Homotopy algorithm for structured matrix-variate Lasso and its application in online graph inference from matrix-variate time series

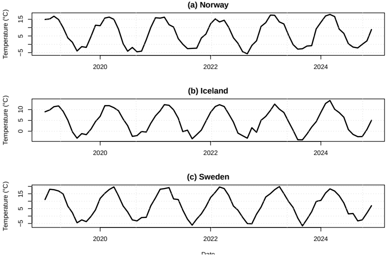
Yiye JIANG

University of Grenoble Alpes, CNRS,
Inria, Grenoble INP, France

January 26, 2026

Table of Contents

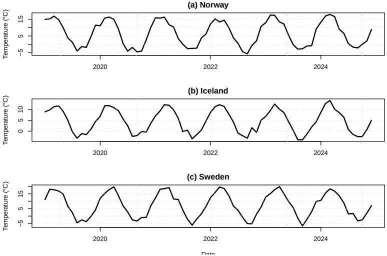
Graph learning in classical setting



- Let $x_t \in \mathbb{R}^3$, $t \in \mathbb{Z}$, denote the 3 temp at time t

$$\text{VAR}(1): \; x_t = A x_{t-1} + b + z_t.$$

Graph learning in classical setting

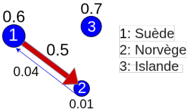


- Let $x_t \in \mathbb{R}^3,\ t \in \mathbb{Z}$, denote the 3 temp at time t

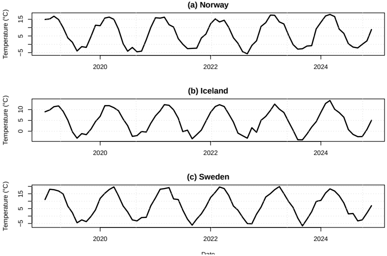
$$\mathrm{VAR}(1):\ x_t = Ax_{t-1} + b + z_t.$$

Numerical example:

$$\begin{pmatrix} x_{1t} \\ {\color{red} x_{2t}} \\ x_{3t} \end{pmatrix} = \begin{pmatrix} 0.6 & 0.04 & 0 \\ {\color{red} 0.5} & 0.01 & 0 \\ 0 & 0 & 0.7 \end{pmatrix} \begin{pmatrix} {\color{red} x_{1t-1}} \\ x_{2t-1} \\ x_{3t-1} \end{pmatrix} + \begin{pmatrix} 15 \\ 12 \\ 10 \end{pmatrix} + z_t$$



Graph learning in classical setting

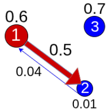


- Let $x_t \in \mathbb{R}^3,\ t \in \mathbb{Z}$, denote the 3 temp at time t

$$\mathrm{VAR}(1):\ x_t = Ax_{t-1} + b + z_t.$$

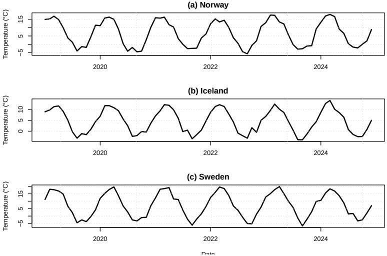
Numerical example:

$$\begin{pmatrix} x_{1t} \\ {\color{red} x_{2t}} \\ x_{3t} \end{pmatrix} = \begin{pmatrix} 0.6 & 0.04 & 0 \\ {\color{red} 0.5} & 0.01 & 0 \\ 0 & 0 & 0.7 \end{pmatrix} \begin{pmatrix} {\color{red} x_{1t-1}} \\ x_{2t-1} \\ x_{3t-1} \end{pmatrix} + \begin{pmatrix} 15 \\ 12 \\ 10 \end{pmatrix} + z_t$$



1: Suède
2: Norvège
3: Islande

Graph learning in classical setting

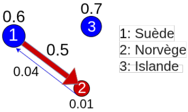


- Let $x_t \in \mathbb{R}^3,\ t \in \mathbb{Z}$, denote the 3 temp at time t

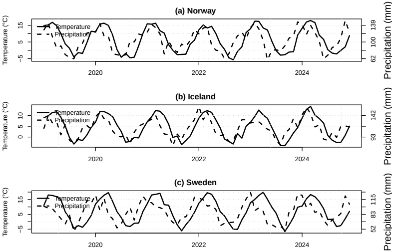
$$\mathrm{VAR}(1):\ x_t = Ax_{t-1} + b + z_t.$$

Numerical example:

$$\begin{pmatrix} x_{1t} \\ {\color{red}x_{2t}} \\ x_{3t} \end{pmatrix} = \begin{pmatrix} 0.6 & 0.04 & 0 \\ {\color{red}0.5} & 0.01 & 0 \\ 0 & 0 & 0.7 \end{pmatrix} \begin{pmatrix} {\color{red}x_{1t-1}} \\ x_{2t-1} \\ x_{3t-1} \end{pmatrix} + \begin{pmatrix} 15 \\ 12 \\ 10 \end{pmatrix} + z_t$$



Matrix-valued time series



- Let $X_t \in \mathbb{R}^{3\times 2}$, $t \in \mathbb{Z}$, denote the temp and precip of 3 countries at time t

$$\mathrm{VAR}(1):\ x_t = Ax_{t-1} + b + z_t. \quad x_t \to X_t? \quad \mathcal{G}(3)?$$

Novel matrix autoregressive (MAR) model

For simplicity, we assume $\mathbb{E}X_t = 0, \forall t$, we propose

$$X_t = A_{\mathrm{N}} X_{t-1} + X_{t-1} A_{\mathrm{F}}^\top + Z_t,$$

where $X_t, Z_t \in \mathbb{R}^{N\times F}$, $A_{\mathrm{N}} \in \mathbb{R}^{N\times N}$, $A_{\mathrm{F}} \in \mathbb{R}^{F\times F}$, N is the nb of sensors, and F is the nb of features.

Novel matrix autoregressive (MAR) model

For simplicity, we assume $\mathbb{E}X_t = 0, \forall t$, we propose

$$X_t = A_{\mathrm{N}} X_{t-1} + X_{t-1} A_{\mathrm{F}}^\top + Z_t,$$

where $X_t, Z_t \in \mathbb{R}^{N \times F}$, $A_{\mathrm{N}} \in \mathbb{R}^{N \times N}$, $A_{\mathrm{F}} \in \mathbb{R}^{F \times F}$, N is the nb of sensors, and F is the nb of features.

$$x_{jt} = A_{\mathrm{N}} x_{jt-1} + z_{jt}, \forall j = 1, \dots, F, x_{jt} = j\text{-th column of } X_t.$$

$A_{\mathrm{N}} \iff$ row/sensor dependence.

Novel matrix autoregressive (MAR) model

For simplicity, we assume $\mathbb{E}X_t = 0, \forall t$, we propose

$$X_t = A_\mathrm{N} X_{t-1} + X_{t-1} A_\mathrm{F}^\top + Z_t,$$

where $X_t, Z_t \in \mathbb{R}^{N \times F}$, $A_\mathrm{N} \in \mathbb{R}^{N \times N}$, $A_\mathrm{F} \in \mathbb{R}^{F \times F}$, N is the nb of sensors, and F is the nb of features.

$$x_{jt} = A_\mathrm{N} x_{jt-1} + z_{jt}, \forall j = 1, ..., F, x_{jt} = j\text{-th column of } X_t.$$

$A_\mathrm{N} \iff$ row/sensor dependence.

$$x_{it} = x_{it-1} A_\mathrm{F}^\top + z_{it}, \forall i = 1, ..., N, x_{it} = i\text{-th row of } X_t.$$

$A_\mathrm{F} \iff$ column/feature dependence.

Novel matrix autoregressive (MAR) model

For simplicity, we assume $\mathbb{E}X_t = 0, \forall t$, we propose

$$X_t = A_{\mathrm{N}} X_{t-1} + X_{t-1} A_{\mathrm{F}}^{\top} + Z_t,$$

where $X_t, Z_t \in \mathbb{R}^{N \times F}$, $A_{\mathrm{N}} \in \mathbb{R}^{N \times N}$, $A_{\mathrm{F}} \in \mathbb{R}^{F \times F}$, N is the nb of sensors, and F is the nb of features.

$$x_{jt} = A_{\mathrm{N}} x_{jt-1} + z_{jt}, \forall j = 1, \dots, F, x_{jt} = j\text{-th column of } X_t.$$

$A_{\mathrm{N}} \iff$ row/sensor dependence.

$$x_{it} = x_{it-1} A_{\mathrm{F}}^{\top} + z_{it}, \forall i = 1, \dots, N, x_{it} = i\text{-th row of } X_t.$$

$A_{\mathrm{F}} \iff$ column/feature dependence.

A_{N} is invariant to features, A_{F} is invariant to sensors.

