Correction matricielle de l'indétermination d'échelle pour l'optimisation alternée

Massyl MOUDOUD Céline MEILLIER Vincent MAZET

ICube, Université de Strasbourg, CNRS (UMR 7357), 300 Bd Sébastien Brant, 67400 Illkirch-Graffenstaden, France

Résumé – L'apprentissage de dictionnaire souffre du problème d'indétermination d'échelle (multiplication d'un terme par une matrice diagonale et de l'autre terme par l'inverse de la matrice). La méthode AMORS répond à ce problème en intégrant un scalaire qui compense l'indétermination. Nous proposons une extension de cette méthode en intégrant une matrice diagonale qui permet de compenser chaque atome du dictionnaire. La méthode proposée est plus rapide et plus stable que AMORS pour l'apprentissage de dictionnaire dans le contexte de la connectivité fonctionnelle en IRMf.

Abstract – Dictionary learning suffers from the problem of scale indetermination (multiplication of one term by a diagonal matrix and the other term by the inverse of the matrix). The AMORS method addresses this problem by integrating a scalar that compensates for the indetermination. We propose an extension to this method by integrating a diagonal matrix allowing to compensate each atom of the dictionary. The proposed method is faster and more stable compared to AMORS for the dictionary learning in the context of functional connectivity in fMRI.

L'apprentissage de dictionnaire est un problème inverse qui apparait dans de nombreuses applications [2, 7]. Il existe une indétermination d'échelle dans la résolution de ce problème. En effet, le problème ne peut être résolu qu'à un facteur d'échelle près. La majorité des méthodes d'apprentissage de dictionnaire [2, 7] ignorent ce facteur d'échelle et se contentent de normaliser la solution *a posteriori* [1, 8].

Thé et. al [6] ont montré que les termes de régularisation du problème sont affectés par le facteur d'échelle. Ils ont présenté la méthode AMORS qui permet de prendre en compte le facteur d'échelle, de trouver sa valeur optimale et de l'utiliser pour ajuster les hyperparamètres de régularisation durant la résolution du problème, ce qui améliore la convergence.

Dans cet article, nous proposons d'étendre cette méthode en faisant en sorte que chaque atome du dictionnaire ait son propre facteur d'échelle. Dans la première partie, nous présentons la preuve de la nécessité de prendre en compte un facteur d'échelle par atome dans les problèmes d'apprentissage de dictionnaire. Ensuite, nous montrons comment tirer avantage de ces facteurs d'échelle pour améliorer la convergence des algorithmes d'optimisation. Enfin, nous présentons une application de notre méthode et un comparatif avec AMORS.

1 Méthode

Nous reprenons ci-après la méthode AMORS [6] en l'adaptant au cas d'un facteur d'échelle par atome du dictionnaire.

1.1 Le modèle d'apprentissage de dictionnaire

On considère le modèle d'apprentissage de dictionnaire

$$\mathbf{C} \approx \mathbf{D}\mathbf{A} \tag{1}$$

avec $\mathbf{C} \in \mathbb{R}^{E\times T}$ une matrice d'observations. L'objectif est d'exprimer cette matrice comme un produit des deux matrices $\mathbf{D} \in \mathbb{D} \subseteq \mathbb{R}^{E\times P}$ et $\mathbf{A} \in \mathbb{A} \subseteq \mathbb{R}^{P\times T}$. On peut voir $\mathbf{D}$ comme un dictionnaire d'atomes et $\mathbf{A}$ comme leurs amplitudes. Les ensembles $\mathbb{D}$ et $\mathbb{A}$ définissent l'espace des solutions qui peuvent être restreintes aux valeurs positives par exemple. Le modèle (1) n'étant pas exact, on cherche à trouver une paire de matrices $(\widehat{\mathbf{D}}, \widehat{\mathbf{A}})$ minimisant l'erreur de reconstruction de $\mathbf{C}$:

$$\widehat{\mathbf{D}}, \widehat{\mathbf{A}} = \underset{\mathbf{D},\mathbf{A}}{\arg\min} \|\mathbf{C} - \mathbf{D}\mathbf{A}\|_F^2 \,. \tag{2}$$

