Régression de faible rang non-paramétrique pour réponses tensorielles

Guillaume RABUSSEAU, Hachem KADRI, François DENIS

Aix*Marseille Université - Laboratoire d'Informatique Fondamentale - Qarma
CMI, 39 rue Frédéric Joliot-Curie, 13453 Marseille cedex 13, France
{guillaume.rabusseau,hachem.kadri,francois.denis}@lif.univ-mrs.fr

Résumé – L'extension de méthodes d'apprentissage statistique aux données tensorielles est un sujet très étudié actuellement. Nous proposons une méthode de régression non-paramétrique de rang faible adaptée aux réponses ayant une structure tensorielle. Pour cela, nous introduisons une notion de noyau reproduisant à valeur tensorielle et nous proposons deux algorithmes de régression basés sur une pénalisation du rang du tenseur de régression. Ces algorithmes sont évalués et comparés à d'autres méthodes de régression multivariée sur des données artificielles.

Abstract – Extending univariate and multivariate methods to tensor-structured data is a challenging task which has recently received a growing interest. We propose a nonparametric reduced-rank regression method adapted to tensor-structured output data. We first generalize reduced-rank regression from vector to tensor-structured response variable. We then develop a novel nonparametric tensor approach relying on the notion of tensor-valued reproducing kernels, and we propose two learning algorithms to solve a rank penalized tensor regression problem. They are evaluated and compared with other multi-output regression methods through simulation study.

1 Introduction

De nombreux travaux récents cherchent à adapter des méthodes d'apprentissage statistique aux tableaux multi-dimensionnels (ou tenseurs). L'extension des méthodes de régression aux données tensorielles constitue un enjeu important. Cet article a un double objectif : définir un modèle de régression capable d'inférer la structure tensorielle des réponses et développer ce modèle dans un cadre non-paramétrique.

La régression à sorties multiples consiste à prédire p variables de réponse à partir de d variables prédictives. Dans ce contexte, appliquer la méthode des moindres carrés ordinaire revient à réaliser p régressions linéaires indépendantes. Le modèle de régression de faible rang (reduced-rank regression model) [3] propose d'apprendre une approximation de rang faible de la matrice de régression, en résolvant le problème suivant :

$$\widehat{\mathbf{W}} = \underset{W \in \mathbb{R}^{d\times p}}{\arg\min} \|\mathbf{X}\mathbf{W} - \mathbf{Y}\|_F^2 \quad \text{t.q. } rang(\mathbf{W}) \leq R, \quad (1)$$

où $R \in \mathbb{N}$, $\mathbf{W}$ est la matrice de régression, $\|\cdot\|_F$ désigne la norme de Frobenius, $\mathbf{X} \in \mathbb{R}^{N\times d}$ et $\mathbf{Y} \in \mathbb{R}^{N\times p}$ désignent resp. les matrices des données d'entrée et de sortie.

Pour résoudre une tache de régression avec sorties tensorielles (e.g., images 3D, signaux spatio-temporelles, etc.), on peut vectoriser les sorties et appliquer une méthode de régression à sortie scalaire ou vectorielle. Mais cette méthode ne permet pas de rendre compte de toutes les dépendances susceptibles d'exister dans des données naturellement structurées comme des tenseurs. Nous montrons dans cet article comment étendre le modèle de régression de rang faible aux réponses tensorielles en introduisant une contrainte de rang sur le tenseur de régression. Nous motivons cette approche en montrant comment cette technique de régularisation impose une collaboration forte entre les fonctions de régression induites par la structure tensorielle des réponses.

Par ailleurs, l'essentiel des travaux sur la régression de rang faible s'est concentré sur les modèles linéaires, avec peu d'extension au cadre non-paramétrique, même dans le cas de réponses vectorielles. Dans cet article, nous proposons de prendre en compte la structure tensorielle des réponses dans le cadre des noyaux à valeur opérateur. Les noyaux à valeur opérateur étendent les noyaux à valeur scalaire au cas des sorties multivariées. Depuis l'article fondateur [4], ces noyaux ont été appliqués avec succès à de nombreux problèmes d'apprentissage automatique. Mais à notre connaissance, ils n'ont jamais été utilisés dans le cadre de sorties à structure tensorielle. Nous introduisons la notion de noyaux à valeur tensorielle et nous montrons que ces noyaux constituent un outil puissant pour la régression non-paramétrique à réponses tensorielles.

2 Définitions et notations

Un tenseur $\mathcal{A} \in \bigotimes_{i=1}^{p} \mathbb{R}^{d_i} = \mathbb{R}^{d_1} \otimes \mathbb{R}^{d_2} \otimes \cdots \otimes \mathbb{R}^{d_p}$ peut être décrit par un tableau $(\mathcal{A}_{i_1 \cdots i_p}) \in \mathbb{R}^{d_1 \times \cdots \times d_p}$ à p dimensions.

