Network Reconstruction From High Dimensional Ordinary Differential Equations

Tianyu Guan

Department of Statistics and Actuarial Science
Simon Fraser University

October 27, 2016
Outline

1. Motivation
2. Literature Review
3. Proposed Approach
4. Theoretical Properties
5. Numerical Experiments
6. Applications
Motivation

The problem:
- Solve problems such as estimating a gene regulatory network from gene expression data measured at discrete time points
- Learn a dynamical system from high-dimensional time-course data

Existing method:
- Estimate the derivatives from noisy observations
- This is challenging and inefficient

Proposed Approach:
- Does not involve derivative estimation
- Can consistently recover the true network structure even in high dimensions
- Has an empirical improvement over competing approaches
A system of ODEs takes the form

\[ X'(t; \theta) \equiv \begin{bmatrix} \frac{dX_1(t; \theta)}{dt} \\ \vdots \\ \frac{dX_p(t; \theta)}{dt} \end{bmatrix} = \begin{bmatrix} f_1(X(t; \theta), \theta) \\ \vdots \\ f_p(X(t; \theta), \theta) \end{bmatrix} \equiv f(X(t; \theta), \theta); \quad t \in [0, 1], \]  

(1)

- \( X(t, \theta) = (X_1(t, \theta), \ldots, X_p(t, \theta))^T \): a set of variables
- \( f = (f_1, \ldots, f_p)^T \): the form maybe known or unknown
- Typically there is also an initial condition of the form \( X(0; \theta) = C \)
- The system is often observed on discrete time points subject to measurement errors
Let $Y_i \in R^p$ be the measurement of the system at time $t_i$ such that

$$Y_i = X(t_i; \theta) + \epsilon_i, \quad i = 1, \ldots, n \quad (2)$$

- $\theta^*$: the true set of parameter values
- $\epsilon_i$: independent measurement errors
Methods that assume a known form of \( f \)
- Gold standard approach
- Two-step collocation methods
- The generalized profiling method

Methods that do not assume the form of \( f \)
Assume $0 = t_1 < t_2 < \ldots < t_n = 1$ and $t_i$ is the $i$th time point

$Y_{ij}$: the observation of the $j$th variable at $t_i$

$\chi(h)$: a nonparametric class of functions on $[0, 1]$

$h$: some smoothing parameter(s)

$Z(\cdot)$: an arbitrary function belonging to $\chi(\cdot)$

$\| \cdot \|_2$: the $l_2$ - norm of a vector or a matrix

$\| f \|$: the $l_2$ - norm of a function $f$ on $[0, 1]$

\[
e \text{g.e. } \| f \|^2 = \int_0^1 f^2(t)dt
\]

Asterisk is used to denote true values, e.g. $\theta^*$ denotes the true value of $\theta$

$\Lambda_{\text{max}}(A)$ and $\Lambda_{\text{min}}(A)$: the maximum and minimum eigenvalues of square matrix $A$
Methods that assume a known form of \( f \)

**Gold standard approach:**

Benson (1979) and Biegler et al. (1986) proposed to estimate the unknown parameter \( \theta^* \) in \( Y_i = X(t_i; \theta) + \epsilon_i \) by solving

\[
\hat{\theta}^{\text{gold}} = \arg \min_{\theta} \sum_{i=1}^{n} \|Y_i - X(t_i; \theta)\|_2^2
\]

subject to \( X'(t; \theta) = f(X(t; \theta), \theta), \quad t \in [0, 1] \).

- \( X(\cdot; \theta) \): a fixed function given \( \theta \), analytic expression may not be available.
Methods that assume a known form of $f$

Gold standard approach:

Advantages of this method: $\hat{\theta}^{\text{gold}}$ has appealing theoretical properties
  - When $\epsilon_i$ are Gaussian, $\hat{\theta}^{\text{gold}}$ is the MLE
  - When $\epsilon_i$ are Gaussian, $\hat{\theta}^{\text{gold}} \sqrt{n}$ - consistent

Drawbacks of this method:
  - Computationally challenging
Two-step collocation methods:

First proposed by Varah (1982):

Step 1. Fit $\hat{X}(\cdot; \theta)$ to observations $Y_1, \ldots, Y_n$

Step 2. Plug $\hat{X}(\cdot; \theta)$ and $\hat{X}'(\cdot; \theta)$ into ODEs and estimate $\theta$

This amounts to solving the optimization problem

$$\hat{\theta}^{TS} = \arg\min_{\theta} \int_0^1 \|\hat{X}'(t; h) - f(\hat{X}(t; h), \theta)\|_2^2 \, dt,$$