Table of Contents

Lasso in classical inference

Given the vectorial observations x_τ, $\tau = 0, ..., t$, assumed follow VAR(1) : (for simplicity, we assume that $\mathbb{E}x_t = 0$)

$$x_t = Ax_{t-1} + z_t.$$

Estimation - classical Lasso

$$\widehat{A} := \arg\min_A \underbrace{\frac{1}{2t}\sum_{\tau=1}^{t} \|x_\tau - Ax_{\tau-1}\|_{\ell_2}^2}_{\substack{\text{data term} \\ \text{(minimize residuals)}}} + \underbrace{\lambda \|A\|_{\ell_1}}_{\substack{\text{penalty term} \\ \text{(encourage sparsity)}}}$$

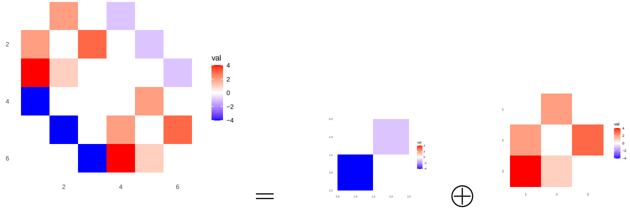
Given matrix observations $X_\tau, \tau = 0, 1, ...,$ assumed follow

$$X_\tau = A_N X_{\tau-1} + X_{\tau-1} A_F^\top + Z_\tau.$$

Vectorial representation:

$$x_\tau = A x_{\tau-1} + z_\tau, \quad \text{where } A = A_\mathrm{F} \oplus A_\mathrm{N}, \quad x_\tau = \mathrm{vec}(X_\tau).$$

Recall Kronecker sum: $A_\mathrm{F} \oplus A_\mathrm{N} = A_\mathrm{F} \otimes I_F + I_N \otimes A_\mathrm{N}$



Vectorial representation:

$$x_t = Ax_{t-1} + z_t, \quad \text{where } A = A_{\mathrm{F}} \oplus A_{\mathrm{N}}, \quad x_t = \mathrm{vec}(X_t).$$

Estimation

$$\widehat{A} := \underset{A = A_{\mathrm{F}} \oplus A_{\mathrm{N}}}{\arg\min} \frac{1}{2t} \sum_{\tau=1}^{t} \|x_\tau - Ax_{\tau-1}\|_{\ell_2}^2 + \lambda \|A_{\mathrm{N}}\|_{\ell_1}$$

Vectorial representation:

$$x_t = Ax_{t-1} + z_t, \quad \text{where } A = A_{\mathrm{F}} \oplus A_{\mathrm{N}}, \quad x_t = \mathrm{vec}(X_t).$$

Estimation

$$\widehat{A} := \operatorname*{arg\,min}_{A \in \mathcal{K}_{\mathcal{G}}} \frac{1}{2t} \sum_{\tau=1}^{t} \|x_\tau - Ax_{\tau-1}\|_{\ell_2}^2 + \lambda \|A_{\mathrm{N}}\|_{\ell_1},$$

where

$$\mathcal{K}_{\mathcal{G}} = \left\{ M \in \mathbb{R}^{NF \times NF} : M = M_{\mathrm{F}} \oplus M_{\mathrm{N}},\ M_{\mathrm{F}} \in \mathbb{R}^{F \times F}, M_{\mathrm{N}} \in \mathbb{R}^{N \times N} \right\}.$$

Vectorial representation:

$$x_t = Ax_{t-1} + z_t, \quad \text{where } A = A_{\mathrm{F}} \oplus A_{\mathrm{N}}, \quad x_t = \mathrm{vec}(X_t).$$

Estimation

$$\widehat{A} := \arg\min_{A \in \mathcal{K}_{\mathcal{G}}} \frac{1}{2t} \sum_{\tau=1}^{t} \|x_\tau - Ax_{\tau-1}\|_{\ell_2}^2 + \lambda \|A_{\mathrm{N}}\|_{\ell_1},$$

where

$$\begin{aligned}\mathcal{K}_{\mathcal{G}} \quad = \quad & \big\{ M \in \mathbb{R}^{NF \times NF} : \exists\, M_{\mathrm{F}} \in \mathbb{R}^{F \times F}, M_{\mathrm{N}} \in \mathbb{R}^{N \times N}, \text{ such that,} \\ & \mathrm{offd}(M) = M_{\mathrm{F}} \oplus M_{\mathrm{N}}, \text{ with, } \mathrm{diag}(M_{\mathrm{F}}) = 0,\ \mathrm{diag}(M_{\mathrm{N}}) = 0, \\ & M_{\mathrm{F}} = M_{\mathrm{F}}^\top,\ M_{\mathrm{N}} = M_{\mathrm{N}}^\top \big\}.\end{aligned}$$

Our framework is independent of the specific structure !

Offline optimization

$$\min_{A \in \mathcal{K}_{\mathcal{G}}} f(A) + \lambda \left\| A_{\mathrm{N}} \right\|_{\ell_1}, \; \textit{where } f(A) = \frac{1}{2t} \sum_{\tau=1}^{t} \| x_\tau - A x_{\tau-1} \|_{\ell_2}^2$$

Proximal gradient descent

$$\begin{aligned} A^{k+1} &= \mathrm{prox}(A^k - \eta^k \nabla f(A^k)), \\ &= \operatorname*{arg\,min}_{A \in \mathcal{K}_{\mathcal{G}}} \frac{1}{2\eta^k} \left\| A - \left(A^k - \eta^k \nabla f(A^k) \right) \right\|_{\ell_2}^2 + \frac{\lambda}{F} \left\| A_{\mathrm{N}} \right\|_{\ell_1}. \end{aligned}$$

Online optimization

$$A(t,\lambda) := \arg\min_{A \in \mathcal{K}_{\mathcal{G}}} \frac{1}{2t} \sum_{\tau=1}^{t} \|x_\tau - A x_{\tau-1}\|_{\ell_2}^2 + \lambda \|A_{\mathrm{N}}\|_{\ell_1}$$

Online inference:

$$\begin{aligned} \text{regularization path}: \; & A(t,\lambda_1) \to A(t,\lambda_2), \\ \text{data path}: \; & A(t,\lambda) \to A(t+1,\lambda). \end{aligned}$$

Table of Contents

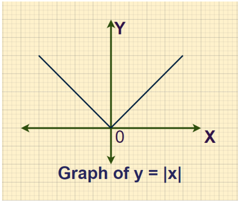
Algorithm for classical Lasso

Regularization path, $\theta(\lambda_1) \to \theta(\lambda_2)$:

$$\theta(\lambda) = \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{2}\|y - X\theta\|_{\ell_2}^2 + \lambda\|\theta\|_{\ell_1},$$

Optimality condition:

$$0 \in X^\top (X\theta(\lambda) - y) + \partial\|\theta(\lambda)\|_{\ell_1}.$$



$$\partial\|\theta(\lambda)\|_{\ell_1} = \left\{ v \in \mathbb{R}^d : \begin{cases} v_i = \operatorname{sgn}(\theta_i(\lambda)) & \text{if } \theta_i(\lambda) \neq 0, \\ v_i \in [-1, 1] & \text{if } \theta_i(\lambda) = 0. \end{cases} \right.$$

Algorithm for classical Lasso

Regularization path, $\theta(\lambda_1) \rightarrow \theta(\lambda_2)$:

$$\theta(\lambda) = \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{2}\|y - X\theta\|_{\ell_2}^2 + \lambda\|\theta\|_{\ell_1},$$

Optimality condition:

$$\mathbf{0} \in X^\top (X\theta(\lambda) - y) + \partial\|\theta(\lambda)\|_{\ell_1}. \qquad (5)$$

Suppose unique solution: $\theta(\lambda) = (\theta_1(\lambda), \mathbf{0})$ at λ, $X = (X_1, X_2)$:

$$(5) \iff \begin{cases} (X_1^\top X_1)^{-1} \left(X_1^\top y - \lambda\,\mathrm{sgn}[\theta_1(\lambda)]\right) = \theta_1(\lambda), \\ \frac{1}{\lambda} X_2^\top (y - X_1\theta_1(\lambda)) \in \partial\|\mathbf{0}\|_{\ell_1}. \end{cases}$$

Recall: $\partial\|\mathbf{0}\|_{\ell_1} = \begin{pmatrix} [-1,1] \\ [-1,1] \\ \dots \\ [-1,1] \end{pmatrix}.$

Optimality condition:

$$\begin{cases} (X_1^\top X_1)^{-1}(X_1^\top y - \lambda \operatorname{sgn}[\theta_1(\lambda)]) = \theta_1(\lambda), \\ \frac{1}{\lambda} X_2^\top (y - X_1\theta_1(\lambda)) \in \partial \|\mathbf{0}\|_{\ell_1}. \end{cases}$$

Optimality condition:

$$\begin{cases} (X_1^\top X_1)^{-1}(X_1^\top y - \lambda \mathrm{sgn}[\theta_1(\lambda)]) = \theta_1(\lambda), \\ \frac{1}{\lambda} X_2^\top (y - X_1\theta_1(\lambda)) \in \partial \|\mathbf{0}\|_{\ell_1}. \end{cases}$$

Solutions $\theta(\lambda')$, $|\lambda' - \lambda| < \epsilon$:

Define $\theta(\lambda') = (\theta_1(\lambda'), \mathbf{0})$, with

$$\theta_1(\lambda') = (X_1^\top X_1)^{-1}(X_1^\top y - \lambda' \mathrm{sgn}[\theta_1(\lambda)]).$$

Optimality condition:

$$\begin{cases} (X_1^\top X_1)^{-1}(X_1^\top y - \lambda \text{sgn}[\theta_1(\lambda)]) = \theta_1(\lambda), \\ \frac{1}{\lambda} X_2^\top (y - X_1\theta_1(\lambda)) \in \partial \|\mathbf{0}\|_{\ell_1}. \end{cases}$$

Solutions $\theta(\lambda')$, $|\lambda' - \lambda| < \epsilon$:

Define $\theta(\lambda') = (\theta_1(\lambda'), \mathbf{0})$, with

$$\theta_1(\lambda') = (X_1^\top X_1)^{-1}(X_1^\top y - \lambda' \text{sgn}[\theta_1(\lambda)]).$$

Proof:

- $\theta_1(\lambda')$ is **linear** in λ', by continuity, $\theta_1(\lambda') \approx \theta_1(\lambda)$. Thus $\theta_1(\lambda) \neq 0 \rightarrow \theta_1(\lambda') \neq 0$ et $sgn[\theta_1(\lambda')] = \text{sgn}[\theta_1(\lambda)]$.

