Analyse discriminante linéaire par distance de Wasserstein

Rémi FLAMARY[1], Marco CUTURI[2], Nicolas COURTY[3], Alain RAKOTOMAMONJY[4]

[1]Laboratoire Lagrange, CNRS, OCA, Université Côte d'Azur, France
[2]Laboratoire CREST, ENSAE, Université Paris Saclay, France
[3]Laboratoire IRISA, Université de Bretagne Sud, France
[4]Laboratoire LITIS, Université de Rouen, France

Résumé – Nous présentons une nouvelle méthode d'analyse discriminante utilisant une version régularisée de la distance de Wasserstein. Le principe de notre approche est de chercher un sous espace qui maximise la séparation des distributions entre classes tout en minimisant l'étalement de chaque classe. Un des avantages de l'utilisation de la distance de Wasserstein est qu'elle est non paramétrique et permet ainsi de discriminer des distributions empiriques dans un cadre non linéairement séparable. Finalement, des simulations numériques sur plusieurs jeux de données montrent l'intérêt de notre approche par rapport à l'état de l'art.

Abstract – We propose in this work a novel discriminant analysis method that builds on a regularized Wasserstein distance. Similar to Fisher discriminant analysis we aim at finding a subspace that maximizes the separation of classes while minimizing the spread of each class. The fact that we use a non parametric Wasserstein distance between empirical distributions allows us to find linear subspaces where the data can be nonlinearly separable. Numerical experiments performed on several datasets show the robustness and interest of our approach with respect to state of the art methods.

1 Introduction

Les techniques de réduction de dimension visent à projeter des données avec des caractéristiques redondantes dans un espace de dimension réduite tout en préservant le maximum d'information. Il existe plusieurs types de méthodes qui peuvent effectuer des transformations linéaires ou non linéaires, et travailler dans un cadre non-supervisé comme l'analyse en composante principale (PCA pour *Principal Component Analysis*) ou dans un cadre supervisé comme l'analyse discriminante de Fisher (FDA pour *Fisher Discriminant Analysis*). Nous nous concentrons dans ce travail sur la famille des méthodes linéaires et supervisées. Dans cette famille, une méthode de référence est la FDA qui, à partir d'un jeu de données $\{(\mathbf{x}_i, y_i)\}_i$, $\mathbf{x}_i \in \mathbb{R}^d$, cherche une projection linéaire qui maximise la discrimination entre les classes. Une limite de la FDA est que le modèle probabiliste gaussien sous-jacent impose que les données soient linéairement séparables. Lorsque ce n'est pas le cas des variantes locales de la FDA ont été proposées, comme par exemple LFDA [1]. Les méthodes d'apprentissage de métrique telles que les plus proches voisins à vaste marge ou *Large Margin Nearest Neighbors* (LMNN) [2] estiment également un sous espace mais à partir des relations locales entre les plus proches voisins, ce qui permet également de trouver des sous-espaces linéaires qui sont non-linéairement discriminant. Plus récemment sont apparues des méthodes basées sur des critères issus de la théorie de l'information. La méthode LSDR [3] cherche un sous-espace qui maximise l'information mutuelle entre les données projetées et les étiquettes et donc la discrimination mais elle nécessite une estimation de densité complexe à régler en pratique. Une autre méthode appelée CEML [4] optimise l'entropie conditionnelle entre les données projetées en évitant l'estimation de densité ce qui la rend potentiellement plus robuste.

Nous proposons dans ce document une méthode qui s'inspire des méthodes globales comme l'analyse de Fisher mais qui permet de trouver des sous-espaces plus complexes à l'instar des méthodes locales. Ces dernières se concentrent sur la préservation des relations locales entre les données, *i.e.* des différences $\|\mathbf{P}\mathbf{x}_i - \mathbf{P}\mathbf{x}_j\|$ avec $\mathbf{P}$ un opérateur de projection linéaire lorsque $\mathbf{x}_i$ est proche de $\mathbf{x}_j$. Une illustration des relations locales est donnée Figure 1 où on voit que les similarités sont calculées à l'aide d'une matrice $\mathbf{T}$ qui encode les relations $T_{i,j}$ entre exemples. L'analyse discriminante linéaire par distance de Wasserstein (WDA) consiste à maximiser les distances de Wasserstein régularisées entre classe tout en minimisant ces distances à l'intérieur des classes. Ce problème s'exprime classiquement sous la forme d'une maximisation d'un ratio :

