Statistical analysis of spatio-temporal and multi-dimensional data from a network of sensors

Yiye JIANG

Université de Bordeaux
Institut de Mathématiques de Bordeaux

Data recorded over a network of sensors such as traffic analysis, brain network analysis, social network, and citation network.

Data recorded over a network of sensors such as traffic analysis, brain network analysis, social network, and citation network.

Data we consider in this thesis:

- Spatio-temporal: Observations along time per sensor (node).

Data recorded over a network of sensors such as traffic analysis, brain network analysis, social network, and citation network.

Data we consider in this thesis:

- Spatio-temporal: Observations along time per sensor (node).
- 3 diverse forms: Observation per sensor (node) per time is scalar/vector/distribution.

Data illustration: scalar observation

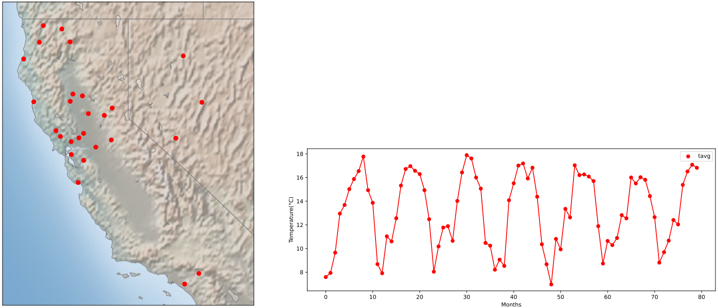


Figure 1: *Monthly climatological records of weather stations in California.* A value $x_{it} \in \mathbb{R}$ is recorded on each station (sensor/node) i, at each time t. In this example, x_{it} is the average temperature.

Data illustration: vectorial observation

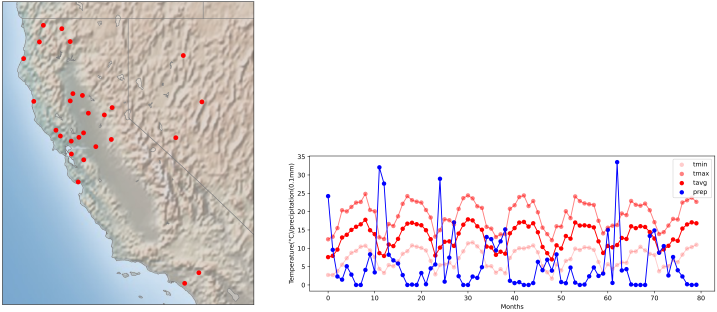


Figure 2: *Monthly climatological records of weather stations in California.* A vector $\mathbf{x}_{it} \in \mathbb{R}^4$ is recorded on each station (sensor/node) i, at each time t. In this example, $\mathbf{x}_{it}$ is the vector of: min/max/avg temperature, and precipitation.

Data illustration: distributional observation

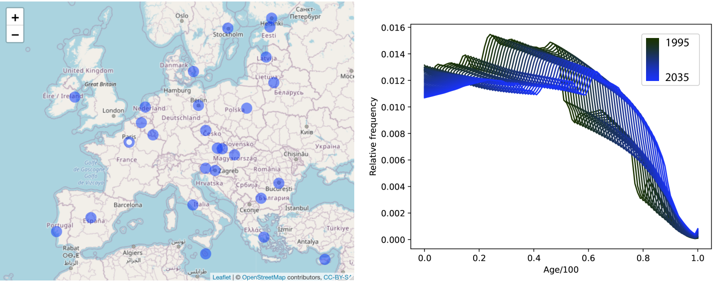


Figure 3: *Annual records of age distributions of EU countries.* A distribution $\boldsymbol{\mu}_{it} \in \mathcal{P}([0,1])$ is recorded on each node i, at each time t. In this example, $\boldsymbol{\mu}_{it}$ is an age distribution. Time is represented by color instead of x-axis. Lighter curves correspond to the distributions from more recent years.

A network of sensor $\rightarrow$ a collection of random processes indexed by nodes.

A network of sensor $\rightarrow$ a collection of random processes indexed by nodes.

- Main goal: identifying the dependency structure between these random processes

A network of sensor $\rightarrow$ a collection of random processes indexed by nodes.

- Main goal: identifying the dependency structure between these random processes $\Longrightarrow$ *graph learning*

A network of sensor $\rightarrow$ a collection of random processes indexed by nodes.

- Main goal: identifying the dependency structure between these random processes $\Longrightarrow$ *graph learning* from vectorial $(\mathbf{x}_{it})_t$ and distributional $(\boldsymbol{\mu}_{it})_t$ processes.

A network of sensor $\rightarrow$ a collection of random processes indexed by nodes.

- Main goal: identifying the dependency structure between these random processes $\Longrightarrow$ *graph learning* from vectorial $(\mathbf{x}_{it})_t$ and distributional $(\boldsymbol{\mu}_{it})_t$ processes.
- Second goal: understanding the predictability of each process, for scalar $\boldsymbol{x}_{it}$ data.

A network of sensor $\rightarrow$ a collection of random processes indexed by nodes.

- Main goal: identifying the dependency structure between these random processes $\Longrightarrow$ *graph learning* from vectorial $(\mathbf{x}_{it})_t$ and distributional $(\boldsymbol{\mu}_{it})_t$ processes.
- Second goal: understanding the predictability of each process, for scalar $\boldsymbol{x}_{it}$ data. $\longrightarrow$ the details are not considered in this presentation.

Causal graph and vector auto-regressive model

For $(x_{it})_t \in \mathbb{R}$, $i = 1, \ldots, N$, the **VAR Models** have been widely adapted in literature to learn their causality (Granger) dependency.

$$\text{VAR}(1): \; \mathbf{x}_t - \mathsf{u} = A(\mathbf{x}_{t-1} - \mathsf{u}) + \mathbf{z}_t, \tag{1}$$

where $\mathbf{x}_t = (x_{1t}, \ldots, x_{Nt})$, $\mathsf{u} = \mathbb{E}\mathbf{x}_t$, and $\mathbf{z}_t$ is white noise.

When VAR (1) is stationary, the sparsity structure of $A \overset{\text{adj. mat.}}{\Longleftrightarrow} \mathcal{G}$.

Contribution of the thesis:

$$\mathcal{G} \text{ of } (x_{it})_t \in \mathbb{R} \to \begin{cases} \mathcal{G} \text{ of } (\mathbf{x}_{it})_t \in \mathbb{R}^F + \text{online inference}, \\ \mathcal{G} \text{ of } (\boldsymbol{\mu}_{it})_t \in \mathcal{W}_2(\mathbb{R}). \end{cases}$$

Table of Contents

Kronecker sum, causal product graph and Matrix AR

$$(\mathbf{x}_{it})_t \in \mathbb{R}^F,\ i = 1, \dots, N \iff (\mathbf{X}_t)_t \in \mathbb{R}^{N \times F},$$

Kronecker sum, causal product graph and Matrix AR

$$(\mathbf{x}_{it})_t \in \mathbb{R}^F,\ i = 1, \ldots, N \iff (\mathbf{X}_t)_t \in \mathbb{R}^{N \times F},$$

where $\mathbf{X}_t$'s *row* $\sim$ *spatial (node-wise)* dim, *col* $\sim$ *feature* dim.

Kronecker sum, causal product graph and Matrix AR

$$(\mathbf{x}_{it})_t \in \mathbb{R}^F,\ i = 1, \dots, N \iff (\mathbf{X}_t)_t \in \mathbb{R}^{N \times F},$$

where $\mathbf{X}_t$'s *row* $\sim$ *spatial (node-wise)* dim, *col* $\sim$ *feature* dim.

We propose the matrix-variate AR for $(\mathbf{X}_t)_t$:

$$\mathbf{x}_t - \mathsf{u} = A(\mathbf{x}_{t-1} - \mathsf{u}) + \mathbf{z}_t,$$

where $\mathbf{x}_t = \mathrm{vec}(\mathbf{X}_t) = (\boldsymbol{x}_{ift})_{i,f}$,

Kronecker sum, causal product graph and Matrix AR

$$(\mathbf{x}_{it})_t \in \mathbb{R}^F,\ i = 1, \dots, N \iff (\mathbf{X}_t)_t \in \mathbb{R}^{N \times F},$$

where $\mathbf{X}_t$'s *row* ∼ *spatial (node-wise)* sim, *col* ∼ *feature* dim.

We propose the matrix-variate AR for $(\mathbf{X}_t)_t$:

$$\mathbf{x}_t - \mathsf{u} = A(\mathbf{x}_{t-1} - \mathsf{u}) + \mathbf{z}_t, \text{ with } A = A_{\mathrm{F}} \oplus A_{\mathrm{N}}, \qquad (1')$$

where $\mathbf{x}_t = \mathrm{vec}(\mathbf{X}_t) = (\boldsymbol{x}_{ift})_{i,f}$, $A_{\mathrm{N}} \in \mathbb{R}^{N \times N}$, $A_{\mathrm{F}} \in \mathbb{R}^{F \times F}$, and

$$A_{\mathrm{F}} \oplus A_{\mathrm{N}} := A_{\mathrm{F}} \otimes I_N + I_F \otimes A_{\mathrm{N}}.$$

Kronecker sum, causal product graph and Matrix AR

$$(\mathbf{x}_{it})_t \in \mathbb{R}^F,\ i = 1, \ldots, N \iff (\mathbf{X}_t)_t \in \mathbb{R}^{N \times F},$$

where $\mathbf{X}_t$'s *row* $\sim$ *spatial (node-wise)* sim, *col* $\sim$ *feature* dim.

