Empirical Bayes and Dynamical Systems Approaches to Clustering Gene Expression Time Series Data

Sara Venkatraman, Sumanta Basu, Andrew G. Clark, Sofie Delbare, Myung-Hee Lee, Martin T. Wells

Cornell University, Department of Statistics and Data Science

Abstract. Time-course gene expression data provides insight into the dynamics of complex biological processes, such as immune response. It is of interest to identify genes with similar temporal expression patterns because such genes are often biologically related. However, this task is complicated by the high dimensionality of genetic datasets and the nonlinearity of gene expression time dynamics. We propose an empirical Bayes approach to estimating ordinary differential equation (ODE) models of gene expression, from which we derive metrics called the Bayesian lead-lag R^2 values that capture similarities in the time dynamics of two genes. The key feature of our method is that it leverages biological databases that document known interactions between genes. This information is used to define informative prior distributions on the ODE model's parameters. Our biologically-informed similarity metrics allow us to recover clusters or networks of functionally-related genes.

Problem setup

- Objective: Cluster genes based on their temporal expression patterns. Similar expressions might be due to genes being *co-regulated* by same transcription factors.
- Statistical challenges:
  - High-dimensional data (thousands of genes) + small sample sizes (number of time points)
  - Time dynamics of gene expression are nonlinear
  - Clustering relies on similarity metrics, which often ignore *a priori* biological network information in literature
- Our approach:
  - Derive similarity metrics from models that explain dynamics of gene expression
  - Incorporate prior biological information (protein interaction networks, pathway databases) into calculation of similarity metrics

Gene expression as a dynamical system

- Ordinary differential equation model of co-regulated genes: For possibly co-regulated genes A and B, [1] proposed modeling temporal expressions $m_A(t)$, $m_B(t)$ as dependent on common signal $p(t)$:

$$\frac{dm_A(t)}{dt} = \alpha_A p(t) + \beta_A - \kappa_A m_A(t) + \varepsilon_t$$
$$\frac{dm_B(t)}{dt} = \alpha_B p(t) + \beta_B - \kappa_B m_B(t) + \varepsilon_t$$

where α_A, α_B measure strength of $p(t)$ in first-order dynamics; κ_A, κ_B are mRNA degradation rates. Solve for $p(t)$ and integrate to get $m_A(t)$ in terms of $m_B(t)$:

$$m_A(t) = c_1 m_B(t) + c_2 \int_0^t m_B + c_3 \int_0^t m_A + c_4 t + c_5 \qquad (1)$$

where, e.g., $c_1 = \alpha_A/\alpha_B$ and $c_2 = \alpha_A \kappa_B/\alpha_B$.

- Linear modeling: Given gene expression measurements $\{m_A(t_i)\}_{i=1}^n$ and $\{m_B(t_i)\}_{i=1}^n$, fit (1) as the linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $\boldsymbol{\beta} = [c_1, c_2, c_3, c_4, c_5]^T$, $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, and

$$\mathbf{Y} = \begin{bmatrix} m_A(t_1) \\ \dots \\ m_A(t_n) \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} m_B(t_1) & \int_0^{t_1} m_B & \int_0^{t_1} m_A & t_1 & 1 \\ \dots & \dots & \dots & \dots & \dots \\ m_B(t_n) & \int_0^{t_n} m_B & \int_0^{t_n} m_A & t_n & 1 \end{bmatrix}.$$

- Similarity metric (lead-lag R^2): According to [1], R^2 from fitting (1) indicates genes A, B may be co-regulated or at least associated:

$$\text{Lead-lag } R^2(A,B) = \frac{\|\mathbf{X}\hat{\boldsymbol{\beta}}_{\text{OLS}} - \overline{Y}\mathbb{1}_n\|^2}{\|\mathbf{Y} - \overline{Y}\mathbb{1}_n\|^2}.$$

Empirical Bayes methodology

- Motivation: Lead-lag R^2 similariity metric from [1] returns many *false positives* - gene pairs with high lead-lag R^2, but no biological relationship.
- Why incorporate prior biological network information into lead-lag R^2:
  - Encourage genes known to be associated to receive higher pairwise scores
  - Filter away gene pairs that are unlikely to be related
  - Identify which genes are uncharacterized, but show similar patterns to those with known functionality
- Part 1 – Prior adjacency matrix: Given a dataset of N genes measured at n time points, use biological databases (e.g. GO, KEGG, Reactome, STRING) to construct $N \times N$ adjacency matrix $\mathbf{W}$:

$$\mathbf{W}_{ij} = \begin{cases} 1, & \text{if genes } i,j \text{ are known to be associated} \\ \text{NA}, & \text{if genes } i,j \text{ have an unknown relationship} \\ 0, & \text{if genes } i,j \text{ are unlikely to be associated.} \end{cases}$$

Next, apply parts 2 and 3 below to each gene pair.