$$\max_{\mathbf{P}\in\Delta} \quad \frac{\sum_{c,c'>c} W_\lambda(\mathbf{P}\mathbf{X}^c, \mathbf{P}\mathbf{X}^{c'})}{\sum_c W_\lambda(\mathbf{P}\mathbf{X}^c, \mathbf{P}\mathbf{X}^c)}, \tag{1}$$

où $\Delta = \{\mathbf{P} = [\mathbf{p_1}, \ldots, \mathbf{p}_p] \quad | \quad \mathbf{p}_i \in \mathbb{R}^d, \|\mathbf{p}_i\|_2 = 1, \mathbf{p}_i^\top \mathbf{p}_j = 0 \ \forall i \neq j\}$ est la variété de Stiefel [5] *i.e.* l'ensemble des matrices orthogonales de taille $d \times p$ et $\mathbf{P}\mathbf{X}^c$ est la matrice des exemples projetés appartenant à la classe c. W_λ est la distance de Wasserstein régularisée qui s'exprime comme $W_\lambda(\mathbf{X}, \mathbf{Z}) = \sum_{i,j} T_{i,j} \|\mathbf{x}_i - \mathbf{z}_j\|_2^2$, avec $T_{i,j}$ encodant les relations entre exemples.

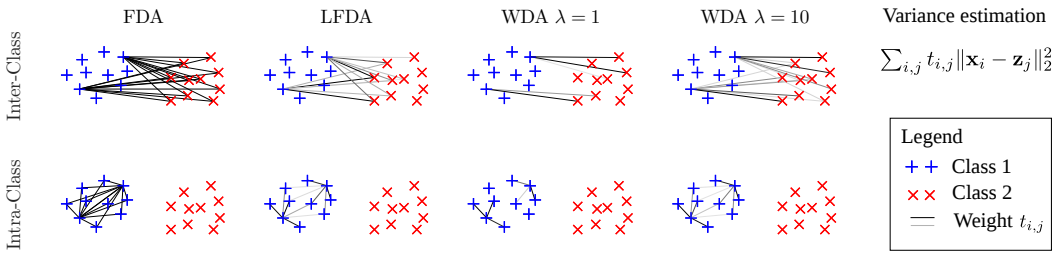

FIGURE 1 – Illustration des relations inter et intra-classes utilisées pour l'analyse de Fisher, l'analyse de Fisher locale et notre proposition.

Ici T s'obtient comme la solution d'un problème de transport optimal sur les exemples projetés [6] (cf section 2). L'intérêt de WDA est de permettre à la fois l'encodage de relations globales et locales grâce à l'utilisation du transport optimal. Comme discuté plus loin et illustré dans la Figure 1, le coefficient associé à la régularisation permet de balancer efficacement entre ces deux informations. La régularisation entropique permet aussi d'éviter pour tout $\lambda > 0$ que le dénominateur du ratio soit nul, évitant par là l'obtention de solutions triviales au problème. Après avoir introduit les notations et la notion de distance de Wasserstein, nous détaillons plus en détails notre méthode ainsi que les simulations numériques.