Optimality condition:

$$\begin{cases} (X_1^\top X_1)^{-1}(X_1^\top y - \lambda \operatorname{sgn}[\theta_1(\lambda)]) = \theta_1(\lambda), \\ \frac{1}{\lambda} X_2^\top (y - X_1\theta_1(\lambda)) \in \partial \|\mathbf{0}\|_{\ell_1}. \end{cases}$$

Solutions $\theta(\lambda')$, $|\lambda' - \lambda| < \epsilon$:

Define $\theta(\lambda') = (\theta_1(\lambda'), \mathbf{0})$, with

$$\theta_1(\lambda') = (X_1^\top X_1)^{-1}(X_1^\top y - \lambda' \operatorname{sgn}[\theta_1(\lambda)]).$$

Proof:

- $\theta_1(\lambda')$ is **linear** in λ', by continuity, $\theta_1(\lambda') \approx \theta_1(\lambda)$. Thus $\theta_1(\lambda) \neq 0 \rightarrow \theta_1(\lambda') \neq 0$ et $sgn[\theta_1(\lambda')] = \operatorname{sgn}[\theta_1(\lambda)]$.

$$\checkmark\, OC1 : (X_1^\top X_1)^{-1} \left(X_1^\top y - \lambda \operatorname{sgn}[\theta_1(\lambda')]\right) = \theta_1(\lambda')$$

Optimality condition:

$$\begin{cases} (X_1^\top X_1)^{-1}(X_1^\top y - \lambda \operatorname{sgn}[\theta_1(\lambda)]) = \theta_1(\lambda), \\ \frac{1}{\lambda} X_2^\top (y - X_1 \theta_1(\lambda)) \in \partial \|\mathbf{0}\|_{\ell_1}. \end{cases}$$

Solutions $\theta(\lambda')$, $|\lambda' - \lambda| < \epsilon$:

Define $\theta(\lambda') = (\theta_1(\lambda'), \mathbf{0})$, with

$$\theta_1(\lambda') = (X_1^\top X_1)^{-1}(X_1^\top y - \lambda' \operatorname{sgn}[\theta_1(\lambda)]).$$

Optimality condition:

$$\begin{cases} (X_1^\top X_1)^{-1}(X_1^\top y - \lambda \operatorname{sgn}[\theta_1(\lambda)]) = \theta_1(\lambda), \\ \frac{1}{\lambda} X_2^\top (y - X_1\theta_1(\lambda)) \in \partial\|\mathbf{0}\|_{\ell_1}. \end{cases}$$

Solutions $\theta(\lambda')$, $|\lambda' - \lambda| < \epsilon$:

Define $\theta(\lambda') = (\theta_1(\lambda'), \mathbf{0})$, with

$$\theta_1(\lambda') = (X_1^\top X_1)^{-1}(X_1^\top y - \lambda' \operatorname{sgn}[\theta_1(\lambda)]).$$

To prove: $v(\lambda') := \frac{1}{\lambda'} X_2^\top (y - X_1\theta_1(\lambda')) \in \partial\|\mathbf{0}\|_{\ell_1} = \begin{pmatrix} [-1,1] \\ [-1,1] \\ \dots \\ [-1,1] \end{pmatrix}.$

Optimality condition:

$$\begin{cases} (X_1^\top X_1)^{-1}(X_1^\top y - \lambda \mathrm{sgn}[\theta_1(\lambda)]) = \theta_1(\lambda), \\ \frac{1}{\lambda} X_2^\top (y - X_1\theta_1(\lambda)) \in \partial \|\mathbf{0}\|_{\ell_1}. \end{cases}$$

Solutions $\theta(\lambda')$, $|\lambda' - \lambda| < \epsilon$:

Define $\theta(\lambda') = (\theta_1(\lambda'), \mathbf{0})$, with

$$\theta_1(\lambda') = (X_1^\top X_1)^{-1}(X_1^\top y - \lambda' \mathrm{sgn}[\theta_1(\lambda)]).$$

To prove: $v(\lambda') := \frac{1}{\lambda'} X_2^\top (y - X_1\theta_1(\lambda')) \in \partial \|\mathbf{0}\|_{\ell_1} = \begin{pmatrix} [-1,1] \\ [-1,1] \\ \dots \\ [-1,1] \end{pmatrix}.$

- $v_i(\lambda) \in [-1, 1]$.
  $v(\lambda')$ is **smooth** in λ', $v_i(\lambda') \approx v_i(\lambda)$. Done !

Optimality condition:

$$\begin{cases} (X_1^\top X_1)^{-1}(X_1^\top y - \lambda \mathrm{sgn}[\theta_1(\lambda)]) = \theta_1(\lambda), \\ \frac{1}{\lambda} X_2^\top (y - X_1\theta_1(\lambda)) \in \partial \|\mathbf{0}\|_{\ell_1}. \end{cases}$$

Solutions $\theta(\lambda')$, $|\lambda' - \lambda| < \epsilon$:

Define $\theta(\lambda') = (\theta_1(\lambda'), \mathbf{0})$, with

$$\theta_1(\lambda') = (X_1^\top X_1)^{-1}(X_1^\top y - \lambda' \mathrm{sgn}[\theta_1(\lambda)]).$$

To prove: $v(\lambda') := \frac{1}{\lambda'} X_2^\top (y - X_1\theta_1(\lambda')) \in \partial \|\mathbf{0}\|_{\ell_1} = \begin{pmatrix} [-1,1] \\ [-1,1] \\ \dots \\ [-1,1] \end{pmatrix}$.

- $v_i(\lambda) \in [-1, 1]$.
  $v(\lambda')$ is **smooth** in λ', $v_i(\lambda') \approx v_i(\lambda)$. Done !?
  $|v_i(\lambda)| = 1, v_i(\lambda') \notin [-1, 1]$.

Optimality condition:

$$\begin{cases} (X_1^\top X_1)^{-1}(X_1^\top y - \lambda \mathrm{sgn}[\theta_1(\lambda)]) = \theta_1(\lambda), \\ \frac{1}{\lambda} X_2^\top (y - X_1\theta_1(\lambda)) \in \partial \|\mathbf{0}\|_{\ell_1}. \end{cases}$$

Solutions $\theta(\lambda')$, $|\lambda' - \lambda| < \epsilon$:

Define $\theta(\lambda') = (\theta_1(\lambda'), \mathbf{0})$, with

$$\theta_1(\lambda') = (X_1^\top X_1)^{-1}(X_1^\top y - \lambda' \mathrm{sgn}[\theta_1(\lambda)]).$$

To prove: $v(\lambda') := \frac{1}{\lambda'} X_2^\top (y - X_1\theta_1(\lambda')) \in \partial \|\mathbf{0}\|_{\ell_1} = \begin{pmatrix} [-1,1] \\ [-1,1] \\ \ldots \\ [-1,1] \end{pmatrix}.$

- $v_i(\lambda) \in [-1, 1]$.
  $v(\lambda')$ is **smooth** in λ', $v_i(\lambda') \approx v_i(\lambda)$. Done !?
  $|v_i(\lambda)| = 1, v_i(\lambda') \notin [-1, 1]$.

Assump.: $\forall i, v_i(\lambda) \in (-1, 1)$, namely, λ is not a critical point.

Optimality condition:

$$\begin{cases} (X_1^\top X_1)^{-1}(X_1^\top y - \lambda \operatorname{sgn}[\theta_1(\lambda)]) = \theta_1(\lambda), \\ \frac{1}{\lambda} X_2^\top (y - X_1\theta_1(\lambda)) \in \partial \|\mathbf{0}\|_{\ell_1}. \end{cases}$$

Solutions $\theta(\lambda')$, $|\lambda' - \lambda| < \epsilon$:

Define $\theta(\lambda') = (\theta_1(\lambda'), \mathbf{0})$, with

$$\theta_1(\lambda') = (X_1^\top X_1)^{-1}(X_1^\top y - \lambda' \operatorname{sgn}[\theta_1(\lambda)]).$$

Piecewise constant support/sign pattern of Lasso solutions

1. The support/sign pattern of Lasso solutions will stay unchanged over a small range of λ.

? ϵ, if $|\lambda_2 - \lambda| < \epsilon$, done !

