Network Reconstruction via Dynamic Systems:
Biological Network Reconstruction using Differential Equation Models

James Henderson and George Michailidis
jbhender@umich.edu, gmichail@umich.edu

University of Michigan STATISTICS

Problem: Network Reconstruction

Gene regulatory networks represent functional relationships among genes as a directed network with edges corresponding to regulatory relationships; *i.e.* an edge $i \to j$ denotes that gene i regulates gene j. **Recovering the regulatory relationships among genes is a fundamental problem in systems biology** with important applications in disease treatment and diagnosis.

Common Approaches

Conditional Independence Models, including *Bayesian Networks*, are based on learning a *directed acyclic graph* encoding conditional independence relations. *However, this approach cannot accomodate biologically relevant cycles such as feedback loops.* **Time Series Models** including *Dynamic Bayesian Networks* and *Vector Autoregressive Models* avoid this limitation by restricting edges to point forward in time. **ODE Models** are perhaps the most widespread class of models in the biological and physical sciences. *ODE-based methods* for network reconstruction from time-series data have been proposed but take an *ad-hoc* approach to inference.

Networks, Dynamic Systems, & Coupling

This poster describes **methodology for time series data, assumed to be noisy observations of an underlying dynamic system**,

$$\dot{x}^{(r)}(t) = \boldsymbol{\alpha} + f(\boldsymbol{x}^{(r)}(t)) + \boldsymbol{u}^{(r)}(t), \quad \boldsymbol{y}^{(r)}(t_k) = \boldsymbol{x}^{(r)}(t_k) + \boldsymbol{\epsilon}_k, \qquad (1)$$

with r indexing inedpendent experiments and $\boldsymbol{u}^{(r)}(t)$ representing a vector of control parameters. **Only the trajectories are observed, not the derivatives.**

Network reconstruction is recast as variable selection for a dynamic system; i.e. determining the important variables in f,

$$i \to j \iff \frac{\partial f_i}{\partial x_j}(\boldsymbol{x}(t)) \not\equiv 0. \qquad (2)$$

Rank potential edges using the **coupling metric**,

$$\rho(i,j) := \int_0^1 \left| \frac{\partial f_i}{\partial x_j}(\boldsymbol{x}) \right| d\boldsymbol{x}, \qquad (3)$$

measuring the strength of the regulatory relationship $i \to j$. Signed coupling metrics $\rho_+(i,j), \rho_-(i,j)$ can be defined similarly by replacing $|\cdot|$ with $(\cdot)_+$ or $(\cdot)_-$ respectively; useful for recovering signed edges.

References

[1] Gugushvili, Shota and Klaassen, Chris A.J. '$\sqrt{n}$-consistent parameter estimation for systems of ordinary differential equations: bypassing numerical integration via smoothing.' In *Bernoulli*, 2012.

[2] Ravikumar, P., Lafferty, J., Liu, H. and Wasserman, L. 'Sparse Additive Models.' In *JRSS:B*, 2009.

ODE Estimation: A Two-Stage Approach

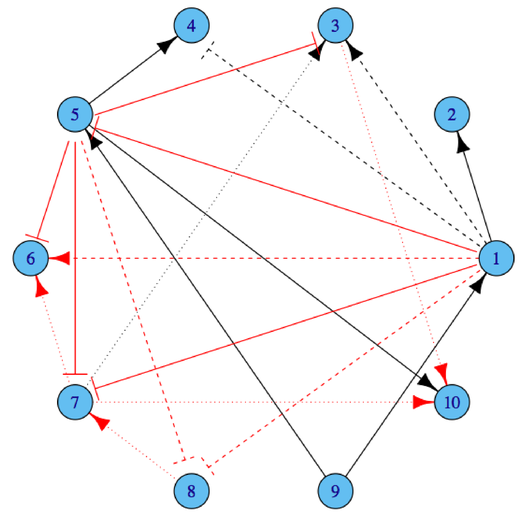
Stage 1: Estimate trajectories and derivatives using smoothing splines

- Avoids computationally intensive numerical integration
- Decouples the ODE estimation into d sub-problems

Stage 2: Fit an additive, non-parametric, approximation to the ODE

- The non-parametric approach allows for flexible nonlinear models and sidesteps potential model misspecification.
- An additive approximation keeps the problem computationally feasible.