2 Définitions et distance de Wasserstein

Soit $\mu = \frac{1}{n}\sum_i \delta_{\mathbf{x}_i}$, $\nu = \frac{1}{m}\sum_i \delta_{\mathbf{z}_i}$ deux mesures empiriques dont les échantillons $\mathbb{R}^d$ sont exprimés comme des matrices $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n]$ et $\mathbf{Z} = [\mathbf{z}_1, \cdots, \mathbf{z}_m]$. La distance Euclidienne au carré entre les échantillons est définie comme $\mathbf{M}_{\mathbf{X},\mathbf{Z}} := [\|\mathbf{x}_i - \mathbf{z}_j\|_2^2]_{ij} \in \mathbb{R}^{n\times m}$. U_{nm} est le polytope des matrices $n \times m$ non-négatives dont les sommes des colonnes et lignes sont égales respectivement à $\mathbf{1}_n/n$ et $\mathbf{1}_m/m$.

$$U_{nm} := \{\mathbf{T} \in \mathbb{R}_+^{n\times m} \ : \ \mathbf{T}\mathbf{1}_m = \mathbf{1}_n/n,\ \mathbf{T}^T\mathbf{1}_n = \mathbf{1}_m/m\}.$$

Wasserstein avec régularisation entropique. Soit $\langle A, B\rangle := \mathrm{tr}(A^T B)$ le produit scalaire matriciel de Frobenius. La distance de Wasserstein régularisée entre μ et ν est définie pour $\lambda \geq 0$ par

$$W_\lambda(\mu,\nu) := W_\lambda(\mathbf{X},\mathbf{Z}) := \langle \mathbf{T}_\lambda, \mathbf{M}_{\mathbf{X},\mathbf{Z}}\rangle, \tag{2}$$

où $\mathbf{T}_\lambda$ est la solution d'un problème de transport optimal avec régularisation entropique de la forme

$$\mathbf{T}_\lambda := \mathrm{argmin}_{\mathbf{T}\in U_{nm}} \ \lambda\langle \mathbf{T}, \mathbf{M}_{\mathbf{XZ}}\rangle - \Omega(\mathbf{T}), \tag{3}$$

$\Omega(\mathbf{T}) := -\sum_{ij} t_{ij}\log(t_{ij})$ est l'entropie de $\mathbf{T}$ vue comme une loi jointe discrète. Le problème ci-dessus a une solution de la forme

$$\mathbf{T} = \mathrm{diag}(\mathbf{u})e^{-\lambda\mathbf{M}}\,\mathrm{diag}(\mathbf{v}) = \mathbf{u}\mathbf{1}_m^T \circ e^{-\lambda\mathbf{M}} \circ \mathbf{1}_n\mathbf{v}^T, \tag{4}$$

où $\circ$ est la multiplication point à point et l'exponentielle est également appliquée point à point.

Les valeurs de $\mathbf{u}$ et $\mathbf{v}$ peuvent être obtenus à l'aide de l'algorithme de Sinkhorn-Knopp [6] où chaque itération correspond à une normalisation itérative de matrice. Les vecteurs sont mis à jour pour une itération k

$$\mathbf{v}^k = \frac{\mathbf{1}_m/m}{\mathbf{K}^T\mathbf{u}^{k-1}}, \quad \mathbf{u}^k = \frac{\mathbf{1}_n/n}{\mathbf{K}\mathbf{v}^k} \tag{5}$$

avec une initialisation de $\mathbf{u}^0 = \mathbf{1}_n$. Ces itérations peuvent être implémentées de manière efficace car elles font appel à des opérations de multiplications matricielles facilement parallélisables.

3 Analyse discriminante linéaire par distance de Wasserstein (WDA)

Problème d'optimisation. Dans la suite du papier nous simplifions les notation en dénotant les données associées à chaque classe c par $\mathbf{X}^c$ En utilisant la définition (2) on peut reformuler le problème WDA par

$$\max_{\mathbf{P}\in\Delta} \left\{ J(\mathbf{P},\mathbf{T}(\mathbf{P})) = \frac{\sum_{c,c'>c}\langle \mathbf{P}^T\mathbf{P}, \mathbf{C}^{c,c'}\rangle}{\sum_c \langle \mathbf{P}^T\mathbf{P}, \mathbf{C}^{c,c}\rangle} \right\} \tag{6}$$

$$\text{s.c. } \mathbf{C}^{c,c'} = \sum_{i,j} T_{i,j}^{c,c'}(\mathbf{x}_i^c - \mathbf{x}_j^{c'})(\mathbf{x}_i^c - \mathbf{x}_j^{c'})^T, \quad \forall c, c'$$

