Transport optimal pour la complétion de données manquantes dans des séries temporelles

Léo LAFFEACH

ENS Rennes

7 octobre 2022

Sommaire

Introduction

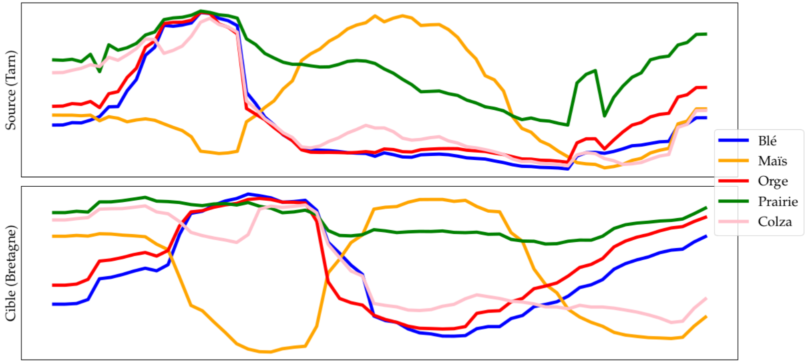


Figure – Moyennes par classe d'occupation du sol d'un indicateur de croissance de végétation pour deux zones géographiques différentes.

Soit **X** et **X**', deux ensembles d'échantillons avec des poids, dans $\mathbb{R}^d$: $\{(x_i, w_i)\}_{i=1}^{n}$ avec $\sum_i w_i = 1$ et $\{(x'_j, w'_j)\}_{j=1}^{n'}$ avec $\sum_j w'_j = 1$.

Transport Optimal

$$\mathbf{OT}(\mathbf{X}, \mathbf{X}') = \underset{\gamma \in \Gamma(\mathbf{w}, \mathbf{w}')}{\arg\min} \langle \mathbf{C}(\mathbf{X}, \mathbf{X}'), \gamma \rangle$$

- $\Gamma(\mathbf{w}, \mathbf{w}') = \{\gamma | \gamma \geq 0, \gamma \mathbb{1}_{n'} = \mathbf{w}, \gamma^\top \mathbb{1}_n = \mathbf{w}'\}$
  est l'ensemble de transports linéaires contraints, de telle sorte que toute la masse de **X** est transportée vers toute la masse de **X**'.
- $\mathbf{C}(\mathbf{X}, \mathbf{X}') = \{d(\mathbf{X}_{ij}, \mathbf{X}'_{i'j'})\}$
  Avec $d(\mathbf{X}_{ij}, \mathbf{X}'_{i'j'})$ est la distance entre deux éléments de **X** et **X**'.

Transport Optimal avec régularisation

Pour rendre le problème précédent différentiable, on peut ajouter une régularisation.

Transport Optimal avec régularisation

$$\mathrm{OT}_\epsilon(\mathbf{X}, \mathbf{X}') = \underset{\gamma \in \Pi(\mathbf{w}, \mathbf{w}')}{\arg\min} \langle \mathbf{C}(\mathbf{X}, \mathbf{X}'), \gamma \rangle + \epsilon h(\gamma)$$

- $\epsilon > 0$.
- $h(\gamma) = \sum_{ij} \gamma_{ij} \log \gamma_{ij}$ est l'entropie négative.

Cependant, dû au terme de l'entropie, $\mathbf{OT}_\epsilon$ n'est plus forcément positif.

OT avec régularisation

Ce qui peut se résoudre via un débiaisement, en soustrayant les termes d'auto-correction.

Divergence de Sinkhorn

$$S_\epsilon(\mathbf{X}, \mathbf{X}') = \mathbf{OT}_\epsilon(\mathbf{X}, \mathbf{X}') - \frac{1}{2}(\mathbf{OT}_\epsilon(\mathbf{X}, \mathbf{X}) + \mathbf{OT}_\epsilon(\mathbf{X}', \mathbf{X}'))$$

Ainsi on a une équation qui est positive, convexe, et qui peut être calculée avec un faible coût additionel comparé à $\mathbf{OT}_\epsilon$

Complétion de données en utilisant le transport optimal

Soit $\Omega = (\omega_{ij})_{ij} \in \{0,1\}^{n \times d}$ un masque binaire qui indique si la donnée est observée ou non, ie : $\omega_{ij} = 1$ (resp. 0) si et seulement si l'entrée (i,j) est observée (resp. manquante).

