Cross-Correlation Analyses

Caleb Lareau

March 21, 2017

Introduction

The type problem proposed is the modeling of the relationship between two time-dependent variables wherein a potential time "lag" may be observed between the correlation of these two variables. For example, one may expect the price of crude oil to be correlated with the currency conversation rate between the Saudi riyal and US dollar. Howevever, the best correlation between these variables is likely observed after some delay. One could imagine that the price of Saudi crude oil plummetting on Tuesday could may be correlated with a depleted riyal/USD conversion rate on Friday. Or vice-versa. For my final project, I propose to examine the statistical estimation procedure and utilities available in `R` to examine time-varying variables in the context of lag. In particular, I hope to explore the utility of these statistical forms in the context of single-cell biological data.

Statistics

For paired time-series data (X_t, Y_t), the cross-correlation is defined as:

$$\rho_{XY}(\tau) = \mathrm{E}[(X_t - \mu_X)\,(Y_{t+\tau} - \mu_Y)]/(\sigma_X \sigma_Y)$$

where the optimal lag, τ^*, can be easily computed as:

$$\tau^* = \arg\max_t \mathrm{abs}(\rho_{XY}(t))$$

While estimation of τ^* is fairly straightforward in the single pair case, I would also be interested in i pairs of time series– (X_t^i, Y_t^i) where we'd want to estimate a singular τ^* that optimizes the correlation over all these time series. Related to the example in the **Introduction**, X_t^i may be the price of crude oil in a given country i whereas Y_t^i may denote the country i's currency / USD conversion ratio. Estimating a singular τ^* may be of interest to determine *overall* what the time lag between foreign crude oil and foreign currency. One can quickly imagine building out a set of estimators for τ^*, which as far as I can tell is largely unexplored in the literature. Examples include computing τ_i^* per pair i and then taking the mean. Alternatively, one could compute τ^* from solving a global optimization problem over the $\rho_{XY}^i(t)$. I can go down this route if appropriate.

Intended Application

In cellular phenotype regulation, regions of DNA called "enhancers" selectively become unbound to histones to control gene expression. In other words, to define what makes a neuron different than a T-cell, variable regions of DNA (enhancers) become open/active, which boost the expression of certain genes and define the processes that define cell identify. Recent technological advancements have enabled us to determine which enhancers are active and genes expressed in any given cell. Additionally, computational methods allow for inference of "pseudotime" ordering of cells along a developmental trajectory. **Figure 1** provides a rough overview of this mechanism.