We propose the matrix-variate AR for $(\mathbf{X}_t)_t$:

$$\mathbf{x}_t - \mathsf{u} = A(\mathbf{x}_{t-1} - \mathsf{u}) + \mathbf{z}_t, \text{ with } A = A_{\mathrm{F}} \oplus A_{\mathrm{N}}, \qquad (1')$$

where $\mathbf{x}_t = \mathrm{vec}(\mathbf{X}_t) = (\boldsymbol{x}_{ift})_{i,f}$, $A_{\mathrm{N}} \in \mathbb{R}^{N \times N}$, $A_{\mathrm{F}} \in \mathbb{R}^{F \times F}$, and

$$A_{\mathrm{F}} \oplus A_{\mathrm{N}} := A_{\mathrm{F}} \otimes I_N + I_F \otimes A_{\mathrm{N}}.$$

KS endows the matrix representation of the vector Model $(1')$:

$$\mathbf{X}_t - \mathsf{U} = A_{\mathrm{N}}(\mathbf{X}_{t-1} - \mathsf{U}) + (\mathbf{X}_{t-1} - \mathsf{U})A_{\mathrm{F}}^\top + \mathbf{Z}_t,$$

$A_{\mathrm{N}} \sim$ spatial dependency, $A_{\mathrm{F}} \sim$ feature dependency.

Kronecker sum, causal product graph and Matrix AR

Moreover, the vector representation implies

$$A \overset{\text{adj. mat.}}{\Longleftrightarrow} \mathcal{G} \text{ of } (\boldsymbol{x}_{ift})_t,$$

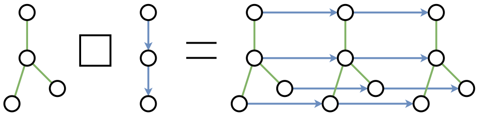
Kronecker sum, causal product graph and Matrix AR

Moreover, the vector representation implies

$$A \overset{\text{adj. mat.}}{\Longleftrightarrow} \mathcal{G} \text{ of } (\boldsymbol{x}_{ift})_t,$$

then the KS structure in A furthermore implies

$$\mathcal{G} = \mathcal{G}_N \square \mathcal{G}_F, \text{ where } \mathcal{G}_N \overset{\text{adj. mat.}}{\Longleftrightarrow} A_\mathrm{N} \text{ and } \mathcal{G}_F \overset{\text{adj. mat.}}{\Longleftrightarrow} A_\mathrm{F}.$$



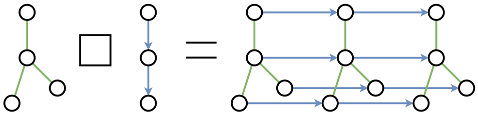
Kronecker sum, causal product graph and Matrix AR

Moreover, the vector representation implies

$$A \overset{\text{adj. mat.}}{\Longleftrightarrow} \mathcal{G} \text{ of } (\boldsymbol{x}_{ift})_t,$$

then the KS structure in A furthermore implies

$$\mathcal{G} = \mathcal{G}_N \square \mathcal{G}_F, \text{ where } \mathcal{G}_N \overset{\text{adj. mat.}}{\Longleftrightarrow} A_{\mathrm{N}} \text{ and } \mathcal{G}_F \overset{\text{adj. mat.}}{\Longleftrightarrow} A_{\mathrm{F}}.$$



Cartesian product of subgraphs. Subgraphs are **retained** in every section of the other dimension.

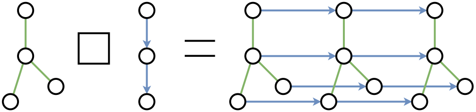
Kronecker sum, causal product graph and Matrix AR

Moreover, the vector representation implies

$$A \overset{\text{adj. mat.}}{\Longleftrightarrow} \mathcal{G} \text{ of } (\boldsymbol{x}_{ift})_t,$$

then the KS structure in A furthermore implies

$$\mathcal{G} = \mathcal{G}_N \square \mathcal{G}_F, \ \text{where } \mathcal{G}_N \overset{\text{adj. mat.}}{\Longleftrightarrow} A_\text{N} \text{ and } \mathcal{G}_F \overset{\text{adj. mat.}}{\Longleftrightarrow} A_\text{F}.$$



Cartesian product of subgraphs. Subgraphs are **retained** in every section of the other dimension. For nodes on right as $(\boldsymbol{x}_{ift})_t$, $\forall$ fixed f, Subgraph of $(\boldsymbol{x}_{ift})_t = \mathcal{G}_N$, $\forall$ fixed i, Subgraph of $(\boldsymbol{x}_{ift})_t = \mathcal{G}_F$.

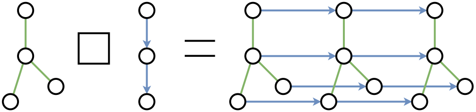
Kronecker sum, causal product graph and Matrix AR

Moreover, the vector representation implies

$$A \overset{\text{adj. mat.}}{\Longleftrightarrow} \mathcal{G} \text{ of } (\boldsymbol{x}_{ift})_t,$$

then the KS structure in A furthermore implies

$$\mathcal{G} = \mathcal{G}_N \square \mathcal{G}_F, \text{ where } \mathcal{G}_N \overset{\text{adj. mat.}}{\Longleftrightarrow} A_{\mathrm{N}} \text{ and } \mathcal{G}_F \overset{\text{adj. mat.}}{\Longleftrightarrow} A_{\mathrm{F}}.$$



Cartesian product of subgraphs. Subgraphs are **retained** in every section of the other dimension. For nodes on right as $(\boldsymbol{x}_{ift})_t$, $\forall$ fixed f, Subgraph of $(\boldsymbol{x}_{ift})_t = \mathcal{G}_N$, $\forall$ fixed i, Subgraph of $(\boldsymbol{x}_{ift})_t = \mathcal{G}_F$.

$\mathcal{G}_N$ = spatial graph of $(\mathbf{x}_{it})_t$, $\mathcal{G}_F$ = feature graph.

Constraint set $\mathcal{K}_{\mathcal{G}}$

Due to the model identifiability and application reasons, we employ a more sophisticated structure for A. The complete MAR(1) is

$$\mathbf{x}_t = A\mathbf{x}_{t-1} + \mathbf{z}_t, \; A \in \mathcal{K}_{\mathcal{G}},$$

where $\mathbf{x}_t = \text{vec}\left(\mathbf{X}_t\right)$, and

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

Constraint set $\mathcal{K}_{\mathcal{G}}$

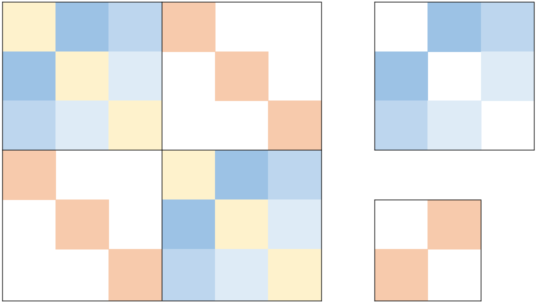


Figure 4: $\mathcal{K}_{\mathcal{G}}$ *for* $N = 3$, $F = 2$. M (left), M_{N} (right upper), M_{F} (right bottom).

Matrix AR(1)

The complete MAR(1) is

$$\mathbf{x}_t = A\mathbf{x}_{t-1} + \mathbf{z}_t,\ A \in \mathcal{K}_{\mathcal{G}},$$

where $\mathbf{x}_t = \mathrm{vec}\left(\mathbf{X}_t\right)$, we assume $\mathbb{E}\mathbf{x}_t = 0$ for now, and

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

Matrix AR(1)

The complete MAR(1) is

$$\mathbf{x}_t = A\mathbf{x}_{t-1} + \mathbf{z}_t, \ A \in \mathcal{K}_{\mathcal{G}},$$

where $\mathbf{x}_t = \text{vec}\,(\mathbf{X}_t)$, we assume $\mathbb{E}\mathbf{x}_t = 0$ for now, and

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

$\mathbf{z}_t \in \mathbb{R}^{NF} \sim \mathrm{IID}\,(0, \Sigma)$ is white noise with a non-singular covariance structure Σ and bounded fourth moments, with $\|A\|_2 < 1$.