Valeurs observées

$$\mathbf{X} = \mathbf{X}^{(obs)} \odot \Omega + \mathbf{NA} \odot (\mathbb{1} - \Omega)$$

- $\mathbf{X}^{(obs)} \in \mathbb{R}^{\mathrm{n} \times \mathrm{d}}$ contient les données observées
- $\odot$ est le produit élément par élément

Valeurs imputées

$$\hat{\mathbf{X}} = \mathbf{X}^{(obs)} \odot \Omega + \hat{\mathbf{X}}^{(imp)} \odot (\mathbb{1} - \Omega)$$

où $\hat{\mathbf{X}} \in \mathbb{R}^{n \times d}$ contient les valeurs imputées.

Complétion de données en utilisant le transport optimal

Algorithm 1: Imputation avec Sinkhorn par lots

Input: $\mathbf{X} \in (\mathbb{R} \cup \{NA\})^{n \times d}, \Omega \in \{0,1\}^{n \times d}, \alpha, \eta, \epsilon > 0, n \geq m > 0,$

Initialisation : pour j = 1, ..., d,

- pour i t.q. $\omega_{ij} = 0, \hat{x}_{ij} \leftarrow \bar{x}_{:j}^{obs} + \epsilon_{ij}$ *avec* $\epsilon \sim \mathcal{N}(0, \eta)$ *et* $\bar{x}_{:j}^{obs}$ correspondant à la moyenne des données observées dans la j-ème variable (données manquantes)
- pour i t.q. $\omega_{ij} = 1, \hat{x}_{ij} \leftarrow x_{ij}$ (entrées observés)

Pour **iter** $= 1, 2, ..., iter_{max}$ **faire**

Echantilloner deux ensembles K et L de m indices

$\mathcal{L}(\hat{\mathbf{X}}_K, \hat{\mathbf{X}}_L) \leftarrow S_\epsilon(\hat{\mathbf{X}}_K, \hat{\mathbf{X}}_L)$

$\hat{\mathbf{X}}_{K \cup L}^{(imp)} \leftarrow \hat{\mathbf{X}}_{K \cup L}^{(imp)} - \alpha \mathbf{RMSprop}(\nabla_{\hat{\mathbf{X}}_{K \cup L}^{(imp)}} \mathcal{L})$

Fin Pour

Output: $\hat{\mathbf{X}}$

Alignement temporel dynamique

x et $x' \in \mathbb{R}^{t \times d}$.

Dynamic Time Warping

$$\mathbf{DTW}(x, x') = \underset{\pi \in \mathcal{A}(T, T')}{\arg\min} \langle \mathbf{C}(x, x'), \pi \rangle$$

- $\mathbf{C}(x, x') = \{d(x_{jk}, x'_{j'k'})\}$, où $d(x_{jk}, x'_{j'k'})$ est la distance entre deux éléments de x et x'.
- $\mathcal{A}(T, T')$ est l'ensemble des alignements admissibles entre x et x'.

Un alignement admissible $\pi \in \mathcal{A}(T, T')$ est une matrice binaire telle que $\pi_{1,1} = \pi_{T,T'} = 1$, et pour chaque couple d'horodatage $(l; m)$ tel que $\pi_{l,m} = 1$, il y a soit $\pi_{l-1,m} = 1$ ou $\pi_{l,m-1} = 1$ ou $\pi_{l-1,m-1} = 1$. Les autres valeurs de π valent 0.

Match-And-Deform

MAD

$$
\begin{aligned}
MAD(\mathbf{X}, \mathbf{X}') &= \underset{\substack{\gamma \in \Gamma(w,w') \\ \pi \in \mathcal{A}(T,T')}}{\arg\min} \langle \mathbf{L}(\mathbf{X},\ \mathbf{X}') \otimes \pi, \gamma \rangle \\
&= \underset{\substack{\gamma \in \Gamma(w,w') \\ \pi \in \mathcal{A}(T,T')}}{\arg\min} \sum_{i,j} \sum_{l,m} d(x_l^i, x_m'^j) \pi_{lm} \gamma_{ij}
\end{aligned}
$$

- $\otimes$ est la multiplication tenseur-matrice
- γ est le plan de transport entre les échantillons
- π est le chemin DTW global qui aligne les horodatages entre $\mathbf{X}$ et $\mathbf{X}'$.
- $\mathbf{L}(\mathbf{X},\ \mathbf{X}')$ est un tenseur en 4 dimensions dont les éléments sont $\boldsymbol{L}_{l,m}^{i,j} = d(x_l^i, x_m'^j)$, avec $d\ :\ \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^+$ une distance.