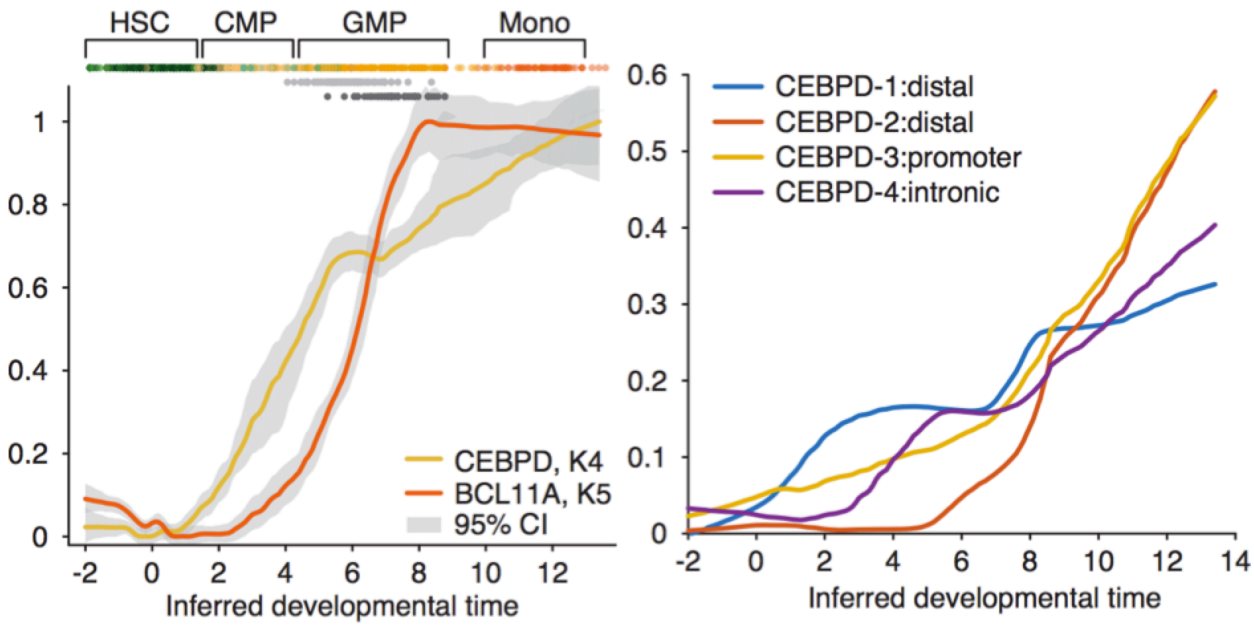

Figure 1: **A graphical overview of single-cell data.** In the plot to the left, a measure of total intensity (for simplicity, think gene expression) of two genes, including CEBPD. This plot shows that this gene is lowly expressed in stem cells (to the left), but highly expressed in monocytes (to the right). Each point represents a single cell where data is observed. In the plot on the right, four different enhancer activity scores are plotted for CEBPD, showing how enhancer activity is correlated, potentially with some lag, with gene expression. Curves represent smoothed values over the individual data point values measured in each cell.

If I wind up presenting this single cell data, the notion of "pseudotime" will be important for me to discuss. The idea is that given a set of single cells, we can use the covariance of their measured features to say how closely does each cell resemble one of two extremes (in this case, stem cells or monocytes). I won't discuss the specifics of these methods, but it suffices to consider that the x-axes in these plots as a longitudinal feature (t) of data. Using the notation above, X_t^j would be the enhancer activity score and Y_t^j would be the gene expression value for the j^{th} enhancer-gene pairing. Moreover, there are natural ways to group these pairings base on transcription factor binding motifs, which I won't discuss here, but give credence to estimating a single τ^* over multiple X, Y pairings. The result of estimating these different τ terms per transcription factor would shed insights into the kinetics of transcription factor activity in establishing differentiated cell states, which is a fundamental open question in systems biology.

As the time-lag effect is hard to conceptualize in **Figure 1**, I'll close this proposal showing a set of more simple examples in **Figure 2**. Here, we see a variable time-lag effect between the two panels. To estimate this lag, I applied the `ccf` function in `R`, shown in **Figure 3**.

I'm optimistic that my proposed topic would incorporate elements of longitudinal data analysis and correlated variables already discussed in class while providing insight into gene regulation and time-varying "lag." Some notable functions/packages that will be examined include the `forecast` (https://github.com/robjhyndman/forecast) package.

