Apprentissage multitâche en grande dimension : classification basée sur les covariances

Cyprien DOZ[1] Malik TIOMOKO[2] Chengfang REN[1] Jean-Philippe OVARLEZ[1,3]

[1]SONDRA, CentraleSupélec, Université Paris-Saclay

[2]Huawei Paris Research Center

[3]DEMR, ONERA, Université Paris-Saclay

Résumé – Cet article étudie le comportement asymptotique du Kernel Least Square Support Vector Machine dans le contexte de l'apprentissage multi-tâches pour des modèles de mélange gaussien en grande dimension avec de nombreux échantillons. La performance asymptotique de l'algorithme, validée sur des données synthétiques et réelles, met en évidence la relation entre les statistiques des données, les covariances en particulier, dans chaque tâche ainsi que les hyperparamètres reliant les tâches entre elles. Plus important encore, l'analyse permet d'améliorer la méthode en optimisant les labels.

Abstract – This article studies the asymptotic behavior of Kernel Least Square Support Vector Machine in the context of Multi Task Learning for Gaussian mixture models of high dimension with numerous samples. The asymptotic performance of the algorithm, validated on both synthetic and real data, sets forth the relation between the statistics of the data, covariances in particular, in each task as well as the hyperparameters relating the tasks together. More importantly the analysis allows for an improvement of the method by optimizing the labels.

1 Introduction

L'apprentissage multi-tâches (ou multi-task learning, MTL) consiste à apprendre simultanément plusieurs tâches de prédiction liées les unes aux autres. L'hypothèse sous-jacente est que les informations communes pertinentes pour la prédiction peuvent être partagées entre ces tâches. Ainsi, l'apprentissage conjoint de ces tâches permettrait une meilleure généralisation que l'apprentissage indépendant de chacune d'entre elles. L'apprentissage multi-tâches a été utilisé avec succès dans diverses applications de l'apprentissage automatique : traitement du langage naturel [1], reconnaissance vocale [2],vision par ordinateur [3] et découverte de médicaments [4].

Si, dans certains cas, de grandes améliorations ont été signalées par rapport à l'apprentissage d'une seule tâche, les praticiens ont également observé des résultats problématiques ; les performances de certaines tâches ayant diminué en raison de l'interférence des tâches. Prévoir quand et pour quelle tâche cela se produit est un défi exacerbé par le manque d'outils analytiques.

Basées sur un ensemble d'outils issus de la théorie des matrices aléatoires, les analyses statistiques en grande dimension ont émergé pour étudier un ensemble d'algorithmes d'apprentissage automatique, notamment le clustering spectral à noyau[5], la machine à vecteur de support des moindres carrés (ou least square support vector machine, LS-SVM) à tâche unique[6], l'apprentissage semi supervisé[7] et, plus intéressant pour cet article, l'apprentissage à tâches multiples[8]. Dans ce dernier, les auteurs étudient le comportement en grande dimension d'un algorithme de machine à vecteurs de support des moindres carrés à apprentissage multitâche (MTL LS-SVM) dans le cas d'un noyau linéaire. L'article met en évidence la relation entre les statistiques de premier ordre des données et les hyperparamètres du modèle. Cependant, le modèle linéaire pour le noyau ne permet pas de présenter les statistiques suffisantes dans le cas d'un algorithme basé sur la covariance. Dans cet article, nous essayons de combler cette lacune en fournissant une analyse théorique et une amélioration dans le cas de la classification des caractéristiques basée sur la covariance qui présente un grand intérêt pour l'apprentissage automatique et le traitement des signaux (classification EEG [9], hyperspectrale [10] et images SAR [11], par exemple).