Soit $I \subseteq [p] = \{1, \cdots, p\}$. On note $\mathcal{A}_{(I)}$ la matrice obtenue à partir de $\mathcal{A}$ en considérant les n-uplets $(i_j)_{j\in I}$ (resp. $(i_j)_{j\in[p]\setminus I}$) comme des indices de lignes (resp. de colonnes). Pour $n \in [p]$, la matricisation de $\mathcal{A}$ selon le mode n est définie par $\mathcal{A}_{(n)} = \mathcal{A}_{(\{n\})}$, et sa vectorisation par $vec(\mathcal{A}) = \mathcal{A}_{([p])}$.

On définit le produit extérieur $\mathcal{A} \otimes \mathcal{B} \in \bigotimes_{i=1}^{p+q} \mathbb{R}^{d_i}$ de $\mathcal{A} \in \bigotimes_{i=1}^{p} \mathbb{R}^{d_i}$ et $\mathcal{B} \in \bigotimes_{i=p+1}^{p+q} \mathbb{R}^{d_i}$ par $(\mathcal{A} \otimes \mathcal{B})_{i_1 \cdots i_p j_1 \cdots j_q} =$

$\mathcal{A}_{i_1\cdots i_p}\mathcal{B}_{j_1\cdots j_q}$, le produit scalaire de $\mathcal{T},\mathcal{U} \in \bigotimes_{i=1}^{p}\mathbb{R}^{d_i}$ par $\langle\mathcal{T},\mathcal{U}\rangle = \sum_{i_1\cdots i_p}\mathcal{T}_{i_1,\cdots i_p}\mathcal{U}_{i_1,\cdots i_p}$ et la norme de Frobenius de $\mathcal{T}$ par $\|\mathcal{T}\|_F^2 = \langle\mathcal{T},\mathcal{T}\rangle$.

Soit $\mathcal{A} \in \mathbb{R}^{I_1}\otimes\cdots\otimes\mathbb{R}^{I_m}\otimes\mathbb{R}^{J_1}\otimes\cdots\otimes\mathbb{R}^{J_n}$ et $\mathcal{B} \in \mathbb{R}^{J_1}\otimes\cdots\otimes\mathbb{R}^{J_n}\otimes\mathbb{R}^{K_1}\otimes\cdots\otimes\mathbb{R}^{K_p}$. Le produit $\mathcal{AB} \in \mathbb{R}^{I_1}\otimes\cdots\otimes\mathbb{R}^{I_m}\otimes\mathbb{R}^{K_1}\otimes\cdots\otimes\mathbb{R}^{K_p}$ est défini par la relation [1] $(\mathcal{AB})_{([m])} = \mathcal{A}_{([m])}\mathcal{B}_{([n])}$. La pseudo-inverse de $\mathcal{AB}$ est définie par $\left((\mathcal{AB})^{+}\right)_{([m])} = \left((\mathcal{AB})_{([m])}\right)^{+}$. Le tenseur identité $\mathcal{I} \in \mathcal{Y}\otimes\mathcal{Y}$ est l'unique tenseur vérifiant $\mathcal{IA} = \mathcal{AI} = \mathcal{A}$ pour tout $\mathcal{A} \in \mathcal{Y}$. Le tenseur $\mathcal{T} \in \mathcal{Y}\otimes\mathcal{Y}$ est dit symétrique (resp. semi-défini positif (s.d.p.)) si sa matricisation $\mathcal{T}_{([p])} \in \mathbb{R}^{d_1\cdots d_p\times d_1\cdots d_p}$ est symétrique (resp. semi-défini positive).

Le rang d'un tenseur $\mathcal{T} \in \bigotimes_{i=1}^{p}\mathbb{R}^{d_i}$ est le plus petit entier R tel que $\mathcal{T} = \sum_{r=1}^{R}\mathbf{v}_r^1\otimes\cdots\otimes\mathbf{v}_r^p$ où $\mathbf{v}_r^i \in \mathbb{R}^{d_i}$ pour tout $i \in [p]$. On note $[\![\mathbf{V}^1,\cdots,\mathbf{V}^p]\!] = \sum_{r=1}^{R}\mathbf{v}_r^1\otimes\cdots\otimes\mathbf{v}_r^p$ où chaque matrice $\mathbf{V}^i \in \mathbb{R}^{d_i\times R}$ a $\mathbf{v}_1^i,\cdots,\mathbf{v}_R^i$ pour colonnes.

