Abstract

The recovery of large-scale gene networks from microarray data has been the subject of many papers in recent years. One of the primary difficulties in this problem is the size of the networks, coupled with the relative scarcity of observations. Using a type of high dimension nonparametric regression we hope to provide an algorithm which will not only allow us to uncover these networks on a global scale, but might also provide some insight into the dynamics of the system. We will demonstrate the power of this approach on several bench mark in silico data sets with added noise.

Journal of Ryan Botts Spring 2009: Reverse Engineering of Gene Regulatory Functions with Sums of Separable Functions

Ryan T. Botts

Department of Mathematics
Ohio University

1 April - 8 April: Identifying Swamps

We wish to discover and document the location of "swamps", that is regions where the gradient of $\|f-g\|^2$ is very near 0.

We consider the case where we have a rank 2 target of the form $g = g_1 + g_2 = \prod_{i=1}^{d} [0,1] + A \prod_{i=1}^{d} [1,0]$ and we fit with a rank 1 of the form $f = \prod_{i=1}^{d} [c,e]$.

If $A = 1$ then we would guess that the f which minimizes $\|f-g\|^2$ would occur when $f = g_1$ or $f = g_2$. As A increases we would expect the minimum to be when $f = g_2$ and a local minima to occur when $f = g_1$.

Surprising Results

In the dimension 2 case with $A = 1$ we may solve to find that there are infinitely many minima $c^2 + e^2 = 1$.

We would think that this problem would be one of finding $[c, e]$ which is some optimal compromise between $[0, 1]$ and $[1, 0]$, and due to the symmetry one might assume that the same $[c, e]$ would produce an optimal fit in other dimensions.

THIS IS NOT THE CASE! For example $f = \prod_{i=1}^{2} [\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}]$ minimizes the square error.

However, in the $d = 3$ case $f = \prod_{i=1}^{3} [\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}]$ is not even a local minima.

What does happen?

The following animations show that as A increases it appears that the minima occurs where expected, however we also observe other local minima.

Notice the difference between in symmetry between the different degrees.

A plot of $\|f - g\|^2$ in the case where $g = \prod_{i=1}^{2} [0, 1] + A \prod_{i=1}^{2} [1, 0]$ and $f = \prod_{i=1}^{2} [c, e]$.

A plot of $\|f - g\|^2$ in the case where

$$g = \prod_{i=1}^{3} [0,1] + A \prod_{i=1}^{3} [1,0] \text{ and } f = \prod_{i=1}^{3} [c,e].$$

A plot of $\|f - g\|^2$ in the case where

$$g = \prod_{i=1}^{8} [0,1] + A \prod_{i=1}^{8} [1,0] \text{ and } f = \prod_{i=1}^{8} [c_1, c_2].$$

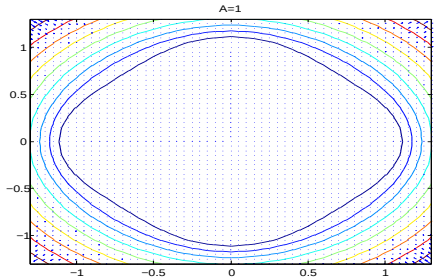
Approximating a rank 2 target with a rank 1

We did confirm the statement from last week that if $f = \prod_{i=1}^{2} [\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}]$ minimizes the square error $g = \prod_{i=1}^{2} [0,1] + \prod_{i=1}^{2} [1,0]$, but does not minimize the least squares error if $d = 3$.

Also worked out the minimum for fitting something of dimension $d > 2$ of the form $g = g_1 + g_2 = \prod_{i=1}^{d} [0,1] + A \prod_{i=1}^{d} [1,0]$ with a symmetric rank 1 tensor of the form $f = \lambda \prod_{i=1}^{d} [c_1, c_2]$, where

$[c_1, c_2]$ is normalized. The minimum is when $f = g_2$, with a local minimum when $f = g_1$ and has a local maximum when $f = \lambda \prod_{i=1}^{d} [c_1, c_2]$ where

$$c_1 = \left(\frac{A^{\frac{2}{d-2}}}{1+A^{\frac{2}{d-2}}} \right)^{\frac{1}{2}}, \; c_2 = \left(\frac{1}{1+A^{\frac{2}{d-2}}} \right)^{\frac{1}{2}} \text{and } \lambda = c_1^d + c_2^d.$$


A=1
1
0.5
0
−0.5
−1
−1
−0.5
0
0.5
1

What is required of interaction identification techniques

Began asking what properties the interaction identification techniques should have.