Optimality condition:

$$\begin{cases} (X_1^\top X_1)^{-1}(X_1^\top y - \lambda \operatorname{sgn}[\theta_1(\lambda)]) = \theta_1(\lambda), \\ \frac{1}{\lambda} X_2^\top (y - X_1\theta_1(\lambda)) \in \partial \|\mathbf{0}\|_{\ell_1}. \end{cases}$$

Solutions $\theta(\lambda')$, $|\lambda' - \lambda| < \epsilon$:

Define $\theta(\lambda') = (\theta_1(\lambda'), \mathbf{0})$, with

$$\theta_1(\lambda') = (X_1^\top X_1)^{-1}(X_1^\top y - \lambda' \operatorname{sgn}[\theta_1(\lambda)]).$$

Piecewise constant support/sign pattern of Lasso solutions

1. The support/sign pattern of Lasso solutions will stay unchanged over a small range of λ.
2. Knowing the support/sign pattern of $\iff$ Knowing the complete solution.

Optimality condition:

$$\begin{cases} (X_1^\top X_1)^{-1}(X_1^\top y - \lambda \operatorname{sgn}[\theta_1(\lambda)]) = \theta_1(\lambda), \\ \frac{1}{\lambda} X_2^\top (y - X_1\theta_1(\lambda)) \in \partial\|\mathbf{0}\|_{\ell_1}. \end{cases}$$

Solutions $\theta(\lambda')$, $|\lambda' - \lambda| < \epsilon$:

Define $\theta(\lambda') = (\theta_1(\lambda'), \mathbf{0})$, with

$$\theta_1(\lambda') = (X_1^\top X_1)^{-1}(X_1^\top y - \lambda' \operatorname{sgn}[\theta_1(\lambda)]).$$

Piecewise constant support/sign pattern of Lasso solutions

1. The support/sign pattern of Lasso solutions will stay unchanged over a small range of λ.
2. Knowing the support/sign pattern of $\iff$ Knowing the complete solution.

? ϵ, if $|\lambda_2 - \lambda| < \epsilon$, done !

Optimality condition:

$$\begin{cases} (X_1^\top X_1)^{-1}(X_1^\top y - \lambda \text{sgn}[\theta_1(\lambda)]) = \theta_1(\lambda), \\ \frac{1}{\lambda} X_2^\top (y - X_1\theta_1(\lambda)) \in \partial \|\mathbf{0}\|_{\ell_1}. \end{cases}$$

Solutions $\theta(\lambda')$, $|\lambda' - \lambda| < \epsilon$:

Define $\theta(\lambda') = (\theta_1(\lambda'), \mathbf{0})$, with

$$\theta_1(\lambda') = (X_1^\top X_1)^{-1}(X_1^\top y - \lambda' \text{sgn}[\theta_1(\lambda)]).$$

- When $\lambda' = \lambda$, $\theta_1(\lambda') = \theta_1(\lambda)$.

Optimality condition:

$$\begin{cases} (X_1^\top X_1)^{-1}(X_1^\top y - \lambda \mathrm{sgn}[\theta_1(\lambda)]) = \theta_1(\lambda), \\ \frac{1}{\lambda} X_2^\top (y - X_1\theta_1(\lambda)) \in \partial \|\mathbf{0}\|_{\ell_1}. \end{cases}$$

Solutions $\theta(\lambda')$, $|\lambda' - \lambda| < \epsilon$:

Define $\theta(\lambda') = (\theta_1(\lambda'), \mathbf{0})$, with

$$\theta_1(\lambda') = (X_1^\top X_1)^{-1}(X_1^\top y - \lambda' \mathrm{sgn}[\theta_1(\lambda)]).$$

- When $\lambda' = \lambda$, $\theta_1(\lambda') = \theta_1(\lambda)$.
- When λ' *just* leaves λ,

$$\theta_1(\lambda') \approx \theta_1(\lambda), sgn[\theta_1(\lambda')] = \mathrm{sgn}[\theta_1(\lambda)], \checkmark OC1.$$

- Until the critical value λ', s.t. for some i, $\theta_{1,i}(\lambda') = 0$, then stays 0, or, changes sign.

Optimality condition:

$$\begin{cases} (X_1^\top X_1)^{-1}(X_1^\top y - \lambda \text{sgn}[\theta_1(\lambda)]) = \theta_1(\lambda), \\ \frac{1}{\lambda} X_2^\top (y - X_1\theta_1(\lambda)) \in \partial \|\mathbf{0}\|_{\ell_1}. \end{cases}$$

Solutions $\theta(\lambda')$, $|\lambda' - \lambda| < \epsilon$:

Define $\theta(\lambda') = (\theta_1(\lambda'), \mathbf{0})$, with

$$\theta_1(\lambda') = (X_1^\top X_1)^{-1}(X_1^\top y - \lambda' \text{sgn}[\theta_1(\lambda)]).$$

- When $\lambda' = \lambda$, $\theta_1(\lambda') = \theta_1(\lambda)$.
- When λ' *just* leaves λ,

$$\theta_1(\lambda') \approx \theta_1(\lambda), sgn[\theta_1(\lambda')] = \text{sgn}[\theta_1(\lambda)], \checkmark OC1.$$

- Until the critical value λ', s.t. for some i, $\theta_{1,i}(\lambda') = 0$, then stays 0, or, changes sign.

  Critical values: $(\max_i\{\lambda' < \lambda : \theta_{1,i}(\lambda') = 0\}, \min_i\{\lambda' > \lambda : \theta_{1,i}(\lambda') = 0\})$

Optimality condition:

$$\begin{cases} (X_1^\top X_1)^{-1}(X_1^\top y - \lambda \mathrm{sgn}[\theta_1(\lambda)]) = \theta_1(\lambda), \\ \frac{1}{\lambda} X_2^\top (y - X_1\theta_1(\lambda)) \in \partial \|\mathbf{0}\|_{\ell_1}. \end{cases}$$

$\theta(\lambda') = (\theta_1(\lambda'), \mathbf{0})$, with

$$\theta_1(\lambda') = (X_1^\top X_1)^{-1}(X_1^\top y - \lambda' \mathrm{sgn}[\theta_1(\lambda)]).$$

Validity interval of λ

Let $\lambda_i, i = 1, ..., d$ be the solutions of following eqs:

$$\begin{cases} \theta_1(\lambda') = \mathbf{0}, \\ \frac{1}{\lambda'} X_2^\top (y - X_1\theta_1(\lambda')) = \mathbf{1}, \\ \frac{1}{\lambda'} X_2^\top (y - X_1\theta_1(\lambda')) = -\mathbf{1}. \end{cases}$$

$\theta_1(\lambda')$ is a Lasso solution, for $\forall \lambda' \in I_\lambda$, with

$$I_\lambda = (\max_i \{\lambda_i : \lambda_i < \lambda\}, \min_j \{\lambda_j : \lambda_j > \lambda\})$$

Optimality condition:

$$\begin{cases} (X_1^\top X_1)^{-1}(X_1^\top y - \lambda \mathrm{sgn}[\theta_1(\lambda)]) = \theta_1(\lambda), \\ \frac{1}{\lambda} X_2^\top (y - X_1\theta_1(\lambda)) \in \partial \|\mathbf{0}\|_{\ell_1}. \end{cases}$$

$\forall \lambda' \in I_\lambda$, $\theta(\lambda') = (\theta_1(\lambda'), \mathbf{0})$, with

$$\theta_1(\lambda') = (X_1^\top X_1)^{-1}(X_1^\top y - \lambda' \mathrm{sgn}[\theta_1(\lambda)]).$$

- Thus if the new $\lambda_2 \in I_\lambda$, done !

Optimality condition:

$$\begin{cases} (X_1^\top X_1)^{-1}(X_1^\top y - \lambda \mathrm{sgn}[\theta_1(\lambda)]) = \theta_1(\lambda), \\ \frac{1}{\lambda} X_2^\top (y - X_1\theta_1(\lambda)) \in \partial \|\mathbf{0}\|_{\ell_1}. \end{cases}$$

$\forall \lambda' \in I_\lambda$, $\theta(\lambda') = (\theta_1(\lambda'), \mathbf{0})$, with

$$\theta_1(\lambda') = (X_1^\top X_1)^{-1}(X_1^\top y - \lambda' \mathrm{sgn}[\theta_1(\lambda)]).$$

- Thus if the new $\lambda_2 \in I_\lambda$, done !
- If not? e.g. $\lambda_2 \geqslant \lambda_c := \min_j\{\lambda_j : \lambda_j > \lambda\}$.