L'erreur de reconstruction est invariante à la multiplication de $\mathbf{A}$ par une matrice diagonale $\mathbf{\Gamma}$ et $\mathbf{D}$ par son inverse :

$$\|\mathbf{C} - \mathbf{D}\mathbf{A}\|_F^2 = \|\mathbf{C} - (\mathbf{D}\mathbf{\Gamma}^{-1})(\mathbf{\Gamma}\mathbf{A})\|_F^2$$
$$\text{avec} \quad \mathbf{\Gamma} \in \mathbb{R}^{P\times P} \text{ et } \quad \mathbf{\Gamma}_{ij} = \begin{cases} \boldsymbol{\gamma}_i \in \mathbb{R}_+^*, & \text{si } i = j \\ 0, & \text{sinon} \end{cases} \tag{3}$$

Cela signifie que si chaque atome du dictionnaire (colonne i de $\mathbf{D}$) est divisé par un facteur $\boldsymbol{\gamma}_i$, alors le problème (2) reste inchangé tant que l'amplitude correspondante dans la ligne i de $\mathbf{A}$ est multipliée par ce même facteur $\boldsymbol{\gamma}_i$.

Le problème (2) est mal posé au sens d'Hadamard. Pour le résoudre, il est courant d'ajouter des termes de régularisation sur $\mathbf{D}$ et $\mathbf{A}$ issus de connaissances *a priori*. Le problème (2) est donc remplacé par ($\boldsymbol{\lambda}, \boldsymbol{\mu} \in {\mathbb{R}_+^*}^P$) :

$$\widehat{\mathbf{D}}, \widehat{\mathbf{A}} = \underset{\mathbf{D},\mathbf{A}}{\arg\min}\ \mathcal{L}(\mathbf{D}, \mathbf{A}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \tag{4}$$
$$\mathcal{L}(\mathbf{D}, \mathbf{A}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \|\mathbf{C} - \mathbf{D}\mathbf{A}\|_F^2 + \boldsymbol{\lambda}^T\mathcal{J}(\mathbf{D}) + \boldsymbol{\mu}^T\mathcal{K}(\mathbf{A})$$
$$\boldsymbol{\lambda}^T\mathcal{J}(\mathbf{D}) + \boldsymbol{\mu}^T\mathcal{K}(\mathbf{A}) = \sum_{i=1}^{P} \left(\boldsymbol{\lambda}_i j(\mathbf{D}_i) + \boldsymbol{\mu}_i k(\mathbf{A}^i)\right)$$

où $\mathbf{D}_i$ est la colonne i de $\mathbf{D}$ et $\mathbf{A}^i$ la ligne i de $\mathbf{A}$. Contrairement à la formulation habituelle où les hyperparamètres de régularisation sont des scalaires, nous choisissons ici le cas plus général où les paramètres $\boldsymbol{\lambda}$ et $\boldsymbol{\mu}$ sont des vecteurs. L'intérêt est d'affecter une régularisation adaptée à chaque atome

Cette recherche a été financée par l'Agence Nationale de la Recherche au titre du projet DynaSTI ANR-22-CE45-0008.

du dictionnaire en réglant plus précisément les connaissances a priori. Par exemple, si un atome est supposé devoir être d'amplitude plus faible que les autres, alors son paramètre de régularisation propre sera augmenté. Il reste toutefois possible d'avoir le même poids de régularisation sur l'ensemble des inconnues en choisissant des vecteurs $\boldsymbol{\lambda}$ et $\boldsymbol{\mu}$ constants.

On impose que les termes de régularisations $\mathcal{J}$ et $\mathcal{K}$ soient séparables en fonction des colonnes de $\mathbf{D}$ et des lignes de $\mathbf{A}$ ce qui est le cas de la majorité des fonctions utilisées (normes ℓ_1, ℓ_2, variation totale anisotrope…) :

$$\mathcal{J}(\mathbf{D}) = \Big[j(\mathbf{D}_i)\Big]_{i\in[1,P]} \quad \mathcal{K}(\mathbf{A}) = \Big[k(\mathbf{A}^i)\Big]_{i\in[1,P]} \tag{5}$$

La fonction $\mathcal{L}$ est minimisée par une optimisation alternée des sous-problèmes sur $\mathbf{D}$ et $\mathbf{A}$:

$$\hat{\mathbf{D}} = \arg\min_{\mathbf{D}} \|\mathbf{C} - \mathbf{D}\mathbf{A}\|_F^2 + \boldsymbol{\lambda}^T \mathcal{J}(\mathbf{D})\,, \tag{6a}$$

$$\hat{\mathbf{A}} = \arg\min_{\mathbf{A}} \|\mathbf{C} - \mathbf{D}\mathbf{A}\|_F^2 + \boldsymbol{\mu}^T \mathcal{K}(\mathbf{A})\,. \tag{6b}$$

Les sous-problèmes sont convexes tant que $\mathcal{J}$ et $\mathcal{K}$ sont convexes et peuvent donc être optimisés avec des méthodes simples. Notons que des contraintes de positivité ou de support peuvent être ajoutées sans compromettre la méthode proposée.

