Un processus ponctuel déterminantal pour la sélection de variables supervisée

Xiaoyi Mai, Rémi Bardenet

Un processus ponctuel déterminantal pour la sélection de variables supervisée

Xiaoyi MAI[1], Rémi BARDENET[1],

[1]Université de Lille, CNRS, UMR 9189 – CRIStAL, 59651 Villeneuve d'Ascq, France
xiaoyi.mai@univ-lille.fr, remi.bardenet@univ-lille.fr

Résumé – La sélection de variables est une étape de réduction de dimension courante en apprentissage. Nous proposons un algorithme randomisé de sélection de variables pour la régression linéaire, basé sur un processus ponctuel déterminantal. À la différence de résultats précédents, notre processus utilise les étiquettes du jeu de données d'entrainement. L'idée-force est d'échantillonner k colonnes de la matrice des variables qui approximent bien le sous-espace de Krylov d'ordre k.

Abstract – Feature selection is a common technique in ML for dimensionality reduction. We propose a randomized feature selection algorithm for linear regression, based on a determinantal point process. Unlike previous work, our process uses the training labels. The key idea is to sample k columns of the data matrix, whose span is close to the Krylov subspace of order k.

1 Introduction

En apprentissage artificiel, la sélection de variables supervisée a pour objectif de réduire le coût de calcul et améliorer la performance de généralisation, en sélectionnant un sous-ensemble de variables considérées comme pertinentes. Puisque trouver le meilleur sous-ensemble de variables est en général un problème combinatoire difficile, infaisable en grande dimension, de nombreux algorithmes approchés ont été proposés. Une famille d'approches consiste à déterminer itérativement le sous-ensemble de variables sélectionnées. Il y a d'abord la méthode ascendante (*forward stepwise selection*), qui commence avec un ensemble vide et ajoute des variables une par une, ou encore la méthode descendante (*backward stepwise selection*) qui part de l'ensemble complet et élimine une variable à chaque itération.

Ces méthodes, bien que largement employées dans la pratique, viennent avec différents risques et inconvénients. La méthode ascendante commence forcément par évaluer chaque variable, ce qui peut guider le processus dans la mauvaise direction en présence de groupes de variables qui sont utiles collectivement mais moins intéressantes individuellement. Ce risque peut être mitigé par la méthode descendante, mais cette dernière court le risque du surapprentissage. Ce problème se produit quand la dimension du modèle appris est trop grande par rapport au nombre de couples entrée-sortie utilisés pour entraîner ce modèle. En conséquence, le modèle obtenu correspond parfaitement aux données d'entraînement, le risque empirique utilisé pour évaluer la qualité des sous-ensembles de variables reste alors minimisé après l'élimination de n'importe quelle variable, avant qu'on ne sorte du régime de surapprentissage. Dans ce papier, nous nous intéressons au cas de l'apprentissage linéaire supervisé par la méthode des moindre carrés. Nous proposons une distribution de probabilité sur les ensembles de variables, un processus ponctuel déterminantal, qui favorise les variables collectivement diverses. En comparaison avec les méthodes ascendante et descendante , notre méthode a l'avantage de prendre en compte l'impact des corrélations entre les variables, et de rester valable dans le régime de surapprentissage.

Les processus ponctuels déterminantaux (DPP) ont été exploités dans plusieurs travaux pour faire la sélection de variables en mode non-supervisé, notamment pour résoudre le problème de sélection de colonnes [2]. Dans un problème de sélection de colonnes, on cherche un sous-ensemble de variables qui décrivent le mieux un nombre de données non-étiquetées, au sens où les variables non-sélectionnées sont bien approximées par des combinaisons linéaires des variables sélectionnées. L'intérêt d'un DPP pour la sélection de variables non-supervisée réside dans sa tendance statistique à favoriser la diversité de variables sélectionnées. Par contre, les DPP sont, à première vue, moins adaptés à la sélection de variables supervisée, pour laquelle le critère principal n'est plus seulement la diversité. Comme expliqué dans [1], en apprentissage supervisé, il arrive que les variables fortement corrélées soient plus utiles que celles qui sont indépendantes.