Ce travail fait suite à la récente analyse spectrale du noyau du produit intérieur pour les caractéristiques basées sur la covariance effectuées pour une application dans le clustering [12]. Dans cet article, nous appliquons des techniques similaires à celles utilisées dans [12] pour gérer le cadre de l'apprentissage multitâche. Ainsi, les principales contributions de l'article sont résumées comme suit :

- Nous étudions théoriquement le comportement asymptotique de la machine à vecteur de support à moindre carré pour l'apprentissage multitâche en prédisant les statistiques du score de décision pour un modèle de caractéristiques basé sur la covariance.
- En conséquence, nous présentons les statistiques suffisantes dans le fonctionnement interne du cadre d'apprentissage multi-tâches et leur relation avec les hyperparamètres du modèle. Nous définissons ensuite un test statistique précis pour assigner une classe à de nouvelles données de test basées sur les statistiques asymptotiques du score.
- Nous fournissons un schéma LSSVM amélioré en optimisant les labels et nous effectuons des expériences sur des ensembles de données réelles et synthétiques pour corroborer tous les résultats précédents.

Notation. $\|.\|_{sp}$ désigne la norme spectrale. $e_m^{[n]} \in \mathbb{R}^n$ est le vecteur canonique de $\mathbb{R}^n$ avec $[e_m^{[n]}]_i = \delta_{mi}$. De plus, $e_{ij}^{[2k]} = e_{2(i-1)+j}^{[2k]}$. Les notations $A \otimes B$ et $A \odot B$ pour les matrices ou vecteurs A, B sont respectivement les produits de Kronecker et de Hadamard. $\mathcal{D}_x$ est la matrice diagonale contenant

sur sa diagonale les éléments du vecteur x. Enfin, $\mathbb{1}_m$ et I_m sont respectivement le vecteur de tous les 1 de dimension m et la matrice identité de dimension $m \times m$. La paire d'indices i, j fait généralement référence à la classe j dans la tâche i.

2 Modèle et hypothèses

Soit $X = [X_1, \ldots, X_k] \in \mathbb{R}^{p \times n}$ la collection de n vecteurs de données indépendants tirés de k "tâches". La tâche i est un problème de classification binaire à partir des échantillons d'apprentissage $X_i = [X_i^{(1)}, X_i^{(2)}] \in \mathbb{R}^{p \times n_i}$ avec $X_i^{(j)} = [x_{i1}, \ldots, x_{in_{ij}}] \in \mathbb{R}^{p \times n_{ij}}$ les n_{ij} vecteurs de classe $j \in \{1, 2\}$ pour la tâche i.

La machine à vecteurs de support des moindres carrés à apprentissage multitâche (MTL LS-SVM) vise à prédire un résultat $y \in \{-1, 1\}$ pour tout vecteur d'entrée $x \in \mathbb{R}^p$. À cette fin, MTL LS-SVM détermine k hyperplans de séparation ; chaque hyperplan caractérise une tâche de classification définie par $\omega_i^\mathsf{T} x + b_i$, avec ω_i le vecteur de direction de l'hyperplan de séparation pour la i-ième tâche de classification et b_i le terme de biais.

Pour intégrer les liens entre les tâches [13], on considère en outre que chaque vecteur ω_i peut être écrit sous la forme $\omega_i = v_i + \omega_0$, où ω_0 contient les informations "communes" entre les tâches et v_i est spécialisé pour chaque tâche. Ainsi, $\omega_i \sim \omega_0$ lorsque les tâches sont similaires tandis que $\omega_i \sim v_i$ lorsque les tâches ne sont pas liées.

MTL LS-SVM est caractérisé par le problème d'optimisation contraint :

$$\min_{\omega_0 \in \mathbb{R}^p, v_i \in \mathbb{R}^p, b \in \mathbb{R}^k} \quad \mathcal{J}(\omega_0, v_i, b)\,, \tag{1}$$

$$\textit{t.q. } \xi_i = y_i - \left(X_i^\mathsf{T} \omega_{i,lin} + b_i\right),\ \forall i \in \{1, \ldots, k\}\,, \tag{2}$$

où $\mathcal{J}(\omega_0, v_i, b) = \dfrac{1}{2\lambda}\|\omega_0\|^2 + \dfrac{1}{2}\displaystyle\sum_{i=1}^{k} \frac{\|v_i\|^2}{\gamma_i} + \sum_{i=1}^{k} \|\xi_i\|^2$.