Optimality condition:

$$\begin{cases} (X_1^\top X_1)^{-1}(X_1^\top y - \lambda \mathrm{sgn}[\theta_1(\lambda)]) = \theta_1(\lambda), \\ \frac{1}{\lambda} X_2^\top (y - X_1\theta_1(\lambda)) \in \partial \|\mathbf{0}\|_{\ell_1}. \end{cases}$$

$\forall \lambda' \in I_\lambda$, $\theta(\lambda') = (\theta_1(\lambda'), \mathbf{0})$, with

$$\theta_1(\lambda') = (X_1^\top X_1)^{-1}(X_1^\top y - \lambda' \mathrm{sgn}[\theta_1(\lambda)]).$$

- Thus if the new $\lambda_2 \in I_\lambda$, done !
- If not? e.g. $\lambda_2 \geqslant \lambda_c := \min_j \{\lambda_j : \lambda_j > \lambda\}$.
  In $(\lambda_c, \lambda_c + \epsilon)$, only 1 entry changes **sparsity** wrt $\theta(\lambda)$!

Validity interval of λ

$$\begin{cases} \theta_1(\lambda') = \mathbf{0}, \theta_{j'} \to 0 \\ \frac{1}{\lambda'} X_2^\top (y - X_1\theta_1(\lambda')) = \mathbf{1}, \\ \frac{1}{\lambda'} X_2^\top (y - X_1\theta_1(\lambda')) = -\mathbf{1}. \end{cases}$$

Optimality condition:

$$\begin{cases} (X_1^\top X_1)^{-1}(X_1^\top y - \lambda \mathrm{sgn}[\theta_1(\lambda)]) = \theta_1(\lambda), \\ \frac{1}{\lambda} X_2^\top (y - X_1\theta_1(\lambda)) \in \partial \|\mathbf{0}\|_{\ell_1}. \end{cases}$$

$\forall \lambda' \in I_\lambda$, $\theta(\lambda') = (\theta_1(\lambda'), \mathbf{0})$, with

$$\theta_1(\lambda') = (X_1^\top X_1)^{-1}(X_1^\top y - \lambda' \mathrm{sgn}[\theta_1(\lambda)]).$$

- Thus if the new $\lambda_2 \in I_\lambda$, done !
- If not? e.g. $\lambda_2 \geqslant \lambda_c := \min_j \{\lambda_j : \lambda_j > \lambda\}$.
  In $(\lambda_c, \lambda_c + \epsilon)$, only 1 entry changes **sparsity** wrt $\theta(\lambda)$!

Validity interval of λ

$$\begin{cases} \theta_1(\lambda') = \mathbf{0} \\ \frac{1}{\lambda'} X_2^\top (y - X_1\theta_1(\lambda')) = \mathbf{1}, 0 \to \theta_{j'} > 0 \\ \frac{1}{\lambda'} X_2^\top (y - X_1\theta_1(\lambda')) = -\mathbf{1}. \end{cases}$$

Optimality condition:

$$\begin{cases} (X_1^\top X_1)^{-1}(X_1^\top y - \lambda \text{sgn}[\theta_1(\lambda)]) = \theta_1(\lambda), \\ \frac{1}{\lambda} X_2^\top (y - X_1\theta_1(\lambda)) \in \partial \|\mathbf{0}\|_{\ell_1}. \end{cases}$$

$\forall \lambda' \in I_\lambda$, $\theta(\lambda') = (\theta_1(\lambda'), \mathbf{0})$, with

$$\theta_1(\lambda') = (X_1^\top X_1)^{-1}(X_1^\top y - \lambda' \text{sgn}[\theta_1(\lambda)]).$$

- Thus if the new $\lambda_2 \in I_\lambda$, done !
- If not? e.g. $\lambda_2 \geqslant \lambda_c := \min_j\{\lambda_j : \lambda_j > \lambda\}$.
  In $(\lambda_c, \lambda_c + \epsilon)$, only 1 entry changes **sparsity** wrt $\theta(\lambda)$!

Validity interval of λ

$$\begin{cases} \theta_1(\lambda') = \mathbf{0} \\ \frac{1}{\lambda'} X_2^\top (y - X_1\theta_1(\lambda')) = \mathbf{1}, \\ \frac{1}{\lambda'} X_2^\top (y - X_1\theta_1(\lambda')) = -\mathbf{1}, 0 \rightarrow \theta_{j'} < 0. \end{cases}$$

Optimality condition:

$$\begin{cases} (X_1^\top X_1)^{-1}(X_1^\top y - \lambda \mathrm{sgn}[\theta_1(\lambda)]) = \theta_1(\lambda), \\ \frac{1}{\lambda} X_2^\top (y - X_1\theta_1(\lambda)) \in \partial \|\mathbf{0}\|_{\ell_1}. \end{cases}$$

$\forall \lambda' \in I_\lambda$, $\theta(\lambda') = (\theta_1(\lambda'), \mathbf{0})$, with

$$\theta_1(\lambda') = (X_1^\top X_1)^{-1}(X_1^\top y - \lambda' \mathrm{sgn}[\theta_1(\lambda)]).$$

- Thus if the new $\lambda_2 \in I_\lambda$, done !
- If not? e.g. $\lambda_2 \geqslant \lambda_c := \min_j\{\lambda_j : \lambda_j > \lambda\}$.
  In $(\lambda_c, \lambda_c + \epsilon)$, only 1 entry changes **sparsity** wrt $\theta(\lambda)$!

Validity interval of λ

$$\begin{cases} \theta_1(\lambda') = \mathbf{0} \\ \frac{1}{\lambda'} X_2^\top (y - X_1\theta_1(\lambda')) = \mathbf{1}, \\ \frac{1}{\lambda'} X_2^\top (y - X_1\theta_1(\lambda')) = -\mathbf{1}, 0 \to \theta_{j'} < 0. \end{cases}$$

j' is known !

Optimality condition:

$$\begin{cases} (X_1^\top X_1)^{-1}(X_1^\top y - \lambda \text{sgn}[\theta_1(\lambda)]) = \theta_1(\lambda), \\ \frac{1}{\lambda} X_2^\top (y - X_1\theta_1(\lambda)) \in \partial \|\mathbf{0}\|_{\ell_1}. \end{cases}$$

- Thus if the new $\lambda_2 \in I_\lambda$, done !
- If not? e.g. $\lambda_2 \geqslant \lambda_c := \min_j\{\lambda_j : \lambda_j > \lambda\}$.

In $(\lambda_c, \lambda_c + \epsilon)$:

Optimality condition:

$$\begin{cases} (X_1^\top X_1)^{-1}(X_1^\top y - \lambda \mathrm{sgn}[\theta_1(\lambda)]) = \theta_1(\lambda), \\ \frac{1}{\lambda} X_2^\top (y - X_1\theta_1(\lambda)) \in \partial \|\mathbf{0}\|_{\ell_1}. \end{cases}$$

- Thus if the new $\lambda_2 \in I_\lambda$, done !
- If not? e.g. $\lambda_2 \geqslant \lambda_c := \min_j \{\lambda_j : \lambda_j > \lambda\}$.

  In $(\lambda_c, \lambda_c + \epsilon)$: new support and sign pattern are both known

Optimality condition:

$$\begin{cases} (X_1^\top X_1)^{-1}(X_1^\top y - \lambda \operatorname{sgn}[\theta_1(\lambda)]) = \theta_1(\lambda), \\ \frac{1}{\lambda} X_2^\top (y - X_1\theta_1(\lambda)) \in \partial\|\mathbf{0}\|_{\ell_1}. \end{cases}$$

- Thus if the new $\lambda_2 \in I_\lambda$, done !
- If not? e.g. $\lambda_2 \geqslant \lambda_c := \min_j\{\lambda_j : \lambda_j > \lambda\}$.

  In $(\lambda_c, \lambda_c + \epsilon)$: new support and sign pattern are both known $\Longrightarrow$ update the OC $\Longrightarrow$ new validity interval $(\lambda_c, ?)$ $\Longrightarrow$ until cover λ_2.

Optimality condition:

$$\begin{cases} (X_1^\top X_1)^{-1}(X_1^\top y - \lambda \text{sgn}[\theta_1(\lambda)]) = \theta_1(\lambda), \\ \frac{1}{\lambda} X_2^\top (y - X_1\theta_1(\lambda)) \in \partial \|\mathbf{0}\|_{\ell_1}. \end{cases}$$

- Thus if the new $\lambda_2 \in I_\lambda$, done !
- If not? e.g. $\lambda_2 \geqslant \lambda_c := \min_j \{\lambda_j : \lambda_j > \lambda\}$.

  In $(\lambda_c, \lambda_c + \epsilon)$: new support and sign pattern are both known $\Longrightarrow$ update the OC $\Longrightarrow$ new validity interval $(\lambda_c, ?)$ $\Longrightarrow$ until cover λ_2.

  Detail: The update of $(X_1^\top X_1)^{-1}$ is 1-rank, fast calculation exists !

Table of Contents

Compute optimality condition wrt structure

$$\widehat{A} := \arg\min_{A \in \mathcal{K}_{\mathcal{G}}} \frac{1}{2t} \sum_{\tau=1}^{t} \|x_\tau - A x_{\tau-1}\|_{\ell_2}^2 + \lambda \|A_{\mathrm{N}}\|_{\ell_1} ,$$

Compute optimality condition wrt structure

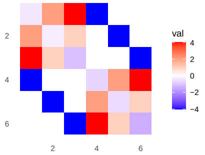
$$\widehat{A} := \arg\min_{A \in \mathcal{K}_{\mathcal{G}}} \frac{1}{2t} \sum_{\tau=1}^{t} \|x_\tau - Ax_{\tau-1}\|_{\ell_2}^2 + \lambda \|A_{\mathrm{N}}\|_{\ell_1},$$

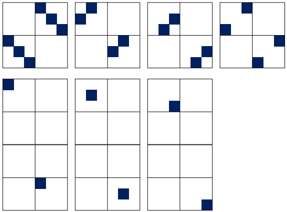
Key idea: Removing the explicit constraint using an orthonormal basis of $\mathcal{K}_{\mathcal{G}}$.