Notre méthode s'appuie d'abord sur le fait que la solution de l'apprentissage supervisé donnée par la méthode des moindres carrés peut être approchée récursivement par une séquence de solutions dans les sous-espaces de Krylov [3]. Nous proposons alors d'échantillonner, par un DPP, des indices de variables qui tendent à approximer un sous-espace de Krylov, dans l'espoir que la solution obtenue avec les variables correspondant aux indices échantillonnés soit proche de celle prenant en compte

toutes les variables, qui vit principalement dans le sous-espace de Krylov.

2 Préliminaires

On rappelle d'abord la méthode des moindres carrés pour l'apprentissage supervisé, avant de faire le lien avec les sous-espaces de Krylov, puis d'introduire les processus ponctuels déterminantaux (DPP).

2.1 Méthode des moindres carrés

À partir des n observations i.i.d. $(\mathbf{x}_i, y_i)$ de couples entrée-sortie, l'apprentissage supervisé cherche à trouver, dans une classe de fonctions f (e.g., les fonctions linéaires), celle qui prédit le mieux la sortie d'un nouveau couple $(\mathbf{x}, y)$ tiré de la même distribution. Cela se fait souvent en minimisant un risque empirique $\sum_{i=1}^n L(f(x_i), y_i)$ défini par une fonction de perte L. Dans ce papier, nous nous intéressons à l'apprentissage par la méthode des moindres carrés, avec les fonctions linéaires $f(\mathbf{x}) = \mathbf{w}^\mathsf{T}\mathbf{x}$ et la perte quadratique $L(a, b) = (a - b)^2$.

Soient une matrice de données $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n] \in \mathbb{R}^{p\times n}$ avec n données $\mathbf{x}_i$ à p variables, et un vecteur de valeurs de sortie $\mathbf{y} \in \mathbb{R}^n$ pour ces n observations. La méthode des moindres carrés estime un vecteur de poids à attribuer aux variables par

$$\mathbf{w}^* \in \operatorname{argmin}_{\mathbf{w}} \|\mathbf{X}^\mathsf{T}\mathbf{w} - \mathbf{y}\|^2.$$

Supposant pour simplifier $\mathbf{X}\mathbf{X}^\mathsf{T}$ inversible, on obtient un système linéaire dont la solution est

$$\mathbf{w}^* = \left(\mathbf{X}\mathbf{X}^\mathsf{T}\right)^{-1}\mathbf{X}\mathbf{y}. \tag{1}$$

L'erreur d'entrainement s'écrit alors

$$\epsilon^2 \triangleq \|\mathbf{X}^\mathsf{T}\mathbf{w}^* - \mathbf{y}\|^2 = \mathbf{y}^\mathsf{T}\left[\mathbf{I}_p - \mathbf{X}^\mathsf{T}\left(\mathbf{X}\mathbf{X}^\mathsf{T}\right)^{-1}\mathbf{X}\right]\mathbf{y}. \tag{2}$$

Enfin, en sélection de variables, on cherche à trouver un sous-ensemble $S \subset [p] \triangleq \{1, \dots, p\}$ tel que l'excès de risque normalisé

$$\mathcal{R}(\mathbf{w}_S) = \frac{\|\mathbf{X}^\mathsf{T}\mathbf{w}_S - \mathbf{y}\|^2 - \|\mathbf{X}^\mathsf{T}\mathbf{w}^* - \mathbf{y}\|^2}{\|\mathbf{y}\|^2 - \|\mathbf{X}^\mathsf{T}\mathbf{w}^* - \mathbf{y}\|^2} \in [0, 1] \tag{3}$$

soit petit, où

$$\mathbf{w}_S = \operatorname{argmin}_{\mathbf{w}\in\mathcal{C}_S} \|\mathbf{X}^\mathsf{T}\mathbf{w} - \mathbf{y}\|^2, \tag{4}$$

avec $\mathcal{C}_S$ le sous-espace engendré par les vecteurs de la base canonique de $\mathbb{R}^p$ dont l'indice est dans S.

