Learning the Dependence Graph of Time Series with Latent Factors

Ali Jalali and Sujay Sanghavi
ICML 2012

Presented by Xin YUAN

August 31, 2012
Outline

1. Problem Setting and Main Idea
2. Main Results
3. Experimental Results
Problem Setting

Consider a system with state vectors $x(t) \in \mathbb{R}^p$ and $u(t) \in \mathbb{R}^r$, for $t \in \mathbb{R}^+$

$$\frac{d}{dt} \begin{bmatrix} x(t) \\ u(t) \end{bmatrix} = \underbrace{\begin{bmatrix} A^* & B^* \\ C^* & D^* \end{bmatrix}}_{A^*} \begin{bmatrix} x(t) \\ u(t) \end{bmatrix} + \frac{d}{dt} w(t),$$  

(1)

where, $w(t) \in \mathbb{R}^{p+r}$ is an independent standard Brownian motion vector and $A^*, B^*, C^*, D^*$ are system parameters.

**Task:** We observe the process $x(t)$ for some time horizon $0 \leq t \leq T$, but not the process $u(\cdot)$. We are interested in learning the matrix $A^*$, which captures the interactions between the observed variables.

We will also be interested in a similar objective for an analogous *discrete time* system with parameter $0 < \eta < \frac{2}{\sigma_{\text{max}}(A^*)}$:

$$\begin{bmatrix} x(n+1) \\ u(n+1) \end{bmatrix} - \begin{bmatrix} x(n) \\ u(n) \end{bmatrix} = \eta \begin{bmatrix} A^* & B^* \\ C^* & D^* \end{bmatrix} \begin{bmatrix} x(n) \\ u(n) \end{bmatrix} + w(n)$$  

(2)
Main idea

In case of without the latent time series, the likelihood:

\[ \mathcal{L}(A) = \frac{1}{2\eta^2} \mathbb{E} \left[ \| x(i+1) - x(i) - \eta Ax(i) \|^2_2 \right]. \]

Lemma 1. For \( x(\cdot) \) generated by (2), the the optimum \( \hat{A} := \max_A \mathcal{L}(A) \) is given by

\[ \hat{A} = A^* + B^* R^*(Q^*)^{-1}. \]

Captures the spurious interactions obtained due to the latent time series.

We want to recover

Sparse

Low Rank

The number of latent time series should be \textit{smaller} than the number of observed variables.

\( Q^* \): covariance matrices of the observed variables

\( P^* \): covariance matrices of the latent variables,

\( R^* \): the cross-covariance between observed and latent variables.
Algorithm

• For the continuous-time system observed up to time $T$:

$$\hat{(A, L)} = \arg \min_{A, L} \frac{1}{2T} \int_{t=0}^{T} \| (A + L)x(t) \|_2^2 \, dt - \frac{1}{T} \int_{t=0}^{T} x(t)^T (A + L)^T dx(t) + \lambda_A \| A \|_1 + \lambda_L \| L \|_* ,$$

(3)

• For the discrete-time system given $n$ samples

$$\hat{(A, L)} = \arg \min_{A, L} \frac{1}{2 \eta^2 n} \sum_{i=0}^{n-1} \| x(i+1) - x(i) - \eta (A + L)x(i) \|_2^2 + \lambda_A \| A \|_1 + \lambda_L \| L \|_* .$$

(4)
Assumptions

(A1) Stable Overall System

for continuous system: \[ D := -\lambda_{\text{max}}\left(\frac{A^* + A^{*T}}{2}\right) > 0 \]

for discrete system: \[ D := \frac{1 - \Sigma_2^{\text{max}}}{\eta} > 0. \]

(A2) Identifiability

s: The maximum number of the non-zero entries in any row or column of A* -incoherent of the low-rank L*, which has rank r, \[ \alpha := 3\sqrt{\frac{\mu r s}{p}} < 1. \]

(A3) Incoherence

The covariance matrices of the observed variables need to satisfy the incoherence conditions same as LASSO

(A4) Regularizers

\[ m = \max\left(\frac{80}{\sqrt{D}}\|B^*\|_{\infty,1}, \sqrt{\|x(0)\|^2 + \|u(0)\|^2 + (\sqrt{\eta} + 1)^2}\right) \]

(A4-1) \[ \frac{\lambda_l}{\lambda_{\text{A}\sqrt{p}}} = \frac{16m(4-\theta)}{\theta \sqrt{D}} \sqrt{\frac{\log\left(\frac{4((s+2r)p + r^2)}{8}\right)}{n\eta}} \]

(A4-2) \[ \frac{\lambda_l}{\lambda_{\text{A}\sqrt{p}}} = \frac{1}{1-\alpha} \left(\frac{3\alpha\sqrt{s}}{4} + \frac{8-\theta s}{\theta(4-\theta)}\right)\left(\frac{\theta\sqrt{p}}{9s\sqrt{s}} + 1 + \frac{1}{2}\right). \]

(A5) Sample Complexity

A larger number of samples, \[ T = n\eta \geq \frac{K s^2}{D^2 \theta^2 C_{\text{min}}^2} \log\left(\frac{4((s + 2r)p + r^2)}{\delta}\right). \]

Main Result

**Theorem 1.** If assumptions (A1)-(A5) are satisfied, then with probability $1 - \delta$, our algorithm outputs a pair $(\hat{A}, \hat{L})$ satisfying

(a) **Subset Support Recovery:** $\text{Supp}(\hat{A}) \subset \text{Supp}(A^*)$.