Dans ce problème, $\lambda, \gamma = (\gamma_1, \ldots, \gamma_k)^T$ sont des paramètres de régularisation positifs et ξ_i mesure l'erreur que chacun des modèles ω_i fait sur les données d'apprentissage X_i. La fonction $\mathcal{J}(\omega_0, v_i, b)$ établit un compromis entre l'erreur de classification d'apprentissage sur toutes les tâches $\sum_{i=1}^k \|\xi_i\|^2$ et la complexité du modèle.

En introduisant le paramètre lagrangien α et en résolvant la formulation duale du problème d'optimisation, la solution de (1) est explicite (voir plus de détails dans [8, Appendix A.1]) et se lit comme suit :

$$\omega_{i,lin} = \left(e_i^{[k]\mathsf{T}} \otimes I_p\right) A\, Z\, \alpha_{lin}\,, \tag{3}$$

où $\alpha_{lin} = Q_{lin}\, y$, avec $y = [y_1^\mathsf{T}, y_2^\mathsf{T}, \ldots, y_k^\mathsf{T}]^\mathsf{T} \in \mathbb{R}^n$ et :

$$Q_{lin} = (Z^\mathsf{T} A Z + I_n)^{-1}\,, A = \left(\mathcal{D}_\gamma + \lambda\, \mathbb{1}_k\, \mathbb{1}_k^\mathsf{T}\right) \otimes I_p\,,$$

$$Z = \begin{pmatrix} X_1 & & \\ & \ddots & \\ & & X_k \end{pmatrix}.$$

Pour une tâche de classification i, nous voulons déterminer si une nouvelle donnée x appartient à la classe $i1$ ou $i2$. Nous définissons la fonction de décision pour tout $x \in \mathbb{R}^p$ pour la tâche i :

$$g_{i,lin}(x) \triangleq \omega_{i,lin}^\mathsf{T}\, x\,, \tag{4}$$

qui, après avoir résolu le problème Lagrangien mentionné ci-dessus, devient

$$g_{i,lin}(x) = \alpha_{lin}^\mathsf{T} Z^\mathsf{T} A\left(e_i^{[k]} \otimes x\right). \tag{5}$$

Le comportement asymptotique de $g_{i,lin}(x)$ a été largement étudié dans [8]. Cette étude ne discrimine toutefois les données que sur la base de leurs moyennes. Lorsqu'il s'agit d'une discrimination de caractéristiques basée sur la covariance, un noyau non linéaire est nécessaire.

La version non linéaire de l'apprentissage multitâche peut être facilement dérivée de la version linéaire. Afin d'éviter que le noyau non linéaire ne soit qu'une simple déformation d'un noyau linéaire, le noyau matriciel aléatoire suivant a été considéré dans [14, 15, 12] :

$$K_{ij} = p^{-1/2}\, f\left(\sqrt{p}\, x_i^\mathsf{T}\, x_j\right)\, \delta_{i \neq j}. \tag{6}$$

où f est une fonction continue dérivable. Pour une tâche unique, la procédure d'entraînement repose sur la matrice noyau $K = \Phi^{ST} = (kp)^{-1/2}\, f\left(\sqrt{kp}\, X^\mathsf{T} X\right)$, où c'est la même fonction f qui est appliqué élément par élément sur $\sqrt{kp}\, X^\mathsf{T} X$. La nouvelle fonction de décision, pour la tâche i, avec le noyau non linéaire est alors semblable à (5) :

$$g_i(x) = \kappa_i(x)^\mathsf{T}\, \alpha = \kappa_i(x)^\mathsf{T}\, (\Phi + I_n)^{-1}\, y\,,$$

où l'on utilise $\kappa_i(x) = (kp)^{-1/2}\, f\left(\sqrt{kp}\, Z^\mathsf{T}\, A\, \left(e_i^{[k]} \otimes x\right)\right)$ et $\Phi = (kp)^{-1/2}\, f\left(\sqrt{kp}\, Z^\mathsf{T}\, A\, Z\right)$ à la place du noyau linéaire $\left(e_i^{[k]} \otimes x\right)^\mathsf{T}\, A\, Z$ et de $Z^\mathsf{T}\, A\, Z$ utilisés dans l'équation (5).