2.2 Sous-espaces de Krylov

Les sous-espaces de Krylov sont souvent utilisés pour résoudre approximativement un système linéaire de la forme $\mathbf{A}\mathbf{w} = \mathbf{b}$. Pour faire le lien avec la méthode des moindres carrés de la Section 2.1, il suffit de prendre $\mathbf{A} = \mathbf{X}\mathbf{X}^\mathsf{T}$ et $\mathbf{b} = \mathbf{X}\mathbf{y}$.

Définition 1 (Sous-espaces de Krylov)**.** *Le sous-espace de Krylov d'ordre k associé une matrice $\mathbf{A} \in \mathbb{R}^{p\times p}$ et un vecteur $\mathbf{b} \in \mathbb{R}^n$ est défini par*

$$\mathcal{K}_k = \operatorname{Vect}\{\mathbf{b}, \mathbf{A}\mathbf{b}, \dots, \mathbf{A}^{k-1}\mathbf{b}\}. \tag{5}$$

En pratique, la solution de $\mathbf{A}\mathbf{w} = \mathbf{b}$ est souvent approchée par une suite de solutions dans des espaces de Krylov d'ordre croissant. Par exemple, considérons

$$\mathbf{w}_k = \operatorname{argmin}_{\mathbf{w}\in\mathcal{K}_k} \|\mathbf{w}^* - \mathbf{w}\|^2_{\mathbf{A}}, \tag{6}$$

où, supposant $\mathbf{A}$ inversible pour simplifier, $\mathbf{w}^* = \mathbf{A}^{-1}\mathbf{b}$ est la solution de $\mathbf{A}\mathbf{w} = \mathbf{b}$. Cette suite est donnée par l'algorithme du gradient conjugué, où on retrouve la solution exacte après le nombre d'itérations égale à celui de valeurs propres distinctes de $\mathbf{A}$. D'après l'analyse de convergence pour la méthode du gradient conjugué [4], on a

$$\frac{\|\mathbf{w}^* - \mathbf{w}_k\|^2_{\mathbf{A}}}{\|\mathbf{w}^*\|^2_{\mathbf{A}}} \leq 2\left(\frac{\sqrt{\lambda_1/\lambda_p} - 1}{\sqrt{\lambda_1/\lambda_p} + 1}\right)^k \triangleq \tau_k \tag{7}$$

où $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ sont les valeurs propres de $\mathbf{A}$.

2.3 Processus ponctuels déterminantaux

Un processus ponctuel est la loi d'un sous-ensemble aléatoire de $[p] \triangleq \{1, \dots, p\}$.

Définition 2 (DPP)**.** *Soit $\mathbf{K} \in \mathbb{R}^{p\times p}$. Un processus ponctuel $\mathcal{S} \subset [p]$ est dit déterminantal de noyau marginal $\mathbf{K}$ si*

$$\forall \mathcal{B} \subset [p], \quad \mathbb{P}(\mathcal{B} \subseteq \mathcal{S}) = \det(\mathbf{K}_\mathcal{B}), \tag{8}$$

où $\mathbf{K}_\mathcal{B} = [\mathbf{K}_{ij}]_{ij\in\mathcal{B}}$, et $\det(\mathbf{K}_\emptyset) = 1$ par convention.

Les DPP ont été introduits par [6] comme modèle en optique quantique électronique, et ont depuis suscité un grand intérêt, particulièrement en apprentissage [7]. Un DPP tend à échantillonner des points divers, avec leur similarités encodées par les éléments croisés de la matrice $\mathbf{K}$ de noyau. En effet, on note que, pour tout $i, j \in [p]$,

$$\begin{aligned}\mathbb{P}(i, j \subseteq \mathcal{S}) =& [\mathbf{K}]_{ii}[\mathbf{K}]_{jj} - [\mathbf{K}]^2_{ij} \\ =& \mathbb{P}(i \subseteq \mathcal{S})\mathbb{P}(j \subseteq \mathcal{S}) - [\mathbf{K}]^2_{ij}.\end{aligned}$$

En particulier, deux points $i, j \in [p]$ avec une grande similarité $[\mathbf{K}]_{ij}$ ont peu de chance d'apparaitre ensemble dans une réalisation de $\mathcal{S}$.