Sparse estimators of A_{N}

MAR(1):

$$\mathbf{x}_t = A\mathbf{x}_{t-1} + \mathbf{z}_t,\ A \in \mathcal{K}_{\mathcal{G}}.$$

Sparse estimators of A_{N}

MAR(1):

$$\mathbf{x}_t = A\mathbf{x}_{t-1} + \mathbf{z}_t,\ A \in \mathcal{K}_{\mathcal{G}}.$$

- In low dimension: $\widehat{A}_{OLS}$ projected onto linear subspace $\mathcal{K}_{\mathcal{G}}$. CLT → nullity test on $\widehat{A}_N$.

Sparse estimators of A_{N}

MAR(1):

$$\mathbf{x}_t = A\mathbf{x}_{t-1} + \mathbf{z}_t, \; A \in \mathcal{K}_{\mathcal{G}}.$$

- In low dimension: $\widehat{A}_{OLS}$ projected onto linear subspace $\mathcal{K}_{\mathcal{G}}$. CLT $\rightarrow$ nullity test on $\widehat{A}_N$.
- In high dimension, we propose the novel Lasso

$$\mathbf{A}(t,\lambda) = \underset{A \in \mathcal{K}_{\mathcal{G}}}{\arg\min} \frac{1}{2t} \sum_{\tau=1}^{t} \|\mathbf{x}_\tau - A\mathbf{x}_{\tau-1}\|_{\ell_2}^2 + \lambda_t F \|A_{\mathrm{N}}\|_{\ell_1},$$

Sparse estimators of A_{N}

MAR(1):

$$\mathbf{x}_t = A\mathbf{x}_{t-1} + \mathbf{z}_t,\ A \in \mathcal{K}_{\mathcal{G}}.$$

- In low dimension: $\widehat{A}_{OLS}$ projected onto linear subspace $\mathcal{K}_{\mathcal{G}}$. CLT → nullity test on $\widehat{A}_N$.
- In high dimension, we propose the novel Lasso

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

  - Off-line: for example, the proximal gradient descend,

Sparse estimators of A_{N}

MAR(1):

$$\mathbf{x}_t = A\mathbf{x}_{t-1} + \mathbf{z}_t, \ A \in \mathcal{K}_{\mathcal{G}}.$$

- In low dimension: $\widehat{A}_{OLS}$ projected onto linear subspace $\mathcal{K}_{\mathcal{G}}$. CLT → nullity test on $\widehat{A}_N$.
- In high dimension, we propose the novel Lasso

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

  - Off-line: for example, the proximal gradient descend,
  - Online: Given $\mathbf{x}_{t+1}$, $\mathbf{A}(t, \lambda_t) \to \mathbf{A}(t+1, \lambda_{t+1})$.

Sparse estimators of A_{N}

MAR(1):

$$\mathbf{x}_t = A\mathbf{x}_{t-1} + \mathbf{z}_t,\ A \in \mathcal{K}_{\mathcal{G}}.$$

- In low dimension: $\widehat{A}_{OLS}$ projected onto linear subspace $\mathcal{K}_{\mathcal{G}}$. CLT → nullity test on $\widehat{A}_N$.
- In high dimension, we propose the novel Lasso

$$\mathbf{A}(t,\lambda) = \underset{A\in\mathcal{K}_{\mathcal{G}}}{\arg\min} \frac{1}{2t}\sum_{\tau=1}^{t} \|\mathbf{x}_\tau - A\mathbf{x}_{\tau-1}\|_{\ell_2}^2 + \lambda_t F \|A_{\mathrm{N}}\|_{\ell_1},$$

  - Off-line: for example, the proximal gradient descend,
  - Online: Given $\mathbf{x}_{t+1}$, $\mathbf{A}(t,\lambda_t) \to \mathbf{A}(t+1,\lambda_{t+1})$.

$\lambda_t \to \lambda_{t+1}$, $\mathbf{A}(t,\lambda_t) \to \mathbf{A}(t,\lambda_{t+1})$, $\mathbf{A}(t,\lambda_{t+1}) \to \mathbf{A}(t+1,\lambda_{t+1})$.

Homotopy algorithms and optimality conditions

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

Algo: $\boldsymbol{\theta}^*(\lambda_1) \rightarrow \boldsymbol{\theta}^*(\lambda_2)$ relies on **Optimality condition** of minimizer $\boldsymbol{\theta}^*$:

$$\frac{\partial L(\theta)}{\partial \theta} = 0 \iff \mathbf{X}^\top(\mathbf{X}\boldsymbol{\theta}^* - \mathbf{y}) + \lambda\mathbf{w} = 0,\ \mathbf{w} = \partial\|\boldsymbol{\theta}^*\|_{\ell_1}.$$

Homotopy algorithms and optimality conditions

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

Algo: $\boldsymbol{\theta}^*(\lambda_1) \to \boldsymbol{\theta}^*(\lambda_2)$ relies on **Optimality condition** of minimizer $\boldsymbol{\theta}^*$:

$$\frac{\partial L(\theta)}{\partial \theta} = 0 \iff \mathbf{X}^\top(\mathbf{X}\boldsymbol{\theta}^* - \mathbf{y}) + \lambda \mathbf{w} = 0,\ \mathbf{w} = \partial\|\boldsymbol{\theta}^*\|_{\ell_1}.$$

Unique $\boldsymbol{\theta}^* = (\boldsymbol{\theta}_1^*, 0)$ at λ, $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_0)$, $\mathbf{w} = (\mathbf{w}_1, \mathbf{w}_0)$:

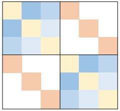
$$\begin{cases} \boldsymbol{\theta}_1^* = (\mathbf{X}_1^\top \mathbf{X}_1)^{-1}(\mathbf{X}_1^\top \mathbf{y} - \lambda \mathbf{w}_1), \\ \lambda \mathbf{w}_0 = \mathbf{y} - \mathbf{X}_0^\top \mathbf{X}_1 \boldsymbol{\theta}_1^*. \end{cases}$$

Continuity $\implies$ (8) is the explicit form of all lasso solutions in a neighbourhood of λ, which ends with the critical values.

Sub-gradients under the structure constraint

$$\min_{A \in \mathcal{K}_{\mathcal{G}}} \frac{1}{2t} \sum_{\tau=1}^{t} \|\mathbf{x}_\tau - A\mathbf{x}_{\tau-1}\|_{\ell_2}^2 + \lambda F \|A_{\mathrm{N}}\|_{\ell_1}$$

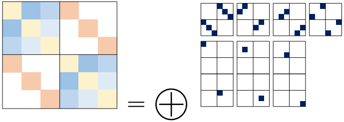
$? \dfrac{\partial L(A)}{\partial A}$ while $A \in \mathcal{K}_{\mathcal{G}}$



Sub-gradients under the structure constraint

$$\min_{A \in \mathcal{K}_{\mathcal{G}}} \frac{1}{2t} \sum_{\tau=1}^{t} \|\mathbf{x}_\tau - A\mathbf{x}_{\tau-1}\|_{\ell_2}^2 + \lambda F \|A_{\mathrm{N}}\|_{\ell_1}$$

$? \dfrac{\partial L(A)}{\partial A}$ while $A \in \mathcal{K}_{\mathcal{G}}$



$$\mathcal{K}_{\mathcal{G}} = \bigoplus_{k \in K} \text{span}\{\widetilde{U}_k\} \Rightarrow A = \sum_{k \in K} \langle \widetilde{U}_k, A^0 \rangle_F \widetilde{U}_k, \text{ where}$$

$$I_F \otimes A_{\mathrm{N}} = \sum_{k \in K_{\mathrm{N}}} \langle \widetilde{U}_k, A^0 \rangle_F \widetilde{U}_k, K_{\mathrm{N}} \subset K.$$

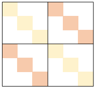
Lasso above becomes

$$\min_{A^0 \in \mathbb{R}^{NF \times NF}} \frac{1}{2t} \sum_{\tau=1}^{t} \left\| \mathbf{x}_\tau - \sum_{k \in K} \langle U_k, A^0 \rangle U_k \mathbf{x}_{\tau-1} \right\|_{\ell_2}^2 + \lambda \left\| \sum_{k \in K_{\mathrm{N}}} \langle U_k, A^0 \rangle U_k \right\|_{\ell_1}$$

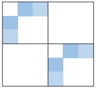
Sub-gradients under the structure constraint

$\frac{\partial L(A^0)}{\partial A^0} = 0 \Longrightarrow$ The optimality condition of $\boldsymbol{A} \in \mathcal{K}_{\mathcal{G}}$:

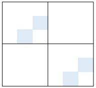
$$\text{Proj}_{\text{DF}}\left(\mathbf{A}\widehat{\mathbf{\Gamma}}_t(0) - \widehat{\mathbf{\Gamma}}_t(1)\right) = 0,$$



$$\text{Proj}_{K_{\text{N}}^1}\left(\mathbf{A}\widehat{\mathbf{\Gamma}}_t(0) - \widehat{\mathbf{\Gamma}}_t(1)\right) + \lambda I_F \otimes \boldsymbol{W}^1 = 0,$$



$$\text{Proj}_{K_{\text{N}}^0}\left(\mathbf{A}\widehat{\mathbf{\Gamma}}_t(0) - \widehat{\mathbf{\Gamma}}_t(1)\right) + \lambda I_F \otimes \boldsymbol{W}^0 = 0,$$



where $\widehat{\mathbf{\Gamma}}_t(0) = \sum_{\tau=1}^{t} \mathbf{x}_{\tau-1}\mathbf{x}_{\tau-1}^\top, \widehat{\mathbf{\Gamma}}_t(1) = \sum_{\tau=1}^{t} \mathbf{x}_{\tau}\mathbf{x}_{\tau-1}^\top$, $\boldsymbol{W}^0$ is the sub-gradient matrix of zero entries in $\boldsymbol{A}_{\mathbf{N}}$, and $\boldsymbol{W}^1$ is the sign matrix of active entries in $\boldsymbol{A}_{\mathbf{N}}$.