$$\text{et } \mathbf{T}^{c,c'} = \mathrm{argmin}_{\mathbf{T}\in U_{n_c n_{c'}}} \lambda\langle \mathbf{T}, \mathbf{M}_{\mathbf{P}\mathbf{X}^c,\mathbf{P}\mathbf{X}^{c'}}\rangle - \Omega(\mathbf{T}), \tag{7}$$

qui peut être reformulé comme

$$\max_{\mathbf{P}\in\Delta} \quad J(\mathbf{P},\mathbf{T}(\mathbf{P})) \tag{8}$$

$$\text{s.t. } \mathbf{T}(\mathbf{P}) = \mathrm{argmin}_{\mathbf{T}\in U_{n_c n_{c'}}} \quad E(\mathbf{T},\mathbf{P}) \tag{9}$$

où $\mathbf{T} = \{\mathbf{T}_{c,c'}\}_{c,c'}$ contient toutes les matrices de transport entre les distributions des classes et E est la somme des coûts $\forall c, c' \leq c$ exprimés équation (7). La fonction objectif peut également être exprimée comme $J(\mathbf{P},\mathbf{T}(\mathbf{P})) = \frac{\langle \mathbf{P}^T\mathbf{P}, \mathbf{C}_b\rangle}{\langle \mathbf{P}^T\mathbf{P}, \mathbf{C}_w\rangle}$ avec $\mathbf{C}_b = \sum_{c,c'>c}\mathbf{C}_{c,c'}$ and $\mathbf{C}_w = \sum_c \mathbf{C}_{c,c}$ des matrices de covariances inter et intra classe qui dépendent de $\mathbf{T}$. Le problème d'optimisation (8)-(9) est un problème d'optimisation bi-niveau qui peut être résolu par descente de gradient [7].

Calcul du gradient. Puisque la fonction $\mathbf{T}(\mathbf{P})$ est lisse et le problème d'optimisation (9) est strictement convexe, le gradient de J peut être calculé à l'aide de la règle de dérivation en chaine suivante :

$$\nabla_{\mathbf{P}} J(\mathbf{P},\mathbf{T}(\mathbf{P})) = \frac{\partial J(\mathbf{P},\mathbf{T})}{\partial \mathbf{P}} + \sum_{c,c'\geq c} \frac{\partial J(\mathbf{P},\mathbf{T})}{\partial \mathbf{T}^{c,c'}} \frac{\partial \mathbf{T}^{c,c'}}{\partial \mathbf{P}}. \tag{10}$$

Le premier terme du gradient (10) suppose que $\mathbf{T}$ est constante et est de la forme (Eq. 94-95 [8])

$$\frac{\partial J(\mathbf{P},\mathbf{T})}{\partial \mathbf{P}} = \mathbf{P}\left(\frac{2}{\sigma_w^2}\mathbf{C}_b - \frac{2\sigma_b^2}{\sigma_w^4}\mathbf{C}_w\right) \tag{11}$$

où $\sigma_w^2 = \langle \mathbf{P}^T\mathbf{P}, \mathbf{C}_w \rangle$ et $\sigma_b^2 = \langle \mathbf{P}^T\mathbf{P}, \mathbf{C}_b \rangle$. Pour le calcul du second terme il faut séparer le cas $c = c'$ et $c \neq c'$. Les dérivées partielles sont de la forme $\frac{\partial J(\mathbf{P},\mathbf{T})}{\partial \mathbf{T}^{c,c'\neq c}} = \text{vec}(\frac{1}{\sigma_b^2}\mathbf{M}_{\mathbf{P}\mathbf{X}^c,\mathbf{P}\mathbf{X}^{c'}})$, $\frac{\partial J(\mathbf{P},\mathbf{T})}{\partial \mathbf{T}^{c,c}} = -\text{vec}(\frac{\sigma_w^2}{\sigma_b^4}\mathbf{M}_{\mathbf{P}\mathbf{X}^c,\mathbf{P}\mathbf{X}^c})$. La difficulté numérique est ici de calculer les jacobiennes $\partial \mathbf{T}^{c,c'}/\partial \mathbf{P}$ car $\mathbf{T}$ est la solution d'un problème d'optimisation et n'a pas de forme analytique. Une approche classique pour calculer le gradient est l'utilisation du théorème des fonctions implicites régulièrement utilisé en apprentissage [9]. Le problème de cette approche est qu'elle nécessite la résolution d'un système linéaire en grande dimension et ne peut pas être implémenté de manière efficace à chaque calcul de gradient. Nous proposons de résoudre ce problème à l'aide d'une seconde approche devenue commune en apprentissage statistique : la différentiation automatique. En adaptant l'approche proposée par [10] (qui dérive un algorithme plus complexe que Sinkhorn). Pour cela nous fixons le nombre d'itérations Sinkhorn à la valeur L choisi a priori. Soit $\mathbf{T}^k(\mathbf{P})$ la solution de l'algorithme du Sinkhorn après k itérations, pour une paire c, c' donnée nous avons