L'un des intérêts computationnels des DPP est qu'ils peuvent être échantillonnés en temps polynomial. Si de plus $\mathbf{K}$ est un noyau de projection de rang k, les échantillons sont de cardinalité k presque sûrement [8]. On parle alors d'un DPP de projection.

3 Méthode proposée

Dans le cadre de la régression, présenté en Section 2.1, on pose $\mathbf{A} = \mathbf{X}\mathbf{X}^\mathsf{T}$, $\mathbf{b} = \mathbf{X}\mathbf{y}$. On propose de tirer un sous-ensemble $\mathcal{S} \subset [p]$ contenant k indices de variables avec

$$\mathbb{P}(\mathcal{S}) = \det[\mathbf{U}_{\mathcal{S},[k]}\mathbf{U}^\mathsf{T}_{\mathcal{S},[k]}], \tag{9}$$

où $\mathbf{U} \in \mathbb{R}^{p\times p}$ est une matrice unitaire telle que

$$\begin{bmatrix}\mathbf{b}, & \mathbf{A}\mathbf{b} & \dots & \mathbf{A}^{p-1}\mathbf{b}\end{bmatrix} = \mathbf{U}\mathbf{R}$$

et $\mathbf{R}$ est une matrice triangulaire supérieure. Le processus ponctuel défini dans (9) est un DPP de projection de noyau

$$\mathbf{K} = \mathbf{U}_{:,[k]}\mathbf{U}_{:,[k]}^{\mathsf{T}},$$

qui tire donc avec probabilité 1 un ensemble de cardinalité k.

Puisque que $\mathbf{U}$ est donnée par la décomposition QR d'une matrice dont les k premiers vecteurs colonnes engendrent le sous-espace $\mathcal{K}_k$ de Krylov (5) d'ordre k, il est facile de voir que $\mathbf{U}_{:,[k]}\mathbf{U}_{:,[k]}^{\mathsf{T}}$ est la matrice de projection sur $\mathcal{K}_k$. Remarquons enfin que

$$\det[\mathbf{U}_{\mathcal{S},[k]}\mathbf{U}_{\mathcal{S},[k]}^{\mathsf{T}}] = \prod_{i=1}^{k}\cos^2(\theta_i(\mathcal{S}))$$

où $\theta_1(\mathcal{S}),\dots,\theta_k(\mathcal{S})$ sont les angles principaux [5] entre $\mathcal{K}_k$ et le sous-espace $\mathcal{C}_\mathcal{S}$ engendré par $\mathbf{e}_i$, $i \in \mathcal{S}$, avec $\mathbf{e}_i$ le i-ième vecteur de la base canonique de $\mathbb{R}^p$. Notre processus tend alors à échantillonner les coordonnées, considérées comme indices de variables sélectionnées, qui permettent d'approximer un sous-espace de Krylov associé à la solution des moindres carrée. Cette méthode d'≪ alignement d'espaces ≫ est directement inspirée du travail de [2], où les auteurs sélectionnent des variables en mode non-supervisé par un DPP de projection permettant de rapprocher le sous-espace engendré par les variables sélectionnées et celui engendré par les vecteurs propres associés aux plus grandes valeurs propres.

On s'attend donc à ce que $\mathbf{w}_k$, le k-ième élément de la suite de Krylov définie dans (6), soit bien approximé par sa projection sur le sous-espace $\mathcal{C}_\mathcal{S}$ avec $\mathcal{S}$ tiré par (9). D'autre part, rappelons qu'on peut approximer la solution classique $\mathbf{w}^*$ des moindres carrés par $\mathbf{w}_k$ avec une garantie d'approximation donnée dans (7). Par conséquent, on tend à avoir petits résidus de $\mathbf{w}^*$ après sa projection sur $\mathcal{C}_\mathcal{S}$. Rappelons qu'on s'intéresse plutôt à minimiser l'excès de risque normalisé $\mathcal{R}(\mathbf{w}_\mathcal{S})$. Comme (1) implique que