1.2 Exploiter le facteur d'échelle

Afin de pouvoir utiliser l'indétermination de facteur d'échelle à notre avantage, on impose l'homogénéité de degré $r > 0$ et $q > 0$ des régularisations [6] :

$$\begin{aligned} j(\tfrac{1}{\boldsymbol{\gamma}_i}\mathbf{D}_i) = \boldsymbol{\gamma}_i^{-q} j(\mathbf{D}_i) \implies \mathcal{J}(\mathbf{D}\boldsymbol{\Gamma}^{-1}) = \boldsymbol{\Gamma}^{-q}\mathcal{J}(\mathbf{D}) \\ k(\boldsymbol{\gamma}_i\mathbf{A}_i) = \boldsymbol{\gamma}_i^{r} k(\mathbf{A}_i) \implies \mathcal{K}(\boldsymbol{\Gamma}\mathbf{A}) = \boldsymbol{\Gamma}^{r}\mathcal{K}(\mathbf{A}) \\ \forall \boldsymbol{\gamma}_i \in \mathbb{R}_+^* \quad \text{et} \quad \forall\, \mathbf{D}\in\mathbb{D}, \mathbf{A}\in\mathbb{A} \end{aligned} \tag{7}$$

D'après (4) et (7), on constate que la fonction de coût $\mathcal{L}$ avec les inconnues $\mathbf{D}$ et $\mathbf{A}$ multipliées par le facteur d'échelle est équivalente à la remise à l'échelle des hyperparamètres de régularisations :

$$\mathcal{L}(\mathbf{D}\boldsymbol{\Gamma}^{-1}, \boldsymbol{\Gamma}\mathbf{A}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \mathcal{L}(\mathbf{D}, \mathbf{A}, \boldsymbol{\Gamma}^{-q}\boldsymbol{\lambda}, \boldsymbol{\Gamma}^{r}\boldsymbol{\mu}) \tag{8}$$

En d'autres termes, le facteur d'échelle change la valeur des hyperparamètres de régularisations. Cela signifie que, si au cours des itérations de l'optimisation alternée, le facteur d'échelle évolue, alors le problème résolu évolue également. La matrice qui augmente en valeur sera plus pénalisée et celle qui diminue le sera moins. Cet effet va donc changer la forme des matrices estimées étant donné qu'elles sont la solution à une autre formulation du problème et non celle souhaitée. Cet effet a pour conséquence de déplacer les minima locaux.

Pour résoudre ce problème, on cherche la matrice diagonale $\boldsymbol{\Gamma}$ qui minimise la fonction de coût $\mathcal{L}$. On définit :

$$\begin{aligned} \mathcal{E}(\boldsymbol{\Gamma}) &= \mathcal{L}(\mathbf{D}\boldsymbol{\Gamma}^{-1}, \boldsymbol{\Gamma}\mathbf{A}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \\ &= \|\mathbf{C} - \mathbf{D}\mathbf{A}\|_F^2 + \sum_{i=1}^{P} \left(\boldsymbol{\lambda}_i \boldsymbol{\gamma}_i^{-q} j(\mathbf{D}_i) + \boldsymbol{\mu}_i \boldsymbol{\gamma}_i^{r} k(\mathbf{A}^i)\right) \end{aligned} \tag{9}$$

Les termes $\epsilon(\boldsymbol{\gamma}_i) = \boldsymbol{\lambda}_i \boldsymbol{\gamma}_i^{-q} j(\mathbf{D}_i) + \boldsymbol{\mu}_i \boldsymbol{\gamma}_i^{r} k(\mathbf{A}^i)$ sont chacun dépendant d'un élément $\boldsymbol{\gamma}_i$ de $\boldsymbol{\Gamma}$ et indépendants les uns des autres. Il est alors possible de minimiser par rapport à $\boldsymbol{\gamma}_i$ la fonction globale $\mathcal{E}(\boldsymbol{\Gamma})$ en minimisant chaque terme $\epsilon(\boldsymbol{\gamma}_i)$.