En outre, une opération de centrage préliminaire est effectuée sur la matrice de données du noyau et sur les labels à l'aide d'une matrice de centrage $P = \left(I_n - n^{-1} \mathbb{1}_n \mathbb{1}_n^\mathsf{T}\right)$. Cette opération de centrage permet d'éliminer les biais et les termes résiduels indésirables dans la dérivation théorique comme dans [7, 16]. Nous travaillerons donc systématiquement avec la matrice à noyau centré :

$$\mathring{k}_i(x) = P\, k_i(x),\ \mathring{\Phi} = P\, \Phi\, P\,.$$

Grâce à cette opération de centrage, la prédiction de la classe de tout nouveau point de données x est alors obtenue à partir du score de classification :

$$\mathring{g}_i(x) = \mathring{k}_i(x)^\mathsf{T} \left(\mathring{\Phi} + I_n\right)^{-1} P\, y, \tag{7}$$

où nous rappelons que $\mathring{k}_i(x)$ et $\mathring{\Phi}$ sont le vecteur et la matrice de noyau centré. Il est maintenant nécessaire de dériver l'expression analytique des performances en étudiant le comportement statistique du score de décision $g_i(x)$. Pour cela, nous devons introduire les hypothèses suivantes sur la distribution des données et sur le taux de croissance de p et n.

Considérons $x_1, \ldots, x_n \in \mathbb{R}^p$ une suite de vecteurs aléatoires gaussiens indépendants. Pour tout $n_{11}, \ldots, n_{k2}$ tels que $n_{11} + \ldots + n_{k2} = n$, supposons que :

$$x_{n_{i(j-1)}+\ldots+n_{ij}} \sim \mathcal{N}(0, p^{-1}\Sigma_{ij}) \text{ avec } \Sigma_{11}, \ldots, \Sigma_{k2} \in \mathbb{R}^{p \times p}.$$

Définissons $\Sigma^0 = n^{-1} \displaystyle\sum_{i,j} n_{ij} \Sigma_{ij}$ et pour chaque tâche i et classe j, $\Sigma_{ij}^0 = \Sigma_{ij} - \Sigma^0$. Les matrices $\Sigma_{11}, \ldots, \Sigma_{k2}$ satisfont en outre aux hypothèses suivantes :

Hypothèse 1. *Lorsque $p \to \infty$, les hypothèses suivantes sont valables :*

1. $\frac{kp}{n} \xrightarrow{\text{a.s.}} c_0 \in (0,\infty)$, $\forall i \in \{1,\dots,k\}$ *et* $j \in \{1,2\}$, $n_{ij}/n = c_{ij}$, $n_{ij}/n = c_{ij}$, $n_{ij}/n = c_{ij}$,
2. $\forall i, i' \in \{1,\dots,k\}$ *et* $j, j' \in \{1,2\}$, $p^{-1}\text{tr}\left(\Sigma^0_{ij}\Sigma^0_{i'j'}\right) = \mathcal{O}(p^{-1/2})$,
3. $\max_{1\le i\le k, 1\le j\le 2} \limsup \|\Sigma_{ij}\|_{sp} < \infty$,

impliquant $(kp)^{-1}\text{tr}\left(\Sigma^0\right)^2 < \infty$ et $(kp)^{-1}\,\text{tr}\left(\Sigma^0\right)^4 < \infty$.

Le choix du noyau a été analysé dans [5, 16], notamment la fonction non linéaire $f(x) = x^2$ est asymptotiquement optimale sous les hypothèses précédentes.

L'objectif de cet article est d'étudier la fonction $\mathring{g}_i$. Cette fonction dépend de la matrice aléatoire Z à travers la matrice $\mathring{Q} = \left(\mathring{\Phi} + zI_p\right)^{-1}$ appelée la résolvante de la matrice $\mathring{\Phi}$.