$$\mathcal{R}(\mathbf{w}_\mathcal{S}) = \frac{\|\mathbf{w}^* - \mathbf{w}_\mathcal{S}\|_\mathbf{A}^2}{\|\mathbf{w}^*\|_\mathbf{A}^2} \tag{10}$$

avec $\mathbf{A} = \mathbf{X}\mathbf{X}^{\mathsf{T}}$ et $\|\mathbf{v}\|_\mathbf{A}^2 = \mathbf{v}^{\mathsf{T}}\mathbf{A}\mathbf{v}$, minimiser l'excès de risque nous amène à minorer $\|\mathbf{w}^* - \mathbf{w}_\mathcal{S}\|_\mathbf{A}^2$. On expliquera dans la section suivante comment un sous-ensemble $\mathcal{S} \subset [p]$ tiré de (9) permet de borner en espérance $\|\mathbf{w}^* - \mathbf{w}_\mathcal{S}\|_\mathbf{A}^2$, donnant ainsi une garantie de performance pour notre méthode.

4 Analyse de performance

On donne dans cette section une borne supérieure sur l'espérance de l'excès de risque $\mathbb{E}[\mathcal{R}(\mathbf{w}_\mathcal{S})]$, qui est valable avec grande probabilité sous (9) pour les problèmes où à la fois la dimension d et le nombre k de variables sélectionnées sont grands. On commence par décomposer $\mathbf{A}^{\frac{1}{2}}\mathbf{w}^*$ comme

$$\mathbf{A}^{\frac{1}{2}}\mathbf{w}^* = \sum_{i=1}^{k-1} c_i\mathbf{q}_i + r_k\mathbf{g}_k$$

où $\mathbf{q}_i$ est le i-ième vecteur colonne de la matrice unitaire $\mathbf{Q}$ donnée par

$$\begin{bmatrix}\mathbf{A}^{\frac{1}{2}}\mathbf{b} & \mathbf{A}^{\frac{3}{2}}\mathbf{b} & \dots & \mathbf{A}^{\frac{2p-1}{2}}\mathbf{b}\end{bmatrix} = \mathbf{Q}\mathbf{R}',$$

avec $\mathbf{R}'$ une matrice triangulaire supérieure, et $\mathbf{g}_k$ un vecteur unitaire orthogonal à tout $\mathbf{q}_i$, $i < k$. Selon la définition (6) de $\mathbf{w}_k$, on peut déduire que $\mathbf{A}^{\frac{1}{2}}\mathbf{w}_k$ est la projection de $\mathbf{A}^{\frac{1}{2}}\mathbf{w}^*$ sur le sous-espace $\mathrm{Vect}\{\mathbf{A}^{\frac{1}{2}}\mathbf{b}, \mathbf{A}^{\frac{3}{2}}\mathbf{b}, \dots, \mathbf{A}^{\frac{2k-1}{2}}\mathbf{b}\}$. C'est-à-dire

$$\begin{aligned}\|\mathbf{w}^* - \mathbf{w}_{k-1}\|_\mathbf{A}^2 =& \mathbf{w}^{*\mathsf{T}}\mathbf{A}^{\frac{1}{2}}\left(\mathbf{I}_p - \mathbf{Q}_{:,[k]}\mathbf{Q}_{:,[k]}^{\mathsf{T}}\right)\mathbf{A}^{\frac{1}{2}}\mathbf{w}^* \\ =& \|\mathbf{w}^*\|_\mathbf{A}^2 - \sum_{i=1}^{k-1} c_i^2 = r_k^2.\end{aligned}$$

Comme $\|\mathbf{w}^*\|_\mathbf{A}^2 = \|\mathbf{y}\|^2 - \epsilon^2$ selon (1) et (2), on a

$$r_k = \sqrt{\|\mathbf{y}\|^2 - \epsilon^2 - \sum_{i=1}^{k-1} c_i^2}.$$

Théorème 1. *Soit $\mathcal{S} \subset [p]$ un sous-ensemble de cardinalité k tiré aléatoirement selon* (9)*. Pour k assez grand, on a, avec grande probabilité sous la loi* (9)*,*

$$\mathbb{E}[\mathcal{R}(\mathbf{w}_\mathcal{S})] \leq 1 - \frac{k r_k^{\frac{2}{k}}\prod_{i=1}^{k-1} c_i^{\frac{2}{k}}}{\epsilon^2}\left(\frac{\prod_{i=1}^{k}\lambda_{p+1-i}}{\prod_{i=1}^{k}\lambda_i}\right)^{\frac{1}{k}}\left(\frac{k!(p-k)!}{p!}\right)^{\frac{1}{k}}. \tag{11}$$