Le produit de Kronecker des vecteurs $\mathbf{a} \in \mathbb{R}^p$ et $\mathbf{b} \in \mathbb{R}^q$ est défini par $\mathbf{a}\otimes_K\mathbf{b} = vec(\mathbf{a}\otimes\mathbf{b})$. Le produit de Khatri-Rao des matrices $\mathbf{A} = [\mathbf{a}_1,\cdots,\mathbf{a}_R] \in \mathbb{R}^{d_1\times R}$ et $\mathbf{B} = [\mathbf{b}_1,\cdots,\mathbf{b}_R] \in \mathbb{R}^{d_2\times R}$ est la matrice $\mathbf{A}\odot\mathbf{B} = [\mathbf{a}_1\otimes_K\mathbf{b}_1,\cdots,\mathbf{a}_R\otimes_K\mathbf{b}_R] \in \mathbb{R}^{d_1d_2\times R}$. Pour une liste de p matrices $\mathbf{U} = (\mathbf{U}^1,\cdots,\mathbf{U}^p)$ où $\mathbf{U}^1 \in \mathbb{R}^{d_1\times R},\cdots,\mathbf{U}^p \in \mathbb{R}^{d_p\times R}$, on note $\mathbf{U}_{\odot} = \mathbf{U}^p\odot\mathbf{U}^{p-1}\odot\cdots\odot\mathbf{U}^1$ et $\mathbf{U}_{\odot^i} = \mathbf{U}^p\odot\cdots\odot\mathbf{U}^{i+1}\odot\mathbf{U}^{i-1}\odot\cdots\odot\mathbf{U}^1$ pour $1 \le i \le p$. En utilisant ces notations, on peut montrer l'identité suivante : $\left([\![\mathbf{U}^1,\cdots,\mathbf{U}^p]\!]\right)_{(i)} = \mathbf{U}^i\mathbf{U}_{\odot^i}^{\top}$.

3 Régression pour réponses tensorielles

3.1 Modèle linéaire

Soit $\mathcal{X} = \mathbb{R}^d$ et $\mathcal{Y} = \mathbb{R}^{d_1\times\cdots\times d_p}$. Étant donné un échantillon $\{(\mathbf{x}^{(n)},\mathcal{Y}^{(n)})\}_{n=1}^{N} \subset \mathcal{X}\times\mathcal{Y}$ tiré suivant le modèle $\mathcal{Y} = f(\mathbf{x}) + \boldsymbol{\mathcal{E}}$ (où $\boldsymbol{\mathcal{E}}$ est le terme d'erreur), nous voulons estimer la fonction $f : \mathcal{X}\to\mathcal{Y}$.

Une approche possible pour résoudre ce problème serait de vectoriser les sorties et d'utiliser la méthode traditionnelle de régression de faible rang. Cependant, la structure de la sortie tensorielle serait perdue pendant l'étape de vectorisation ; les dépendances linéaires entre les composantes de la sortie seraient prises en compte, mais pas les *dépendances d'ordre supérieur* que l'on pourrait trouver dans une sortie tensorielle (e.g. les dépendances entre les lignes ou les colonnes d'une réponse matricielle).

Pour prendre en compte ces dépendances, la méthode de régression de faible rang pour réponses tensorielles utilise une contrainte de rang faible sur le tenseur de régression en résolvant le problème d'optimisation

$$\widehat{\mathcal{W}} = \underset{\mathcal{W}\in\mathbb{R}^{d\times d_1\times\cdots\times d_p}}{\arg\min} \|\mathbf{X}\mathcal{W} - \mathcal{Y}\|_F^2 \quad \text{t.q. } rang(\mathcal{W}) \le R, \quad (2)$$

pour un R donné, où $\mathbf{X} \in \mathbb{R}^{N\times d}$ est la matrice des entrées et $\mathcal{Y} \in \mathbb{R}^{N\times d_1\times\cdots\times d_p}$ le tenseur des sorties (i.e., $\mathcal{Y}_{n:\cdots:} = \mathcal{Y}^{(n)}$).