MAD pour l'imputation de données dans des séries temporelles

Algorithm 2: Imputation avec MAD par lots

Input: $\mathbf{X} \in (\mathbb{R} \cup \{NA\})^{n \times t \times d}, \cdot \in \{0,1\}^{n \times t \times d}, \mathbf{X}' \in (\mathbb{R} \cup \{NA\})^{n' \times t' \times d}, \cdot' \in \{0,1\}^{n' \times t' \times d}, \alpha, \eta, \epsilon > 0, n \geq m > 0,$

Initialisation : for k = 1, ..., d,

- pour i pour j t.q. $\omega_{ijk} = 0, \hat{x}_{ijk} \leftarrow \bar{x}^{obs}_{::k} + \epsilon_{ijk}$ *avec* $\epsilon \sim \mathcal{N}(0, \eta)$ *et* $\bar{x}^{obs}_{::k}$ correspondant à la moyenne des données observées dans la k-ème variable (données manquantes)
- pour i pour j t.q. $\omega_{ijk} = 1, \hat{x}_{ijk} \leftarrow x_{ijk}$ (données observées)

faire la même chose pour la donnée cible.

Pour iter $= 1, 2, ..., iter_{max}$ **faire**

Echantilloner deux ensembles K et L de m indices

$\mathcal{L}(\hat{\mathbf{X}}_K, \hat{\mathbf{X}}'_L) \leftarrow \mathbf{MAD}(\hat{\mathbf{X}}_K, \hat{\mathbf{X}}'_L)$

$\hat{\mathbf{X}}_K^{(imp)} \leftarrow \hat{\mathbf{X}}_K^{(imp)} - \alpha \mathbf{RMSprop}(\nabla_{\hat{\mathbf{X}}_K^{(imp)}} \mathcal{L})$

$\hat{\mathbf{X}'}_L^{(imp)} \leftarrow \hat{\mathbf{X}'}_L^{(imp)} - \alpha \mathbf{RMSprop}(\nabla_{\hat{\mathbf{X}'}_L^{(imp)}} \mathcal{L})$

Fin Pour

Output: $\hat{\mathbb{X}}, \hat{\mathbb{X}}'$

MAD pour l'imputation de données dans des séries temporelles

Algorithm 3: Imputation avec MAD et Sinkhorn par lots

Input: $\mathbf{X} \in (\mathbb{R} \cup \{NA\})^{n \times t \times d}, \cdot \in \{0,1\}^{n \times t \times d}, \mathbf{X}' \in (\mathbb{R} \cup \{NA\})^{n' \times t' \times d}, \cdot' \in \{0,1\}^{n' \times t' \times d}, \alpha, \eta, \epsilon > 0, n \geq m > 0,$

Initialisation : for k = 1, ..., d,

- pour i pour j t.q. $\omega_{ijk} = 0, \hat{x}_{ijk} \leftarrow \bar{x}_{::k}^{obs} + \epsilon_{ijk}$ *avec* $\epsilon \sim \mathcal{N}(0, \eta)$ *et* $\bar{x}_{::k}^{obs}$ correspondant à la moyenne des données observées dans la k-ème variable (données manquantes)
- pour i pour j t.q. $\omega_{ijk} = 1, \hat{x}_{ijk} \leftarrow x_{ijk}$ (données observées)

faire la même chose pour la donnée cible.

