Cross Correlation Analyses

Caleb Lareau

March 20, 2017

Introduction

The basic problem that is examined is the modeling of the relationship between two time-dependent variables wherein a potential "lag" may be observed. 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. In this example, the price of crude oil plummetting one day could may be correlated with a depleted riyal/USD conversion rate after another 2-3 days. 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. I will explore the utility of these stastical forms in the context of single-cell biological data.

Statistics

For continuous functions f and g, the cross-correlation is defined as:

$$(f \star g)(\tau) \stackrel{\text{def}}{=} \int_{-\infty}^{\infty} f(t)\ g(t+\tau)\, dt$$

where the optimal lag is defined as:

$$\tau_{\text{delay}} = \arg\max_t((f \star g)(t))$$

In the context of real data, f and g will not be continuous, but discrete functions, and the cross-correlation can similarly be computed.

Intended Application

In gene regulation, regions of DNA called "enhancers"

Figure 2

Links to the BST245 Course

- Longitudinal data analysis
- Correlated correlated variables

R Packages/functions discussed

- `ccf`
- `forecast` (https://github.com/robjhyndman/forecast)

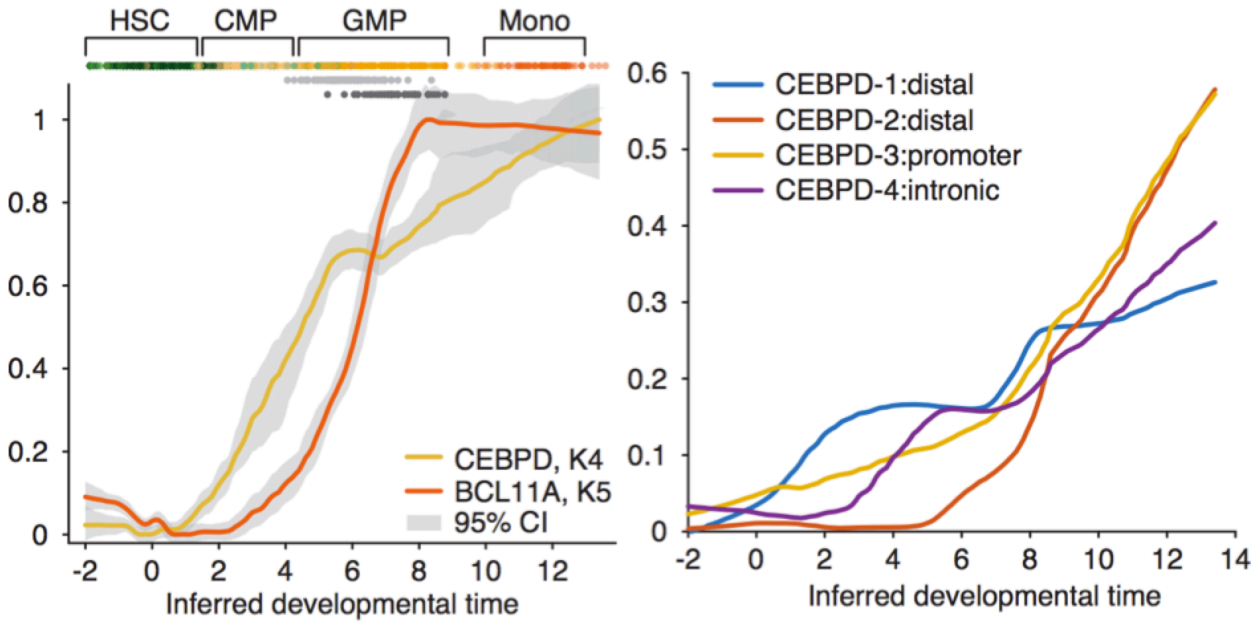

Figure 1: **A graphical overview of single-cell data.** In the plot to the left, a measure of total intensity (for simplicity, think gene expression) of two genes, including CEBPD. This plot shows that this gene is lowly expressed in stem cells (to the left), but highly expressed in monocytes (to the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