(4a)

where

$$\hat{X}(\cdot; h) = \arg\min_{Z(\cdot) \in X(h)} \sum_{i=1}^n \|Y_i - Z(t_i)\|_2^2.$$

(4b)
Methods that assume a known form of $f$

Two-step collocation methods:

First proposed by Varah (1982):

*Step 1.* Fit $\hat{X}(:,\theta)$ to observations $Y_1, \ldots, Y_n$

*Step 2.* Plug $\hat{X}(:,\theta)$ and $\hat{X}'(:,\theta)$ into ODEs and estimate $\theta$

This amounts to solving the optimization problem

$$\hat{\theta}^{TS} = \arg\min_\theta \int_0^1 \|\hat{X}'(t; h) - f(\hat{X}(t; h), \theta)\|_2^2 \, dt,$$

(4a)

where

$$\hat{X}(::; h) = \arg\min_{Z(\cdot) \in \mathcal{X}(h)} \sum_{i=1}^n \|Y_i - Z(t_i)\|_2^2.$$

(4b)
Methods that assume a known form of $f$

**Two-step collocation methods:**

Advantages of this method:
- Estimate $\theta$ and $X$ separately
- It is computationally efficient

Drawbacks of this method:
- Estimating derivatives from noisy observations is challenging and inefficient
- The properties of $\hat{\theta}^{TS}$ rely heavily on $\hat{X}(\cdot; h)$
- $\sqrt{n}$-consistency has only been shown for certain values of the smoothing parameter $h$ that are hard to choose in practice
Methods that assume a known form of \( f \)

**Two-step collocation methods:**

Dattner and Klaassen (2015) proposed an improvement to (4) for a special case of (1). Assume \( f_i(X(t), \theta) \) in (1) is a linear function of \( \theta \), which leads to

\[
\begin{bmatrix}
\frac{dX_1(t)}{dt} \\
\vdots \\
\frac{dX_p(t)}{dt}
\end{bmatrix} =
\begin{bmatrix}
g_1^T(X(t))\theta \\
\vdots \\
g_p^T(X(t))\theta
\end{bmatrix} \equiv g(X(t))\theta; \quad t \in [0, 1],
\]

where \( g(X(t)) \) is a known function of \( X(t) \). Integrating both sides of (5) gives

\[
X(t) = \left\{ \int_0^t g(X(u)) \, du \right\} \theta + C,
\]

(6)
Methods that assume a known form of $f$

Two-step collocation methods:

where $C \equiv X(0; \theta)$. The unknown parameter $\theta^*$ is estimated by solving

$$
\hat{\theta}^{LM} = \arg \min_{\theta} \int_0^1 \left\| \hat{X}(t; h) - \left\{ \int_0^t g(\hat{X}(u; h)) \, du \right\} \theta - C \right\|^2 dt,
$$

where

$$
\hat{X}(\cdot; h) = \arg \min_{Z(\cdot) \in X(h)} \sum_{i=1}^n \| Y_i - Z(t_i) \|^2_2.
$$

- This approach requires an estimate of the integral rather than the derivative
- This has profound effects on the asymptotic behaviour of the estimator $\hat{\theta}^{LM}$
- $\sqrt{n}$-consistency of $\hat{\theta}^{LM}$ has been established under mild conditions
- The choice of smoothing parameter $h$ is less crucial
Methods that assume a known form of $f$

Two-step collocation methods:

Brunel et al. (2014) and Hall and Ma (2014) adopt the idea to provide alternatives of (4a)

$$\int_0^1 f(X(t), \theta)\phi(t)dt + \int_0^1 X(t)\phi'(t)dt = 0.$$ 

to provide alternatives of (4a)

$$\hat{\theta}^V = \arg\min_{\theta} \frac{1}{L} \sum_{l=1}^L \left\| \int_0^1 f(\hat{X}(t; h), \theta)\phi_l(t)dt + \int_0^1 \hat{X}(t; h)\phi'_l(t)dt \right\|_2^2,$$
Methods that assume a known form of $f$

The generalized profiling method:

Ramsay et al. (2007) proposed the following method:

$$\hat{\theta}_{\lambda}^{GP} = \arg \min_{\theta} \sum_{i=1}^{n} \| Y_i - \hat{X}(t_i; h, \theta) \|_2^2,$$

where

$$\hat{X}(\cdot; h, \theta) = \arg \min_{Z(\cdot) \in \mathcal{X}(h)} \frac{1}{n} \sum_{i=1}^{n} \| Y_i - Z(t_i) \|_2^2 + \lambda \int_{0}^{1} \| Z'(t) - f(Z(t), \theta) \|_2^2 \, dt.$$  