Comme expliqué à la fin de Section 3, on arrive à borner $\mathbb{E}[\mathcal{R}(\mathbf{w}_\mathcal{S})]$ en s'appuyant sur la tendance statistique de notre méthode à minimiser $\|\mathbf{w}^* - \mathbf{w}_\mathcal{S}\|_\mathbf{A}^2$. Nous donnons ici quelques éléments de preuve. Notons d'abord que

$$\det[\mathbf{U}_{\mathcal{S},[k]}\mathbf{U}_{\mathcal{S},[k]}^{\mathsf{T}}] = \frac{\det\left[\mathbf{M}_{\mathcal{S},[k]}\mathbf{M}_{\mathcal{S},[k]}^{\mathsf{T}}\right]}{\det[\mathbf{M}_{:,[k]}^{\mathsf{T}}\mathbf{M}_{:,[k]}]}$$

où $\mathbf{M} = \begin{bmatrix}\mathbf{b}, & \mathbf{A}\mathbf{b} & \dots & \mathbf{A}^{p-1}\mathbf{b}\end{bmatrix}$. Observons aussi que

$$[\mathbf{M}]_{:,[k]} = \mathbf{A}^{\frac{1}{2}}\mathbf{H}_k\begin{bmatrix}\mathbf{0}_{k-1} & [\boldsymbol{\eta}]_1 \\ [\mathbf{R}']_{k-1,k-1} & [\boldsymbol{\eta}]_{2:k}\end{bmatrix}$$

avec $\mathbf{H}_k = [\mathbf{g}_k, \mathbf{Q}_{:,[k-1]}]$ et $\boldsymbol{\eta}$ un vecteur dans $\mathbb{R}^k$. On a alors

$$\mathbb{P}(\mathcal{S}) = \det[\mathbf{U}_{\mathcal{S},[k]}\mathbf{U}_{\mathcal{S},[k]}^{\mathsf{T}}] = \frac{\det[\mathbf{A}_{\mathcal{S},\mathcal{S}}]}{\det[\mathbf{H}_k^{\mathsf{T}}\mathbf{A}\mathbf{H}_k]}\prod_{i=1}^{k}\cos^2(\phi_i(\mathcal{S}))$$

où $\phi_1(\mathcal{S}),\dots,\phi_k(\mathcal{S})$ sont les angles principaux entre les sous-espaces $\mathcal{H}_k$ et $\tilde{\mathcal{C}}_\mathcal{S}$ avec $\mathcal{H}_k$ engendré par les vecteurs colonnes de $\mathbf{H}_k$ et $\tilde{\mathcal{C}}_\mathcal{S}$ par $\mathbf{A}^{\frac{1}{2}}\mathbf{e}_i$, $i \in \mathcal{S}$. Remarquons de plus que

$$\|\mathbf{w}^* - \mathbf{w}_\mathcal{S}\|_\mathbf{A}^2 = \min_{\mathbf{w}\in\tilde{\mathcal{C}}_\mathcal{S}} \|\mathbf{A}^{\frac{1}{2}}\mathbf{w}^* - \mathbf{w}\|^2,$$

sans oublier que $\mathbf{w}^* \in \mathcal{H}_k$. On conclut alors que pour minimiser $\|\mathbf{w}^* - \mathbf{w}_\mathcal{S}\|_\mathbf{A}^2$, on a intérêt à rapprocher les sous-espaces $\mathcal{H}_k$ et $\tilde{\mathcal{C}}_\mathcal{S}$. C'est précisément ce que fait notre méthode, qui tend à minimiser les angles principaux entre ces deux sous-espaces, avec plus ou moins de succès en fonction du spectre de $\mathbf{A}$.