3 Résultats asymptotiques

Sous l'hypothèse 1, la performance de l'algorithme MTL LS-SVM pour le modèle de caractéristiques basé sur la covariance dépendra des hyperparamètres par l'intermédiaire de la matrice $\mathcal{A}$, de la matrice des données X par l'intermédiaire de la matrice $\mathcal{T}$ et du nombre d'échantillons n_{ij} par l'intermédiaire de $\mathcal{P}_c$ définies de la manière suivante :

$$\mathcal{A} = \mathcal{D}_\gamma + \lambda\, \mathbb{1}_k\, \mathbb{1}_k^\mathsf{T}, \quad \mathcal{P}_c = \frac{1}{c_0}\left(\mathcal{D}_c - c\, c^\mathsf{T}\right),$$

$$\mathcal{T} = \sum_{i,a=1}^{k}\sum_{j,b=1}^{2} \frac{1}{\sqrt{kp}}\,\text{tr}\left(\Sigma_{ij}\Sigma_{ab}\right) e_{ij}e_{ab}^\mathsf{T}.$$

Les trois matrices $\mathcal{A}$, $\mathcal{T}$ et $\mathcal{P}_c$ sont combinées dans la matrice principale :

$$\Gamma = \left(I_{2k} + \tau\,\bar{T}\mathcal{P}_c\right)^{-1},\ \bar{T} = \left(\mathcal{A}^{\odot 2} \otimes \mathbb{1}_2\mathbb{1}_2^\mathsf{T}\right) \odot \mathcal{T}, \quad (8)$$

avec τ solution de l'équation $\tau = -\dfrac{1}{1 + c_0\omega^2\tau}$.

En relation avec les statistiques suffisantes $\bar{T}$, nous définissons $\tilde{T}^{(ij)}$ (défini pour chaque classe j dans la tâche i) qui interviendra dans la variance du score :

$$\tilde{T}^{(ij)} = (\mathcal{A}_{\cdot i}^{\odot 2}\mathcal{A}_{\cdot i}^{\odot 2^\mathsf{T}} \otimes \mathbb{1}_2\mathbb{1}_2^\mathsf{T}) \odot \tilde{t},$$

$$\tilde{t} = \sum_{a,b,a',b'} \frac{1}{kp}\text{tr}\left(\Sigma_{ab}\Sigma_{ij}\Sigma_{a'b'}\Sigma_{ij}\right) e_{ab}^{[2k]}{e_{a'b'}^{[2k]}}^\mathsf{T}.$$

En outre, étant donné que $x_{i1},\dots,x_{in_i}$ sont des vecteurs de données i.i.d., nous imposons des scores égaux $y_{i1} = \dots = y_{in_i}$ au sein de chaque classe. Ainsi, nous pouvons réduire le vecteur de score total $y \in \mathbb{R}^n$ sous la forme $y = \left[\tilde{y}_{11}\mathbb{1}_{n_{11}}^\mathsf{T},\dots,\tilde{y}_{k2}\mathbb{1}_{n_{k2}}^\mathsf{T}\right]^\mathsf{T}$ pour $\tilde{y} = [\tilde{y}_{11},\dots,\tilde{y}_{k2}]^\mathsf{T} \in \mathbb{R}^{2k}$.

Théorème 1. *Avec les hypothèses 1 et les notations introduites précédemment, pour un vecteur gaussien aléatoire x avec $\mathbb{E}[x] = 0_p$ et $\text{Cov}[x] = \Sigma_{ij}$, on obtient :*