Adaptive tuning of lambda

$$\mathbf{A}(t, \lambda_t) \rightarrow \mathbf{A}(t, \lambda_{t+1}), \mathbf{A}(t, \lambda_{t+1}) \rightarrow \mathbf{A}(t+1, \lambda_{t+1})$$

$\lambda_t \rightarrow \lambda_{t+1}$:
Monti et al. (2018); Garrigues and Ghaoui (2008) propose an adaptive tuning method, in our notations:

$$f_{t+1}(\lambda) = \frac{1}{2}\|\mathbf{x}_{t+1} - \mathbf{A}(t, \lambda)\mathbf{x}_t\|_{\ell_2}^2,$$

and updating rule:

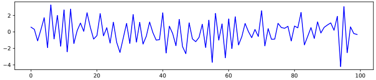
$$\lambda_{t+1} = \lambda_t - \eta \frac{\mathrm{d} f_{t+1}(\lambda)}{\mathrm{d}\lambda}|_{\lambda=\lambda_t},$$

$\frac{\mathrm{d}\mathbf{A}(t,\lambda)}{\mathrm{d}\lambda}|_{\lambda=\lambda_t}$ can be calculated from the optimality condition of $\mathbf{A}(t, \lambda_t)$.

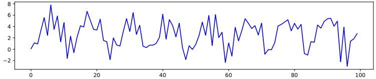
Online graph and trend learning

Come back to the assumption:

$\mathbb{E}(\mathbf{x}_\tau)_\tau = 0, \ \forall \tau \Rightarrow \frac{1}{t}\sum_{\tau=1}^{t} \|\mathbf{x}_\tau - A\mathbf{x}_{\tau-1}\|_{\ell_2}^2 .$



However, raw data $\mathbb{E}(\mathbf{x}_\tau)_\tau = \mathrm{b}_\tau$, that is, a trend is present.



Offline: Detrend $\mathbf{x}_\tau - \widehat{\mathrm{b}}_\tau \Rightarrow$ is forbidden online.

Online graph and trend learning

Augmented data model:

$$\begin{cases} \mathbf{x}_t = \mathsf{b}_t + \mathbf{x}'_t, \rightarrow \text{ Observations} \\ \mathbf{x}'_t = \mathbf{A}\mathbf{x}'_{t-1} + \mathbf{z}_t, \rightarrow \text{ underlying stationary process.} \end{cases}$$

In particular, we consider periodic trend of period M:

$$\mathbf{b}_t = \mathbf{b}_m,\ m = 0, ..., M-1,\ m = t \bmod M.$$

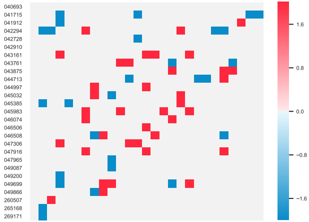
Augmented structured matrix Lasso:

$$\underset{A \in \mathcal{K}_{\mathcal{G}}, \mathsf{b}_m}{\arg\min} \frac{1}{2t} \sum_{m=0}^{M-1} \sum_{\tau \in I_{m,t}} \|(\mathbf{x}_\tau - \mathsf{b}_m) - A(\mathbf{x}_{\tau-1} - \mathsf{b}_{m-1})\|_{\ell_2}^2 + \lambda_t F \|A_{\mathrm{N}}\|_{\ell_1},$$

where $I_{m,t} = \{\tau = 1, ..., t : \tau \bmod M = m\}$.

Detrend + graph estimation simultaneously
$\Longrightarrow$ Online graph learning on raw data

Climatology data



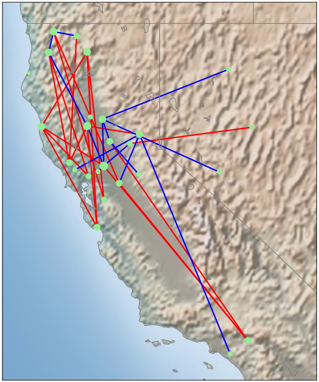
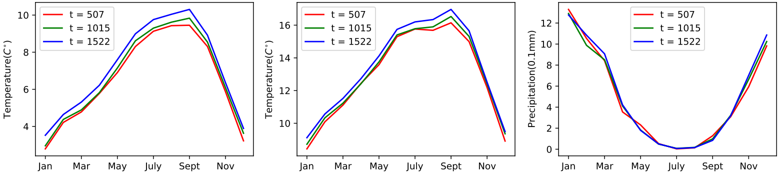


Figure 5: *California weather graph.* Graph Adjacency matrix (left), visualization on the map (right) using sensor coordinates. The nodes with bigger sizes connect with more nodes.

Climatology data



Minimal temperature Average temperature Precipitation

Figure 6: *Estimated trends along years.* On the left, middle, right are the estimated trends at different years of a certain station for the 3 features. Experiment settings: $N = 27$, $F = 4$, $M = 12$.

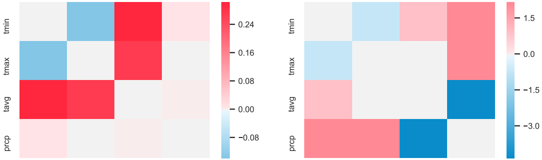
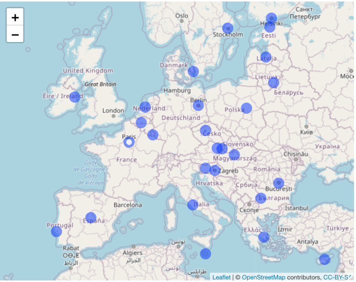


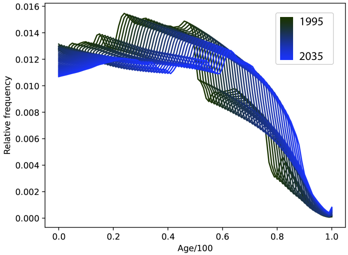
Figure 7: *Updated feature graph at $t = 1522$.* Projected OLS (left), and Lasso (right). Experiment settings: $N = 27$, $F = 4$, $M = 12$.

Note that $t = 1522 < \#params = 1761$.

Table of Contents

Recall the data setting:





Learn $\mathcal{G}$ of $(\boldsymbol{\mu}_{it})_t \in \mathcal{W}_2(\mathbb{R}), i = 1, \ldots N$ with a multivariate distributional AR model.

Random probability measures in Wasserstein space

$$\mathcal{W}_2(\mathbb{R}) = \left\{ \mu \in \mathcal{P}(\mathbb{R}) \middle| \int_{\mathbb{R}} x^2 d\mu(x) < \infty \right\},$$

endowed with the 2-Wasserstein distance

$$d_W(\mu, \gamma)^2 = \inf_{\pi \in \Pi(\mu,\gamma)} \int_{\mathbb{R} \times \mathbb{R}} (x_1 - x_2)^2 d\pi(x_1, x_2)$$

Random probability measures in Wasserstein space

$$\mathcal{W}_2(\mathbb{R}) = \left\{ \mu \in \mathcal{P}(\mathbb{R}) \middle| \int_{\mathbb{R}} x^2 d\mu(x) < \infty \right\},$$

endowed with the 2-Wasserstein distance

$$d_W(\mu, \gamma)^2 = \inf_{\pi \in \Pi(\mu,\gamma)} \int_{\mathbb{R}\times\mathbb{R}} (x_1 - x_2)^2 d\pi(x_1, x_2)$$
$$\overset{\|\cdot\|^2 + \mathbb{R}}{=} \int_0^1 \left(F_\mu^{-1}(u) - F_\gamma^{-1}(u) \right)^2 du,$$

where $F_\mu^{-1}(u), F_\gamma^{-1}(u)$ are the quantile functions of μ and γ.