5 Discussion et simulation

Dans la sélection de variables supervisée, il est souvent proposé de sélectionner des variables qui sont alignées avec la réponse et peu corrélées entre elles. L'article [1] a remis en question cette intuition en démontrant, avec comme exemple des modèles de mélange gaussien, que les variables fortement corrélées peuvent être celles qui aident le plus à distinguer les classes. Pour comprendre cet effet bénéfique de sélectionner des variables corrélées, considérons un modèle de mélange gaussien $\mathbf{x}|y \sim \mathcal{N}(y\boldsymbol{\mu}, \boldsymbol{\Sigma})$, pour un problème de classification binaire avec

$$p(y = \pm 1) = 1/2.$$

Il peut être montré que [9], étant donné un sous-ensemble $\mathcal{S} \subset [p]$, la performance bayésienne (oracle) de classification, c'est-à-dire $\max_f \mathbb{P}(f(\mathbf{x}) = y)$, est une fonction croissante de $\boldsymbol{\mu}_{\mathcal{S}}^{\mathsf{T}} \boldsymbol{\Sigma}_{\mathcal{S},\mathcal{S}}^{-1} \boldsymbol{\mu}_{\mathcal{S}}$. S'intéressant aux scénarios "non-structurés" où les corrélations entre les variables sont indépendantes de leurs alignements avec la réponse, on suppose une loi aléatoire sur $\boldsymbol{\Sigma}$, qui est de forme

$$\begin{bmatrix} \frac{1}{\rho k}\mathbf{Z}\mathbf{Z}^{\mathsf{T}} & \mathbf{0}_{k,p-k} \\ \mathbf{0}_{k,p-k} & \mathbf{I}_{p-k} \end{bmatrix}, \tag{12}$$

où $\mathbf{Z} \in \mathbb{R}^{k \times \rho k}$ avec $[\mathbf{Z}]_{ij} \sim \mathcal{N}(0,1)$ i.i.d., et $\boldsymbol{\mu}$ déterministe. On a alors, d'un côté, un groupe de variables (les k premières) corrélées entre elles avec un niveau de corrélation déterminé par ρ, et d'un autre côté, les $p-k$ dernières variables qui sont indépendantes de toutes les autres. La loi de Wishart inverse indique que

$$\mathbb{E}\left[\boldsymbol{\mu}_{\mathcal{S}}^{\mathsf{T}} \boldsymbol{\Sigma}_{\mathcal{S},\mathcal{S}}^{-1} \boldsymbol{\mu}_{\mathcal{S}}\right] = \frac{\rho}{\rho - 1} \|\boldsymbol{\mu}_{\mathcal{S} \cap [k]}\|^2 + \|\boldsymbol{\mu}_{\mathcal{S} \setminus (\mathcal{S} \cap [k])}\|^2.$$

En autres termes, la performance est positivement impactée par l'alignement de variables sélectionnées avec les valeurs de sortie et le niveau de corrélation entre elles, ici caractérisé par ρ.

Si ρ est proche de 1, une classification quasi-parfaite est réalisée avec un petit nombre k de variables, même si elles ne sont que modérément corrélées avec la réponse. Ce genre de situations est mal géré par la méthode ascendante, qui commence par comparer les éléments de $\mathbf{X}\mathbf{y}$ afin de trouver la variable la plus corrélée avec la réponse. En conséquent, elle peut passer à côté du groupe de variables intéressant si elles sont moins alignées avec la sortie que les autres. Ce problème peut être contourné par la méthode descendante qui est toutefois inapplicable dans le régime de surapprentissage comme expliqué dans l'introduction. Notre méthode offre une alternative pour faire face aux corrélations intéressantes entre les variables, avec un processus bien défini dans le régime de surapprentissage. Cela dit, le fait d'utiliser peu de données dans ce régime peut introduire trop de bruit pour que notre méthode soit efficace à extraire des variables pertinentes à cause d'une mauvaise estimation des corrélations entre les variables par $\mathbf{X}\mathbf{X}^{\mathsf{T}}$. Dans ce cas, on propose d'adopter une version semi-supervisée de notre méthode en remplacant $\mathbf{X}\mathbf{X}^{\mathsf{T}}$ par $\mathbf{X}\mathbf{X}^{\mathsf{T}} + \tilde{\mathbf{X}}\tilde{\mathbf{X}}^{\mathsf{T}}$ où $\tilde{\mathbf{X}}$ est la matrice de données non-étiquetées, normalement disponibles en grande quantité. Pour une comparaison équitable, on inclut aussi les données non-étiquetées dans les méthodes ascendante et descendante de la même manière. Les courbes de performance de classification sur les données non-étiquetées sous modèle de mélange gaussien sont tracées en Figure 1, où on observe un avantage consistant de notre méthode.