$$E[\mathring{g}_i(x)] - m_{ij} \to 0, \quad Var[g_i(x)] - \sigma_{ij}^2 \to 0$$

où $m = [m_{11},\dots,m_{k2}]^\mathsf{T}$, $m = (I_{2k} - \Gamma)\,\tilde{y}$ *et* $\sigma_{ij}^2 = \dfrac{2\,\tau^2}{c_0}\,\tilde{y}^\mathsf{T}\,\Gamma\,\mathcal{K}^{(ij)}\,\Gamma\tilde{y}$ *avec* $\mathcal{K}^{(ij)} = c_0\,\mathcal{P}_c\,\tilde{T}^{(ij)}\,\mathcal{P}_c + \mathcal{D}_{c\odot\tilde{v}} - 2\,c\,(c^\mathsf{T} \odot \tilde{v}^\mathsf{T}) + (c^\mathsf{T}\,\tilde{v})\,c\,c^\mathsf{T} + \mathcal{C}\,\mathcal{P}_c$ *avec* $\tilde{v} = [\tilde{v}_{11},\dots,\tilde{v}_{k2}]$, $\tilde{v}_{ab} = \dfrac{1}{kp}\,\bar{T}^2_{(ij)(ab)}$ *et* $\mathcal{C} = \dfrac{c^\mathsf{T}\,\tilde{v}\,\omega^2\,\tau^2}{1 - \omega^2\,c_0\,\tau^2}$ *et* $\omega = \sqrt{2(kp)^{-1}\,\text{tr}\left(\Sigma^0\right)^2}$.

De même que dans [8, Theorem 4], le theorème 1 indique que les statistiques (asymptotiques) des scores de classification $g_i(x)$, pour $1 \le i \le k$, se réduisent à une simple fonction de vecteurs et de matrices déterministes à $2k$-dimensions. En particulier, les statistiques dépendent de la covariance des données Σ_{ij} par l'intermédiaire de la matrice $\mathcal{T}$ à la différence de [8] où les statistiques suffisantes étaient les moyennes des données, et les hyperparamètres $\lambda, \gamma_1, \dots, \gamma_k$ principalement par l'intermédiaire de la matrice Γ à $2k$ dimensions.

Ainsi, le théorème 1 présente $\mathcal{T}$ comme la statistique suffisante de l'apprentissage multitâche pour les données de caractéristiques basées sur la covariance.

En fournissant les statistiques asymptotiques de premier ordre et de second ordre du score de décision, on peut concevoir un test de décision pour décider de l'allocation de toute nouvelle donnée x dans la tâche i comme suit :

$$y(x) = \arg\max_j \sigma_{ij}^{-1}(x - m_{ij})^2\,,$$

De plus, le résultat asymptotique du théorème 1 dépend des labels $\tilde{y}$. On peut alors trouver le label[1] $\tilde{y}^\star = \arg\max_{\tilde{y}\in\mathbb{R}^{2k}} \dfrac{(m_{i1} - m_{i2})^2}{\sigma_{i1}^2}$ donné par

$$\tilde{y}^\star = \arg\max_{\tilde{y}\in\mathbb{R}^{2k}} \frac{\left\|\tilde{y}^\mathsf{T}\mathcal{D}_c\mathcal{T}(e_{i1}^{[2k]} - e_{i2}^{[2k]})\right\|^2}{\tilde{y}^\mathsf{T}\mathcal{K}^{(i1)}\tilde{y}}\,,$$

$$= {\mathcal{K}^{(i1)}}^{-1}(I_{2k} - \Gamma)(e_{i1}^{[2k]} - e_{i2}^{[2k]})\,.$$

4 Simulations et Applications

4.1 Application aux données synthétiques

Nous considérons le cadre suivant à deux tâches (k = 2) et deux classes (m = 2) : pour la tâche 1, $x_{1l}^{(j)} \sim \mathcal{N}(0_p, \Sigma_{1j})$ et pour la tâche 2, $x_{1l}^{(j)} \sim \mathcal{N}(0_p, \Sigma_{2j})$. Dans la figure 1, nous étudions l'impact de nombreuses tâches ($k > 2$) dans le cadre d'une classification binaire, mettant ainsi l'accent sur les effets du transfert négatif et sa correction par l'optimisation du label d'entrée. La figure illustre que notre paramétrage proposé évite le transfert négatif, puisque l'erreur de classification de MTL n'augmente jamais avec le nombre de tâches à la différence du schéma non optimisé qui souffre sévèrement du transfert négatif.

