Abstract

Many regulation networks and control systems may be modeled using systems of ordinary differential equations. These systems are usually unknown, however it is possible to approximate them using multivariate regression. We will use sums of separable functions in the regression models. Through the analysis of these models one can then identify the contributions of individual components to the overall network. Here we will outline one such regression model using sums of separable functions and then develop techniques for analyzing these models. We will discuss and prove several results regarding the sources of error and the performance of these routines.

Learning and Analysis of Regulatory Networks with Sums of Separable Functions

Ryan T. Botts

Department of Mathematics
Ohio University

Introduction to gene regulatory networks

Genes contain information to perform functions, but do not actually perform them.

Genes code for mRNA, their physical expression, which go into cytoplasm and are converted into proteins to perform specific functions.

Question: What determines when and how the genes will be expressed?

Def: A gene regulatory network (GRN) is a networks of interacting genes which are each regulated by the other genes in the system as well as control parameters, e.g. metabolite levels.

Increasing the levels one gene may cause others to increase (promotion) or decrease (inhibition) or some combination of the two.

Represent d genes at time t as
$\mathbf{x(t)} = (x_1(t), x_2(t), \ldots x_d(t))$.

Represent k control parameters at time t as
$\mathbf{z(t)} = (z_1(t), z_2(t), \ldots z_k(t))$.

Only model the gene expression levels as a system of first order differential equations:

$$\dot{\mathbf{x}} = \mathbf{g}(\mathbf{x}, \mathbf{z}). \tag{1}$$

This is a deterministic system as knowing $\mathbf{g}$ would entirely determine how the genes interact.

Usually $\mathbf{g}$ is unknown.

Our model does not distinguish between control parameters and gene expression levels, so we write $\mathbf{g}(\mathbf{x})$.

Frequently one can perform microarray experiments to obtain data of the form $\{\mathbf{x}^j\}_{j=1}^N$ where $\mathbf{x}^j = (x_1(t_j), x_2(t_j), \ldots x_d(t_j))$.

Method

1. Use finite differences to approximate each $\mathbf{g}$

   $$\mathbf{g}(\mathbf{x}(t_j)) = \dot{\mathbf{x}}(t_j) \approx \Big(\mathbf{x}(t_{j+1}) - \mathbf{x}(t_{j-1})\Big) \,/\, \Big(t_{j+1} - t_{j-1}\Big) \,.$$

2. These give us a list of data $D = \{(\mathbf{x}^j, \mathbf{y}_j)\}_{j=1}^N$

3. Fit data using an alternating least squares (ALS) routine with sums of separable functions. (Will fit each component in $\mathbf{y}$ independently.)

4. Develop new techniques to identify these interactions from the approximation to $\mathbf{g}$.

Project Goals

- Identify the sources of error in the reg. model.
- Develop tools for identifying and measuring interactions from the regression model.
- Determine what effects the accuracy of these measures.
- Determine if these models may be used to predict steady states or any other future states.

Summary of Regression with Sums of Separable Functions

Define the pseudo-norm generated by the data-driven pseudo-inner product (extends to inner products over data)

$$\langle f, h \rangle = \frac{1}{N} \sum_{j=1}^{N} f(\mathbf{x}^j) h(\mathbf{x}^j) \,. \tag{2}$$

We will use **sums of separable functions** of the form

$$f(\mathbf{x}) = \sum_{l=1}^{r} s_l \prod_{i=1}^{d} f_i^l(x_i). \tag{3}$$

Currently we assume that $f_i^l(x_i) = \sum_{k=1}^{M} c_k^{i,l} \phi_k(x_i)$, for some orthogonal set $\{\phi_k\}_{k=1}^{M}$.

We want to minimize
$\|D - f\|^2 = \|D - \sum_{l=1}^{r} s_l \prod_{i=1}^{d} f_i^l(x_i)\|^2$.

Nonlinear optimization so we use an alternating least squares (ALS) approach. Alternate through each direction x_i independently.

Yields the normal equations $\mathbb{A}\mathbf{z} = \mathbf{b}$ in the first iteration. Where

$$A(k,l;k',l') = \frac{1}{N}\sum_{j=1}^{N}\left(\phi_k(x_1^j)p_j^l\right)\left(\phi_{k'}(x_1^j)p_j^{l'}\right),$$

and

$$b(k,l) = \frac{1}{N}\sum_{j=1}^{N} p_j^l \phi_k(x_1^j) y_j\,,$$

where $p_j^l = s_l \prod_{i=2}^{d} f_i^l(x_i^j)$.

Solve this and update f_1 and then proceed to next direction.

Not necessarily optimal solution, but after many iterations previous work shows that it is usually a good approximation.

Techniques for identifying interactions

First we must identify the properties of the g_i that would allow us to conclude that there is a significant interaction, i.e. strong interaction at one point or over a region.

Methods for measuring and identifying interactions, label the effect of the jth component on the ith component as $r_{i,j}$.

- $r_{i,j} = \max_{x \in [a,b]} F_{i,j}(x) - \min_{x \in [a,b]} F_{i,j}(x)$ where

$$F_{i,j}(x) = \frac{\int_{[a,b]} \cdots \int_{[a,b]} \left(f_i(\mathbf{x})\right)^2 dx_1 \ldots dx_{j-1} dx_{j+1} \ldots dx_d}{(b-a)^{d-1}}.$$

- $r_{i,j} = \max\limits_{k=1,\ldots N} \left(F_{i,j}(x_j^k)\right) - \min\limits_{k=1,\ldots N} \left(F_{i,j}(x_j^k)\right)$ where

$$F_{i,j}(x) = \frac{1}{N} \sum_{k=1}^{N} \left(f_i(x_1^k, \ldots, x_{j-1}^k, x, x_{j+1}^k, \ldots, x_d^k)\right)^2 .$$

- Other partially discrete methods.

- Some combination $r_{i,j} = \dfrac{1}{N} \displaystyle\sum_{k=1}^{N} \frac{\partial}{\partial x_j} f(\mathbf{x})$ and $r_{i,j} = \frac{1}{N} \sum_{k=1}^{N} (\frac{\partial}{\partial x_j} f(\mathbf{x}))^2$.

Current Work

- Identify how the methods for identifying interactions would behave were the system known, e.g. can we distinguish between primary and secondary interactions.

- Identify how these methods would behave when we have approximations to the true model.

- Distinguish between the sources of error in the model resulting from:

- The structure of the functions we are approximating with.
- Only knowing values of the function at data points.

- Identify the sources of error in the regression model.
  - Identify the role of the rank and the number of ϕ_i.
  - If each individual g is rank 1 then the ALS will identify a unique f.

  - The same cannot be said if g is rank 2, but perhaps under certain conditions it might.

- Compare the results of this method to other methods on both real and simulated data.

- Test how well the approximations might be used to predict future states of the system.