Reasoning over pathways and multi-omic data: Model concepts

David Merrell
dmerrell@cs.wisc.edu

July 31, 2020

Abstract

Our goal is to construct a Bayesian model that allows us to reason about (i) pathways and (ii) multi-omic data. We take inspiration from PARADIGM, but want a tool that (a) can integrate more kinds of data; (b) treats pathway activation as a first-class model variable, and models the pathway activations jointly; (c) is more computationally efficient; (d) is informed by modern probabilistic modeling tools. In this document I brainstorm some model ideas.

It turns out that my model ideas share some common features. Each model does two things:

- Each assigns a precise meaning to pathways.
- Each tells a mathematical story that connects pathways to observed (and unobserved) data.

1 "Steady State Diffusion" Model

- Overview
  - This model is based on a couple of core premises:
    - (1) Pathways describe the *dynamics* of a system that evolves over time.
    - (2) Our data are measured after the system has reached a *steady state*.

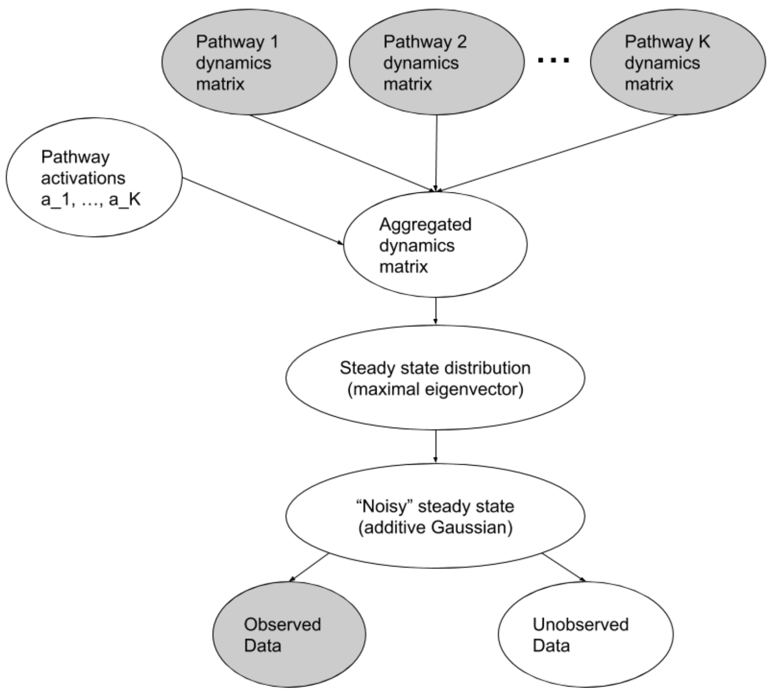

Figure 1: Schematic of the steady state diffusion concept.

- This "steady state" formulation solves a tough conceptual issue: biological systems are dynamic, but our measurements are static (assuming we don't have time series data).
- Premise (2) may or may not be correct. But I can't think of any better way to relate the data to a dynamical system.

- Model Details
  - Pathway $\leftrightarrow$ dynamical system
    - Let there be K pathways represented by adjacency matrices $A_1, \ldots, A_K$.
      - Let the pathways include d DNA, RNA, and protein-level entities, just as in PARADIGM.
      - They may include other entities as well (e.g., abstract processes or phenotypes like "apoptosis"). Again, just as in PARADIGM.
    - For each pathway A_k let there be a *pathway activation* variable $\alpha_k \geq 0$.
- Inference on this model
  - Variational Bayes; ADVI

2 "Hierarchical Precision Matrix" Model

- Overview
  - A pathway depicts a set of independencies and correlations between variables.
    - The directed network structure implies a set of conditional independencies between variables, just as in a directed graphical model.
    - The promoter/inhibitor relationships imply positive and negative correlations between variables, respectively.
  - A precision matrix (inverse covariance matrix) is perhaps the most straightforward way to capture all of these relationships (*under an assumption of normality*).
  - The idea is to translate pathways into a precision matrix, and assume a patient's data is drawn from the corresponding multivariate normal distribution.
- Model Details

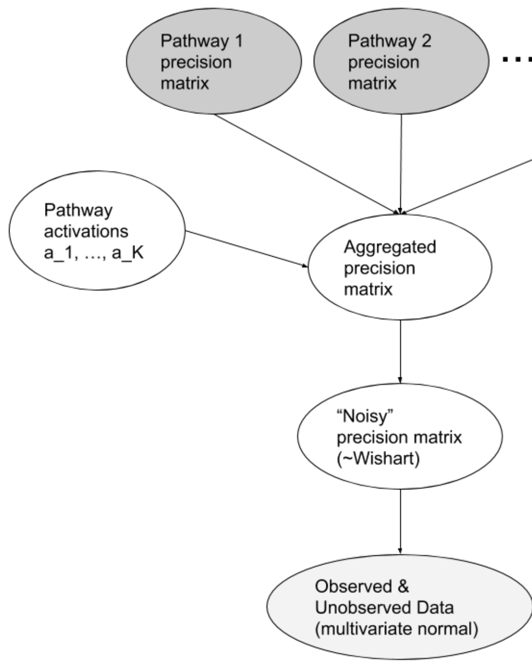

Figure 2: Schematic of the hierarchical precision matrix concept.

- Standardize the data (make each variable's marginal distribution $\mathcal{N}(0,1)$)
- Let there be K pathways, represented by directed graphs.
  * Let the "pathways" include d DNA, RNA, and protein-level entities just as in PARADIGM
  * They may include other entities as well (e.g., processes or phenotypes like "apoptosis"). Again, just as in PARADIGM.
- For each pathway, construct a precision matrix Ω_k.
  * There exists a straightforward way to do this if all the variables are continuous/normally distributed.
    · Initiate the precision matrix with zeros: $\Omega_k \leftarrow 0$.
    · Eliminate all 1-cycles and 2-cycles in the directed graph (k-cycles are okay, for $k > 2$.)
    · Assume each variable X is a linear combination of its parents Y. For promoter parents the coefficient is positive. For suppressor parents the coefficient is negative.
    · For each variable $X \sim \mathcal{N}(a^\top Y, \sigma^2)$, update Ω_k with the following rule:

$$\Omega_k \leftarrow \Omega_k + \frac{1}{\sigma^2} \begin{bmatrix} 1 & a^\top \\ a & aa^\top \end{bmatrix}$$

    (More accurately, the rule updates a *submatrix* of the full $d \times d$ precision matrix Ω_k. The upper/left dimension corresponds to X; the lower/right dimensions correspond to the parents Y.)
    · Remarkably, this produces a consistent joint distribution even in the presence of k-cycles ($k > 2$). (NEED TO PROVE/DISPROVE)
  * If there are discrete variables, then it gets more complicated. I'm still thinking about ways to handle them.
- For each pathway, let there be an *activation* variable $a_k \in (0,1)$.
- Construct an aggregate precision matrix Ω_{agg} from the pathway-specific Ω_ks:

$$\Omega_{agg} = \frac{1}{K+1} \left(I + \sum_k \Omega_k a_k \right)$$

- Treat Ω_{agg} as the parameter for a Wishart distribution; draw a final precision matrix $\hat{\Omega}$ from it:

$$\hat{\Omega} \sim W_n(\Omega_{agg}, d)$$

- Assume that the patient's observed and unobserved data are distributed by a multivariate Gaussian, conditioned on $\hat{\Omega}$:

$$X \sim \mathcal{N}(0, \hat{\Omega})$$

- Inference on this model
  - Variational Bayes; ADVI