Décorrélation adaptative pour la prédiction en grande dimension

Florian Hébert [1] & Mathieu Emily [2] & David Causeur [3]

[1] *florian.hebert@agrocampus-ouest.fr*
[2] *mathieu.emily@agrocampus-ouest.fr*
[3] *david.causeur@agrocampus-ouest.fr*

Agrocampus Ouest (IRMAR, UMR CNRS 6625), 65 rue de Saint-Brieuc, 35000 Rennes

Résumé. Dans les procédures de tests en grande dimension, la prise en compte ou non de la dépendance donne lieu à de nombreux développements méthodologiques et discussions, notamment sur l'impact de la décorrélation des statistiques de tests. Pourtant, dans une optique d'estimation d'un modèle pour la prédiction, la question de la décorrélation de grands profils de variables prédictrices n'est pas abordée dans les mêmes termes, bien que de nombreuses études comparatives aient rapporté la supériorité de méthodes de prédiction dites naïves, au sens où elles ignorent la dépendance. Sous l'hypothèse classique en analyse linéaire discriminante d'un mélange de lois gaussiennes, nous montrons que pour une structure de dépendance des prédicteurs donnée, les performances de classification ignorant ou non cette dépendance peuvent être très variables et opposées selon la forme du signal d'association entre les prédicteurs et la classe. Afin de minimiser le risque maximal d'erreur de classification, nous proposons donc une prise en compte adaptative de la dépendance et montrons sur des simulations que les performances de la règle de classification proposée sont généralement au moins aussi bonnes que la meilleure des règles parmi celles ignorant la dépendance ou au contraire basées sur une décorrélation des prédicteurs.

Mots-clés. Classification supervisée, analyse discriminante, dépendance.

Abstract. In large-scale significance analysis, ignoring dependence or not is a core issue, leading to many recent results about the impact of decorrelating the pointwise test statistics. Yet, for the estimation of a prediction model, decorrelating large profiles of predicting variables is not as clearly questioned, although many comparative studies have reported the superiority of so-called naive methods, ignoring dependence. Under the usual Gaussian mixture model assumption of Linear Discriminant Analysis, we show that, for a given dependence structure, the classification performance of methods ignoring or not dependence may be markedly different, according to the pattern of the association signal between the predicting variables and the response. In order to minimize the largest probability of misclassification, we propose a method handling adaptively the dependence. A simulation study shows that the performance of the present method is at least as good as the best of methods ignoring dependence or based on a complete decorrelation of the predicting variables.

Keywords. Classification, discriminant analysis, dependence.

1 Introduction

On considère ici le problème de la classification supervisée d'individus en deux groupes sous les hypothèses classiques de mélange de lois Gaussiennes pour le profil des prédicteurs. Conditionnellement à la variable réponse Y, prenant la valeur 0 ou 1, le vecteur des prédicteurs $\boldsymbol{X} = (X_1, \ldots, X_p)'$ est distribué selon une loi normale multivariée: $\boldsymbol{X}|Y = j \sim \mathcal{N}_p(\boldsymbol{\mu_j}, \boldsymbol{\Sigma})$, $j = 0$ ou 1. Lorsque le nombre d'individus n de l'échantillon d'apprentissage est plus grand que p, le score linéaire discriminant de Fisher peut être vu comme l'estimation par la méthode des moments du score linéaire optimal dit de Bayes:

$$S_F(\boldsymbol{X}) = (\hat{\boldsymbol{\mu}}_\mathbf{1} - \hat{\boldsymbol{\mu}}_\mathbf{0})'\widehat{\boldsymbol{\Sigma}}^{-1}\left(\boldsymbol{X} - \frac{1}{2}(\hat{\boldsymbol{\mu}}_\mathbf{0} + \hat{\boldsymbol{\mu}}_\mathbf{1})\right).$$

Du fait de l'optimalité de la règle de Bayes dont elle est dérivée, la règle de classification $\mathcal{L}_F(\boldsymbol{X}) = \mathbf{1}_{\{S_F(\boldsymbol{X})\geq 0\}}$ de Fisher est parmi les plus populaires. En grande dimension, lorsque $p \gg n$, plusieurs extensions du score de Fisher existent, estimées sous une hypothèse de sparsité de ses coefficients ou en susbstituant à l'estimation empirique de $\boldsymbol{\Sigma}$ dans l'expression de $S_F(\boldsymbol{X})$ une estimation de plein rang, de type *James-Stein* notamment [1].