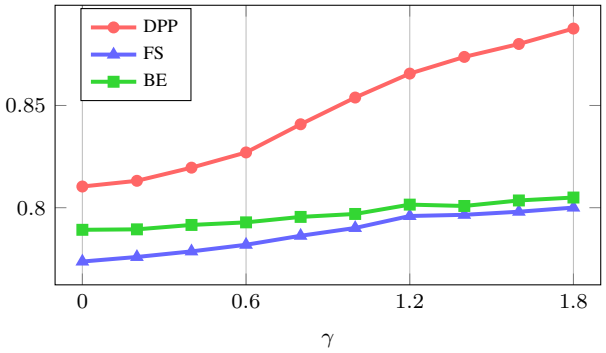

FIGURE 1 – Exactitude de classification sur les données non-étiquetées, obtenue avec 40 données étiquetées, 8000 non-étiquetée et 20 variables sélectionnées, sous le modèle de mélange gaussien $\mathbf{x} \sim \mathcal{N}(y\boldsymbol{\mu}, \boldsymbol{\Sigma})$ où $\mathbb{P}(y = \pm 1) = 1/2$, $\boldsymbol{\mu} = [\mathbf{1}_{20}^{\mathsf{T}}, \gamma \mathbf{1}_{60}^{\mathsf{T}}]^{\mathsf{T}}$, $\boldsymbol{\Sigma}_{[20],[80]\setminus[20]} = \mathbf{0}_{20,60}$, $\boldsymbol{\Sigma}_{[20],[20]} = \frac{1}{20}\mathbf{Z}_1\mathbf{Z}_1^{\mathsf{T}}$ avec $\mathbf{Z}_1 \in \mathbb{R}^{20\times 20}$ de i.i.d. $[\mathbf{Z}_1]_{ij} \sim \mathcal{N}(0,1)$ et $\boldsymbol{\Sigma}_{[80]\setminus[20],[80]\setminus[20]} = \frac{1}{180}\mathbf{Z}_2\mathbf{Z}_2^{\mathsf{T}}$ avec $\mathbf{Z}_1 \in \mathbb{R}^{60\times 180}$ de i.i.d. $[\mathbf{Z}_2]_{ij} \sim \mathcal{N}(0,1)$. Résultats moyennés sur 1000 réalisations.

Références

[1] I., Guyon et A. Elisseeff. *An Introduction to Variable and Feature Selection*. Journal of machine learning research, 2003.

[2] A. Belhadji, R. Bardenet et P. Chainais. *A Determinantal Point Process for Column Subset Selection*. Journal of machine learning research, 2018.

[3] J. Liesen et Z. Strakos. *Krylov Subspace Methods : Principles and Analysis*. Oxford University Press, 2013.

[4] J.R. Shewchuk *An Introduction to the Conjugate Gradient Method without the Agonizing Pain*. Carnegie-Mellon University, 1994.

[5] P. Zhu et A.V. Knyazev. *Angles between Subspaces and Their Tangents*. Journal of Numerical Mathematics, 2013.

[6] O.Macchi. *The Coincidence Approach to Stochastic Point Processes*. Cambridge University Press, 1975.

[7] A. Kulesza et B. Taskar. *Determinantal Point Processes for Machine Learning*. Foundations and Trends in Machine Learning, 2012.

[8] J.B. Hough, M. Krishnapur, Y. Peres et B. Virág. *Determinantal Processes and Independence*. Probability Surveys, 2006.

[9] G.J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley & Sons, 2005.