Performance: *E. Coli* 2 SubNetwork



Edge and sign recovery on the E. coli 2 subnetwork. *Line type is used to indicate edge recovery among the top 15-ranked edges: solid lines are true positives (10), dashed lines are false negatives (5), and dotted lines are false positives (5). Blunt-arrows indicate negatives on the data-generating network, while negatives on the estimated network are colored red. The area under the ROC curve and Precision-Recall curve are .87 and .68, respectively. All 15 signs are recovered correctly.*

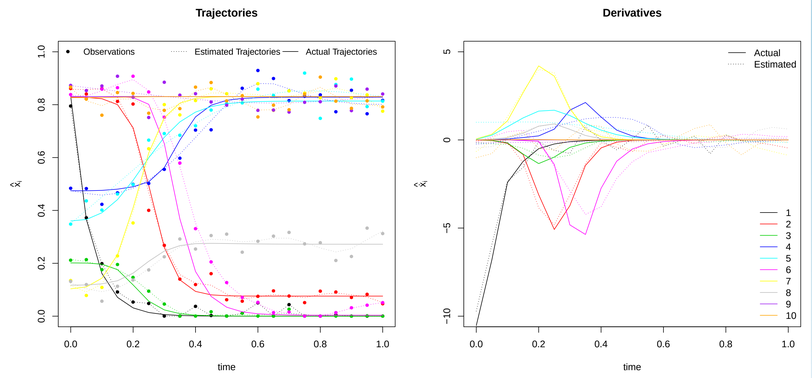
Tuning Parameters

There are 3 sets of tuning parameters to be selected:

- λ_0 controls the smoothness of the first-stage estimates,
- λ_1 controls the smoothness of the additive components,
- λ_2 controls the sparsity of each additive model.

Leave-one-experiment-out cross validation is used to select $\lambda_2 = \lambda_{2i}$ while $\lambda_0 = \lambda_{0,ir}$ is selected by GCV. Here $\lambda_1 = .76$ is fixed.

Stage 1: Smooth

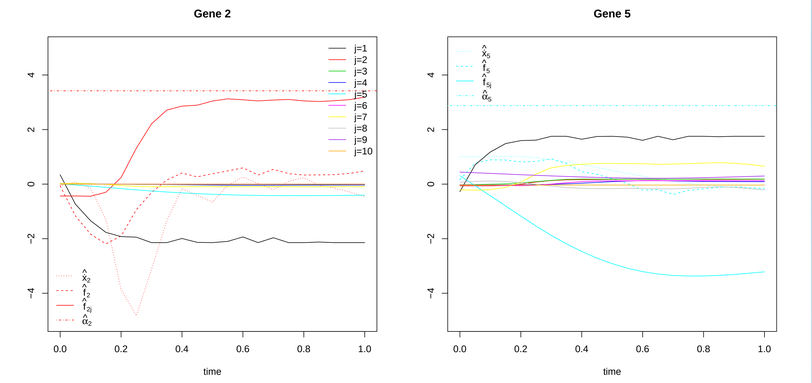


Trajectories and derivatives after a knockout of gene 1.
For components $i = 1, ..., p$ and experimental realizations $r = 1, ..., R$,

$$\hat{x}_i^{(r)}(t) = \arg\min_{x \in W_2^2} n^{-1} \sum_{k=1}^{n} [y_i^{(r)}(t_k) - x(t_k)]^2 + \lambda_{0ir} \int_0^1 [\ddot{x}(t)]^2 dt, \quad (4)$$

and $\hat{\dot{x}}_i(t) = \frac{d}{dt}\hat{x}_i(t)$.

Stage 2: Fit an Additive Model



The estimated differential equation (dashed line) for gene 2 (left) and gene 5 (right) following a knockout of gene 1. The estimates are sums of the additive components (solid lines). The dotted lines show the derivative estimate from the first stage.
Estimate the right-hand side function f using $\hat{f}_i = \arg\min M_{in}^{(r)}(f)$,

$$M_{in}(f) = ||\hat{\dot{x}}_i - \sum_{j=1}^{p} f_{ij}(\hat{x}_i)||_n^2 + J(\lambda_1, \lambda_{2i}), \qquad (5)$$

$$J(\lambda_1, \lambda_2) := \sum_{j=1}^{p} \lambda_1 \int [\ddot{f}_{ij}(x)]^2 dx + \lambda_2 \sum_{j=1}^{p} \sqrt{\int [f_{ij}(x)]^2 dx}. \qquad (6)$$

For knockout time-series constrain $f_{ii}(x_i) = \alpha 1_{[r=i]} + \beta x_i$. The minimization is carried out using a modified version of the sparse-backfitting algorithm (Ravikumar, 2009).