4.2 Application à la classification d'images PolSAR

Nous nous penchons ensuite sur la classification d'images de radar polarimétrique à synthèse d'ouverture (PolSAR). Dans la classification des images PolSAR, la discrimination entre les différentes classes se fait sur la base des matrices de covariance.

[1] puisque σ_{ij}^2 n'est pas indépendant de la classe j. La prise en compte de la classe j conduirait à un problème d'optimisation non convexe.

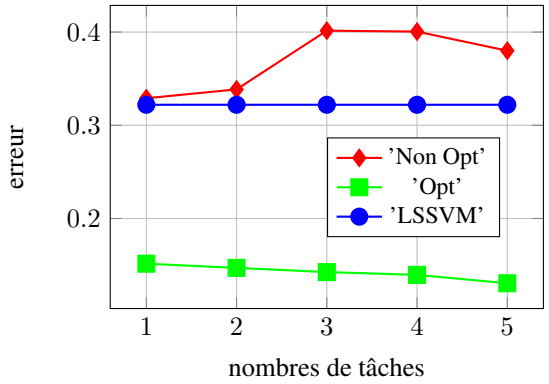

Figure 1: Erreur de classification pour un nombre croissant de tâches, label optimisé vs. label classique. La tâche unique est représentée par "LSSVM" de paramètres suivants : $\Sigma_{ij} = \alpha_{ij}\tau(\beta_{ij})$ avec $\alpha_{i1} = 4$, $\alpha_{i2} = 5$, $\beta_{i1} = 0.3$, $\beta_{i2} = 0.2$ $p = 100$. Même nombre d'échantillons par tâche et par classe, $\gamma = \mathbb{1}_k$, $\lambda = 1$.

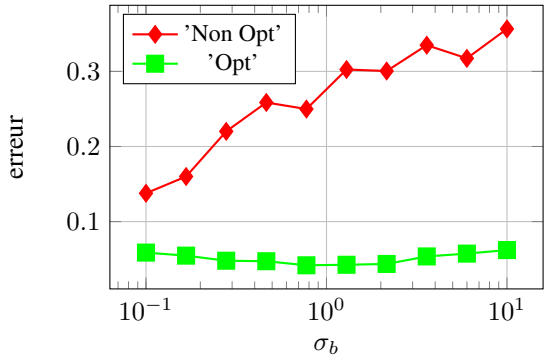

Figure 2: Erreur de classification en fonction de l'écart-type σ_b du bruit ajouté aux tâches sources. Apprentissage par transfert pour un ensemble de 4 tâches représentant la même image pour 4 bandes spectrales différentes. $\gamma = \mathbb{1}_k$, $\lambda = 1$. Bruit gaussien de variance σ_b^2 ajouté à chaque échantillon $x_{il}^{(j)}$ de la classe j de la tâche i dans la source, $c = \begin{bmatrix} 0.16\mathbb{1}_6 & 0.01\mathbb{1}_2 \end{bmatrix}$, $\gamma = \mathbb{1}_4$, $\lambda = 1$ et $p = 216$. $1,481$ échantillons de test utilisés pour évaluer l'erreur de classification.

Nous utilisons le même prétraitement que décrit dans [17] pour vectoriser les images SAR[2] en un vecteur à 216 dimensions pour chaque pixel. Nous considérons le problème du transfert de la classification des images PolSAR d'une bande de fréquence à une autre. Pour ce faire, nous considérons quatre bandes spectrales adjacentes obtenues en subdivisant la gamme de fréquences (70 MHz) en quatre bandes de fréquences également espacées. La classification dans la première bande est considérée comme la tâche cible tandis que les autres bandes sont considérées comme trois sources pour "aider" la classification des tâches. Nous ajoutons progressivement un bruit gaussien à la tâche de données source et comparons le noyau LSSVM avec des labels classiques à celui optimisé. Cette comparaison est illustrée à la figure 2.