- Part 2 – Define normal-inverse gamma prior (Zellner's g-prior) on $\boldsymbol{\beta}$:
  - Parameters c_1, c_2 in (1) link expressions of genes A and B
  - Therefore: place prior distributions of non-zero mean on c_1, c_2 if the two genes are known to be associated. Otherwise, use priors with mean zero.
  - Normal-inverse gamma prior on $(\boldsymbol{\beta}, \sigma^2)$:

$$\boldsymbol{\beta}|\sigma^2 \sim N(\boldsymbol{\beta}_0, \sigma^2 \mathbf{V}_0), \quad \sigma^2 \sim \Gamma^{-1}(a,b)$$

  where $\boldsymbol{\beta}_0 \in \mathbb{R}^5$, $\mathbf{V}_0$ is positive semi-definite, and $a, b > 0$.
  - Set $\boldsymbol{\beta}_0$ using prior adjacency matrix $\mathbf{W}$:

$$\boldsymbol{\beta}_0 = \begin{cases} [1,1,0,0,0]^T, & \text{if } \mathbf{W}_{ij} = 1 \\ [0,0,0,0,0]^T, & \text{if } \mathbf{W}_{ij} = 0 \text{ or NA.} \end{cases}$$

  Or, set $\boldsymbol{\beta}_0 = [\xi, \xi, 0, 0, 0]^T$ when $\mathbf{W}_{ij} = 1$, where ξ is chosen adaptively.
  - Set $\mathbf{V}_0$ according to *Zellner's g-prior*: $\mathbf{V}_0 = g(\mathbf{X}^T\mathbf{X})^{-1}$ where $g > 0$.
  - Under Zellner's g-prior, posterior mean of $\boldsymbol{\beta}$ is:

$$\boldsymbol{\beta}_* = \mathbb{E}(\boldsymbol{\beta}|\mathbf{Y}) = \frac{1}{1+g}\boldsymbol{\beta}_0 + \frac{g}{1+g}\hat{\boldsymbol{\beta}}_{\text{OLS}}, \qquad (2)$$

  where $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ is the least-squares estimator $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$.
  - $\boldsymbol{\beta}_*$ is a weighted average of prior information ($\boldsymbol{\beta}_0$) and the data ($\hat{\boldsymbol{\beta}}_{\text{OLS}}$). Admissible and consistent estimator of $\boldsymbol{\beta}$.
  - See right for data-driven selection of g parameter
- Part 3 – Compute Bayesian lead-lag R^2: Standard definition of R^2 for least-squares regression may yield $R^2 > 1$ for Bayesian models. So we use:

$$\text{Bayesian lead-lag } R^2(A,B) = \frac{\widehat{\text{Var}}(\mathbf{X}\boldsymbol{\beta}_*)}{\widehat{\text{Var}}(\mathbf{X}\boldsymbol{\beta}_*) + \widehat{\text{Var}}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_*)}$$

Automatic balance between priors and data

- Parameter g in Zellner's g-prior balances prior information and data in posterior regression coefficients $\boldsymbol{\beta}_*$ in (2)
- Data-driven selection of g:
  - No analytical solutions to $g_* = \text{argmin}_{g>0} \|\mathbf{Y} - \hat{\mathbf{Y}}\|^2$ where $\hat{\mathbf{Y}} = \mathbf{X}\boldsymbol{\beta}_*$
  - Instead, minimize *Stein's unbiased risk estimate* $\delta_0(\mathbf{Y})$, an unbiased estimator of $\|\hat{\mathbf{Y}} - \mathbf{X}\boldsymbol{\beta}\|^2$. If estimating $\boldsymbol{\beta}$ by $\boldsymbol{\beta}_*$ in (2), $\delta_0(\mathbf{Y})$ is:

$$\delta_0(\mathbf{Y}) = \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}_*\|^2 + \left(\frac{2gp}{1+g} - n\right)\hat{\sigma}^2,$$

  where $\hat{\sigma}^2 = \|\mathbf{Y} - \hat{\mathbf{Y}}_{\text{OLS}}\|^2/(n-p)$, and $\mathbf{X}$ is $n \times p$.