1. Les indices utilisés dans la matricisation seront claires dans chaque contexte d'utilisation.

Cette méthode offre l'avantage de réduire le nombre de paramètres du modèle de $dd_1\cdots d_p$ à $R(d + d_1 + \cdots + d_p)$, et elle permet de capturer non seulement les dépendances entre les composantes de la sortie, mais également les dépendances entre toutes ses fibres (e.g. entre les lignes et les colonnes d'une sortie matricielle). Une analyse plus fine montre que cette régularisation encourage une forte collaboration entre les fonctions de régression associées aux sous-tenseurs de la réponse. En effet, la Proposition 1 montre que la contrainte de rang sur le tenseur $\mathcal{W}$ implique une décomposition de la fonction de régression en deux composantes (voir Figure 1, gauche) : (i) une fonction G associe toute entrée $\mathbf{x} \in \mathcal{X} = \mathbb{R}^d$ à R tenseurs de rang 1 dans $\mathbb{R}^{d_2\times\cdots\times d_p}$, (ii) chaque sous-tenseurs $\mathcal{Y}_{j:\cdots:}$ de la réponse est obtenu par combinaisons linéaires de ces R tenseurs via un opérateur linéaire $\mathbf{A} \in \mathbb{R}^{d_1\times R}$. Les dépendances linéaires entre la sortie et l'entrée sont pris en compte par G, tandis que la structure de la réponse tensorielle est prise en compte par l'opérateur $\mathbf{A}$. Pour aller plus loin, soit $f_j : \mathcal{X}\to\mathbb{R}^{d_2\times\cdots\times d_p}$ la fonction de régression pour le sous-tenseur $\mathcal{Y}_{j:\cdots:}$; ces fonctions f_j sont des combinaisons linéaires des mêmes R *fonctions de rang 1* (i.e. des fonctions dont l'image ne contient que des tenseurs de rang 1), i.e. l'espace $span(\{f_1,\cdots,f_p\})$ est généré par R fonctions de rang 1 dans $\mathcal{X}\to\mathbb{R}^{d_2\times\cdots\times d_p}$ (voir Figure 1 (droite) pour une illustration dans le cas d'une réponse tensorielle d'ordre 3).

Proposition 1. *Soit $\mathcal{W}$ la solution de (2). La fonction $f : \mathcal{X}\to\mathcal{Y}$ définie par $f(\mathbf{x}) = \mathbf{x}\mathcal{W}$ se décompose en $f(\mathbf{x}) = \mathbf{A}G(\mathbf{x})$ où $\mathbf{A} \in \mathbb{R}^{d_1\times R}$ et G associe à chaque $\mathbf{x}$ un élément de $\left(\mathbb{R}^{d_2}\times\cdots\times\mathbb{R}^{d_p}\right)^R$, i.e. R tenseurs de rang 1 dans $\mathbb{R}^{d_2\times\cdots\times d_p}$.*

Cette proposition, ainsi que la discussion précédente, reste valide après permutation de la réponse. Par exemple, si l'on considère les fonctions $g_j : \mathcal{X}\to\mathbb{R}^{d_1\times d_3\times\cdots\times d_p}$ associées aux sous-tenseurs $\mathcal{Y}_{:j:\cdots:}$, f se décompose en $f(\mathbf{x}) = \mathbf{A}G(\mathbf{x})$ où $\mathbf{A}$ est maintenant une matrice de taille $d_2\times R$.

3.2 Noyaux à valeur tensorielle

Dans le cas général des noyaux à valeur opérateur, la fonction noyau est à valeur dans l'espace des opérateurs linéaires bornés de $\mathcal{Y}$ vers $\mathcal{Y}$. Lorsque l'espace de sortie $\mathcal{Y} = \mathbb{R}^{d_1\times\cdots\times d_p}$ est un produit tensoriel d'espaces vectoriels, l'espace des opérateur linéaires bornés de $\mathcal{Y}$ est isomorphe à l'espace $\mathcal{Y}\otimes\mathcal{Y} = \mathbb{R}^{d_1\times\cdots\times d_p\times d_1\cdots\times d_p}$. Nous appelons une fonction $K : \mathcal{X}\times\mathcal{X}\to\mathcal{Y}\otimes\mathcal{Y}$ un *noyau à valeur tensorielle*. Un tel noyau est *symétrique* si $\left(K(\mathbf{w},\mathbf{z})\right)_{([p])} = \left(K(\mathbf{z},\mathbf{w})\right)_{([p])}^{\top}$ $\forall\mathbf{w},\mathbf{z} \in \mathcal{X}$; il est *semi-défini positif* s'il est symétrique et si $\forall N \in \mathbb{N}, \mathbf{x}_1,\ldots,\mathbf{x}_N \in \mathcal{X}$, le tenseur de Gram $\mathbf{K} \in \mathbb{R}^N\otimes\mathcal{Y}\otimes\mathbb{R}^N\otimes\mathcal{Y}$, défini par $\mathbf{K}_{i:\cdots:j:\cdots:} = K(\mathbf{x}_i,\mathbf{x}_j)$, est un tenseur semi-défini positif.

Puisque $\mathcal{Y}$ est isomorphe à $\mathbb{R}^{d_1d_2\cdots d_p}$, K pourrait être simplement considéré comme un noyau à valeur matricielle. Cependant, ce serait équivalent à vectoriser les réponses tensorielles et, comme expliqué dans la Section 3.1, la structure tensorielle de l'espace de sortie serait perdue.