On distingue trois cas dans le calcul de la dérivée de $\epsilon(\boldsymbol{\gamma}_i)$:

1. Si $j(\mathbf{D}_i) \neq 0$ **et** $k(\mathbf{A}^i) \neq 0$ alors :

   $$\frac{\partial \mathcal{E}}{\partial \boldsymbol{\gamma}_i} = \frac{d\epsilon(\boldsymbol{\gamma}_i)}{d\boldsymbol{\gamma}_i} = -q\boldsymbol{\lambda}_i \boldsymbol{\gamma}_i^{-q-1} j(\mathbf{D}_i) + r\boldsymbol{\mu}_i \boldsymbol{\gamma}_i^{r-1} k(\mathbf{A}^i) \tag{10}$$

   Dans ce cas $\epsilon(\boldsymbol{\gamma}_i)$ est strictement convexe et son minimum est :

   $$\hat{\boldsymbol{\gamma}}_i = \left(\frac{q\boldsymbol{\lambda}_i j(\mathbf{D}_i)}{r\boldsymbol{\mu}_i k(\mathbf{A}^i)}\right)^{\frac{1}{r+q}} \tag{11}$$

2. Si $j(\mathbf{D}_i) = 0$ **et** $k(\mathbf{A}^i) = 0$, alors $\boldsymbol{\gamma}_i$ peut prendre une valeur arbitraire.

3. Si $j(\mathbf{D}_i) = 0$ **ou** $k(\mathbf{A}^i) = 0$, mais pas les deux, alors c'est un cas dégénéré qui n'a pas de solution. Il peut apparaitre pendant une estimation alternée de $\mathbf{D}$ et $\mathbf{A}$. Un atome du dictionnaire peut avoir une activation nulle mais quand même exister dans $\mathbf{D}$ ou vice-versa. Nous pouvons négliger ce cas et assigner une valeur arbitraire à $\boldsymbol{\gamma}_i$ car dès l'estimation suivante d'une des deux matrices, on retombe dans le cas 2.

L'algorithme 1 s'inspire de la méthode AMORS [6] pour incorporer la connaissance du facteur d'échelle optimal dans la procédure d'optimisation alternée. Après chaque estimation de l'une des deux matrices $\mathbf{D}$ et $\mathbf{A}$, le facteur d'échelle optimal est calculé pour chaque atome du dictionnaire avec (9). Lors de l'estimation suivante, le sous-problème (6a) ou (6b) est résolu en pondérant l'autre matrice (qui reste constante) par le facteur d'échelle optimal. Dans AMORS, c'est l'hyperparamètre de régularisation qui est pondéré. En théorie, selon (8), c'est équivalent. Cependant, nous avons constaté des problèmes numériques avec cette approche du fait de l'apparition de valeurs extrêmes à la limite de la précision machine.

Algorithme 1 : Minimisation alternée avec matrice de correction d'échelle optimale.

Entrées : $\mathbf{D}^{[0]}$, $\boldsymbol{\lambda}$, $\boldsymbol{\mu}$
$\boldsymbol{\Gamma} \leftarrow \mathbf{I}_P$;
$k \leftarrow 0$;
tant que *pas convergé* **faire**
  $\mathbf{A} \leftarrow \arg\min_{\mathbf{A}} \|\mathbf{C} - \mathbf{D}\boldsymbol{\Gamma}^{-1}\mathbf{A}\|_F^2 + \boldsymbol{\mu}^T\mathcal{K}(\mathbf{A})$;
  $\widetilde{\boldsymbol{\Gamma}} \leftarrow \arg\min_{\boldsymbol{\Gamma}} \mathcal{L}(\mathbf{D}\boldsymbol{\Gamma}^{-1}, \boldsymbol{\Gamma}\mathbf{A}, \boldsymbol{\lambda}, \boldsymbol{\mu})$;
  si *k = 0* **et** $\max_i \Big(\log(\widetilde{\boldsymbol{\Gamma}}_{ii}) - \log(\boldsymbol{\Gamma}_{ii})\Big) > \epsilon$ **alors**
    $\boldsymbol{\Gamma} \leftarrow \widetilde{\boldsymbol{\Gamma}}$;
    Retourner à la ligne : 4
  fin
  $\mathbf{D} \leftarrow \arg\min_{\mathbf{D}} \|\mathbf{C} - \mathbf{D}\widetilde{\boldsymbol{\Gamma}}\mathbf{A}\|_F^2 + \boldsymbol{\lambda}^T\mathcal{J}(\mathbf{D})$;
  $\boldsymbol{\Gamma} \leftarrow \arg\min_{\boldsymbol{\Gamma}} \mathcal{L}(\mathbf{D}\boldsymbol{\Gamma}^{-1}, \boldsymbol{\Gamma}\mathbf{A}, \boldsymbol{\lambda}, \boldsymbol{\mu})$;
  $k \leftarrow k + 1$;
fin
retourner $(\mathbf{D}, \mathbf{A})$