(10)
Methods that do not assume the form of $f$

Henderson and Michailidis, 2014; Wu et al., 2014 assume the right-hand side of (1) is additive

$$X_j'(t) = \theta_{j0} + \sum_{k=1}^{p} f_{jk}(X_k(t)), \quad \theta_{j0} \in \mathbb{R}. \quad (11)$$

Approximate the unknown $f_{jk}$ with a truncated basis expansion

$$f_{jk}(a_k) = \psi(a_k)^T \theta_{jk} + \delta_{jk}(a_k), \quad \theta_{jk} \in \mathbb{R}^M, \quad (12)$$

The additive ODEs can be written as

$$X_j'(t) = \theta_{j0} + \sum_{k=1}^{p} \psi(X_k(t))^T \theta_{jk} + \sum_{k=1}^{p} \delta_{jk}(X_k(t)), \quad j = 1, \ldots, p. \quad (13)$$

where $\delta_{jk}(a_k)$ denotes the residual
Methods that do not assume the form of $f$

they propose to solve optimization problems of the form

\[
\hat{\theta}_j^{NP} = \arg \min_{\theta_0 \in \mathbb{R}, \theta_j \in \mathbb{R}^M} \int_0^1 \left\| \hat{X}_j'(t; h) - \theta_0 - \sum_{k=1}^{p} \psi \left( \hat{X}_k(t; h) \right)^T \theta_j \right\|^2 dt \\
+ \lambda_n \sum_{k=1}^{p} \left[ \int_0^1 \left\{ \psi \left( \hat{X}_k(t; h) \right)^T \theta_j \right\}^2 dt \right]^{1/2},
\]

(14a)

for $j = 1, \ldots, p$, where

\[
\hat{X}(\cdot; h) = \arg \min_{Z(\cdot) \in \mathcal{X}(h)} \sum_{i=1}^{n} \| Y_i - Z(t_i) \|^2_2.
\]

(14b)
Use a finite basis $\psi(\cdot)$ to approximate the additive components $f_{jk}$

$$f_{jk}(a_k) = \psi(a_k)^T \theta_{jk} + \delta_{jk}(a_k), \quad \theta_{jk} \in \mathbb{R}^M$$

- $\psi(x) = (\psi_1(x), \ldots, \psi_M(x))^T$ is a finite basis
- $\delta_{jk}(a_k)$ is residual

Review of equation (13)

$$X_j'(t) = \theta_{j0} + \sum_{k=1}^{p} \psi(X_k(t))^T \theta_{jk} + \sum_{k=1}^{p} \delta_{jk}(X_k(t)), \quad j = 1, \ldots, p.$$  

Integrate equation (13) yields

$$X_j(t) = X_j(0) + \theta_{j0} t + \sum_{k=1}^{p} \Psi_k(t)^T \theta_{jk} + \sum_{k=1}^{p} \int_{0}^{t} \delta_{jk}(X_k(u)) \, du,$$

(15)

- $\Psi_k(t) = (\Psi_{k1}(t), \ldots, \Psi_{kM}(t)) = \int_{0}^{t} \psi(X_k(u)) du, \quad k = 1, \ldots, p$
- $\Psi_0(t) = t$
Their method is called *Graph Reconstruction via Additive Differential Equations* (GRADE) which solves the following problem for $j = 1, \ldots, p$

\[
\hat{\theta}_j = \arg \min_{C_{j0} \in \mathbb{R}, \theta_{j0} \in \mathbb{R}, \theta_{j1}, \ldots, \theta_{jp} \in \mathbb{R}^M} \frac{1}{2n} \sum_{i=1}^{n} \left\{ Y_{ij} - C_{j0} - \theta_{j0} \hat{\Psi}_0 (t_i) - \sum_{k=1}^{p} \theta_{jk}^T \hat{\Psi}_k (t_i) \right\}^2 \\
+ \lambda_{n,j} \sum_{k=1}^{p} \left[ \frac{1}{n} \sum_{i=1}^{n} \left\{ \theta_{jk}^T \hat{\Psi}_k (t_i) \right\}^2 \right]^{1/2},
\]

(17a)

where

\[
\hat{X} (\cdot; h) = \arg \min_{Z(\cdot) \in \mathcal{X}(h)} \sum_{i=1}^{n} \| Y_i - Z(t_i) \|^2_2,
\]

(17b)

and

\[
\hat{\Psi}_0 (t) = t; \quad \hat{\Psi}_k (t) = \int_{0}^{t} \psi \left( \hat{X}_k (u; h) \right) \, du, \quad k = 1, \ldots, p.
\]

(17c)
- \( \lambda_{n,j} \) is a non-negative sparsity-inducing tuning parameter.