Pour iter $= 1, 2, ..., iter_{max}$ **faire**

Echantilloner quatre ensembles K, L, K' et L' de m indices

$\mathcal{L}(\hat{\mathbf{X}}_K, \hat{\mathbf{X}}'_L) \leftarrow \mathbf{MAD}(\hat{\mathbf{X}}_K, \hat{\mathbf{X}}'_L) + S_\epsilon(\hat{\mathbf{X}}'_{K'}, \hat{\mathbf{X}}'_{L'})$

$\hat{\mathbf{X}}_K^{(imp)} \leftarrow \hat{\mathbf{X}}_K^{(imp)} - \alpha \mathbf{RMSprop}(\nabla_{\hat{\mathbf{X}}_K^{(imp)}} \mathcal{L})$

$\hat{\mathbf{X}'}_{K' \cup L \cup L'}^{(imp)} \leftarrow \hat{\mathbf{X}'}_{K' \cup L \cup L'}^{(imp)} - \alpha \mathbf{RMSprop}(\nabla_{\hat{\mathbf{X}'}_{K' \cup L \cup L'}^{(imp)}} \mathcal{L})$

Fin Pour

Output: $\hat{\mathbb{X}}, \hat{\mathbb{X}}'$

Racine de l'Erreur Quadratique Moyenne

On considère la racine de l'erreur quadratique moyenne comme mesure pour évaluer la complétion.

Racine de l'Erreur Quadratique Moyenne

$$\sqrt{\frac{1}{m_0} \sum_{(i,j)|\omega_{ij}=0} (x_{i,j}^{true} - \hat{x}_{i,j})^2}$$

Génération synthétique de données manquantes

- **Manquantes au hasard** :
  Choisir des couples $(j, k) \in \{0, 1, ..., t\} \times \{0, 1, ..., d\}$ et de les considérer comme manquant pour tout $i \in \{0, 1, ..., n\}$.
- **Manquantes avec un biais** :
  Choisir des horodatages et d'y supprimer les données pour tout $i \in \{0, 1, ..., n\}$.

Manquantes au hasard

Pourcentage de données manquantes	Problèmes	Moyenne	Sinkhorn	MAD	MAD + Sinkhorn
50%	DK1 $\rightarrow$ FR1	1.2469	1.2528	1.1015	**1.0843**
	DK1 $\rightarrow$ FR2	1.1294	1.1305	1.0366	**1.0345**
	DK1 $\rightarrow$ AT1	1.1920	1.1944	**1.0343**	1.0346
	FR1 $\rightarrow$ DK1	1.1862	1.1894	1.0537	**1.0501**
	FR1 $\rightarrow$ FR2	1.1294	1.1305	1.0540	**1.0525**
60%	DK1 $\rightarrow$ FR1	1.2575	1.2614	1.1128	**1.1061**
	DK1 $\rightarrow$ FR2	1.1134	1.1141	1.0407	**1.0390**
	DK1 $\rightarrow$ AT1	1.1699	1.1716	1.0321	**1.0286**
	FR1 $\rightarrow$ DK1	1.1772	1.1791	**1.0567**	1.0568
	FR1 $\rightarrow$ FR2	1.1134	1.1141	1.0486	**1.0480**
70%	DK1 $\rightarrow$ FR1	1.2577	1.2599	1.1036	**1.1004**
	DK1 $\rightarrow$ FR2	1.0620	1.0627	1.0109	**1.0055**
	DK1 $\rightarrow$ AT1	1.1894	1.1904	1.0645	**1.0624**
	FR1 $\rightarrow$ DK1	1.2021	1.2031	1.1075	**1.1038**
	FR1 $\rightarrow$ FR2	1.0620	1.0627	1.0303	**1.0299**
80%	DK1 $\rightarrow$ FR1	1.2614	1.2626	1.1256	**1.1233**
	DK1 $\rightarrow$ FR2	1.1067	1.1072	1.0305	**1.0298**
	DK1 $\rightarrow$ AT1	1.2003	1.2016	1.0820	**1.0806**
	FR1 $\rightarrow$ DK1	1.2267	1.2311	1.1165	**1.1151**
	FR1 $\rightarrow$ FR2	1.1067	1.1072	1.0748	**1.0746**
90%	DK1 $\rightarrow$ FR1	1.2635	1.2605	1.1405	**1.1389**
	DK1 $\rightarrow$ FR2	1.0932	1.0937	1.0441	**1.0436**
	DK1 $\rightarrow$ AT1	1.1977	1.1961	**1.0863**	1.0900
	FR1 $\rightarrow$ DK1	1.2266	1.2247	1.1252	**1.1240**
	FR1 $\rightarrow$ FR2	1.0932	1.0937	**1.0549**	1.0588