$$\mathbf{T}^k(\mathbf{P}) = \text{diag}(\mathbf{u}^k) e^{-\lambda \mathbf{M}} \text{diag}(\mathbf{v}^k),$$

où $\mathbf{M}$ est la matrice de distances induite par $\mathbf{P}$. $\mathbf{T}^L(\mathbf{P})$ peut être dérivé directement par

$$\frac{\partial \mathbf{T}^k}{\partial \mathbf{P}} = \frac{\partial [\mathbf{u}^k \mathbf{1}_m^T]}{\partial \mathbf{P}} \circ e^{-\lambda \mathbf{M}} \circ \mathbf{1}_n {\mathbf{v}^k}^T + \mathbf{u}^k \mathbf{1}_m^T \circ \frac{\partial e^{-\lambda \mathbf{M}}}{\partial \mathbf{P}} \circ \mathbf{1}_n {\mathbf{v}^k}^T + \mathbf{u}^k \mathbf{1}_m^T \circ e^{-\lambda \mathbf{M}} \circ \frac{\partial [\mathbf{1}_n {\mathbf{v}^k}^T]}{\partial \mathbf{P}}.$$

Il est important de noter ici que le calcul complet fait apparaître une récurrence car $\mathbf{u}^k$ dépend de $\mathbf{v}^k$ qui dépend de $\mathbf{u}^{k-1}$. Il est donc possible de calculer les Jacobiennes le long des itérations de Sinkhorn à partir de l'équation (5).

Résolution numérique. Nous proposons de résoudre le problème d'optimisation de WDA à l'aide d'une méthode de descente de gradient dans la variété de Stiefel. Le gradient dans l'espace Euclidien ambiant peut en effet être calculé à chaque itération mais la matrice de projection $\mathbf{P}$ doit rester dans le manifold de Stiefel. Il existe pour résoudre ce problème de nombreux algorithmes basés sur une projection du gradient qui sont implémentés dans la boite à outils Manopt [11].

Relation avec l'analyse de Fisher. Il est intéressant de noter que lorsque le terme de régularisation tend vers l'infini les matrices $\mathbf{T}$ deviennent constantes et chaque exemple est transporté sur tous les autres. Dans ce cas là, WDA est exactement équivalent à l'analyse de Fisher. FDA est donc un cas particulier de notre approche où les relations locales sont oubliées au profit d'une distance globale.

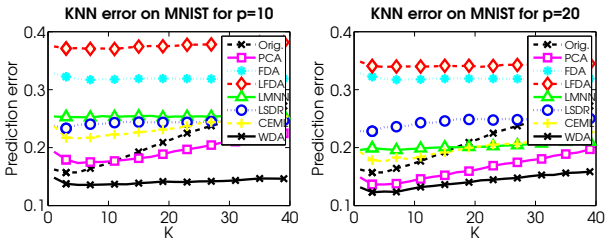

FIGURE 2 – Erreur de prédiction moyenne sur les données MNIST avec $p = 10$ (gauche) et $p = 20$ (droite).