Compute optimality condition wrt structure

$$\widehat{A} := \arg\min_{A \in \mathcal{K}_{\mathcal{G}}} \frac{1}{2t} \sum_{\tau=1}^{t} \|x_\tau - A x_{\tau-1}\|_{\ell_2}^2 + \lambda \|A_{\mathrm{N}}\|_{\ell_1} ,$$

Key idea: Removing the explicit constraint using an orthonormal basis of $\mathcal{K}_{\mathcal{G}}$. $\mathcal{K}_{\mathcal{G}}$ is a linear space of dim = nb of *unique values.*



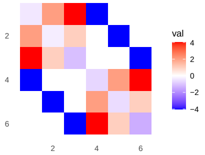


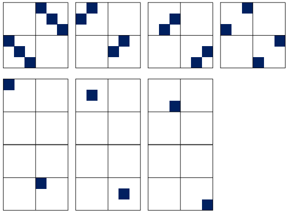
$\forall A \in \mathcal{K}_{\mathcal{G}}, \exists A^0 \in \mathbb{R}^{NF \times NF}, \ s.t. \operatorname{Proj}_{\mathcal{G}}(A^0) = \sum_{k \in K} \left\langle U_k, A^0 \right\rangle_{\mathbf{F}} U_k.$

Compute optimality condition wrt structure

$$\widehat{A} := \underset{A \in \mathcal{K}_{\mathcal{G}}}{\arg\min} \frac{1}{2t} \sum_{\tau=1}^{t} \|x_\tau - A x_{\tau-1}\|_{\ell_2}^2 + \lambda \|A_{\mathrm{N}}\|_{\ell_1} ,$$

Key idea: Removing the explicit constraint using an orthonormal basis of $\mathcal{K}_{\mathcal{G}}$. $\mathcal{K}_{\mathcal{G}}$ is a linear space of dim = nb of *unique values.*



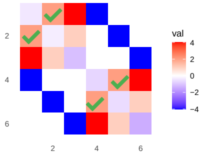


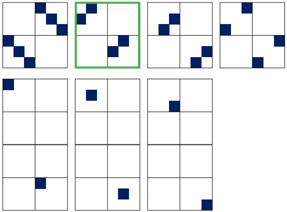
$\forall A \in \mathcal{K}_{\mathcal{G}}, \exists A^0 \in \mathbb{R}^{NF \times NF}$, $s.t.\, \mathrm{Proj}_{\mathcal{G}}(A^0) = \sum_{k \in K} \left\langle U_k, A^0 \right\rangle_{\mathbf{F}} U_k$.
Moreover, $A_k = c_k \left\langle U_k, A^0 \right\rangle_{\mathbf{F}}$: a unique corresponding U_k, known c.

Compute optimality condition wrt structure

$$\widehat{A} := \underset{A \in \mathcal{K}_{\mathcal{G}}}{\arg\min} \frac{1}{2t} \sum_{\tau=1}^{t} \| x_\tau - A x_{\tau-1} \|_{\ell_2}^2 + \lambda \| A_{\mathrm{N}} \|_{\ell_1} ,$$

Key idea: Removing the explicit constraint using an orthonormal basis of $\mathcal{K}_{\mathcal{G}}$. $\mathcal{K}_{\mathcal{G}}$ is a linear space of dim = nb of *unique values.*





$\forall A \in \mathcal{K}_{\mathcal{G}}, \exists A^0 \in \mathbb{R}^{NF \times NF}, \; s.t. \, \mathrm{Proj}_{\mathcal{G}}(A^0) = \sum_{k \in K} \left\langle U_k, A^0 \right\rangle_{\mathbf{F}} U_k.$

Moreover, $A_k = c_k \left\langle U_k, A^0 \right\rangle_{\mathbf{F}}$: a unique corresponding U_k, known c_k.

Compute optimality condition wrt structure

$$\widehat{A} := \arg\min_{A\in\mathcal{K}_{\mathcal{G}}} \frac{1}{2t}\sum_{\tau=1}^{t} \|x_\tau - Ax_{\tau-1}\|_{\ell_2}^2 + \lambda \|A_{\mathrm{N}}\|_{\ell_1},$$

Removed explicit constraint:

$$\widehat{A}^0 := \arg\min_{A^0} \frac{1}{2t}\sum_{\tau=1}^{t} \left\| x_\tau - \sum_{k\in K} \left\langle U_k, A^0 \right\rangle_{\mathbf{F}} U_k x_{\tau-1} \right\|_{\ell_2}^2 + \lambda \sum_{k\in K_{\mathrm{N}}} \left| \left\langle U_k, A^0 \right\rangle_{\mathbf{F}} U_k \right|,$$

where K_{N} corresponds to the values of A imposed penalty.

Compute optimality condition wrt structure

$$\widehat{A} := \arg\min_{A \in \mathcal{K}_{\mathcal{G}}} \frac{1}{2t} \sum_{\tau=1}^{t} \|x_\tau - A x_{\tau-1}\|_{\ell_2}^2 + \lambda \|A_{\mathrm{N}}\|_{\ell_1},$$

Removed explicit constraint:

$$\widehat{A}^0 := \arg\min_{A^0} \frac{1}{2t} \sum_{\tau=1}^{t} \left\| x_\tau - \sum_{k \in K} \left\langle U_k, A^0 \right\rangle_{\mathsf{F}} U_k x_{\tau-1} \right\|_{\ell_2}^2 + \lambda \sum_{k \in K_{\mathrm{N}}} \left| \left\langle U_k, A^0 \right\rangle_{\mathsf{F}} U_k \right|,$$

where K_{N} corresponds to the values of A imposed penalty.

$\widehat{A} = \mathrm{Proj}_{\mathcal{G}}(\widehat{A}^0)$. $\frac{\partial L}{\partial A^0}$ is easy.

Optimality condition

$$0 \in \frac{\partial L}{\partial A^0} = \sum_{k,k' \in K} \langle U_k, U_{k'} \widehat{\Gamma}_t(0) \rangle \langle U_{k'}, A^0 \rangle U_k - \sum_{k \in K} \langle U_k, \widehat{\Gamma}_t(1) \rangle U_k + \lambda \sum_{k \in K_{\mathrm{N}}} \partial \left| \langle U_k, A^0 \rangle \right| U_k,$$

where $\widehat{\Gamma}_t(0) = \sum x_{\tau-1} x_{\tau-1}^\top$, $\widehat{\Gamma}_t(1) = \sum x_{\tau-1} x_\tau^\top$.

Use $A_k = c_k \left\langle U_k, A^0 \right\rangle_{\mathbf{F}}$!

Optimality condition

$$\begin{cases} \left[\Gamma_1^*\right]^{-1}(\gamma_1^* - \lambda w) = a_1^*, \\ \frac{1}{\lambda}(\gamma_0 - \Gamma_0 a_1^*) \in \partial \|\mathbf{0}\|_{\ell_1}, \end{cases}$$

where

- $a_1^* = (A_{k_1}^*/c_{k_1}, \ldots, A_{k_L}^*/c_{k_L})$: all scaled **active** entries of A^*,

Optimality condition

$$\begin{cases} [\Gamma_1^*]^{-1}(\gamma_1^* - \lambda w) = a_1^*, \\ \frac{1}{\lambda}(\gamma_0 - \Gamma_0 a_1^*) \in \partial \|\mathbf{0}\|_{\ell_1}, \end{cases}$$

where

- $a_1^* = (A_{k_1}^*/c_{k_1}, \ldots, A_{k_L}^*/c_{k_L})$: all scaled **active** entries of A^*,
- Matrix $[\Gamma]_{k,k'} := \langle U_k, U_{k'}\widehat{\Gamma}_t(0)\rangle$, vector $[\gamma]_k := \langle U_k, \widehat{\Gamma}_t(1)\rangle$,

Optimality condition

$$\begin{cases} [\Gamma_1^*]^{-1}(\gamma_1^* - \lambda w) = a_1^*, \\ \frac{1}{\lambda}(\gamma_0 - \Gamma_0 a_1^*) \in \partial\|\mathbf{0}\|_{\ell_1}, \end{cases}$$

where

- $a_1^* = (A_{k_1}^*/c_{k_1}, ..., A_{k_L}^*/c_{k_L})$: all scaled **active** entries of A^*,
- Matrix $[\Gamma]_{k,k'} := \langle U_k, U_{k'}\widehat{\Gamma}_t(0)\rangle$, vector $[\gamma]_k := \langle U_k, \widehat{\Gamma}_t(1)\rangle$, $\Gamma_1^* = [\Gamma]_{K_1,K_1}, \Gamma_0 = [\Gamma]_{K_0,K_1}$; $\gamma_1^* = [\gamma]_{K_1}, \gamma_0 = [\gamma]_{K_0}$,
- $w = (w_{k_1}, ..., w_{k_L})$:
  $$\begin{cases} w_{k_l} = sgn(A_{k_l}^*), & \text{if } A_{k_l}^* \text{ is from } A_{\mathrm{N}}^* \text{ (Penalized param.)}, \\ 0, & o.w. \end{cases}$$