Il reste la question de l'initialisation du facteur d'échelle. Nous reprenons la méthode proposée dans [6]. Tout d'abord, le dictionnaire $\mathbf{D}$ est initialisé à une valeur définie par l'utilisateur et la matrice des facteurs d'échelle $\boldsymbol{\Gamma}$ est initialisée à l'identité. Les activations $\mathbf{A}$ sont ensuite estimées grâce à

l'équation (6b), puis la matrice des facteurs d'échelle est mise à jour. Cette opération est répétée tant que le facteur d'échelle change significativement, c'est-à-dire tant que :

$$\max_i \left(\log(\widetilde{\mathbf{\Gamma}}_{ii}) - \log(\mathbf{\Gamma}_{ii}) \right) < \epsilon \,. \tag{12}$$

Concernant le cas où un atome est à zéro et qu'on tombe dans le cas 2 ou 3 du calcul du facteur d'échelle optimal, on choisit de garder la valeur de $\boldsymbol{\gamma}_i$ précédente qui a fait que cet atome est estimé à zéro. Ce choix garantit la stabilité de la méthode, surtout pendant l'initialisation du facteur d'échelle.

Il faut bien noter que l'algorithme 1 ne résout pas le facteur d'échelle : il y aura toujours une indétermination de la solution à un facteur près par atome du dictionnaire. Notre méthode propose d'annuler l'effet du facteur d'échelle sur la résolution du problème. On garantit que les deux matrices sont régularisées au niveau souhaité, peu importe l'évolution du facteur d'échelle au cours des itérations. Grâce à la prise en compte du facteur d'échelle par atome du dictionnaire, nous allons encore plus loin et garantissons que tous les atomes sont pénalisés aux niveaux souhaités, même si les atomes évoluent vers différentes échelles de valeurs.

2 Application

La méthode proposée est testée pour la résolution du problème de dictionnaire lié à l'extraction d'unités de connectivité fonctionnel (UCF) dans l'étude de la dynamique de la connectivité fonctionnelle en IRMf [4, 5]. L'objectif est de décomposer les corrélations dans $\mathbf{C}$ en réseaux cérébraux définis dans $\mathbf{D}$ dont l'activité temporelle est estimée dans $\mathbf{A}$. Chaque ligne de $\mathbf{C}$ représente l'évolution dans le temps de la corrélation entre deux régions du cerveau. Dans nos travaux sur les UCF [4], les sous-problèmes (6a) et (6b) s'écrivent :

$$\min_{\mathbf{D}\in\mathbb{R}_+} \|\mathbf{C} - \mathbf{DA}\|_F^2 + \sum_{i=1}^{P} \left(\boldsymbol{\lambda}_i \|\mathbf{D}_i\|_2^2 \right) + \mathcal{I}_{\widetilde{\mathbf{D}}}(\mathbf{D}) \tag{13a}$$

$$\min_{\mathbf{A}\in\mathbb{R}_+} \|\mathbf{C} - \mathbf{DA}\|_F^2 + \sum_{i=1}^{P} \left(\boldsymbol{\mu}_i \|\mathbf{A}^i\|_1 + \boldsymbol{\nu}_i TV(\mathbf{A}^i) \right) \tag{13b}$$

Dans le sous-problème (13a), la pénalisation ℓ_2 définit le terme $\mathcal{J}$. Elle évite l'apparition de valeurs aberrantes dans le dictionnaire. Il y a aussi un ensemble de contraintes d'égalité avec le terme $\mathcal{I}_{\widetilde{\mathbf{D}}}(\mathbf{D})$. Elles sont définies par la matrice binaire $\widetilde{\mathbf{D}}$ qui impose la structure des réseaux (atomes du dictionnaire), c.-à-d. $\mathbf{D}_{ij} = 0$ si $\widetilde{\mathbf{D}}_{ij} = 0$. Elle est fournie par les biologistes qui définissent les réseaux d'intérêts pour l'étude.