First we see that if $\dot{x}_i = g_i(\mathbf{x})$ is independent of some x_j then $\frac{\partial \dot{x}}{x_j} = 0$.

This is true if and only if

$$r_{i,j} = \frac{1}{N} \sum_{k=1}^{N} (\frac{\partial}{\partial x_j} f(\mathbf{x}))^2 = 0\,. \tag{1}$$

However if

$r_{i,j} = \max_{x \in [a,b]} F_{i,j}(x) - \min_{x \in [a,b]} F_{i,j}(x) = 0$ where

$$F_{i,j}(x) = \frac{1}{(b-a)^{d-1}} \tag{2}$$

$$\int_{[a,b]} \int_{[a,b]} \dots \int_{[a,b]} (f_i(\mathbf{x}))^2 \, dx_1 \dots dx_{j-1} dx_{j+1} \dots dx_d \, . \tag{3}$$

then it is only necessary that $F_{i,j}(x)$ is constant, but it does not necessarily mean that $\frac{\partial \dot{x}}{\partial x_j} = 0$.

Symmetric minima

We now wish to consider whether there might be any asymmetric local minima in the problem of fitting a symmetric rank 2 tensor with a rank 1 tensor.

Proposition 1 *Let* $g = A \prod_{i=1}^{d} [a_1, a_2] + \prod_{i=1}^{d} [0, a_2]$ *and let* $f = \prod_{i=1}^{d} [c_1^{(i)}, c_2^{(i)}]$. *If* f *is a local minima of* $\|f - g\|$ *then f is symmetric.*

If $a_2 = 0$ then by previous work we may conclude that the only minima to this problem occur when f is exactly one of the summands in g and the true minima is when f is exactly the larger of the two summands.

Derivatives of dependent variables

We wish to detect the difference between secondary and primary interactions.

Our system has the form

$$\dot{x}_i = g_i(x_1, \ldots, x_d) \qquad \text{for } i = 1, 2, \ldots d, \tag{4}$$

where the x_i are implicit functions of time and hence dependent variables.

The total derivative of a multivariate function of dependent variables is defined to be

$$\frac{dg}{dt}(x_1, \ldots, x_d) = \sum_{j=1}^{d} \frac{\partial g}{\partial x_j} \dot{x}_j \tag{5}$$

$$= \nabla g(\mathbf{x}) \cdot \dot{\mathbf{x}}\,. \tag{6}$$

Select an arbitrary g_i and call it g. Using this we find that a Taylor expansion of g about t_0 gives

$$g\left(\mathbf{x}(t_0 + h)\right) = g(\mathbf{x}(t_0)) + h\nabla g \cdot \dot{\mathbf{x}}(t_0) + \frac{h^2}{2}\frac{d}{dt}\left(\nabla g \cdot \dot{\mathbf{x}}(t_0)\right) + \mathcal{O}(h^3)$$

We hope this might lead to some way to identify secondary interactions.

Direct and indirect interactions

What we can learn if we knew the true regulatory function. Set $g = g_i$ for a specific component i.

Writing $\mathbf{x_h} = \mathbf{x}(t_0 + h)$ and $\mathbf{x_0} = \mathbf{x}(t_0)$ the Taylor expansion of g is

$$\begin{aligned} g(\mathbf{x_h}) &= g(\mathbf{x_0}) + h\frac{d}{dt}g(\mathbf{x_0}) + \frac{1}{2}h^2\frac{d^2}{dt^2}g(\mathbf{x_0}) + \mathcal{O}(h^3) \\ &\approx g(\mathbf{x_0}) + h\nabla g(\mathbf{x}) \cdot \dot{\mathbf{x}}_\mathbf{0} + \frac{1}{2}h^2\left(\dot{\mathbf{x}}_\mathbf{0}^t H_g \dot{\mathbf{x}}_\mathbf{0} + \nabla g(\mathbf{x_0}) \cdot J_\mathbf{g}\dot{\mathbf{x}}_\mathbf{0}\right) \end{aligned}$$

All secondary interactions and mixed effects are in the second term and primary interactions in the first.