Optimality condition

$$\begin{cases} [\Gamma_1^*]^{-1}(\gamma_1^* - \lambda w) = a_1^*, \\ \frac{1}{\lambda}(\gamma_0 - \Gamma_0 a_1^*) \in \partial\|\mathbf{0}\|_{\ell_1}, \end{cases}$$

where

- $a_1^* = (A_{k_1}^*/c_{k_1}, ..., A_{k_L}^*/c_{k_L})$: all scaled **active** entries of A^*,
- Matrix $[\Gamma]_{k,k'} := \langle U_k, U_{k'}\widehat{\Gamma}_t(0)\rangle$, vector $[\gamma]_k := \langle U_k, \widehat{\Gamma}_t(1)\rangle$, $\Gamma_1^* = [\Gamma]_{K_1,K_1}, \Gamma_0 = [\Gamma]_{K_0,K_1}$; $\gamma_1^* = [\gamma]_{K_1}, \gamma_0 = [\gamma]_{K_0}$,
- $w = (w_{k_1}, ..., w_{k_L})$:

$$\begin{cases} w_{k_l} = sgn(A_{k_l}^*), & \textit{if } A_{k_l}^* \text{ is from } A_{\mathrm{N}}^* \text{ (Penalized param.)}, \\ 0, & o.w. \end{cases}$$

The two LHS of the OC are **smooth** functions in λ.

Optimality condition

$$\begin{cases} [\Gamma_1^*]^{-1} (\gamma_1^* - \lambda w) = a_1^*, \\ \frac{1}{\lambda}(\gamma_0 - \Gamma_0 a_1^*) \in \partial \|\mathbf{0}\|_{\ell_1}, \end{cases}$$

where

- $a_1^* = (A_{k_1}^*/c_{k_1}, ..., A_{k_L}^*/c_{k_L})$: all scaled **active** entries of A^*,
- Matrix $[\Gamma]_{k,k'} := \langle U_k, U_{k'}\widehat{\Gamma}_t(0)\rangle$, vector $[\gamma]_k := \langle U_k, \widehat{\Gamma}_t(1)\rangle$, $\Gamma_1^* = [\Gamma]_{K_1,K_1}, \Gamma_0 = [\Gamma]_{K_0,K_1}$; $\gamma_1^* = [\gamma]_{K_1}, \gamma_0 = [\gamma]_{K_0}$,
- $w = (w_{k_1}, ..., w_{k_L})$:
  $$\begin{cases} w_{k_l} = sgn(A_{k_l}^*), & \text{if } A_{k_l}^* \text{ is from } A_{\mathrm{N}}^* \text{ (Penalized param.)}, \\ 0, & o.w. \end{cases}$$

The two LHS of the OC are **smooth** functions in λ.

Define $a_1(\lambda') = [\Gamma_1^*]^{-1} (\gamma_1^* - \lambda' w)$. Same principle works !

Table of Contents

Data path:

$$\min_{\theta \in \mathbb{R}^d} \frac{1}{2}\|y - X\theta\|_{\ell_2}^2 + \frac{1}{2}(\mu y_{t+1} - \mu x_{t+1}^\top \theta)^2 + (t+1)\lambda\|\theta\|_{\ell_1},$$

The two LHS of OC are piece-wise **smooth** functions of $\mu, \mu \in [0,1]$, see Garrigues and Ghaoui (2008). Same principle !

$$\min_{A \in \mathcal{K}_{\mathcal{G}}} \frac{1}{2t} \sum_{\tau=1}^{t} \|x_\tau - A x_{\tau-1}\|_{\ell_2}^2 + (t+1)\lambda \|A_{\mathrm{N}}\|_{\ell_1} + \frac{1}{2} \sum_{i=1}^{NF} \mu_i \left(x_{t+1,i} - A_{i,:} x_t\right)^2$$

Always smooth in μ_i !

Data path:

$$\min_{\theta \in \mathbb{R}^d} \frac{1}{2}\|y - X\theta\|_{\ell_2}^2 + \frac{1}{2}(\mu y_{t+1} - \mu x_{t+1}^\top \theta)^2 + (t+1)\lambda\|\theta\|_{\ell_1},$$

The two LHS of OC are piece-wise **smooth** functions of $\mu, \mu \in [0,1]$, see Garrigues and Ghaoui (2008). Same principle !

$$\min_{A \in \mathcal{K}_\mathcal{G}} \frac{1}{2t} \sum_{\tau=1}^{t} \|x_\tau - A x_{\tau-1}\|_{\ell_2}^2 + (t+1)\lambda \|A_\mathrm{N}\|_{\ell_1} + \frac{1}{2} \sum_{i=1}^{NF} \color{red}{\mu_i} \left(x_{t+1,i} - A_{i,:} x_t\right)^2$$

Always smooth in μ_i !

Details: if only one μ is used, the update of old OC is NF-rank, no fast calculation !

Table of Contents

Augmented model

Matrix autoregressive model, $\mathbb{E}x_t = 0$,:

$$x_t = Ax_{t-1} + z_t, \text{ with } A \in \mathcal{K}_{\mathcal{G}}, \ x_t = \text{vec}(X_t),$$

Augmented model

Matrix autoregressive model, $\mathbb{E}x_t = 0$,:

$$x_t = Ax_{t-1} + z_t,\ \text{with } A \in \mathcal{K}_{\mathcal{G}},\ x_t = \text{vec}(X_t),$$

Matrix autoregressive model:

$$x_t = b_t + Ax_{t-1} + z_t,\ \text{with } A \in \mathcal{K}_{\mathcal{G}},\ b_t \text{ periodic, e.g. } 12(\text{month}).$$

Augmented model

Matrix autoregressive model, $\mathbb{E}x_t = 0$,:

$$x_t = Ax_{t-1} + z_t, \text{ with } A \in \mathcal{K}_{\mathcal{G}}, \ x_t = \text{vec}(X_t),$$

Matrix autoregressive model:

$$x_t = b_t + Ax_{t-1} + z_t, \text{ with } A \in \mathcal{K}_{\mathcal{G}}, \ b_t \text{ periodic, e.g. } 12(\text{month}).$$

Estimation:

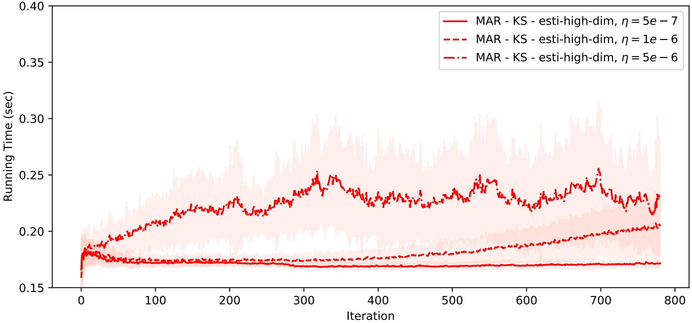
$$\widehat{A}, \widehat{b}_\tau := \underset{A \in \mathcal{K}_{\mathcal{G}}, b_\tau}{\arg\min} \frac{1}{2t} \sum_{\tau=1}^{t} \|x_\tau - b_\tau - Ax_{\tau-1}\|_{\ell_2}^2 + \lambda \|A_{\mathrm{N}}\|_{\ell_1}$$

Adapted algorithms !

Table of Contents

Table of Contents

Running time



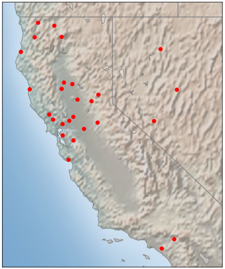
$N = 20$, $F = 5$, $M = 12$, number of model parameters $= 1500$. The accelerated proximal gradient descent needs more than **3 secs**.

Other simulation results

	Existing MAR	Proposed MAR
Prediction performance		✓ (KS-based formula)
Availability for small dataset		✓ (Lasso penalty)
Applicable in graph learning		✓ (KS + Lasso penalty)
Online inference		✓ (Homotopy algorithms)

Table of Contents

California weather TS



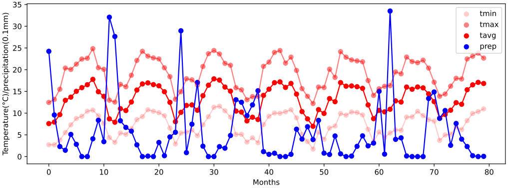


Figure 1: N = 27 (stations), F= 4 (weather metrics), M = 12 (months), T = 1523 (months).

Proposed model:

$$x_t = b_t + A x_{t-1} + z_t, \text{ with } A \in \mathcal{K}_{\mathcal{G}}(A_{\mathrm{N}}, A_{\mathrm{F}}), \ b_t \text{ periodic in t},$$

where $x_t = \mathrm{vec}(X_t)$, with X_t : matrix of 27×4.