4 Simulations numériques

Dans cette section, nous illustrons les performances de notre approche et les comparons à d'autres méthodes de l'état de l'art, notamment sur des jeux de donnés classiques issus du site UCI *machine learning repository*. Nous avons comparé WDA à des méthodes classiques de réduction de dimension telles que l'analyse en composante principales et l'analyse de Fisher, à des méthodes de réduction de dimension locales telles que LFDA et LMNN et finalement à des méthodes d'apprentissage de métrique basées sur l'information mutuelle (LSDR, CEML).

MNIST. Les premières simulations sont effectuées sur un jeu de données classique contenant des caractères manuscrits (10 classes) dans des images de taille 28×28 donc de dimension 784. Notre objectif est ici d'illustrer la robustesse de notre approche lorsque l'on a accès à un nombre limité d'exemples d'apprentissage. Nous avons donc tiré $n = 1000$ exemples d'apprentissage. On peut voir les performances des diverses méthodes en fonction du k d'un kNN pour une projection en dimension 10 et 20 sur la Figure 2. Les performances sont moyennées sur 20 tirages aléatoires des données d'apprentissage. On peut voir sur les courbes de performance que WDA permet de trouver de très bons sous-espaces qui mènent à la meilleure qualité de classification.

Dans un soucis d'interprétation, nous avons utilisé la méthode classique de projection non linéaire t-SNE [12] pour projeter les données de l'espace $p = 10$ vers une visualisation 2D donnée dans la Figure 3. On remarque que les classes sont en effet mieux séparées par WDA que les autres méthodes de réduction de dimension.

Données UCI. Nous testons maintenant notre approche sur plusieurs jeux de donnés obtenus à partir de *UCI*. Pour évaluer la capacité des méthodes à retrouver un sous-espace discriminant, 100 variables de bruit Gaussien ont été ajoutées aux données comme proposé dans [13] (sauf pour Isolet, USPS, MNIST et Caltech). Pour chaque méthode la taille du sous espace p et le k du classifieur kNN ont été trouvés par validation croisée. Nous avons comparé les 8 méthodes décrites en début de section mais n'avons reporté par manque de place uniquement que les méthodes les plus récentes dans la Table 1. Les erreurs moyennes ne sont pas comparables entre jeux de don-

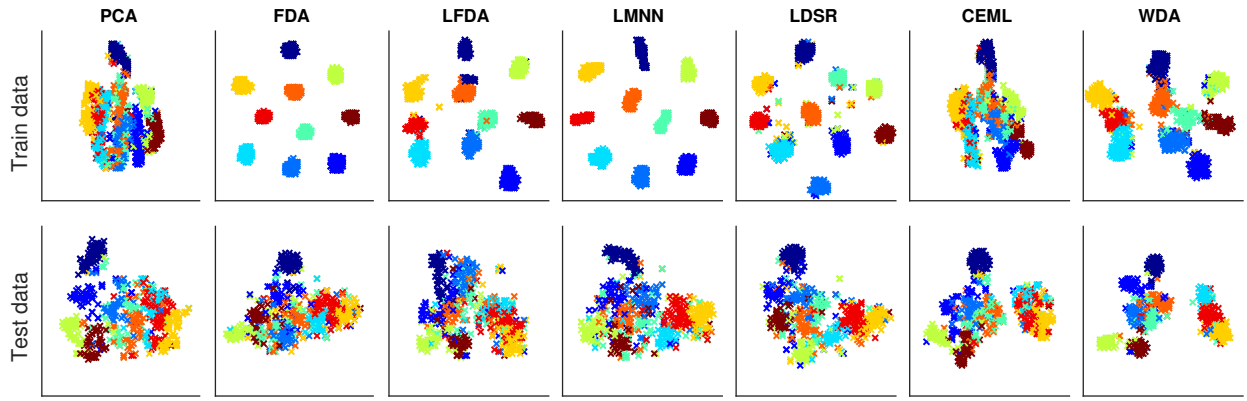

FIGURE 3 – 2D tSNE sur les échantillons MNIST projetés dans un espace de dimension $p = 10$ pour différentes approches. Ensembles d'apprentissage (haut), de test (down).