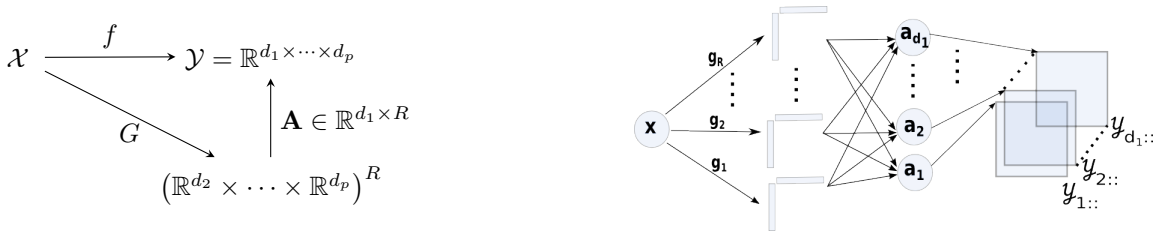

FIGURE 1 – (gauche) Décomposition d'une fonction f de faible rang où G associe à chaque $\mathbf{x}$ R tenseurs de rang 1. (droite) Collaboration entre les fonction de régression associées aux sous-tenseurs d'une réponse tensorielle d'ordre 3 .

Le théorème du représentant pour espace de Hilbert à noyau reproduisant (EHNR) à valeur vectorielle peut être reformuler comme suit pour les EHNR à valeur tensorielle.

Théorème 2. *Soit K un noyau à valeur tensorielle semi-défini positif et $\mathcal{F}$ l'EHNR associé. La solution $\widehat{f} \in \mathcal{F}$ de*

$$\widehat{f} = \arg\min_{f \in \mathcal{F}} \sum_{n=1}^{N} \ell(\mathcal{Y}^{(n)}, f(\mathbf{x}^{(n)})) + \lambda \|f\|_{\mathcal{F}}^2$$

s'écrit sous la forme $\widehat{f}(\cdot) = \sum_{n=1}^{N} K(\mathbf{x}^{(n)}, \cdot)\mathcal{C}^{(n)}$, où chaque $\mathcal{C}^{(n)} \in \mathcal{Y}$ est un tenseur d'ordre p.

Dans cet article, nous considérons uniquement des *noyaux séparables* de la forme $K(\cdot,\cdot) = k(\cdot,\cdot)\mathcal{T}$ où $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ est un noyau à valeur scalaire et $\mathcal{T} \in \mathcal{Y} \otimes \mathcal{Y}$ est un tenseur représentant les dépendances multilinéaires entre les composantes de la réponse. Il est facile de vérifier dans ce cas que K est s.d.p si et seulement si k est s.d.p et $\mathcal{T}$ est symétrique.

A l'instar des noyaux à valeur opérateur, les noyaux à valeur tensorielle permettent l'apprentissage non-paramétrique de fonctions de régression multivariées et la modélisation de dépendances non-linéaires entre les entrées et les sorties. De plus, en explicitant la structure tensorielle de la sortie, les noyaux à valeur tensorielle vont nous permettre d'ajouter une contrainte de rang tensoriel faible sur la fonction non-linéaire de régression, qui encouragera l'apprentissage d'interactions d'ordre supérieur dans la réponse tensorielle.

3.3 Modèle non-paramétrique

Soit $K(\cdot,\cdot) = k(\cdot,\cdot)\mathcal{T}$ un noyau à valeur tensorielle s.d.p. et séparable. Une façon d'apprendre la fonction f de manière non paramétrique consiste à l'estimer dans l'EHNR $\mathcal{F}_K$ associé au noyau K en résolvant le problème d'optimisation suivant :

$$\hat{f} = \arg\min_{f \in \mathcal{F}_K} \sum_{n=1}^{N} \|\mathcal{Y}^{(n)} - f(\mathbf{x}^{(n)})\|_F^2 + \lambda \|f\|_{\mathcal{F}_K}^2. \qquad (3)$$

De par le théorème du représentant, le minimiseur de (3) sécrit $\hat{f}(\cdot) = \sum_{n=1}^{N} k(\cdot, \mathbf{x}^{(n)}) \cdot \mathcal{T}\mathcal{C}^{(n)}$ où chaque $\mathcal{C}^{(n)} \in \mathcal{Y}$. En utilisant la notation du produit entre tenseurs introduite précédemment, et puisque $\mathcal{T}$ est symétrique, le terme d'attache aux données $\sum_{n=1}^{N} \|\mathcal{Y}^{(n)} - f(\mathbf{x}^{(n)})\|_F^2$ dans (3) se réécrit en $\|\mathbf{K}\mathcal{C}\mathcal{T} - \mathcal{Y}\|_F^2$, où $\mathbf{K} \in \mathbb{R}^{N \times N}$ est la matrice de Gram du noyau $k(\cdot,\cdot)$, et $\mathcal{C} \in \mathbb{R}^N \otimes \mathcal{Y} = \mathbb{R}^{N \times d_1 \times \cdots \times d_p}$ (resp. $\mathcal{Y} \in \mathbb{R}^N \otimes \mathcal{Y}$) est le tenseur obtenu en empilant les tenseurs $\mathcal{C}^{(n)}$ (resp. $\mathcal{Y}^{(n)}$) suivant le premier mode.

