf{x}_4)$$

Key quantity: joint probability distribution over nodes: $P(\mathbf{x}) = P(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$
The graph specifies a factorization of this joint pdf: $P(\mathbf{x}) = \prod_i P(\mathbf{x}_i|\mathrm{pa}_i)$
Semantics: Given its parents, each node is *conditionally independent* from its *non-descendents*
Definition: A is *conditionally independent* from B given C if $P(A, B|C) = P(A|C)P(B|C)$ for all A, B, and C s.t. $P(C) \neq 0$.

Linear-Gaussian State-space models (SSMs)



Output equation:
$\mathbf{y}_t = C\mathbf{x}_t + D\mathbf{u}_t + \mathbf{v}_t$

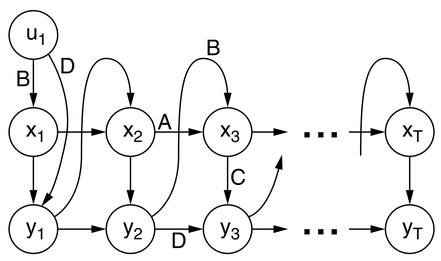
State dynamics equation:
$\mathbf{x}_t = A\mathbf{x}_{t-1} + B\mathbf{u}_t + \mathbf{w}_t$

$$p(\mathbf{x}_{1:T}, \mathbf{y}_{1:T}|\mathbf{u}_{1:T}) = p(\mathbf{x}_1|\mathbf{u}_1)p(\mathbf{y}_1|\mathbf{x}_1, \mathbf{u}_1)\prod_{t=2}^{T} p(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{u}_t)p(\mathbf{y}_t|\mathbf{x}_t, \mathbf{u}_t)$$

Here $\mathbf{x}_t$, $\mathbf{u}_t$ and $\mathbf{y}_t$ are real-valued vectors and $\mathbf{v}$ and $\mathbf{w}$ are uncorrelated zero-mean Gaussian noise vectors.

- A.K.A. stochastic Linear Dynamical Systems, Kalman filter models: These are just continuous-state versions of HMMs.
- Forward–backward algorithm ≡ Kalman smoothing

State-Space Models with Feedback



Output equation:
$\mathbf{y}_t = C\mathbf{x}_t + D\mathbf{y}_{t-1} + \mathbf{v}_t$

State dynamics equation:
$\mathbf{x}_t = A\mathbf{x}_{t-1} + B\mathbf{y}_{t-1} + \mathbf{w}_t$

Key Concept: $\mathbf{y}_t$ represents the measured gene expression level at time step t and $\mathbf{x}_t$ models the many unmeasured (hidden) factors such as

- genes that have not be included in the microarray,
- levels of regulatory proteins,
- the effects of mRNA and protein degradation, etc.

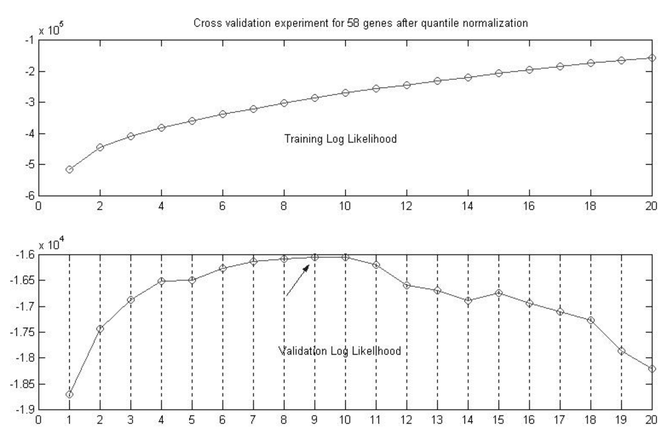
Our Approach

- Let $\theta = \{A, B, C, D, R\}$ be the parameters of the model (R models noise covariance).
- Elements of matrix $[CB + D]$ represent all gene-gene interactions
- Exact Bayesian inference would give us $p(\theta|\mathcal{D})$, which tells us confidence in each parameter and can be used to infer model structure.
- Unfortunately, exact inference is computationally intractable.
- Classical approach uses cross-validation and bootstrapping (Rangel et al., 2004).
- Can also use variational approximations to approximate Bayesian inference in state-space models (Beal, 2003; Beal et al., 2004).

Microarray Data

- Model system of T-cell activation
- Jurkat cells treated with PMA and ionomycin
- Timecourse of gene expression for 88 genes at 10 time points
- 34 'technical' replicates of each profile
- Second experiment with 10 'technical' replicates
- 58 genes in common after removing genes that were poorly reproduced
- Data scaled using *Quantile Normalization*, assuming common distribution of intensities across replicates

Model selection: cross validation to determine number of hidden states



Bootstrap for Parameter Confidence Intervals

Denote a generic element of the matrix $CB + D$ by θ.