Theorem 1. *The value of g which minimizes Stein's unbiased risk estimate is:*

$$g_* = \frac{\|\hat{\mathbf{Y}}_{\text{OLS}} - \mathbf{X}\boldsymbol{\beta}_0\|^2 - p\hat{\sigma}^2}{p\hat{\sigma}^2}.$$

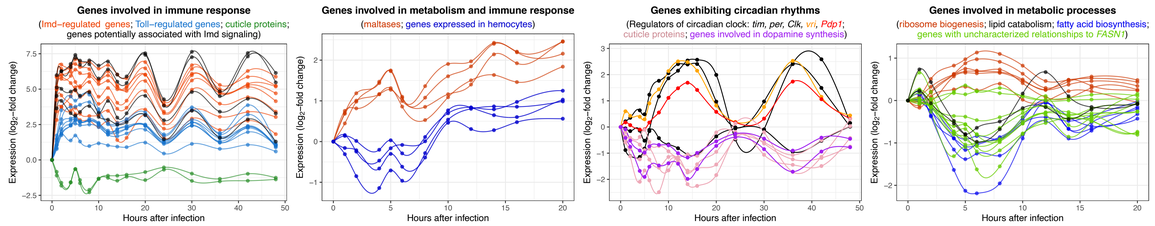
Theorem 2. *When setting $\boldsymbol{\beta}_0 = [\xi, \xi, 0, 0, 0]^T$ in the $\mathbf{W}_{ij} = 1$ case, the values of ξ and g that minimize Stein's unbiased risk estimate are:*

$$\xi_* = \frac{\mathbf{Y}^T\mathbf{X}_{12}}{\|\mathbf{X}_{12}\|^2}, \qquad g_* = \frac{\|\hat{\mathbf{Y}}_{\text{OLS}}\|^2\|\mathbf{X}_{12}\|^2 - (\mathbf{Y}^T\mathbf{X}_{12})^2}{\|\mathbf{X}_{12}\|^2 p\hat{\sigma}^2} - 1,$$

where $\mathbf{X}_{12}$ is the element-wise sum of first two columns of $\mathbf{X}$.

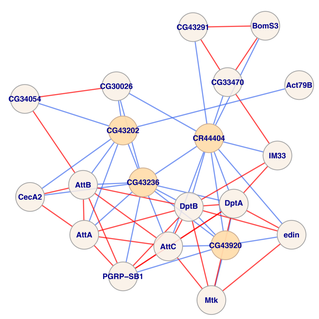
Results on real gene expression data

- Data: provided by [2]; contains expressions of 12,657 genes in *Drosophila melanogaster* (fruit fly) at $n = 21$ time points immediately following an induced immune response. Reduce to a set of $N = 1735$ genes that are differentially-expressed (DE) or are associated with DE genes.
- Analysis: Hierarchical clustering applied to the Bayesian lead-lag R^2 similarity matrix
  - Prior adjacency matrix $\mathbf{W}$ constructed using the STRING database. STRING provides each gene pair a score in $[0,1]$ indicating likelihood of association.
  - All clusters significantly enriched for specific biological functions, according to Gene Ontology (GO) analysis
  - Shown below: clusters recover known interplay between immune response and metabolism, and suggest roles for numerous uncharacterized genes



Network reconstruction

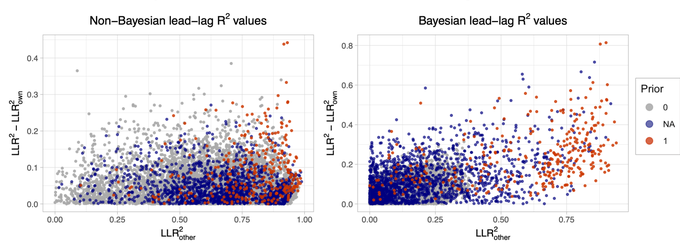
Neighbors of genes *CG43202, CR44404, CG43236, CG43920*. Edge = Bayesian lead-lag $R^2 > 0.9$. Blue edges: previously unknown ($\mathbf{W}_{ij} = \text{NA}$). Red edges: known ($\mathbf{W}_{ij} = 1$).



Comparing distributions of lead-lag R^2 and Bayesian lead-lag R^2 values

In addition to the Bayesian lead-lag R^2 (LLR^2), we also compute:

- "LLR^2_{other}": from the model $m_A(t) = c_1 m_B(t) + c_2 \int_0^t m_B(s)ds + c_5$. Indicates variation in gene A due just to gene B.
- "LLR^2_{own}": from the model $m_A(t) = c_3 \int_0^t m_A(s)ds + c_4 t + c_5$. Indicates variation in gene A due to its own past + linear time trend.
- If genes A, B are truly associated, LLR^2_{other} and LLR^2 – LLR^2_{other} should be large
- Shown below: Bayesian method shifts distribution of lead-lag R^2; fewer false positives



References

[1] Lorenzo Farina, Alberto De Santis, Samanta Salvucci, Giorgio Morelli, and Ida Ruberti. Embedding mRNA stability in correlation analysis of time-series gene expression data. *PLoS Computational Biology*, 4(8), 2008.

[2] Florencia Schlamp, Sofie Delbare, Angela M Early, Martin T Wells, Sumanta Basu, and Andrew G Clark. Dense time-course gene expression profiling of the drosophila melanogaster innate immune response. 2020.