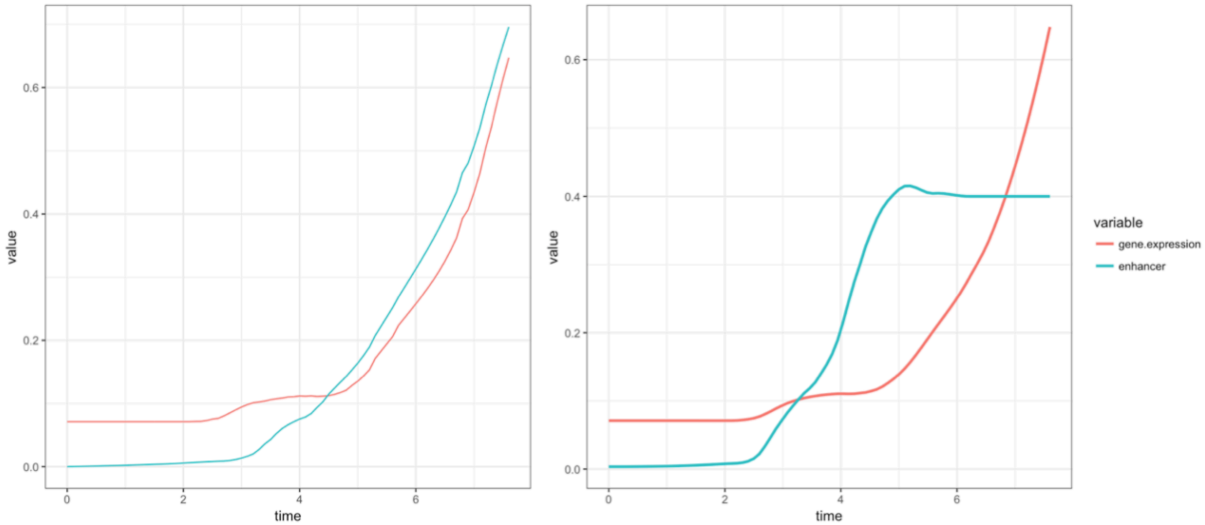

Figure 2: **To examples with varying enhancer-gene expression scores.** In the plot to the left, a gene-enhancer relationship shows a fast-activation. In the plot to the right, the enhancer becomes active but the regulatory process underlying the gene is relatively slow (i.e. a lag is observed.

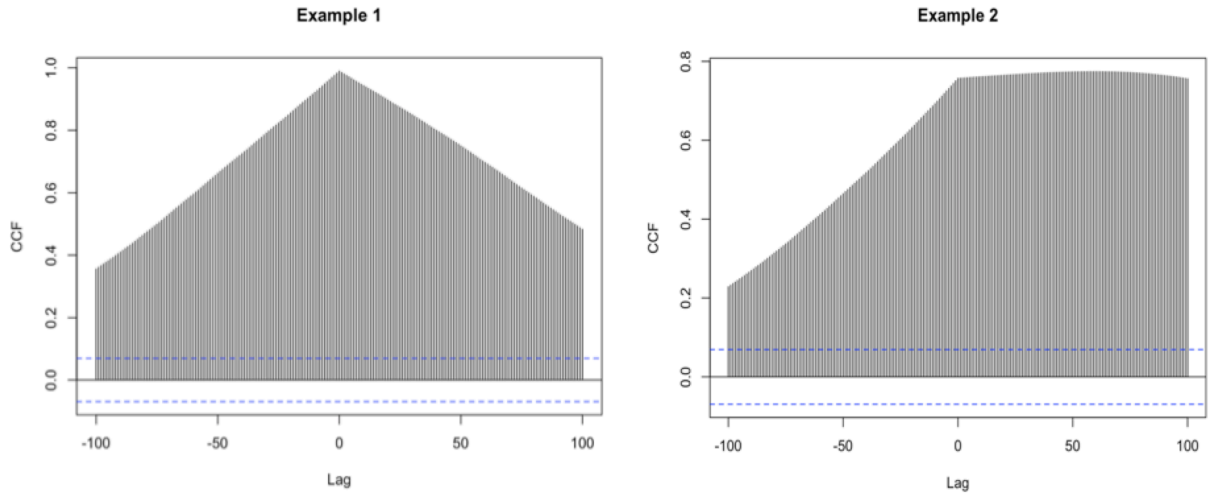

Figure 3: **Output plots from the `ccf` function**. The plot on the left verifies that the tighest correlation between the variables in Figure 2 (left) occurs with no lag in the variables. However, the maximized correlation for the example on the right occurs with a lag of > 50 units.