- Calculate estimates for the unknown matrices A, B, C, D from the full dataset with replicates using the EM algorithm. From the estimates $\hat{B}, \hat{C}, \hat{D}$, compute $\widehat{\theta}$, the estimate of the given element of $CB + D$.
- Generate N_B independent Bootstrap samples $\mathbf{Y}_1^*, \mathbf{Y}_2^*, ..., \mathbf{Y}_{N_B}^*$ from the original data by resampling from complete time series replicates
- For each bootstrap sample compute bootstrap replicates of the parameters using the EM algorithm on each Bootstrap sample $\mathbf{Y}_i^*, i = 1, 2, ..., N_B$. This yields Bootstrap estimates of the parameters $\{\hat{A}_1^*, \hat{B}_1^*, \hat{C}_1^*, \hat{D}_1^*\}$, ... ,$\{\hat{A}_{N_B}^*, \hat{B}_{N_B}^*, \hat{C}_{N_B}^*, \hat{D}_{N_B}^*\}$.
- From $\{\hat{B}_1^*, \hat{C}_1^*, \hat{D}_1^*\}$, $\{\hat{B}_2^*, \hat{C}_2^*, \hat{D}_2^*\}$, ... , $\{\hat{B}_{N_B}^*, \hat{C}_{N_B}^*, \hat{D}_{N_B}^*\}$, compute the corresponding Bootstrap estimates of the parameter of interest, $\widehat{\theta}_1^*, ..., \widehat{\theta}_{N_B}^*$.
- For the given parameter θ, estimate the distribution of $\widehat{\theta} - \theta$ by the empirical distribution of the values

$$\left\{\widehat{\theta}_j^* - \widehat{\theta} : j = 1, 2, ..., N_B\right\}.$$

Using quantiles of this latter empirical distribution to approximate corresponding quantiles of the distribution of $\widehat{\theta} - \theta$, compute an estimated confidence interval on the parameter θ.

- Test the null hypothesis that the selected parameter is 0 by rejecting the null hypothesis if the confidence interval computed in step 4 does not contain the value 0.
- Repeat previous two steps for each element of $CB + D$. Elements for which zero is between the upper and lower bounds will take the value zero. We obtain a network connectivity matrix in which zeros indicate the absence of a connection, and non-zero elements indicate the presence of a connection.

Variational Bayesian Learning Approach

Let the latent variables be $\mathbf{x}$, data $\mathbf{y}$ and the parameters θ.
We can lower bound the marginal likelihood (using Jensen's inequality):

$$\begin{aligned}\ln p(\mathbf{y}|m) &= \ln \int p(\mathbf{y}, \mathbf{x}, \theta|m)\, d\mathbf{x}\, d\theta \\ &= \ln \int q(\mathbf{x}, \theta)\frac{p(\mathbf{y}, \mathbf{x}, \theta|m)}{q(\mathbf{x}, \theta)}\, d\mathbf{x}\, d\theta \\ &\geq \int q(\mathbf{x}, \theta) \ln \frac{p(\mathbf{y}, \mathbf{x}, \theta|m)}{q(\mathbf{x}, \theta)}\, d\mathbf{x}\, d\theta.\end{aligned}$$

Use a simpler, factorised approximation to $q(\mathbf{x}, \theta) \approx q_{\mathbf{x}}(\mathbf{x})q_\theta(\theta)$:

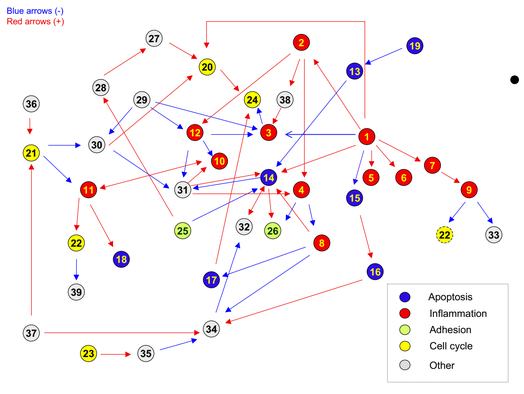
$$\begin{aligned}\ln p(\mathbf{y}|m) &\geq \int q_{\mathbf{x}}(\mathbf{x})q_\theta(\theta) \ln \frac{p(\mathbf{y}, \mathbf{x}, \theta|m)}{q_{\mathbf{x}}(\mathbf{x})q_\theta(\theta)}\, d\mathbf{x}\, d\theta \\ &= \mathcal{F}_m(q_{\mathbf{x}}(\mathbf{x}), q_\theta(\theta), \mathbf{y}).\end{aligned}$$

Maximizing this lower bound, $\mathcal{F}_m$, leads to **EM-like** iterative updates. $-\mathcal{F}_m$ is analogous to a variational free energy