Concernant (13b), $\mathcal{K}$ est composé de deux régularisations : une contrainte de parcimonie de type ℓ_1 et une régularisation de variation totale [3] sur les lignes de $\mathbf{A}$. Cette pénalisation combinée sur $\mathbf{A}$ n'empêche pas l'utilisation de notre méthode car la contrainte ℓ_1 et la variation totale sont toutes les deux homogènes d'ordre $r = 1$. On peut donc les regrouper dans une seule pénalisation en contrôlant l'influence relative des deux termes et l'influence globale avec deux hyperparamètres. Notons qu'il n'aurait pas été possible d'appliquer deux pénalisations homogènes d'ordres différents sur la même matrice, variation totale et ℓ_2 par exemple, car dans ce cas la solution explicite du facteur d'échelle n'est plus valide.

Les sous-problèmes (13a) et (13b) sont convexes, ils peuvent être résolus avec n'importe quelle méthode d'optimisation sous contraintes. Dans notre précédent travail [4], l'algorithme ADMM (*alternating direction method of mutliplier*s) a été utilisé. Dans ce travail, nous préférons utiliser l'algorithme FISTA (*fast iterative shrinkage thresholding algorithm*) qui donne la même solution que ADMM mais converge plus rapidement et avec une plus faible complexité algorithmique.

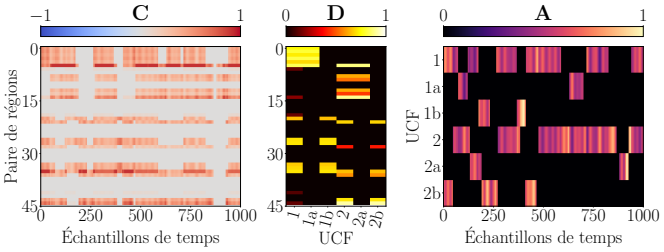
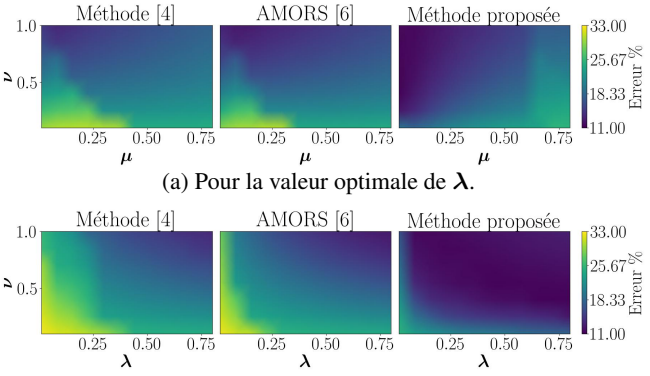


FIGURE 1 : Jeu de données du modèle direct $\mathbf{C} = \mathbf{DA}$ avec $(E, T) = (45, 1000)$ et $P = 6$ atomes.



(a) Pour la valeur optimale de $\boldsymbol{\lambda}$.

(b) Pour la valeur optimale de $\boldsymbol{\mu}$.

FIGURE 2 : Erreur de reconstruction de $\mathbf{C}$ en fonction des couples $(\boldsymbol{\nu}, \boldsymbol{\mu})$ (a) et $(\boldsymbol{\nu}, \boldsymbol{\lambda})$ (b).

3 Résultats

Dans cette section, nous comparons la méthode proposée à l'algorithme AMORS [6] et à notre précédente méthode [4] (optimisation alternée classique sans correction du facteur d'échelle, mais dans laquelle l'algorithme ADMM est remplacé par FISTA). Nous utilisons le jeu de données synthétiques figure 1, généré selon la procédure présentée dans [4]. Ces données sont créées selon le modèle direct (1) et fournissent donc la vérité terrain pour $\mathbf{C}$, $\mathbf{D}$ et $\mathbf{A}$.