5 Conclusion

À travers l'exemple de l'apprentissage multi-tâches appliqué aux données de caractéristiques basées sur la covariance, cet article démontre la capacité de la théorie des matrices aléatoires à prédire et à améliorer les performances des schémas avancés d'apprentissage automatique et de traitement des signaux. Dans ce contexte de modèle de mélange gaussien, on peut supposer l'optimalité des méthodes des moindres carrés, ce qui ouvre la possibilité de prouver que l'approche MTL LS-SVM est probablement proche de l'optimalité, même dans un contexte de données réelles. Les statistiques suffisantes dans ce cadre peuvent donc être utilisées pour dériver des limites de la théorie de l'information sur l'apprentissage par transfert pour un modèle basé sur des caractéristiques de covariance.

[2] acquises en bande X par le système ONERA-SAR à Brétigny, France.

References

[1] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the 25th international conference on Machine learning*, pp. 160–167, 2008.

[2] L. Deng, G. Hinton, and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: An overview," in *2013 IEEE international conference on acoustics, speech and signal processing*, pp. 8599–8603, IEEE, 2013.

[3] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015.

[4] B. Ramsundar, S. Kearnes, P. Riley, D. Webster, D. Konerding, and V. Pande, "Massively multitask networks for drug discovery," *arXiv preprint arXiv:1502.02072*, 2015.

[5] R. Couillet, F. Benaych-Georges, *et al.*, "Kernel spectral clustering of large dimensional data," *Electronic Journal of Statistics*, vol. 10, no. 1, pp. 1393–1454, 2016.

[6] Z. Liao and R. Couillet, "A large dimensional analysis of least squares support vector machines," *IEEE Transactions on Signal Processing*, vol. 67, no. 4, pp. 1065–1074, 2019.

[7] X. Mai and R. Couillet, "A random matrix analysis and improvement of semi-supervised learning for large dimensional data," *The Journal of Machine Learning Research*, vol. 19, no. 1, pp. 3074–3100, 2018.

[8] M. Tiomoko, R. Couillet, and H. Tiomoko, "Large dimensional analysis and improvement of multi task learning," *arXiv preprint arXiv:2009.01591*, 2020.

[9] J. Richiardi, S. Achard, H. Bunke, and D. Van De Ville, "Machine learning with brain graphs: predictive modeling approaches for functional imaging in systems neuroscience," *IEEE Signal Processing Magazine*, vol. 30, no. 3, pp. 58–70, 2013.

[10] C.-I. Chang, *Hyperspectral imaging: techniques for spectral detection and classification*, vol. 1. Springer Science & Business Media, 2003.

[11] Y. Zhou, H. Wang, F. Xu, and Y.-Q. Jin, "Polarimetric sar image classification using deep convolutional neural networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 12, pp. 1935–1939, 2016.

[12] R. Couillet and A. Kammoun, "Random matrix improved subspace clustering," in *2016 50th Asilomar Conference on Signals, Systems and Computers*, pp. 90–94, IEEE, 2016.

[13] S. Xu, X. An, X. Qiao, and L. Zhu, "Multi-task least-squares support vector machines," *Multimedia tools and applications*, vol. 71, no. 2, pp. 699–715, 2014.

[14] X. Cheng and A. Singer, "The spectrum of random inner-product kernel matrices," *Random Matrices: Theory and Applications*, vol. 2, no. 04, p. 1350010, 2013.

[15] Z. Fan and A. Montanari, "The spectral norm of random inner-product kernel matrices," *Probability Theory and Related Fields*, vol. 173, no. 1, pp. 27–85, 2019.

[16] A. Kammoun and R. Couillet, "Subspace kernel spectral clustering of large dimensional data," *(submitted to) Annals of Applied Probability*, 2017.

[17] A. Mian, J.-P. Ovarlez, A. M. Atto, and G. Ginolhac, "Design of new wavelet packets adapted to high-resolution sar images with an application to target detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 6, pp. 3919–3932, 2019.