Random probability measures in Wasserstein space

$$\mathcal{W}_2(\mathbb{R}) = \left\{ \mu \in \mathcal{P}(\mathbb{R}) \middle| \int_{\mathbb{R}} x^2 d\mu(x) < \infty \right\},$$

endowed with the 2-Wasserstein distance

$$d_W(\mu, \gamma)^2 = \inf_{\pi \in \Pi(\mu,\gamma)} \int_{\mathbb{R}\times\mathbb{R}} (x_1 - x_2)^2 d\pi(x_1, x_2)$$
$$\overset{\|\cdot\|^2 + \mathbb{R}}{=} \int_0^1 \left(F_\mu^{-1}(u) - F_\gamma^{-1}(u) \right)^2 du,$$

where $F_\mu^{-1}(u), F_\gamma^{-1}(u)$ are the quantile functions of μ and γ.

$\mathcal{W}_2$ *is not linear space.* Chen et al. (2021); Zhang et al. (2021); Zhu and Müller (2021) extended the univariate AR model

$$\boldsymbol{x}_t - u = \alpha(\boldsymbol{x}_{t-1} - u) + \boldsymbol{\epsilon}_t,$$

by relying on the notion of *Tangent space* in $\mathcal{W}_2$.

Enable again linear methods - Tangent space

$\mathcal{W}_2 := \mathcal{W}_2(\mathbb{R})$ has a pseudo-Riemannian structure (Ambrosio et al., 2008).

Let $\gamma \in \mathcal{W}_2$ be an atomless measure (that is it possesses a continuous cdf F_γ), the tangent space at γ is defined as

$$\mathrm{Tan}_\gamma = \overline{\{t(T_\gamma^\mu - id) : \mu \in \mathcal{W}_2,\ t > 0\}}^{\mathcal{L}^2_\gamma},$$

where $T_\gamma^\mu = F_\mu^{-1} \circ F_\gamma$ is the optimal map, that pushes γ forward to μ.

Enable again linear methods - Tangent space

$\mathcal{W}_2 := \mathcal{W}_2(\mathbb{R})$ has a pseudo-Riemannian structure (Ambrosio et al., 2008).

Let $\gamma \in \mathcal{W}_2$ be an atomless measure (that is it possesses a continuous cdf F_γ), the tangent space at γ is defined as

$$\mathrm{Tan}_\gamma = \overline{\{t(T_\gamma^\mu - id) : \mu \in \mathcal{W}_2,\ t > 0\}}^{\mathcal{L}^2_\gamma},$$

where $T_\gamma^\mu = F_\mu^{-1} \circ F_\gamma$ is the optimal map, that pushes γ forward to μ. Tan_γ is endowed with the inner product $\langle \cdot, \cdot \rangle_\gamma$ defined by

$$\langle f, g \rangle_\gamma := \int_{\mathbb{R}} f(x) g(x)\, d\gamma(x),\ f, g \in \mathcal{L}^2_\gamma(\mathbb{R}),$$

and the induced norm $\| \cdot \|_\gamma$.

Enable again linear methods - Tangent space

$$\mathrm{Tan}_\gamma = \overline{\{t(T_\gamma^\mu - id) : \mu \in \mathcal{W}_2,\ t > 0\}}^{\mathcal{L}^2_\gamma},$$

where $T_\gamma^\mu = F_\mu^{-1} \circ F_\gamma$ is the optimal map, that pushes γ forward to μ.

Definition

The logarithmic map $\mathrm{Log}_\gamma : \mathcal{W}_2 \to \mathrm{Tan}_\gamma$ is defined as

$$\mathrm{Log}_\gamma \mu = T_\gamma^\mu - id.$$

The exponential map $\mathrm{Exp}_\gamma : \mathrm{Tan}_\gamma \to \mathcal{W}_2$ is defined as

$$\mathrm{Exp}_\gamma g = (g + id)\#\gamma,$$

where $T\#\mu$ is the measure pushforwarded by function T, defined as $[T\#\mu](A) = \mu(\{x : T(x) \in A\})$.

Line segment in Tangent space = geodesic in $\mathcal{W}_2$

The geodesic (McCann's interpolant) between γ and μ

$$\mathrm{Exp}_\gamma[\alpha(T_\gamma^\mu - id)], \; \alpha : 0 \to 1,$$

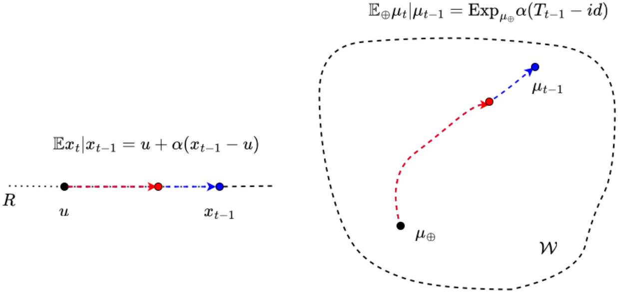
Line segment in Tangent space = geodesic in $\mathcal{W}_2$

The geodesic (McCann's interpolant) between γ and μ

$$
\begin{aligned}
&\mathrm{Exp}_\gamma[\alpha(T_\gamma^\mu - id)],\ \alpha : 0 \to 1, \\
&= [\alpha(T_\gamma^\mu - id) + id]\#\gamma
\end{aligned}
$$

Related work: Univariate Wasserstein AR model

Chen et al. (2021); Zhang et al. (2021); Zhu and Müller (2021) proposed to interpret the regression operation geometrically.

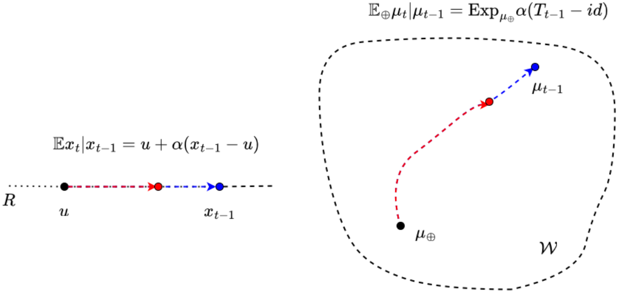


Let $\boldsymbol{\mu}$ be a random measure from $(\Omega, \mathcal{F}, \mathbb{P})$ to $\mathcal{W}_2$

$$\text{(Fréchet mean)} \quad \mathbb{E}_{\oplus}\boldsymbol{\mu} = \arg\min_{\nu \in \mathcal{W}_2} \mathbb{E}\left[d_W^2(\boldsymbol{\mu}, \nu)\right].$$

Related work: Univariate Wasserstein AR model

Chen et al. (2021); Zhang et al. (2021); Zhu and Müller (2021) proposed to interpret the regression operation geometrically.



Let $\boldsymbol{\mu}$ be a random measure from $(\Omega, \mathcal{F}, \mathbb{P})$ to $\mathcal{W}_2$

(conditional Fréchet mean) $\mathbb{E}_{\oplus}\boldsymbol{\mu}|\boldsymbol{\gamma} = \underset{\nu \in \mathcal{W}_2}{\arg\min}\, \mathbb{E}\left[d_W^2(\boldsymbol{\mu}, \nu)|\boldsymbol{\gamma}\right].$

Multivariate Wasserstein AR model

Multivariate regression operation (VAR(1)):
for any fixed $i = 1, \dots N$

$$\mathbb{E}\boldsymbol{x}_{it}|\boldsymbol{x}_{j,t-1} = u_i + \sum_{j=1}^{N} A_{ij}(\boldsymbol{x}_{j,t-1} - u_j) \Rightarrow \begin{cases} \boldsymbol{T}_{1,t-1} - id & \in \mathrm{Tan}_{\mu_{1,\oplus}} \\ \boldsymbol{T}_{2,t-1} - id & \in \mathrm{Tan}_{\mu_{2,\oplus}} \\ & \vdots \end{cases}$$

Multivariate Wasserstein AR model

Multivariate regression operation (VAR(1)):
for any fixed $i = 1, \dots N$

$$\mathbb{E}\boldsymbol{x}_{it}|\boldsymbol{x}_{j,t-1} = u_i + \sum_{j=1}^{N} A_{ij}(\boldsymbol{x}_{j,t-1} - u_j) \Rightarrow \begin{cases} \boldsymbol{T}_{1,t-1} - id & \in \mathrm{Tan}_{\mu_{1,\oplus}} \\ \boldsymbol{T}_{2,t-1} - id & \in \mathrm{Tan}_{\mu_{2,\oplus}} \\ & \vdots \end{cases}$$

$\iff$

$$\begin{cases} \text{Center} & \tilde{\boldsymbol{x}}_{it} = \boldsymbol{x}_{it} - u_i, \quad \xrightarrow{\text{ref pt}} \mathbb{E}\tilde{\boldsymbol{x}}_{it} = 0, \\ \text{Push} & \mathbb{E}\tilde{\boldsymbol{x}}_{it}|\tilde{\boldsymbol{x}}_{j,t-1} = 0 + \sum_{j=1}^{N} A_{ij}\tilde{\boldsymbol{x}}_{jt}, \end{cases}$$