La structure de $\mathbf{D}$ étant connue dans ce problème grâce à $\widetilde{\mathbf{D}}$, les algorithmes sont initialisés avec

$$\mathbf{D}_i^{[0]} = \sum_{t=1}^{T} \mathbf{C}_t \circ \widetilde{\mathbf{D}}_i, \tag{14}$$

c.-à-d. la moyenne temporelle de la matrice d'observations bruitée masquée par $\widetilde{\mathbf{D}}$ ($\circ$ est le produit élément par élément). Pour assurer la bonne convergence, l'optimisation alternée est calculée avec 500 itérations et chaque sous-problème effectue 500 itérations de FISTA.

L'effet des hyperparamètres de régularisations est étudié en exécutant les algorithmes pour des triplets d'hyperparamètres pris dans $\boldsymbol{\Theta} = (\boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\nu}) \in]0, 0.8] \times]0, 0.8] \times]0, 1]$ pour un total de 1008 combinaisons. L'erreur de reconstruction de $\mathbf{C}$,

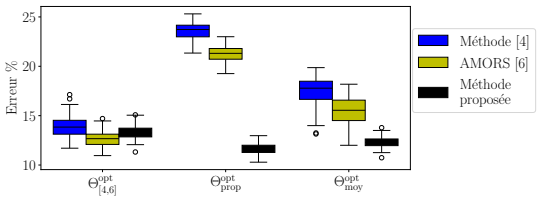


FIGURE 3 : Distribution de l'erreur de reconstruction de $\mathbf{C}$ normaliser en pourcentage pour les méthodes [4], AMORS [6] et la méthode proposée, avec trois jeux d'hyperamètres.

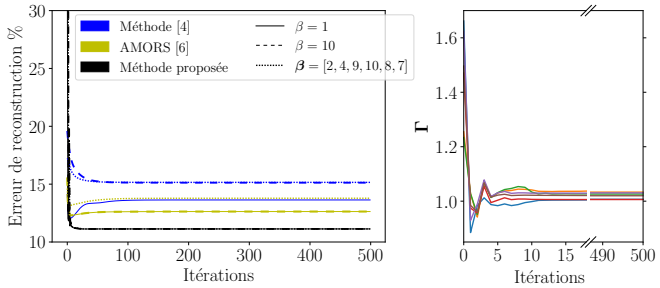


FIGURE 4 : (a) Courbes de convergences de l'erreur de reconstruction de la matrice $\mathbf{C}$ estimée par les trois méthodes : classique [4], AMORS ainsi que la méthode proposée,(b) évolution du facteur d'échelle optimal par atome du dictionnaire estimé par la méthode proposée.

normalisée en pourcentage, pour les trois algorithmes en fonction des couples $(\boldsymbol{\nu}, \boldsymbol{\mu})$ et $(\boldsymbol{\nu}, \boldsymbol{\lambda})$ est affichée dans les figures 2a et 2b pour la meilleure valeur du paramètre restant.

Nous constatons que AMORS a des résultats très similaires à l'optimisation alternée classique. La méthode proposée minimise l'erreur pour une grande plage d'hyperparamètres. Elle est en moyenne meilleure que les deux autres méthodes.

La stabilité d'estimation des méthodes est étudiée en générant 100 matrices $\mathbf{A}$ différentes pour un dictionnaire $\mathbf{D}$ fixé. Pour chaque matrice $\mathbf{C}$ résultante, les trois méthodes sont utilisées pour estimer la décomposition. Trois jeux d'hyperparamètres sont considérés : le triplet $\boldsymbol{\Theta}^{\text{opt}}_{[4,6]}$ minimisant l'erreur de $\mathbf{C}$ pour l'optimisation alternée et AMORS (qui se sont avérées avoir le même minimiseur sur ce jeu de données), celui minimisant l'erreur de la méthode proposée $\boldsymbol{\Theta}^{\text{opt}}_{\text{prop}}$ ainsi que la moyenne de ces minimiseurs $\boldsymbol{\Theta}^{\text{opt}}_{\text{moy}}$.

La distribution de l'erreur pour chaque méthode et jeu de paramètres est affichée figure 3. Dans la majorité des cas, notre méthode a une erreur inférieure et sa variance est plus faible, à l'exception du cas $\boldsymbol{\Theta}^{\text{opt}}_{[4,6]}$ où AMORS a de meilleurs résultats. De plus, notre méthode est plus stable à une variation des valeurs des hyperparamètres. Pour l'estimation de $\mathbf{D}$ et $\mathbf{A}$, l'erreur est significativement plus faible, y compris pour $\boldsymbol{\Theta}^{\text{opt}}_{[4,6]}$.