Tout comme dans le cas linéaire, la régularisation dans (3) ne tient pas compte de la structure tensorielle de la sortie. Pour remédier à ce problème, nous considérons au lieu de (3), le problème de minimisation sous contrainte de rang faible suivant, où le rang du tenseur $\mathcal{C}\mathcal{T}$ est fixé à un entier R donné. Pour éviter le sur-apprentissage et améliorer la robustesse au bruit, une régularisation sur la norme de Frobenius de $\mathcal{C}\mathcal{T}$ est ajoutée :

$$\min_{\mathcal{C} \in \mathbb{R}^N \otimes \mathcal{Y}} \frac{1}{2}\|\mathbf{K}\mathcal{C}\mathcal{T} - \mathcal{Y}\|_F^2 + \frac{\gamma}{2}\|\mathcal{C}\mathcal{T}\|_F^2 \text{ t.q. } rang(\mathcal{C}\mathcal{T}) \leq R. \qquad (4)$$

La fonction de régression $f : \mathcal{X} \to \mathcal{Y}$ s'écrit alors $f(\cdot) = \sum_{n=1}^{N} k(\mathbf{x}^{(n)}, \cdot)\mathcal{T}\mathcal{C}^{(n)}$. La proposition 3 montre que cette fonction f peut se décomposer comme dans la figure 1, ce qui suggère que la régression de rang tensoriel faible dans un EHNR étend naturellement le modèle linéaire présenté précédemment.

Proposition 3. *Soit $\mathcal{C} \in \mathbb{R}^{N \times d_1 \times \cdots \times d_p}$ la solution de (4). La fonction $f : \mathcal{X} \to \mathcal{Y}$ se décompose en $f(\mathbf{x}) = \mathbf{A}G(\mathbf{x})$ où $\mathbf{A} \in \mathbb{R}^{d_1 \times R}$ et G associe un élément de $\left(\mathbb{R}^{d_2} \times \cdots \times \mathbb{R}^{d_p}\right)^R$ à chaque $\mathbf{x}$.*

Tout comme dans le cas linéaire, cette proposition est toujours valide après permutation de la réponse. Le rôle de la matrice $\mathbf{A}$ dans la décomposition $f(\cdot) = \mathbf{A}G(\cdot)$ est le même que dans le cas linéaire, mais la fonction G encode maintenant les dépendances *non-linéaires* entre l'entrée et les R tenseurs de rang 1 utilisés pour reconstruire la réponse tensorielle.

3.4 Algorithmes

Nous proposons deux algorithmes de régression non-paramétrique pour réponses tensorielles. Dans le cas où le noyau à valeur tensorielle est séparable, (4) devient

$$\min_{\mathcal{C} \in \mathbb{R}^N \otimes \mathcal{Y}} \frac{1}{2}\|\mathbf{K}\mathcal{C} - \mathcal{Y}\|_F^2 + \frac{\gamma}{2}\|\mathcal{C}\|_F^2 \text{ t.q. } rang(\mathcal{C}) \leq R. \qquad (5)$$

Algorithme 1 approxime une solution de ce problème en utilisant la méthode des moindres carrés alternée (ALS) [2] sur les facteurs de la décomposition du tenseur de régression $\mathcal{C} = [\![\mathbf{U}^1, \cdots, \mathbf{U}^{p+1}]\!]$ qui découle de la contrainte de rang faible.