Multivariate Wasserstein AR model

Multivariate regression operation (VAR(1)):
for any fixed $i = 1, \dots N$

$$\mathbb{E}\boldsymbol{x}_{it}|\boldsymbol{x}_{j,t-1} = u_i + \sum_{j=1}^{N} A_{ij}(\boldsymbol{x}_{j,t-1} - u_j) \Rightarrow \begin{cases} \boldsymbol{T}_{1,t-1} - id & \in \mathrm{Tan}_{\mu_{1,\oplus}} \\ \boldsymbol{T}_{2,t-1} - id & \in \mathrm{Tan}_{\mu_{2,\oplus}} \\ & \vdots \end{cases}$$

$$\iff$$

$$\begin{cases} \text{Center} & \tilde{\boldsymbol{x}}_{it} = \boldsymbol{x}_{it} - u_i, \quad \xrightarrow{\text{ref pt}} \mathbb{E}\tilde{\boldsymbol{x}}_{it} = 0, \\ \text{Push} & \mathbb{E}\tilde{\boldsymbol{x}}_{it}|\tilde{\boldsymbol{x}}_{j,t-1} = 0 + \sum_{j=1}^{N} A_{ij}\tilde{\boldsymbol{x}}_{jt}, \end{cases}$$

$$\Longrightarrow \begin{cases} \text{Center} & \tilde{\boldsymbol{\mu}}_{it} = ? \xrightarrow{\text{ref pt}} \mathbb{E}_{\oplus}\tilde{\boldsymbol{\mu}}_{it} = c \\ \text{Push} & \mathbb{E}_{\oplus}\tilde{\boldsymbol{\mu}}_{it}|\tilde{\boldsymbol{\mu}}_{j,t-1} = \mathrm{Exp}_c\left(\sum_{j=1}^{N} A_{ij}(\tilde{\boldsymbol{T}}_{j,t-1} - id)\right) \end{cases}$$

Center a random measure $\boldsymbol{\mu}$, s.t. $\mathbb{E}_{\oplus}\boldsymbol{\mu} = U(0,1)$

Zhu and Müller (2021) proposed a notion of addition for two increasing functions:

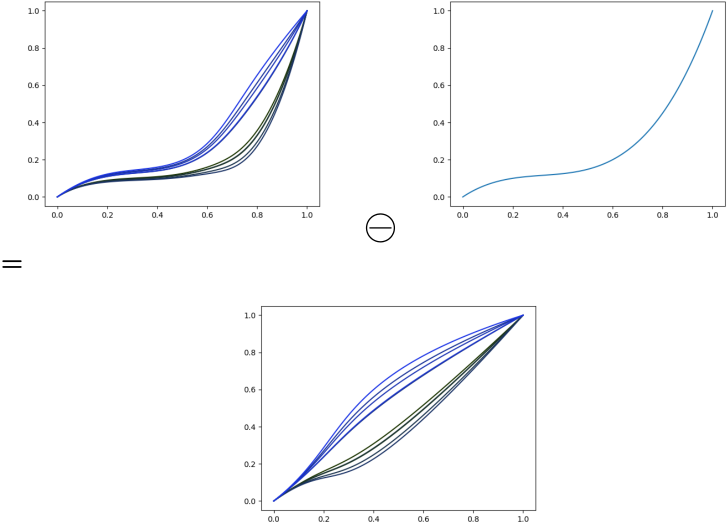
$$g \oplus f := g \circ f \implies g \ominus f := g \circ f^{-1},$$

where $^{-1}$ are the left continuous inverse.
For $\boldsymbol{\mu}$, its centered measure $\widetilde{\boldsymbol{\mu}}$ is defined by the quantile function

$$\widetilde{\boldsymbol{F}}_{\mu}^{-1} = \boldsymbol{F}_{\mu}^{-1} \ominus F_{\oplus}^{-1},$$

where $\boldsymbol{F}_{\mu}^{-1}, F_{\oplus}^{-1}$ et $\widetilde{\boldsymbol{F}}_{\mu}^{-1}$ are respectively quantile functions of $\boldsymbol{\mu}, \mu_{\oplus}$, and $\widetilde{\boldsymbol{\mu}}$.

Center a random measure $\boldsymbol{\mu}$, s.t. $\mathbb{E}_{\oplus}\boldsymbol{\mu} = U(0,1)$



Wasserstein multivariate AR Model

$$\widetilde{\boldsymbol{\mu}}_{it} = \boldsymbol{\epsilon}_{it} \# \operatorname{Exp}_{Leb}\left(\sum_{j=1}^{N} A_{ij}(\widetilde{\boldsymbol{F}}_{j,t-1} - id)\right),$$

where $\{\boldsymbol{\epsilon}_{it}\}_{i,t}$ are i.i.d. random increasing functions, $\boldsymbol{\epsilon}_{it}$ is almost surely independent of $\boldsymbol{\mu}_{j,t-1}$, $i, j = 1, \ldots, N$, for all $t \in \mathbb{Z}$, and

$$\mathbb{E}\left[\boldsymbol{\epsilon}_{it}(x)\right] = x,\ x \in [0,1].$$

Wasserstein multivariate AR Model

$$\widetilde{\boldsymbol{\mu}}_{it} = \boldsymbol{\epsilon}_{it} \# \operatorname{Exp}_{Leb}\left(\sum_{j=1}^{N} A_{ij}(\widetilde{\boldsymbol{F}}_{j,t-1} - id)\right),$$

where $\{\boldsymbol{\epsilon}_{it}\}_{i,t}$ are i.i.d. random increasing functions, $\boldsymbol{\epsilon}_{it}$ is almost surely independent of $\boldsymbol{\mu}_{j,t-1}$, $i, j = 1, \ldots, N$, for all $t \in \mathbb{Z}$, and

$$\mathbb{E}\left[\boldsymbol{\epsilon}_{it}(x)\right] = x,\ x \in [0, 1].$$

Assumption

- *All* $\boldsymbol{\mu}_{it}$, $t \in \mathbb{Z}$, $i = 1, \ldots, N$ *are supported on* $[0, 1]$.
- *(*N*-**simplex**)* $\sum_{j=1}^{N} A_{ij} \leqslant 1$ *and* $0 \leqslant A_{ij} \leqslant 1$.

Wasserstein multivariate AR Model

$$\widetilde{\boldsymbol{\mu}}_{it} = \boldsymbol{\epsilon}_{it} \# \operatorname{Exp}_{Leb} \left(\sum_{j=1}^{N} A_{ij} (\widetilde{\boldsymbol{F}}_{j,t-1} - id) \right),$$

where $\{\boldsymbol{\epsilon}_{it}\}_{i,t}$ are i.i.d. random increasing functions, $\boldsymbol{\epsilon}_{it}$ is almost surely independent of $\boldsymbol{\mu}_{j,t-1}$, $i, j = 1, \ldots, N$, for all $t \in \mathbb{Z}$, and

$$\mathbb{E}\left[\boldsymbol{\epsilon}_{it}(x)\right] = x,\ x \in [0, 1].$$

Assumption

- *All $\boldsymbol{\mu}_{it}$, $t \in \mathbb{Z}$, $i = 1, \ldots, N$ are supported on $[0, 1]$.*
- *(N-**simplex**) $\sum_{j=1}^{N} A_{ij} \leqslant 1$ and $0 \leqslant A_{ij} \leqslant 1$.*

Quantile function representation

$$\widetilde{\boldsymbol{F}}_{i,t}^{-1} = \boldsymbol{\epsilon}_{i,t} \circ \left[\sum_{j=1}^{N} A_{ij} \left(\widetilde{\boldsymbol{F}}_{j,t-1}^{-1} - id \right) + id \right],$$

Wasserstein multivariate AR Model

$$\widetilde{\boldsymbol{\mu}}_{it} = \boldsymbol{\epsilon}_{it} \# \operatorname{Exp}_{Leb}\left(\sum_{j=1}^{N} A_{ij}(\widetilde{\boldsymbol{F}}_{j,t-1} - id)\right),$$

where $\{\boldsymbol{\epsilon}_{it}\}_{i,t}$ are i.i.d. random increasing functions, $\boldsymbol{\epsilon}_{it}$ is almost surely independent of $\boldsymbol{\mu}_{j,t-1}$, $i, j = 1, \ldots, N$, for all $t \in \mathbb{Z}$, and

$$\mathbb{E}\left[\boldsymbol{\epsilon}_{it}(x)\right] = x,\ x \in [0,1].$$

Assumption

- *All $\boldsymbol{\mu}_{it}$, $t \in \mathbb{Z}$, $i = 1, \ldots, N$ are supported on $[0,1]$.*
- *(N-**simplex**) $\sum_{j=1}^{N} A_{ij} \leqslant 1$ and $0 \leqslant A_{ij} \leqslant 1$.*

Quantile function representation

$$\widetilde{\boldsymbol{F}}_{i,t}^{-1} = \boldsymbol{\epsilon}_{i,t} \circ \left[\sum_{j=1}^{N} A_{ij}\left(\widetilde{\boldsymbol{F}}_{j,t-1}^{-1} - id\right) + id\right], \quad A \iff \mathcal{G}$$

Existence, uniqueness, and stationarity

$$\widetilde{\boldsymbol{F}}_{i,t}^{-1} = \boldsymbol{\epsilon}_{i,t} \circ \left[\sum_{j=1}^{N} A_{ij} \left(\widetilde{\boldsymbol{F}}_{j,t-1}^{-1} - id \right) + id \right] \tag{8}$$

Admissible as a time series model: existence, uniqueness and stationarity of series $(\widetilde{\boldsymbol{F}}_{i,t}^{-1})_t, i = 1, \ldots N$.