- If the true function \( f_{jk}^* \) is non-zero, we say that the \( k \)th variable \( X^*_j \) is a true regulator of \( X^* \).

- Let \( S_j \equiv k : \|f_{jk}^*\|_2 \neq 0, k = 1, \ldots, p \) denote the set of true regulators.

- The estimated index set of regulators be \( \hat{S}_j \equiv k : \|\hat{\theta}_{jk}\|_2 \neq 0, k = 1, \ldots, p \).

- We then reconstruct the network using \( \hat{S}_j, j = 1, \ldots, p \).
In some studies, time-course data is collected from multiple samples, or experiments

- Let $R$ denote the total number of experiments
- $Y^{(r)}$ be the observations in the $r$th experiment
- Assume the same ODE system (13) applies across all experiments with the same true parameter $\theta^*_j$
- Allow a different set of initial values for each experiment
Assume that each experiment consists of measurements on the same set of time points. This leads us to modify (17) as follows:

\[
\hat{\theta}_j = \arg\min_{C^{(r)}_0 \in \mathbb{R}, \theta_0 \in \mathbb{R}, \theta_j_1, \ldots, \theta_j_p \in \mathbb{R}^M} \frac{1}{2Rn} \sum_{r=1}^{R} \sum_{i=1}^{n} \left\{ Y_{ij}^{(r)} - C^{(r)}_{j0} - \theta_0 \hat{\Psi}_0(t_i) - \sum_{k=1}^{p} \theta_{jk} \hat{\Psi}_k^{(r)}(t_i) \right\}^2 \\
+ \lambda_n \sum_{k=1}^{p} \left[ \frac{1}{Rn} \sum_{r=1}^{R} \sum_{i=1}^{n} \left\{ \theta_{jk}^T \hat{\Psi}_k^{(r)}(t_i) \right\}^2 \right]^{1/2},
\]

where

\[
\hat{X}^{(r)}(\cdot; h) = \arg\min_{Z(\cdot) \in \mathcal{X}(h)} \sum_{i=1}^{n} \| Y_{i}^{(r)} - Z(t_i) \|_2^2, \ r = 1, \ldots, R, \\
\hat{\Psi}_0(t) = t; \ \hat{\Psi}_k^{(r)}(t) = \int_{0}^{t} \psi(\hat{X}_k^{(r)}(u; h)) \, du, \ k = 1, \ldots, p.
\]
Remark 1. To facilitate the comparison of GRADE (17) with other methods, we introduce an intermediate variable,

\[
\tilde{X}_j(t; h, \theta) \equiv C_{j0} + \theta_{j0} t + \sum_{k=1}^{p} \theta_{jk}^T \Psi_k(t),
\]

(19)

- In gold standard (3), the ODE system (1) is strictly satisfied due to the constraint in (3b)
- In two-step procedure the smoothing estimate \( \hat{X}(\cdot; h) \) does not satisfy (1)
- GRADE stands in between:
  - \( \hat{X}(\cdot; h) \) is solely based on the observations, while
  - intermediate estimate \( \tilde{X}(\cdot; h) \) is calculated by plugging \( \hat{X}(\cdot; h) \) into the additive ODE (13)
Proposition 1.

Suppose that Assumptions 1–6 and S1–S4 in the supplementary material hold, and let

\[ h_n \propto n^{-1/(2\beta_1+1)}, \quad M \propto n^{\frac{1}{2\beta_2+1}} \left( \frac{2\beta_1}{2\beta_1+1} - \alpha \right), \quad \text{and} \quad \lambda_n \propto n^{-\frac{2\beta_2-1}{4\beta_2+2}} \left( \frac{2\beta_1}{2\beta_1+1} - \alpha \right) + 2\gamma, \]

where \( 0 < \alpha < \frac{2\beta_1}{2\beta_1+1} \), \( 0 < \gamma < H_2(\beta_1, \beta_2, \alpha) \), and \( H_2(\beta_1, \beta_2, \alpha) \) is a constant that depends only on \( \beta_1, \beta_2 \) and \( \alpha \). Then as \( n \) increases, the proposed procedure (17) correctly recovers the true graph, i.e., \( \hat{S}_j = S_j \) for all \( j = 1, \ldots, p \), with probability converging to 1, if \( s = O(n^\gamma) \) and \( p \exp \left( -C_4 n^{\alpha}/\sigma^2 \right) = o(1) \) for some constant \( C_4 \).
Numerical Experiments