L'algorithme 2, quant à lui, apprend conjointement la fonction de régression et la partie tensorielle du noyau. Il permet d'approximer la solution du problème

$$\begin{aligned} \min_{\substack{\mathcal{C} \in \mathbb{R}^N \otimes \mathcal{Y}, \\ \mathcal{T} \in \mathcal{Y} \otimes \mathcal{Y}}} & \frac{1}{2}\|\mathbf{K}\mathcal{C}\mathcal{T} - \mathcal{Y}\|_F^2 + \frac{\gamma}{2}\|\mathcal{C}\mathcal{T}\|_F^2 \\ & \text{t.q. } rang(\mathcal{C}\mathcal{T}) \leq R, \mathcal{T} \text{ est s.d.p.}, \end{aligned} \qquad (6)$$

Algorithm 1 `TVK-RRR-I`

Input: Matrice de Gram $\mathbf{K} \in \mathbb{R}^N \times \mathbb{R}^N$, tenseurs des sorties $\mathcal{Y} \in \mathbb{R}^N \otimes \mathcal{Y}$, paramètre de régularisation R.
Output: $\mathcal{C} \in \mathbb{R}^N \otimes \mathcal{Y}$
Initialiser aléatoirement $\mathbf{U}^1 \in \mathbb{R}^N$ et $\mathbf{U}^{i+1} \in R^{d_i \times R}$ pour $i \in [p]$
repeat
$\mathbf{U}^1 \leftarrow ((\mathbf{KK} + \gamma\mathbf{I}))^+ \left(\mathbf{K}\mathcal{Y}\right)_{(1)} \mathbf{U}_{\odot^1} \left(\mathbf{U}_{\odot^1}^\top \mathbf{U}_{\odot^1}\right)^+$
for $i = 2, \cdots, p+1$ **do**
$\widetilde{\mathbf{U}} \leftarrow \left((\mathbf{KK} + \gamma\mathbf{I})\mathbf{U}^1, \mathbf{U}^2, \cdots, \mathbf{U}^{p+1}\right)$
$\mathbf{U}^i \leftarrow \left(\mathbf{K}\mathcal{Y}\right)_{(i)} \mathbf{U}_{\odot^i} \left(\widetilde{\mathbf{U}}_{\odot^i}^\top \mathbf{U}_{\odot^i}\right)^+$
end for
$\mathcal{C} \leftarrow [\![\mathbf{U}^1, \cdots, \mathbf{U}^{p+1}]\!]$
until convergence de $\mathcal{C}$

Algorithm 2 `TVK-RRR-T`

Input: Matrice de Gram $\mathbf{K} \in \mathbb{R}^N \times \mathbb{R}^N$, tenseurs des sorties $\mathcal{Y} \in \mathbb{R}^N \otimes \mathcal{Y}$, paramètre de régularisation R.
Output: $\mathcal{C} \in \mathbb{R}^N \otimes \mathcal{Y}$, $\mathcal{T} \in \mathcal{Y} \otimes \mathcal{Y}$ un tenseur s.d.p.
Initialiser aléatoirement $\mathbf{V}^i \in R^{d_i \times R}$ pour chaque $1 \leq i \leq p$
repeat
$\mathcal{T} \leftarrow [\![\mathbf{V}^1, \cdots, \mathbf{V}^p, \mathbf{V}^1, \cdots, \mathbf{V}^p]\!]$
$\mathcal{C} \leftarrow (\mathbf{KK} + \gamma\mathbf{I})^+ \mathbf{K}\mathcal{Y}\mathcal{T}(\mathcal{T}\mathcal{T})^+$
for $i = 1, \cdots, p$ **do**
$\mathbf{V}^i \leftarrow \arg\min_{\mathbf{V}^i} \frac{1}{2}\|\mathbf{K}\mathcal{C}\mathcal{T} - \mathcal{Y}\|^2 + \frac{\gamma}{2}\|\mathcal{C}\mathcal{T}\|_F^2$ (résolu par descente de gradient).
$\mathcal{T} \leftarrow [\![\mathbf{V}^1, \cdots, \mathbf{V}^p, \mathbf{V}^1, \cdots, \mathbf{V}^p]\!]$
end for
until convergence de $\mathcal{C}$ et $\mathcal{T}$

en utilisant ALS combinée avec une descente de gradient pour résoudre le problème de minimisation par rapport à $\mathcal{T}$.

4 Expérimentations

Sur différents jeux de données synthétiques, nous comparons notre méthode avec les méthodes suivantes : Multilinear Multitask Learning (`MLMTL`, une méthode multilinéaire utilisant une relaxation convexe de la régularisation sur le rang tensoriel [6]), Matrix-Valued Kernel Ridge Regression (`MVK-RR`, qui revient à effectuer une régression ridge à noyau indépendante pour chaque composante de la réponse [4]), Kernelized Reduced-Rank Ridge Regression (`K-RRR`, qui est équivalent à `TVK-RRR-I` après vectorisation de la réponse [5]). Nos algorithmes sont implémentés avec la *Tensor Toolbox* [1].