Existence, uniqueness, and stationarity

$$\widetilde{\boldsymbol{F}}_{i,t}^{-1} = \boldsymbol{\epsilon}_{i,t} \circ \left[\sum_{j=1}^{N} A_{ij} \left(\widetilde{\boldsymbol{F}}_{j,t-1}^{-1} - id \right) + id \right]$$

Admissible as a time series model: existence, uniqueness and stationarity of series $(\widetilde{\boldsymbol{F}}_{i,t}^{-1})_t, i = 1, \dots N$.

Theoretical results:

Under two classical conditions, we have proved:

• Iterated random function system (8) admits uniquely one solution in the metric space

$$(\mathcal{T}, \| \cdot \|_{Leb})^{\otimes N}, \mathcal{T} = \{F_\mu^{-1} | \mu \in \mathcal{W}_2(\mathbb{R})\}$$

• The unique solution is stationary (2nd order) in the Hilbert space

$$(\mathcal{T}, \langle , \rangle_{Leb})^{\otimes N}, \mathcal{T} = \{F_\mu^{-1} | \mu \in \mathcal{W}_2(\mathbb{R})\}$$

according to a proper definition for functional TS.

Constrained least-square estimation

For the auto-regressive model

$$\widetilde{\boldsymbol{F}}_{i,t}^{-1} = \boldsymbol{\epsilon}_{i,t} \circ \left[\sum_{j=1}^{N} A_{ij} \left(\widetilde{\boldsymbol{F}}_{j,t-1}^{-1} - id \right) + id \right],$$

given the centered observations $\widetilde{\boldsymbol{F}}_t^{-1}$, $t = 0, 1, \ldots, T$, we propose

$$\widetilde{\boldsymbol{A}}_{i:} = \underset{A_{i:} \in B_+^1}{\arg\min} \frac{1}{T} \sum_{t=1}^{T} \left\| \widetilde{\boldsymbol{F}}_{i,t}^{-1} - \sum_{j=1}^{N} A_{ij} \left(\widetilde{\boldsymbol{F}}_{j,t-1}^{-1} - id \right) - id \right\|_{Leb}^{2},$$

where B_+^1 is the constraint set of N-simplex.

Constrained least-square estimation

For the auto-regressive model

$$\widetilde{\boldsymbol{F}}_{i,t}^{-1} = \boldsymbol{\epsilon}_{i,t} \circ \left[\sum_{j=1}^{N} A_{ij} \left(\widetilde{\boldsymbol{F}}_{j,t-1}^{-1} - id \right) + id \right],$$

given the centered observations $\widehat{\boldsymbol{F}}_t^{-1}$, $t = 0, 1, \ldots, T$, we propose

$$\widehat{\boldsymbol{A}}_{i:} = \underset{A_{i:} \in B_+^1}{\arg\min} \frac{1}{T} \sum_{t=1}^{T} \left\| \widehat{\boldsymbol{F}}_{i,t}^{-1} - \sum_{j=1}^{N} A_{ij} \left(\widehat{\boldsymbol{F}}_{j,t-1}^{-1} - id \right) - id \right\|_{Leb}^{2},$$

where B_+^1 is the constraint set of N-simplex.

Constrained least-square estimation

For the auto-regressive model

$$\widetilde{\boldsymbol{F}}_{i,t}^{-1} = \boldsymbol{\epsilon}_{i,t} \circ \left[\sum_{j=1}^{N} A_{ij} \left(\widetilde{\boldsymbol{F}}_{j,t-1}^{-1} - id \right) + id \right],$$

given the centered observations $\widehat{\boldsymbol{F}}_t^{-1}$, $t = 0, 1, \ldots, T$, we propose

$$\widehat{\boldsymbol{A}}_{i:} = \underset{A_{i:} \in B_+^1}{\arg\min} \frac{1}{T} \sum_{t=1}^{T} \left\| \widehat{\boldsymbol{F}}_{i,t}^{-1} - \sum_{j=1}^{N} A_{ij} \left(\widehat{\boldsymbol{F}}_{j,t-1}^{-1} - id \right) - id \right\|_{Leb}^2,$$

where B_+^1 is the constraint set of N-simplex.

Theoretical result:

$$\widehat{\boldsymbol{A}} \xrightarrow{p} A, \text{ as } T \to +\infty.$$

Age distribution of countries

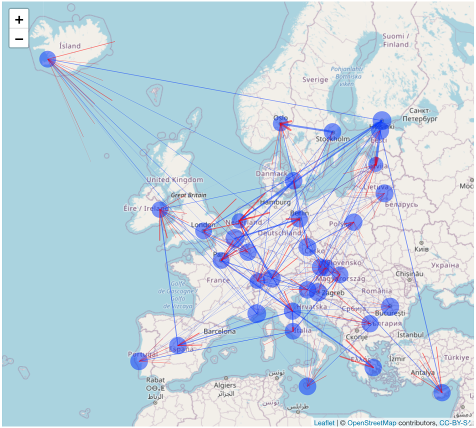


Figure 8: *Inferred age structure graph.* The non-zero coefficients A_{ij} are represented by the weighted directed edges from node j to node i. Thicker arrow corresponds to larger weights. The blue circles around nodes represent the weights of self-loop.

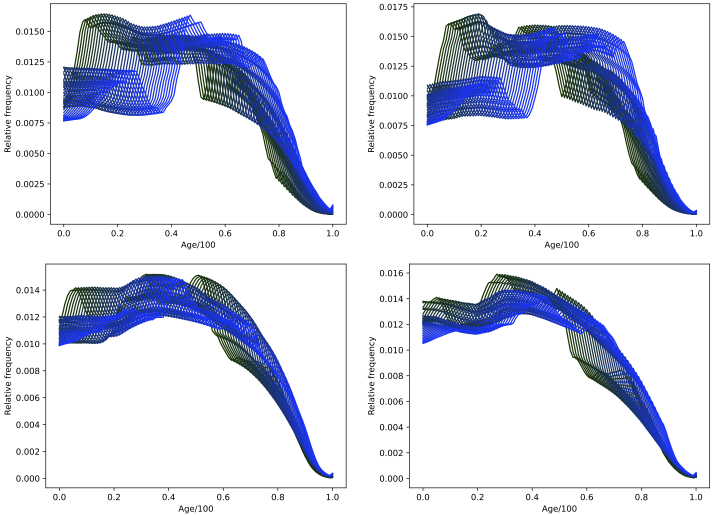


Figure 9: *Evolution of age structure from* 1995 *to* 2035 *(projected). Estonia (top left), Latvia(top right), Sweden (bottom left) versus Norway (bottom right).*

Age distribution of countries

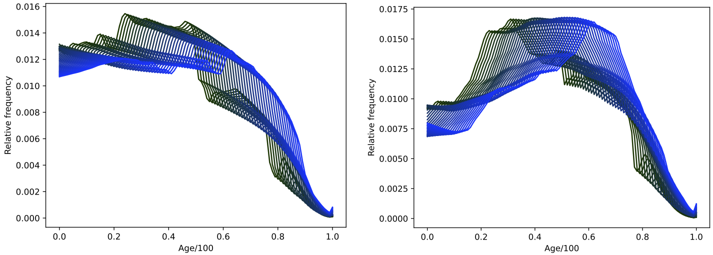


Figure 10: *Evolution of age structure from* 1995 *to* 2035 *(projected) of France (left) versus Italy (right).*

Goal: given $(\mathbf{x}_{nt})_t \in \mathbb{R},\ n \in \mathcal{N} = \{1, \ldots, N\}$, finding the highly predictable series $i \in I \subset \mathcal{N}$, such that their observations $\boldsymbol{x}_{it}$ can be reconstructed accurately by the past and present obs of other series $\boldsymbol{x}_{j\tau},\ j \in I^c, t - H \leqslant \tau \leqslant t$ in real time.

Goal: given $(\mathbf{x}_{nt})_t \in \mathbb{R},\ n \in \mathcal{N} = \{1, \ldots, N\}$, finding the highly predictable series $i \in I \subset \mathcal{N}$, such that their observations $\boldsymbol{x}_{it}$ can be reconstructed accurately by the past and present obs of other series $\boldsymbol{x}_{j\tau},\ j \in I^c, t - H \leqslant \tau \leqslant t$ in real time.

$\mathcal{G}$ can be given, e.g. geographical graph of the weather stations.