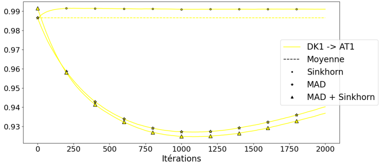
Table – Valeurs de RMSE pour différents pourcentages de données manquantes à 1000 itérations

Manquantes avec un biais

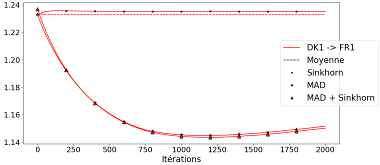
Pourcentage de données manquantes	Problèmes	Moyenne	Sinkhorn	MAD	MAD + Sinkhorn
50%	DK1 $\to$ FR1	1.1238	1.1313	**1.0579**	1.0604
	DK1 $\to$ FR2	1.0147	1.0167	**0.9488**	0.9510
	DK1 $\to$ AT1	0.9867	0.9911	0.9273	**0.9248**
	FR1 $\to$ DK1	1.1694	1.1725	1.1412	**1.1390**
	FR1 $\to$ FR2	1.0147	1.0167	0.9885	**0.9477**
60%	DK1 $\to$ FR1	1.2083	1.2144	1.1503	**1.1139**
	DK1 $\to$ FR2	1.0465	1.0478	0.9877	**0.9853**
	DK1 $\to$ AT1	0.9534	0.9574	**0.9424**	0.9445
	FR1 $\to$ DK1	**1.1259**	1.1286	1.1374	1.1390
	FR1 $\to$ FR2	**1.0465**	1.0478	1.1040	1.1013
70%	DK1 $\to$ FR1	1.2415	1.2454	1.1111	**1.1087**
	DK1 $\to$ FR2	0.9793	0.9810	0.9772	**0.9767**
	DK1 $\to$ AT1	0.9734	0.9757	0.9538	**0.9536**
	FR1 $\to$ DK1	1.2088	1.2104	**1.1429**	1.1505
	FR1 $\to$ FR2	**0.9793**	0.9810	0.9902	1.0016
80%	DK1 $\to$ FR1	1.2400	1.2429	1.1791	**1.1710**
	DK1 $\to$ FR2	1.0273	1.0283	1.0234	**1.0223**
	DK1 $\to$ AT1	1.0578	1.0587	0.9984	**0.9978**
	FR1 $\to$ DK1	1.2590	1.2599	1.2187	**1.2180**
	FR1 $\to$ FR2	1.0273	1.0283	**1.0122**	1.0140
90%	DK1 $\to$ FR1	1.2329	1.2352	1.1453	**1.1442**
	DK1 $\to$ FR2	**1.0631**	1.0639	1.1570	1.1580
	DK1 $\to$ AT1	**1.0473**	1.0484	1.0538	1.0542
	FR1 $\to$ DK1	1.2394	1.2402	**1.1556**	1.1626
	FR1 $\to$ FR2	**1.0631**	1.0639	1.1294	1.1296

Table – Valeurs de RMSE pour différents pourcentages de données manquantes à 1000 itérations

Evolution de la RMSE en fonction du nombre d'itérations



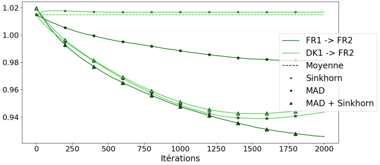
(a) 50% de données manquantes



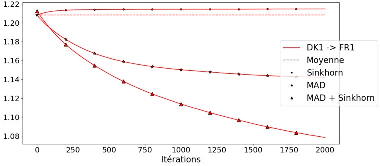
(b) 90% de données manquantes

Figure – RMSE en fonction du nombre d'itérations

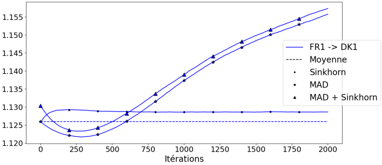
Evolution de la RMSE en fonction du nombre d'itérations



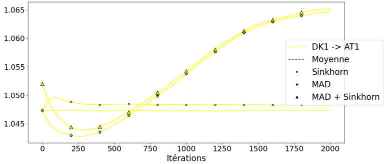
(a) 50% de données manquantes



(b) 60% de données manquantes



(c) 60% de données manquantes



(d) 90% de données manquantes

Figure – RMSE en fonction du nombre d'itérations

Conclusion