Soit $\mathcal{X} = \mathbb{R}^{10}$ et $\mathcal{Y} = \mathbb{R}^{5\times5\times5}$. Pour un noyau séparable à valeur tensorielle avec noyau scalaire $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ et tenseur $\mathcal{T} \in \mathcal{Y} \otimes \mathcal{Y}$, 100 fonctions de bases $k(\mathbf{b}_i, \cdot)$ sont générées en tirant chaque $\mathbf{b}_i \in \mathcal{X}$ suivant $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Pour un tenseurs $\mathcal{C} \in \mathbb{R}^{100} \otimes \mathcal{Y}$ donné, les données sont générées suivant la relation $\mathcal{Y}^{(n)} = \sum_{i=1}^{100} k(\mathbf{b}_i, \mathbf{x}^{(n)})\mathcal{T}\mathcal{C}_{i:::} + \boldsymbol{\xi}$, où $\mathbf{x}^{(n)}$ suit la loi $\mathcal{N}(\mathbf{0}, \mathbf{I})$ et $\boldsymbol{\xi}$ suit la loi $\mathcal{N}(0, 0.1)$. Trois jeux de données sont générés pour différentes caractéristiques du noyau à valeur tensorielle : (i) $k(\cdot,\cdot) = \langle\cdot,\cdot\rangle$ est le noyau linéaire, $\mathcal{T} = \mathcal{I}$ et $\mathcal{C}$ a rang 10 ; (ii) $k(\cdot,\cdot) = \exp(-\gamma\|\cdot - \cdot\|^2)$ est le noyau RBF avec $\gamma = 0.01$, $\mathcal{T} = \mathcal{I}$ et $\mathcal{C}$ a rang 10 ; (iii) $k(\cdot,\cdot)$ est le noyau RBF, $\gamma = 0.01$, $\mathcal{T}$ a rang 5 et $\mathcal{C}$ a rang 100 ($\mathcal{C}$ et $\mathcal{T}$ sont générés aléatoirement parmi les tenseurs du rang désiré).

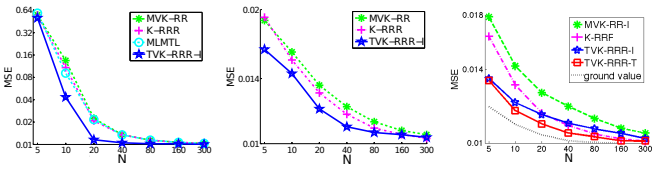

FIGURE 2 – Comparaisons sur données synthétiques. (gauche) (i) : noyau linéaire, $\mathcal{T} = \mathcal{I}$, $rang(\mathcal{C}) = 10$. (milieu) (ii) : noyau RBF, $\mathcal{T} = \mathcal{I}$, $rang(\mathcal{C}) = 10$. (droite) (iii) : noyau RBF, $rang(\mathcal{T}) = 5$, $rang(\mathcal{C}) = 100$ (échelles logarithmiques)

Les algorithmes d'apprentissage sont comparés pour différentes tailles du jeu de données d'apprentissage, 20 expériences sont réalisées pour chaque taille. Le rang R et le paramètre de régularisation γ sont choisis par validation croisée. La moyenne sur les 20 expériences de la MSE (erreur quadratique moyenne normalisée) sur les données de test est reportée dans la Figure 2, où l'on constate que nos algorithmes permettent d'obtenir de meilleurs performances de prédiction.

5 Conclusion

Nous avons montré que les dépendances linéaires d'ordre supérieur d'une réponse tensorielle peuvent être prises en compte par une régularisation sur le rang du tenseur de régression. Nous avons introduit la notion de noyaux à valeur tensorielle, offrant une nouvelle perspective pour la résolution non paramétrique des problèmes de régression avec réponses tensorielles. Notre méthode permet à la fois de modéliser des dépendances non-linéaires entre entrées et sorties, et de prendre en compte les dépendances d'ordre supérieur dans les sorties tensorielles.

Références

[1] B. Bader and T. Kolda. Algorithm 862 : MATLAB tensor classes for fast algorithm prototyping. *ACM Transactions on Mathematical Software*, 32(4) :635–653, 2006.

[2] P. Comon, X. Luciani, and A. L. F. De Almeida. Tensor decompositions, alternating least squares and other tales. *Jour. Chemometrics*, 23 :393–405, 2009.

[3] A. J. Izenman. Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, 5(2) :248–264, 1975.

[4] C. A. Micchelli and M. Pontil. On learning vector-valued functions. *Neural Computation*, 17 :177–204, 2005.

[5] A. Mukherjee and J. Zhu. Reduced rank ridge regression and its kernel extensions. *Statistical analysis and data mining*, 4(6) :612–622, 2011.

[6] B. Romera-Paredes, M. H. Aung, N. Bianchi-Berthouze, and M. Pontil. Multilinear multitask learning. In *ICML'2013*, pages 1444–1452, 2013.