3. Results

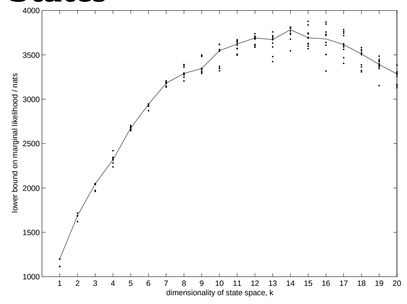
Classical Approach: Inferred Regulatory Networks

Red arrows (+), Blue arrows (−)



- Directed graph representing gene-gene interactions with a confidence level on individual connections equal to 99.66%.
- Some key genes: FYB (1), IL3Rα (2), CD 69 (3), TRAF5 (4), IL4Rα (5), GATA binding protein 3 (6), IL-2Rγ (7), chemokine receptor CX3CR1 (9), interleukin-16 (11), Jun B (13), Caspase 8 (14), Clusterin (15), Caspase 7 (18), survival of motor neuron 1 (19), Cyclin A2 (20), CDC2 (21), PCNA (22), Integrin alpha-M (26), MCL-1 (31).

VB Approach: Inferring the Number of Hidden States



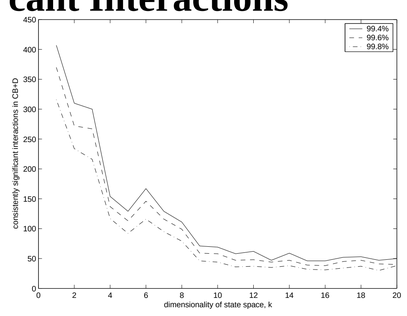
Variation of $\mathcal{F}$, the lower bound on the marginal likelihood, with hidden state dimension k for 10 random initialisations of VBEM.

We can use this lower bound to infer/select the number of hidden states.

VB Approach: Inferring Regulatory Networks

- We examined the gene-gene influences represented by elements of the matrix $[CB + D]$.
- The VB algorithm provides us with approximate posterior distributions for the parameters B, C and D.
- Using the posterior distributions for these parameters we compute the distribution of each of the elements in the combined matrix $[CB + D]$.
- Significant interactions correspond to the zero point being $> n$ standard deviations from the posterior mean for that entry (use Z statistic).

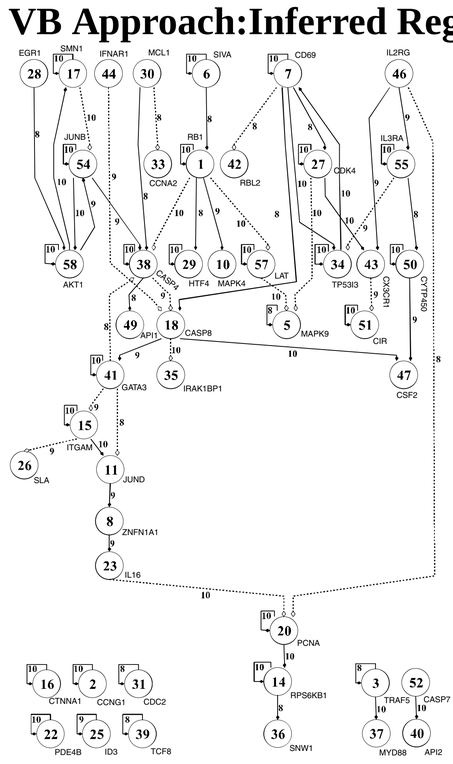
VB Approach: Inferring the Number of Significant Interactions



The number of significant interactions that are repeated in all 10 runs of VBEM at each value of k.

The 3 plots correspond to different significance levels.

VB Approach:Inferred Regulatory Networks



- Gene-gene interactions present in $\geq 80\%$ of the VB state-space models out of 10 random seeds and $k = 14$ at a confidence level of 99.8%.
- The number inside each node is the gene identity
- Numbers on the edges represent the number of models from 10 different random seeds in which the interaction is supported at this confidence level.
- Dotted lines are negative interactions, and continuous lines represent positive interactions.
- Transcriptional networks in T cell activation → testable hypotheses.

4. Conclusions

- Graphical models and Bayesian methods can be used for a variety of modelling problems in bioinformatics.
- These allow large-scale statistical models to be learned and sources of noise and uncertainty to be included in a principled manner
- We have looked at one problem domain: inferring genetic regulatory networks — a simple graphical model (state-space models) can be used.
- State-space models allow hidden variables to be included.
- Bayesian "Occam's Razor" prunes networks to be sparse.
- Models produce plausible biological hypotheses which can be experimentally validated

Future Work

A framework to build on with future work:

- incorporating biologically plausible nonlinearities
- adding prior knowledge (especially in the form of constraints on positive and negative interactions)
- combining gene and protein expression data with metabolomic data
- making and testing knockout and overexpresson predictions
- well defined model systems
- basic difficulty: usually not enough data...

Acknowledgements

MJB is generously supported by the Center of Excellence in Bioinformatics and Life Sciences at SUNY Buffalo NY. CR acknowledges support from the Keck Graduate Institute of Applied Life Sciences.