Learned graph A_{N}

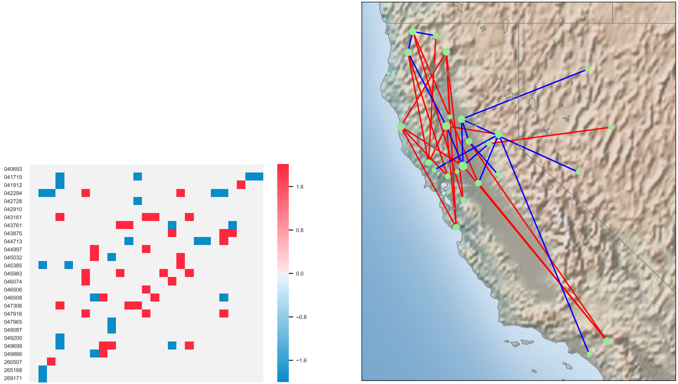


Figure 2: *Dependency between stations.* Estimation of A_{N} (left) using all $\mathrm{T} = 1523$ (months), retrieved graph overlapped on the map (right).

Learned graph A_{F}

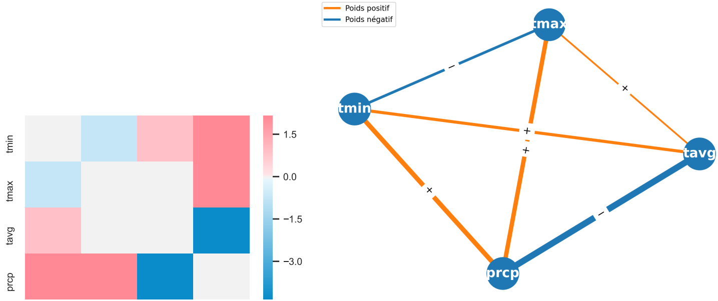


Figure 3: *Dependency between weather metrics.* Estimation of A_{F} (left) using all $\mathrm{T} = 1523$ (months), retrieved graph overlapped on the map (right).

Table of Contents

Contribution:

- A new MAR model and the estimation methods (offline/online).

[1] Y.Jiang and J.Bigot, Wasserstein auto-regressive models for modeling multivariate distributional time series, *to appear, Journal of time series analysis*.

Contribution:

- A new MAR model and the estimation methods (offline/online).
- Online graph learning (GL) from matrix-variate TS, which is a data type not yet considered in the GL.

[1] Y.Jiang and J.Bigot, Wasserstein auto-regressive models for modeling multivariate distributional time series, *to appear, Journal of time series analysis*.

Contribution:

- A new MAR model and the estimation methods (offline/online).
- Online graph learning (GL) from matrix-variate TS, which is a data type not yet considered in the GL.
- New homotopy algorithms. **Online inference:**

$$\begin{aligned} \text{regularization path} &: A(t, \lambda_1) \rightarrow A(t, \lambda_2), \\ \text{data path} &: A(t, \lambda) \rightarrow A(t+1, \lambda), \\ \text{automatic tuning} &: \lambda_t \rightarrow \lambda_{t+1}, \end{aligned}$$

[1] Y.Jiang and J.Bigot, Wasserstein auto-regressive models for modeling multivariate distributional time series, *to appear, Journal of time series analysis*.

Contribution:

- A new MAR model and the estimation methods (offline/online).
- Online graph learning (GL) from matrix-variate TS, which is a data type not yet considered in the GL.
- New homotopy algorithms. **Online inference:**

$$\begin{aligned} \text{regularization path} &: A(t, \lambda_1) \rightarrow A(t, \lambda_2), \\ \text{data path} &: A(t, \lambda) \rightarrow A(t+1, \lambda), \\ \text{automatic tuning} &: \lambda_t \rightarrow \lambda_{t+1}, \end{aligned}$$

Perspectives:

GL from TS of other data natures, e.g. TS of probability measures[1], mixed types.

[1] Y.Jiang and J.Bigot, Wasserstein auto-regressive models for modeling multivariate distributional time series, *to appear, Journal of time series analysis*.

Thank your for your attention !

Jiang Yiye, Jérémie Bigot, and Sofian Maabout. "Online graph topology learning from matrix-valued time series." Computational Statistics & Data Analysis 202 (2025): 108065.

Offline optimization

$$
\begin{aligned}
A^{k+1} &= \mathrm{prox}(A^k - \eta^k \nabla f(A^k)), \\
&= \arg\min_{A \in \mathcal{K}_{\mathcal{G}}} \frac{1}{2\eta^k} \left\| A - \left(A^k - \eta^k \nabla f(A^k)\right) \right\|_{\ell_2}^2 + \lambda \left\| A_{\mathrm{N}} \right\|_{\ell_1} \\
&= \arg\min_{A \in \mathcal{K}_{\mathcal{G}}} \frac{1}{2\eta^k} \left\| A - \mathrm{Proj}_{\mathcal{G}} \left(A^k - \eta^k \nabla f(A^k)\right) \right\|_{\ell_2}^2 + \lambda \left\| A_{\mathrm{N}} \right\|_{\ell_1}
\end{aligned}
$$

$$
\Longleftrightarrow \begin{cases}
A_{\mathrm{N}}^{k+1} = \arg\min_{A_{\mathrm{N}}} \left\| A_{\mathrm{N}} - \mathrm{Proj}_{\mathcal{G}_{\mathrm{N}}} \left(A^k - \eta^k \nabla f(A^k)\right) \right\|_{\ell_2}^2 \\
\qquad\qquad + 2\eta^k \frac{\lambda}{F} \left\| A_{\mathrm{N}} \right\|_{\ell_1}, \\
A_{\mathrm{F}}^{k+1} = \mathrm{Proj}_{\mathcal{G}_{\mathrm{F}}} \left(A^k - \eta^k \nabla f(A^k)\right), \\
\mathrm{diag}(A^{k+1}) = \left(A^k - \eta^k \nabla f(A^k)\right),
\end{cases}
$$

Adaptive tuning of λ

Introduce the empirical objective function (Monti et al., 2018):

$$f_{t+1}(\lambda) = \frac{1}{2}\|x_{t+1} - A(t,\lambda)x_t\|_{\ell_2}^2.$$

Updating rule:

$$\lambda_{t+1} = \lambda_t - \eta \frac{\mathrm{d}f_{t+1}(\lambda)}{\mathrm{d}\lambda}\Big|_{\lambda=\lambda_t},$$

where η is the step size.

Comparison with other AR models

The existing MAR models are all bi-multiplication / Kronecker product based, with the first model proposed in Chen et al. 2021 [1]:

$$X_t = A_{\mathrm{N}} X_{t-1} A_{\mathrm{F}}^{\top} + Z_t \iff \mathrm{vec}(X_t) = (A_{\mathrm{F}} \otimes A_{\mathrm{N}})\mathrm{vec}(X_{t-1}) + \mathrm{vec}(Z_t) \qquad (1)$$

[1] Chen, Rong, Han Xiao, and Dan Yang. "Autoregressive models for matrix-valued time series." Journal of Econometrics 222.1 (2021): 539-560.

Comparison with other AR models

The existing MAR models are all bi-multiplication / Kronecker product based, with the first model proposed in Chen et al. 2021 [1]:

$$X_t = A_{\mathrm{N}} X_{t-1} A_{\mathrm{F}}^{\top} + Z_t \iff \mathrm{vec}(X_t) = (A_{\mathrm{F}} \otimes A_{\mathrm{N}}) \mathrm{vec}(X_{t-1}) + \mathrm{vec}(Z_t) \qquad (1)$$

Competitors: 3 estimators in Chen et al. 2021, VAR(1) with LS estimator.

[1] Chen, Rong, Han Xiao, and Dan Yang. "Autoregressive models for matrix-valued time series." Journal of Econometrics 222.1 (2021): 539-560.

Comparison with other AR models

The existing MAR models are all bi-multiplication / Kronecker product based, with the first model proposed in Chen et al. 2021 [1]:

$$X_t = A_{\mathrm{N}} X_{t-1} A_{\mathrm{F}}^\top + Z_t \iff \mathrm{vec}(X_t) = (A_{\mathrm{F}} \otimes A_{\mathrm{N}})\mathrm{vec}(X_{t-1}) + \mathrm{vec}(Z_t) \tag{1}$$

Competitors: 3 estimators in Chen et al. 2021, VAR(1) with LS estimator.

- Online procedure
  - for us, apply directly on the TS as previously.
  - for VAR and MAR in (1), offline detrending + resolving batch pb at each time step

[1] Chen, Rong, Han Xiao, and Dan Yang. "Autoregressive models for matrix-valued time series." Journal of Econometrics 222.1 (2021): 539-560.

Pierre Garrigues and Laurent Ghaoui. An homotopy algorithm for the lasso with online observations. *Advances in neural information processing systems*, 21:489–496, 2008.

Ricardo P Monti, Christoforos Anagnostopoulos, and Giovanni Montana. Adaptive regularization for lasso models in the context of nonstationary data streams. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 11(5): 237–247, 2018.