(b) **Error Bounds:**

$$\|\hat{A} - A^*\|_{\infty} \leq \nu \lambda_A \quad \text{and} \quad \|\hat{L} - L^*\|_2 \leq \frac{\rho_0}{1 - 5\rho_0} \|L^*\|_2.$$

(c) **Exact Signed Support Recovery:** If additionally we have that the smallest magnitude $A_{\min}$ of a non-zero element of $A^*$ satisfies $A_{\min} > \nu \lambda_A$, then we obtain full signed-support recovery $\text{Sign}(\hat{A}) = \text{Sign}(A^*)$.

$$\nu := \frac{\alpha \theta}{2D_{\max}} + \frac{(8-\theta)\sqrt{s}}{C_{\min}(4-\theta)} \quad \text{and} \quad \rho_0 := \min \left(\frac{\alpha}{4}, \frac{\theta \alpha \lambda_A}{5D_{\max} \|L^*\|_2} \right)$$
Illustrative Example

Latent variable $j$: \[
\frac{dx_j}{dt} = -x_j(t) + \frac{dw_j}{dt}
\]

Observed variable $i$: \[
\frac{dx_i}{dt} = -x_i(t) + x_j(i)(t) + \frac{dw_i}{dt}
\]

$A^* = -I_p \times_p$, $C^* = 0$, and $D^* = -I_r \times_r$.

$Q^* = \frac{1}{2}(I + BB^T)$ and $R^* = B^*T$

$T \geq Ks^3 \log \frac{4(1+2r)p+4r^2}{\delta}$

$L^* = \frac{r}{p+r} BB^T$

$\nu = \frac{3r}{4\sqrt{p}} + \frac{25\sqrt{s}}{7}$

$\rho_0 = \frac{1}{5 + \frac{32\sqrt{p}}{3r\lambda_A}}$

$\|A^* - \hat{A}\|_\infty \leq \left(\frac{3r}{4\sqrt{p}} + \frac{25\sqrt{s}}{7}\right) \lambda_A$

$\|L^* - \hat{L}\|_2 \leq \frac{3r}{32\sqrt{p}} \lambda_A$. 
Experiment Results (1)

• Synthetic Data

We generate the data according to the continuous time model. The solution to the first order system can be written as

\[
\begin{bmatrix}
  x(t) \\
  u(t)
\end{bmatrix} = e^{A^*(t-t_0)} \begin{bmatrix}
  x(t_0) \\
  u(t_0)
\end{bmatrix} + \int_{t_0}^{t} e^{A^*(t-\tau)}dw(\tau),
\]

where, \( e^{A^*} = I + A^* + \frac{1}{2} A^*^2 + \ldots \) is a generalization of the exponential function to matrices. We sub-sample this system at points \( t_i = \eta i \) for \( i = 1, 2, \ldots, n \), that is

\[
\begin{bmatrix}
  x(i) \\
  u(i)
\end{bmatrix} = e^{\eta A} \begin{bmatrix}
  x(i-1) \\
  u(i-1)
\end{bmatrix} + \int_{\eta(i-1)}^{\eta i} e^{A(\eta i-\tau)}dw(\tau)
\]

Figure 1: Probability of success in recovering the true signed support of $A^*$ versus the control parameter $\Theta$ with $p = 200$, $r = 10$ and $s = 20$ for different values of $\eta$ in (a), and, with $p = 200$, $s = 20$ and $\eta = 0.01$ for different number of latent time series $r$ in (b), and, with $p = 200$, $r = 10$ and fixed $\eta = 0.01$ for different sparsity sizes $s$ in (c). Notice that Fig. (c) is plotted versus $\Theta \times s$ which means $\eta \eta$ scales with $s^2$ not $s^3$. This means our theoretical result can be tightened.

$$\Theta = \frac{s^3 \log \left( (s + 2r)p + r^2 \right)}{\eta m}$$
Experiment Results (2)

• Stock Market Data:
The end-of-the-day closing stock prices for 50 different companies in the period of May 17, 2010 - May 13, 2011 (255 business days).

\[ \text{Figure 2. Comparison of the stock dependencies recovered by Pure LASSO (Bento et al., 2010) and our algorithm. This shows that there are latent factors affecting large number of stocks.} \]
Figure 3. Prediction error and model sparsity versus the ratio of the training/testing sample sizes for prediction of the stock price. Prediction error is measured using mean squared error and the model sparsity is the number of non-zero entries divided by the size of $\hat{A}$. 