In this simulation, we compare GRADE with NeRDS (Henderson and Michailidis, 2014) and SA-ODE (Wu et al., 2014) described in (14). We consider the following system of additive ODEs, for $k = 1, \ldots, 5$:

\[
\begin{align*}
X'_{2k-1}(t) &= \theta_{2k-1,0} + \psi(X_{2k-1}(t))^T \theta_{2k-1,2k-1} + \psi(X_{2k}(t))^T \theta_{2k-1,2k}, \\
X'_{2k}(t) &= \theta_{2k,0} + \psi(X_{2k-1}(t))^T \theta_{2k,2k-1} + \psi(X_{2k}(t))^T \theta_{2k,2k}
\end{align*}
\]

where $\psi(x) = (x, x^2, x^3)^T$ is the cubic monomial basis. The parameters and initial conditions are chosen so that the solution trajectories are identifiable under an additive model (Buja et al., 1989). See supplementary material for details.
Figure 1: Performance of network recovery methods on the system of additive ODEs in (26), averaged over 400 simulations. The four curves represent SA-ODE (---), NeRDS (--), and GRADE without (—) and with (—) the additional smoothing penalty in (17a) used by NeRDS. Each point on the curves corresponds to average performance for a given sparsity tuning parameter $\lambda_n$ in (14a) or (17a). The symbols indicate the sparsity tuning parameter $\lambda_n$ selected using BIC (SA-ODE, •, and GRADE, • and *) or gcv (NeRDS, *).
Applications

Application to *in silico* gene expression data

- GeneNetWeaver (GNW) provides an *in silico* benchmark for assessing the performance of network recovery methods.
- GNW is based upon real gene regulatory networks of yeast and E. coli.
- It extracts sub-networks from the yeast or E. coli gene regulatory networks, and assigns a system of ODEs to the extracted network.
- This system of ODEs is non-additive, and includes unobserved variables.
- Therefore, the assumptions of GRADE are VIOLATED in the GNW data.
Applications

- We investigate ten networks from GNW, of which five have 10 nodes and five have 100 nodes.

- For each network, GNW provides one set of noiseless gene expression data consisting of \( R \) perturbation experiments where the trajectories are measured at \( n = 21 \) evenly-spaced time points in \([0, 1]\).

- Here \( R = 10 \) for the five 10-node networks and \( R = 100 \) for the five 100-node networks.

- We add independent \( N(0, 0.0252) \) measurement errors to the data at each time point.
## Results

### Table 1: Area Under ROC Curves for NeRDS and GRADE

<table>
<thead>
<tr>
<th></th>
<th>NeRDS</th>
<th>GRADE</th>
<th>NeRDS</th>
<th>GRADE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$p = 10$</td>
<td>$p = 100$</td>
<td>$p = 10$</td>
<td>$p = 100$</td>
</tr>
<tr>
<td>Ecoli1</td>
<td>0.450 (0.438, 0.462)</td>
<td><strong>0.545</strong> (0.534, 0.557)</td>
<td>0.624 (0.622, 0.627)</td>
<td><strong>0.670</strong> (0.667, 0.673)</td>
</tr>
<tr>
<td>Ecoli2</td>
<td>0.512 (0.502, 0.523)</td>
<td><strong>0.643</strong> (0.634, 0.653)</td>
<td>0.637 (0.635, 0.640)</td>
<td><strong>0.653</strong> (0.650, 0.656)</td>
</tr>
<tr>
<td>Yeast1</td>
<td>0.486 (0.476, 0.495)</td>
<td><strong>0.679</strong> (0.666, 0.691)</td>
<td>0.610 (0.607, 0.612)</td>
<td><strong>0.636</strong> (0.635, 0.638)</td>
</tr>
<tr>
<td>Yeast2</td>
<td>0.525 (0.518, 0.532)</td>
<td><strong>0.607</strong> (0.600, 0.613)</td>
<td>0.568 (0.566, 0.569)</td>
<td><strong>0.584</strong> (0.582, 0.585)</td>
</tr>
<tr>
<td>Yeast3</td>
<td>0.467 (0.460, 0.474)</td>
<td><strong>0.576</strong> (0.566, 0.587)</td>
<td><strong>0.617</strong> (0.616, 0.619)</td>
<td>0.567 (0.566, 0.568)</td>
</tr>
</tbody>
</table>

The average area under the curves and 90% confidence intervals, over 100 simulated data sets. Networks and data generating mechanisms are described in Section 6.1. Boldface indicates the method with larger AUC.
Thank you!