Sparse, stable gene regulatory network recovery

Arwen Meister

Gene regulatory networks

- gene expression regulation allow cells to dynamically control protein levels
- certain genes code for protein that regulate other genes, giving rise to a network with regulatory genes as nodes and regulatory relationships as edges
- want to infer gene regulatory network from experimental data
- network should be sparse (each regulator has only a few targets)
- should also have stable equilibria (cell recovers from small perturbations)

Dynamical systems model

- model cell state as a time-varying vector \( x(t) \in \mathbb{R}^n \) of gene expression levels
- \( x(t) \) evolves according to \( dx/dt = A(x(t)) \), where \( A \) is a smooth nonlinear function on \( \mathbb{R}^n \)
- equilibrium points \( \mu \) with \( A(\mu) = 0 \) correspond to basic cell types like embryonic stem cell or liver cell
- Taylor expand \( A \) about an equilibrium \( \mu \):
  \[
  \frac{d\Delta x}{dt} = A(x) - T(x - \mu) = \mu - e^{T(t - \mu)}
  \]
  \( T \) is the \( n \times n \) Jacobian matrix of \( A \) at \( \mu \) and \( x \) is a perturbation of \( \mu \)
- matrix \( T \) models regulatory network at equilibrium
  - \( T_{ij} > 0 \) if gene \( j \) up-regulates gene \( i \)
  - \( T_{ij} < 0 \) for down-regulation
  - \( T \) reflects self-regulation and degradation of gene products

Convex modeling

To recover the network matrix \( T \) from noisy measurements \( x(t) \), we must solve
\[
\begin{align*}
\text{minimize} & \quad \|x(t) - \mu - e^{T(t - \mu)}\|_2 \\
\text{subject to} & \quad \text{card}(T + \gamma I) \leq k \\
& \quad PT + T^TP < 0
\end{align*}
\]
with variables \( T \in \mathbb{R}^{n \times n} \), \( P \in \mathbb{S}^n \), and data \( t, \gamma, \mu, x_0, x(t) \in \mathbb{R}^n \)
- objective ensures that \( T \) is consistent with the data \( x(t) \)
- first constraint enforces sparsity (taking the degradation rate \( \gamma \) into account on the diagonal)
- second enforces stability of the equilibrium \( \mu \), which is equivalent to the existence of a Lyapunov matrix \( P \) with \( PT + T^TP < 0 \)

Problem data

- problem data comes from expression measurements shortly after gene “knockdown,” where expression level of one gene is reduced to a fixed level
- modeling a knockdown as a small perturbation and subsequent evolution as an exponential trajectory is only approximate
- test on simulated six-gene subnetwork in embryonic stem cell, where network matrix \( T_{\text{True}} \) is known
- generate data by knocking down each gene in turn and letting others evolve
- sample \( x(t_1), x(t_2) \) for small \( t_1, t_2 \); add 10% Gaussian noise

Basic Recovery

- minimize \( \|x(t) - x_0\|/e^{T(t - \mu)}\|_2 \) without constraints (Fig. 1)
- successful recovery from clean data (though not quite sparse or stable)
- fails with noisy data: diagonals are not recovered, large positive eigenvalues

Enforcing sparsity and stability

- sparsity (Fig. 2)
  - regularization can improve both sparsity and diagonal recovery
  - tune parameter \( \lambda \) with leave-one-out cross validation
  - choose \( \lambda = 0.1 \); good tradeoff between accuracy and sparsity
- stability (Fig. 3)
  - iterative heuristic: solve alternately in \( T \) and \( P \), starting with \( P = I \)
  - iterates are always feasible, but may not converge to the solution
  - no non-heuristic stopping criteria, so terminate when \( \|T_k - T_{k-1}\| \leq \text{tol} \)
  - final objective value about the same as without stability constraint
  - \( T_{\text{True}} \) has a higher objective value (model is only approximate)
  - final maximum eigenvalue close to that of \( T_{\text{True}} \)

Nonlinear approximations of the exponential

- instead of using the linearization \( e^{T(t)} \approx I + tT \) to form a convex objective; we can consider more accurate convex approximations of \( e^{T(t)} \)
- Cayley transform \( C_T = (I + \frac{1}{2}T)^{-1}(I - \frac{1}{2}T) \) is quadratically accurate and yields convex objective \( \|I - (I - T/2)z_0 - (I + T/2)z_0/2\|_2 \)
- refine estimate \( \hat{T} \) by minimizing \( \|x(t) - e^{T(t)}x(0)\|_2 \) with \( \delta \in \mathbb{R}^{n \times n} \) and setting \( T = \hat{T} + \delta \)
- (linearize \( e^{T(t)\delta} \) to get a convex problem)
- impressive results on samples from an exponential trajectory (Fig. 4)
- no better than linear model for knockdown data

Conclusion

- successful recovery from noisy data using the linear model with \( \ell_1 \)-regularization and iterative enforcement stability constraint
- recovered solution satisfies sparsity and stability constraints, with good off-diagonal and reasonable diagonals (Fig. 3)
- sophisticated approximations of the exponential are not helpful, since the knockdowns do not really follow an exponential model
- constraints not only impart desired properties but also improve noisy recovery through regularization