Goal: given $(\mathbf{x}_{nt})_t \in \mathbb{R},\ n \in \mathcal{N} = \{1, \dots, N\}$, finding the highly predictable series $i \in I \subset \mathcal{N}$, such that their observations $\boldsymbol{x}_{it}$ can be reconstructed accurately by the past and present obs of other series $\boldsymbol{x}_{j\tau},\ j \in I^c, t - H \leqslant \tau \leqslant t$ in real time.

$\mathcal{G}$ can be given, e.g. geographical graph of the weather stations.

We use 3 prediction methods to evaluate the node predictability: *kernel ridge regression, linear regression, and neural networks*.

$\longrightarrow$ 3 ranking procedures.

In this thesis, we provided new statistical tools for analyzing spatio-temporal and multi-dimensional data. In particular, we extended the classical VAR(1) model for the complex data types: matrix-variate and distributional data in the way to serve graph learning.

In this thesis, we provided new statistical tools for analyzing spatio-temporal and multi-dimensional data. In particular, we extended the classical VAR(1) model for the complex data types: matrix-variate and distributional data in the way to serve graph learning.

Future works:

These two works introduce a more general topic: **object data analysis**. Especially, the 2nd work demonstrates one important way to perform the analysis, that is to view data points as **random objects in a metric space**.

Graph itself is also an important data object. Among others, it is adopted to represent the brain functional connectivity.

Graph itself is also an important data object. Among others, it is adopted to represent the brain functional connectivity.

G simple, undirected, weighted (bounded), N nodes

$\iff L_N \in$ a subspace of $\mathbb{R}^{N\times N}$, endowed with e.g. $\|\cdot\|_{\mathbf{F}}$.

Graph itself is also an important data object. Among others, it is adopted to represent the brain functional connectivity.

G simple, undirected, weighted (bounded), N nodes

$\iff L_N \in$ a subspace of $\mathbb{R}^{N\times N}$, endowed with e.g. $\|\cdot\|_{\mathbf{F}}$.

Already available graph-valued models:

- Network regression with Euclidean predictors (Zhou and Müller, 2021): $\mathbb{E}_{\oplus}\boldsymbol{G}|\mathbf{x}$. The model is applied to study the evolution of brain connectivity wrt age
- Two sample tests (Ginestet et al., 2017): $\boldsymbol{G}_i \overset{i.i.d.}{\sim} \boldsymbol{G}_1$, $\boldsymbol{G}_j \overset{i.i.d.}{\sim} \boldsymbol{G}_2 \to \mathbb{E}_{\oplus}\boldsymbol{G}_1 ? = \mathbb{E}_{\oplus}\boldsymbol{G}_2$. The model is applied to study the impact of gender on brain connectivity.

Graph itself is also an important data object. Among others, it is adopted to represent the brain functional connectivity.

G simple, undirected, weighted (bounded), N nodes

$\iff L_N \in$ a subspace of $\mathbb{R}^{N\times N}$, endowed with e.g. $\|\cdot\|_{\mathbf{F}}$.

Already available graph-valued models:

- Network regression with Euclidean predictors (Zhou and Müller, 2021): $\mathbb{E}_{\oplus}\boldsymbol{G}|\mathbf{x}$. The model is applied to study the evolution of brain connectivity wrt age
- Two sample tests (Ginestet et al., 2017): $\boldsymbol{G}_i \overset{i.i.d.}{\sim} \boldsymbol{G}_1, \boldsymbol{G}_j \overset{i.i.d.}{\sim} \boldsymbol{G}_2 \to \mathbb{E}_{\oplus}\boldsymbol{G}_1 ?= \mathbb{E}_{\oplus}\boldsymbol{G}_2$. The model is applied to study the impact of gender on brain connectivity.

More generalized models to be developed, e.g. Network (functional) regression with network predictors.

L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

Y. Chen, Z. Lin, and H.-G. Müller. Wasserstein regression. *Journal of the American Statistical Association*, pages 1–14, 2021.

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

C. E. Ginestet, J. Li, P. Balachandran, S. Rosenberg, and E. D. Kolaczyk. Hypothesis testing for network data in functional neuroimaging. *The Annals of Applied Statistics*, pages 725–750, 2017.

A. K. Gupta and D. K. Nagar. *Matrix variate distributions*. Chapman and Hall/CRC, 2018.

R. P. Monti, C. Anagnostopoulos, and G. 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.

C. Zhang, P. Kokoszka, and A. Petersen. Wasserstein autoregressive models for density time series. *Journal of Time Series Analysis*, 2021.

Y. Zhou and H.-G. Müller. Dynamic network regression. *arXiv preprint arXiv:2109.02981*, 2021.

C. Zhu and H.-G. Müller. Autoregressive optimal transport models. *arXiv preprint arXiv:2105.05439*, 2021.

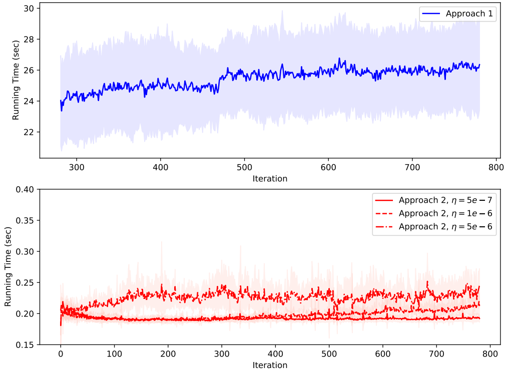


Figure 11: *Running time of each online update.* The red curves are the mean running time of the high-dimensional procedure, taken over 10 simulations each. The blue curve is the mean running time of the low-dimensional procedure, taken over the same 30 simulations. The shaded areas represent the corresponding one standard deviations. Other simulation settings: $N = 20$, $F = 5$, $M = 12$, number of model parameters $= 1500$.

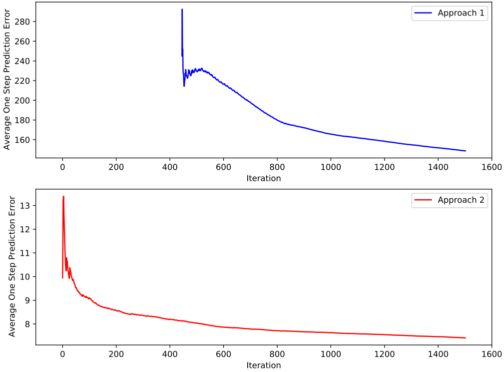


Figure 12: *Average one step prediction error of raw time series.* Projected OLS (top), and Lasso (bottom).

Homotopy algorithms and optimality conditions

Algo: $t_1 \to t_2$:

$$\theta^* = \arg\min_{\theta \in \mathbb{R}^d} L(\theta),\ L(\theta) = \frac{1}{2} \left\| \begin{pmatrix} \mathbf{y} \\ \mu y_{t+1} \end{pmatrix} - \begin{pmatrix} \mathbf{X} \\ \mu \mathbf{x}_{t+1}^\top \end{pmatrix} \theta \right\|_{\ell_2}^2 + \lambda \|\theta\|_{\ell_1},$$

$\frac{\partial L}{\partial \theta} = 0 \Longrightarrow$ Optimality condition $\Longrightarrow$ Homotopy algorithm.

Enable again linear methods - towards tangent space

$$d_W(\mu,\gamma)^2 = \inf_{\pi\in\Pi(\mu,\gamma)} \int_{\mathbb{R}\times\mathbb{R}} (x_1 - x_2)^2 d\pi(x_1, x_2)$$

When γ is an atomless measure, that is F_γ is continue, we have π^* exists uniquely and is induced by a function $T_\gamma^\mu : \mathbb{R} \to \mathbb{R}$, such that

$$T_\gamma^\mu \# \gamma = \mu$$

where $[T_\gamma^\mu \# \gamma](A) = \gamma(\{x : T_\gamma^\mu(x) \in A\}),\ A \subset R$. T_γ^μ is called optimal transport map. Furthermore,

$$T_\gamma^\mu(x) = F_\mu^{-1} \circ F_\gamma(x).$$

Characterization of the difference between μ, γ:

$$d_W(\mu,\gamma) \Longrightarrow T_\gamma^\mu(x) \in \mathcal{L}_\gamma^2(\mathbb{R}).$$

Adaptive tuning of lambda

Updating rule can be interpreted as the steps in the **projected stochastic gradient descent** derived for the batch problem

$$\lambda_n^* = \arg\min_{\lambda \geqslant 0} \frac{1}{2n} \sum_{t=1}^{n} \|\mathbf{x}_{t+1} - \mathbf{A}(t,\lambda)\mathbf{x}_t\|_{\ell_2}^2,$$

which is the average one step prediction error.

KS/KP as a common practice

vec$(\mathbf{X}_i)+$ **vector model + KP/KS imposed in parameters** is a common practice to extend vector models to matrix-variate data in literature. For example, Gupta and Nagar (2018) proposed a matrix-variate Normal distribution as:

$$\text{vec}(\mathbf{X}_i) \overset{iid}{\sim} \mathcal{N}(\text{u}, \Sigma), \text{ where } \Sigma = \Sigma_1 \otimes \Sigma_2,$$

where $\otimes$ is the Kronecker product.