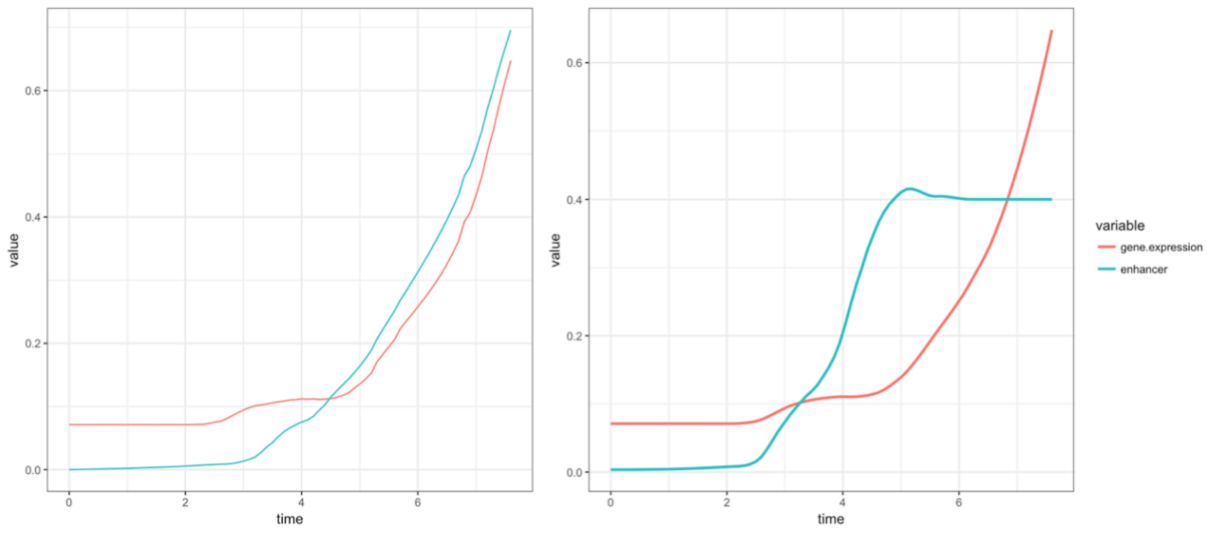

Figure 2: **To examples with varying enhancer-gene expression scores.** In the plot to the left, a gene-enhancer relationship shows a fast-activation. In the plot to the right, the enhancer becomes active but the regulatory process underlying the gene is relatively slow (i.e. a lag is observed).

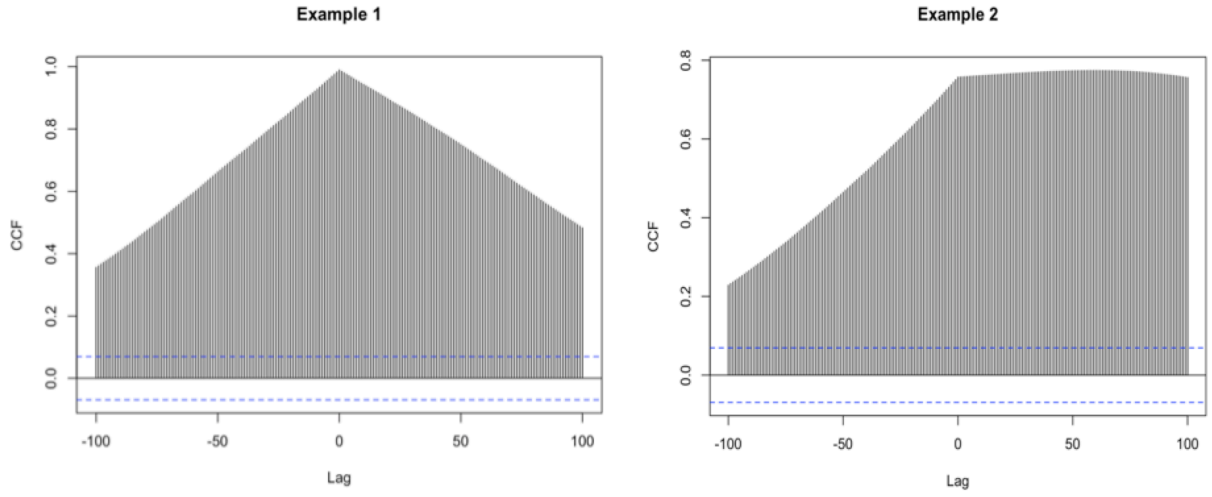

Figure 3: **Output plots from the `ccf` function as applied to data in Figure 2**. The plot on the left verifies that the tighest correlation between the variables in Figure 2 (left) occurs with no lag in the variables. However, the maximized correlation for the example on the right occurs with a lag of > 50 units.