Datasets	LFDA	LMNN	LSDR	CEML	WDA
wines	29.89	32.81	32.81	**15.34**	16.91
iris	24.00	21.67	37.93	**20.87**	**20.87**
glass	59.48	54.25	50.85	**34.86**	45.99
vehicles	48.35	**40.84**	51.86	48.46	51.13
credit	18.44	23.73	24.71	17.65	**17.39**
ionosphere	28.10	30.80	31.08	22.87	**20.40**
isolet	13.96	**11.13**	13.33	30.19	14.41
usps	12.76	**6.05**	8.77	10.15	6.50
mnist	29.92	13.95	26.53	24.68	**13.07**
caltechpca	18.19	11.55	36.08	13.65	**11.45**
Rang moyen (/8)	5.2	3.4	5.7	3.5	2.2

TABLE 1 – Erreur de test moyenne pour 20 tirages sur les données UCI. Les performances ont été obtenues pour 8 méthodes mais seules les plus récentes ont été conservées par soucis de place. La meilleur méthode est désignée en gras et les erreurs de test soulignées ont été trouvées statistiquement équivalentes par un test du signrank de p-val=0.05.

nées mais nous donnons aussi les rangs moyens de chaque méthode de réduction de dimension. WDA est la méthode avec les meilleurs performances avec un rang moyen de 2.2 sur 8 méthodes. Elle est suivie par les méthodes LMNN et CEML qui permettent également de trouver un sous espace avec discrimination non-linéaire.

5 Conclusion

Nous avons présenté dans ce document une nouvelle méthode d'analyse discriminante qui utilise la distance de Wasserstein régularisée pour trouver un sous espace discriminant. Nous avons proposé une approche pratique pour calculer le gradient de la distance de Wasserstein basée sur la différentiation automatique de l'algorithme du Sinkhorn et avons illustré les performances et la robustesse de notre méthode sur plusieurs jeux de données, concluant à l'intérêt pratique de cette nouvelle méthode.

Références

[1] Masashi Sugiyama. Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. *The Journal of Machine Learning Research*, 8 :1027–1061, 2007.

[2] Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *The Journal of Machine Learning Research*, 10 :207–244, 2009.

[3] Taiji Suzuki and Masashi Sugiyama. Sufficient dimension reduction via squared-loss mutual information estimation. *Neural computation*, 25(3) :725–758, 2013.

[4] Luis G Sanchez Giraldo and Jose C Principe. Information theoretic learning with infinitely divisible kernels. In *Proceedings of the first International Conference on Representation Learning (ICLR)*, pages 1–8, 2013.

[5] P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

[6] M. Cuturi. Sinkhorn distances : Lightspeed computation of optimal transport. In *NIPS*, pages 2292–2300, 2013.

[7] Benoît Colson, Patrice Marcotte, and Gilles Savard. An overview of bilevel optimization. *Annals of operations research*, 153(1) :235–256, 2007.

[8] Kaare Brandt Petersen, Michael Syskind Pedersen, et al. The matrix cookbook. *Technical University of Denmark*, 7 :15, 2008.

[9] Yoshua Bengio. Gradient-based optimization of hyperparameters. *Neural computation*, 12(8) :1889–1900, 2000.

[10] Nicolas Bonneel, Gabriel Peyré, and Marco Cuturi. Wasserstein barycentric coordinates : Histogram regression using optimal transport. *ACM Transactions on Graphics*, 35(4), 2016.

[11] Nicolas Boumal, Bamdev Mishra, P-A Absil, and Rodolphe Sepulchre. Manopt, a matlab toolbox for optimization on manifolds. *The Journal of Machine Learning Research*, 15(1) :1455–1459, 2014.

[12] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(2579-2605) :85, 2008.

[13] Voot Tangkaratt, Hiroaki Sasaki, and Masashi Sugiyama. Direct estimation of the derivative of quadratic mutual information with application in supervised dimension reduction. *arXiv preprint arXiv :1508.01019*, 2015.