L'influence de l'échelle de la matrice d'initialisation $\mathbf{D}^{[0]}$ sur la convergence est évaluée. Pour ce faire, l'initialisation de (14) est multipliée par un facteur $\beta = 10$ (scalaire) et $\boldsymbol{\beta} = [2, 4, 9, 10, 8, 7]^T$ où l'échelle de chaque atome est modifiée. Ces deux cas sont comparés au cas $\beta = 1$ (14). Pour ces trois conditions, la valeur des hyperparamètres correspond au $\boldsymbol{\Theta}^{\text{opt}}$ de chaque méthode déterminé dans le cas $\beta = 1$ (*cf.* figure 2) Les courbes de convergence de l'erreur de reconstruction de $\mathbf{C}$ en pourcentage sont affichées figure 4 (a). Notre méthode garde la même vitesse de convergence et n'est pas affectée par l'initialisation contrairement à la méthode [4] et à AMORS qui convergent toutes les deux vers une moins bonne solution en fonction de l'initialisation.

Enfin, la figure 4 (b) montre l'évolution du facteur d'échelle optimal estimé par notre méthode pour chaque atome du dictionnaire. On voit clairement que chaque atome converge vers un facteur différent, confortant la nécessité d'optimiser le facteur d'échelle par atome du dictionnaire et non pour toute la matrice comme dans AMORS.

4 Conclusion

Nous avons proposé une extension de la méthode AMORS pour prendre en compte l'indétermination de facteur d'échelle par atome dans les problèmes d'apprentissage de dictionnaire. Cela permet de réduire le risque de tomber dans des cas dégénérés où un atome converge vers de très grandes valeurs et écrase les autres du fait de la régularisation.

La méthode proposée a été comparée à AMORS et à l'optimisation alternée classique sur des données synthétiques issues de l'étude de la connectivité fonctionnelle en IRMf. Les résultats ont montré que notre méthode améliore la convergence et obtient de meilleurs résultats. Enfin, notre méthode n'est pas affectée par l'échelle de l'initialisation. Un cas pratique de cette dernière propriété est le démélange hyperspectral [3]. Si les différents spectres initiaux des *end-members* sont estimés avec des instruments différents alors le résultat n'est pas affecté par le réglage de ces instruments.

Une perspective à ce travail est d'évaluer notre méthode sur des opérateurs bilinéaires autres que le produit matriciel, notamment le produit de convolution pour la résolution du problème de déconvolution.

Le code pour reproduire les résultats présentés ainsi que la génération des données des simulations est disponible à l'adresse : https://github.com/massylmoudoud/DynaSTI_Gretsi2025.

Références

[1] D. BENACHIR, Y. DEVILLE, S. HOSSEINI, M. S. KAROUI et A. HAMEURLAIN. "Hyperspectral image unmixing by non-negative matrix factorization initialized with modified independent component analysis". In : *WHISPERS*. 2013.

[2] P. COMON et C. JUTTEN, éd. *Handbook of blind source separation. Independent component analysis and applications*. Academic Press, 2010.

[3] M.-D. IORDACHE, J. M. BIOUCAS-DIAS et A. PLAZA. "Total Variation Spatial Regularization for Sparse Hyperspectral Unmixing". In : *IEEE Trans. Geosci. Remote Sens.* 50.11 (2012), p. 4484-4502.

[4] M. MOUDOUD, C. MEILLIER, M. SOURTY et V. MAZET. "Spatio-Temporal Model for Dynamic Functional Connectivity in Resting State fMRI Analysis". In : *EUSIPCO*. 2024.

[5] V. PORTMANN, C. MEILLIER et V. MAZET. "Analyse de la dynamique spatio-temporelle de la connectivité fonctionnelle cérébrale : données synthétiques et modélisation". In : GRETSI, 2023.

[6] S. THÉ, É. THIÉBAUT, L. DENIS et F. SOULEZ. "Exploiting the scaling indetermination of bi-linear models in inverse problems". In : *EUSIPCO*. 2020.

[7] I. TOSIC et P. FROSSARD. "Dictionary Learning". In : *IEEE Signal Process Mag.* (2011).

[8] H. YANG et C. SEOIGHE. "Impact of the Choice of Normalization Method on Molecular Cancer Class Discovery Using Nonnegative Matrix Factorization". In : *PLOS ONE* (2016).