Using finite differences we may find the sum of all primary interactions:

$$\frac{g(\mathbf{x_h}) - g(\mathbf{x_0})}{2h} \approx \nabla g(\mathbf{x_0}) \cdot \dot{\mathbf{x}}_0$$

The sum of the mixed effects and secondary interactions may be approximated using

$$\frac{g(\mathbf{x_h}) + g(\mathbf{x_{-h}}) - 2g(\mathbf{x_0})}{h^2} \approx \dot{\mathbf{x}}_{\mathbf{0}}^t H_g \dot{\mathbf{x}}_{\mathbf{0}} + \nabla g(\mathbf{x_0}) \cdot J_{\mathbf{g}} \dot{\mathbf{x}}_{\mathbf{0}}$$

Cannot recover H_g and $J_{\mathbf{g}}$ from only the finite differences.

$$\|\mathbf{f} - \mathbf{g}\|_{1,2}{}^2 = \|\mathbf{f} - \mathbf{g}\|_2{}^2 + \|J(\mathbf{f}) - J(\mathbf{g})\|_2{}^2$$

Fitting the Jacobian

We see that

$$\frac{\mathbf{x}_{k+1} - 2\mathbf{x}_k + \mathbf{x}_{k-1}}{h^2} \approx \ddot{\mathbf{x}} = \frac{d}{dt}\mathbf{g}(\mathbf{x}) = J(\mathbf{g})\dot{\mathbf{x}}\,. \qquad (7)$$

Pick a direction to fit call it $g = g_i$, and similarly for f, we attempt to minimize

$$\|f - g\|_{1,2}{}^2 = \|f - g\|_2{}^2 + \|\nabla f \cdot \dot{\mathbf{x}} - \nabla g \cdot \dot{\mathbf{x}}\|_2{}^2\,. \qquad (8)$$

Suppose we have a collection of data $\dot{\mathcal{X}} = \{(\mathbf{x}^j, \mathbf{y}^j, \mathbf{z}^j)\}_{j=1}^N$ where $\mathbf{y}^j \approx \mathbf{g}(\mathbf{x}^j)$ and $\mathbf{z}^j \approx J(\mathbf{g}(\mathbf{x}^j))\dot{\mathbf{x}}^j$

Recall that to minimize the first term in the ALS we had the normal equations $\mathbb{A}_1\mathbf{z} = \mathbf{b_1}$

$$b_1(k,l) = \frac{1}{N}\sum_{j=1}^{N} p_j^l \phi_k(x_1^j) y_j$$

$$A_1(k,l;k',l') = \frac{1}{N}\sum_{j=1}^{N} \left(\phi_k(x_1^j)p_j^l\right)\left(\phi_{k'}(x_1^j)p_j^{l'}\right) .$$

with $p_j^l = s_l \prod_{i=2}^{d} f_i^l(x_i^j)$. Setting $c_k^l = z(k,l)$ minimizes the error in the first direction.

Minimizing the Jacobian Error Term

Again pick a direction to minimize, say the first.

Set

$$q_{i,j}^{l} = s_l y^j \frac{\partial}{\partial x_i} \left(\prod_{k=2}^{d} f_k^l(x_k^i) \right) , \tag{9}$$

where $q_{1,j}^{l} = s_l y^j \prod_{k=2}^{d} f_k^l(x_k^j)$.

Obtain normal equations $\mathbb{A}_2\mathbf{z} = \mathbf{b}_2$ where

$$b_2(k,l) = \frac{1}{N}\sum_{j=1}^{N} z^j \left(q_{1,j}^l \phi_k'(x_1^j) + \sum_{i=2}^{d} q_{i,j}^l \phi_k(x_1^j) \right)$$

$$A_2(k,l;k',l') = \frac{1}{N}\sum_{j=1}^{N} \left(q_{1,j}^l \phi_k'(x_1^j) + \sum_{i=2}^{d} q_{i,j}^l \phi_k(x_1^j) \right) \cdot \left(q_{1,j}^{l'} \phi_{k'}'(x_1^j) + \sum_{i=2}^{d} q_{i,j}^{l'} \phi_{k'}(x_1^j) \right).$$

New Normal Equations

We may now solve

$$(\mathbb{A}_1 + \mathbb{A}_2)\,\mathbf{z} = \mathbf{b}_1 + \mathbf{b}_2 \tag{10}$$

repeat this process in each direction several times to minimize

$$\|f - g\|_*{}^2 = \|f - g\|_2{}^2 + \|\nabla f \cdot \dot{\mathbf{x}} - \nabla g \cdot \dot{\mathbf{x}}\|_2{}^2 \,. \tag{11}$$

We may also weight these terms unequally by minimizing

$$\|f - g\|_*{}^2 = \lambda_1 \|f - g\|_2{}^2 + \lambda_2 \|\nabla f \cdot \dot{\mathbf{x}} - \nabla g \cdot \dot{\mathbf{x}}\|_2{}^2 \tag{12}$$

Accomplish this by solving the normal equations

$$(\lambda_1 \mathbb{A}_1 + \lambda_2 \mathbb{A}_2)\,\mathbf{z} = \mathbf{b}_1 + \mathbf{b}_2 \tag{13}$$