Dans ce même contexte de grande dimension, une approche alternative consiste à supposer que $\boldsymbol{\Sigma}$ est diagonale, ignorant ainsi la dépendance intra-classe des variables prédictrices. Ainsi, dans une étude comparative de méthodes d'apprentissage statistique en grande dimension pour des problématiques de prédiction à partir de profils d'expression de gènes, [4] met en avant la supériorité de la règle dite "Naïve Bayes" dans laquelle $\widehat{\boldsymbol{\Sigma}}$ est remplacée par la matrice diagonale des variances empiriques des composantes de $\boldsymbol{X}$, par rapport à de nombreuses méthodes s'appuyant explicitement ou non sur une estimation de $\boldsymbol{\Sigma}$:

$$\mathcal{L}_{NB}(\boldsymbol{X}) = \mathbf{1}_{\{S_{NB}(\boldsymbol{X})\geq 0\}}, \qquad \text{avec } S_{NB}(\boldsymbol{X}) = (\hat{\boldsymbol{\mu}}_\mathbf{1} - \hat{\boldsymbol{\mu}}_\mathbf{0})'\widehat{\boldsymbol{D}}^{-1}\left(\boldsymbol{X} - \frac{1}{2}(\hat{\boldsymbol{\mu}}_\mathbf{0} + \hat{\boldsymbol{\mu}}_\mathbf{1})\right).$$

où $\widehat{\boldsymbol{D}} = \mathrm{diag}(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_p^2)$, avec $\hat{\sigma}_i^2$ la variance empirique de X_i. Dans une approche plus analytique, [3] revient sur ces résultats et montre que la règle "Naïve Bayes" peut effectivement avoir un risque d'erreur de classement non-asymptotique plus faible, notamment dans un cadre de grande dimension. Ces deux règles ont été modifiées par [1], en introduisant une étape préliminaire de sélection de variables puis en remplaçant les estimateurs empiriques par des estimateurs biaisés de type James-Stein. Dans une étude comparative, [1] montre que les règles ainsi obtenues sont plus performantes que leurs versions initiales.

Choisir d'utiliser la règle de Fisher ou la règle "Naïve Bayes" revient à choisir de prendre en compte ou non la dépendance entre les prédicteurs $X_i, 1 \leq i \leq p$. Nous

montrons dans la suite que la supériorité d'une règle par rapport à l'autre ne dépend pas seulement de la structure de dépendance : en effet, pour une structure de dépendance donnée, les performances relatives de ces méthodes dépendent également des positions des prédicteurs X_i réellement associés à Y. Nous proposons ensuite une approche basée sur une prise en compte adaptative de la dépendance.

2 Méthode proposée

Soit $\widehat{\boldsymbol{R}}$ la matrice de corrélation intra-classe du profil des prédicteurs, $\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{D}}^{1/2}\widehat{\boldsymbol{R}}\widehat{\boldsymbol{D}}^{1/2}$, et $(\boldsymbol{U}, \boldsymbol{\Lambda})$ la décomposition en valeurs propres de $\widehat{\boldsymbol{R}}$: $\widehat{\boldsymbol{R}} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}'$ où $\boldsymbol{\Lambda}$ est la matrice diagonale des valeurs propres de $\widehat{\boldsymbol{R}}$, et $\boldsymbol{U}$ est la matrice des vecteurs propres orthonormés correspondants. En notant $\boldsymbol{\gamma} = \boldsymbol{U}'\widehat{\boldsymbol{D}}^{-1/2}(\hat{\boldsymbol{\mu}}_{\boldsymbol{1}} - \hat{\boldsymbol{\mu}}_{\boldsymbol{0}})$ et $\boldsymbol{\mathcal{Z}} = \boldsymbol{U}'\widehat{\boldsymbol{D}}^{-1/2}\left(\boldsymbol{X} - \frac{1}{2}(\hat{\boldsymbol{\mu}}_{\boldsymbol{0}} + \hat{\boldsymbol{\mu}}_{\boldsymbol{1}})\right)$, il vient alors que les deux scores discriminants $S_F(\boldsymbol{X})$ et $S_{NB}(\boldsymbol{X})$ peuvent être vus comme des combinaisons linéaires des produits des coordonnées de $\boldsymbol{\mathcal{Z}}$ et $\boldsymbol{\gamma}$:

$$S_F(\boldsymbol{X}) = \sum_{i=1}^{p} \frac{\gamma_i Z_i}{\lambda_i}, \; S_{NB}(\boldsymbol{X}) = \sum_{i=1}^{p} \gamma_i Z_i.$$

Nous proposons d'élargir le champ des perspectives pour la classification à la classe des règles s'écrivant :

$$\mathcal{L}_h(\boldsymbol{X}) = \mathbf{1}_{\{S_h(\boldsymbol{X}) \geq 0\}}, \qquad \text{avec } S_h(\boldsymbol{X}) = \sum_{i=1}^{p} h_i \gamma_i Z_i.$$

La recherche du prédicteur optimal en condition non-asymptotique dans cette classe conduit à choisir les coefficients h_i maximisant le rapport de la variance inter-groupes sur la variance intra-groupe de $S_h(\boldsymbol{X})$. En notant $\boldsymbol{\mathcal{U}} = (\gamma_1 Z_1, \ldots, \gamma_p Z_p)'$, le vecteur $\boldsymbol{h}$ recherché est:

$$\boldsymbol{h} = \mathrm{Var}(\boldsymbol{\mathcal{U}})^{-1}(\mathbb{E}[\boldsymbol{\mathcal{U}}|Y = 1] - \mathbb{E}[\boldsymbol{\mathcal{U}}|Y = 0]).$$

L'expression ci-dessus fait intervenir les moments d'ordre 1 et 2 de la distribution des vecteurs propres d'une matrice distribuée selon une loi de Wishart, que nous approchons par des méthodes de Monte-Carlo.

3 Résultats

On rapporte ici un extrait des résultats d'une étude comparative par simulation en dimension modérée, afin de ne pas pénaliser la règle de Fisher par l'instabilité numérique de l'estimation de $\boldsymbol{\Sigma}$. Des vecteurs de dimension $p = 50$ sont simulés selon la loi normale

multivariée de vecteur moyen $\mathbf{0}$ et de matrice de corrélation $\mathbf{\Sigma}$ représentée en figure 3. La variable de classe Y est ensuite simulée selon le modèle:

$$\text{logit}(\mathbb{P}[Y = 1|\boldsymbol{X} = \boldsymbol{x}]) = \beta_0 + \boldsymbol{x}'\boldsymbol{\beta}$$

où $\boldsymbol{X}$ est un profil de prédicteurs. 50 individus sont sélectionnés dans chaque classe pour construire l'échantillon d'apprentissage. Chaque méthode est ensuite évaluée sur un échantillon de validation de 100 000 individus générés selon le même procédé. Dans la suite, nous considérons des vecteurs de paramètres $\boldsymbol{\beta}$ ayant un grand nombre de coordonnées nulles. L'ensemble des coordonnées non nulles de $\boldsymbol{\beta}$ est noté $\mathcal{I}$; celles-ci sont toutes choisies égales à une même valeur contrôlant l'intensité de l'association entre Y et $\boldsymbol{X}$. Pour chaque valeur, 1000 échantillons d'apprentissage et de validation sont simulés. Le taux de mauvais classement moyen est ensuite calculé sur les 1000 échantillons de validation afin de comparer les différentes méthodes.

Les trois méthodes sont comparées dans quatre situations différentes. Les résultats sont représentés sur la figure 3. Les deux premières situations sont favorables à la règle de Fisher, dont les performances sont meilleures que celles de la règle "Naïve Bayes", tandis que les deux autres situations sont favorables à la règle "Naïve Bayes". Dans chaque situation, les performances de l'approche adaptative proposée sont au moins aussi bonnes que la meilleure des deux approches précédentes.

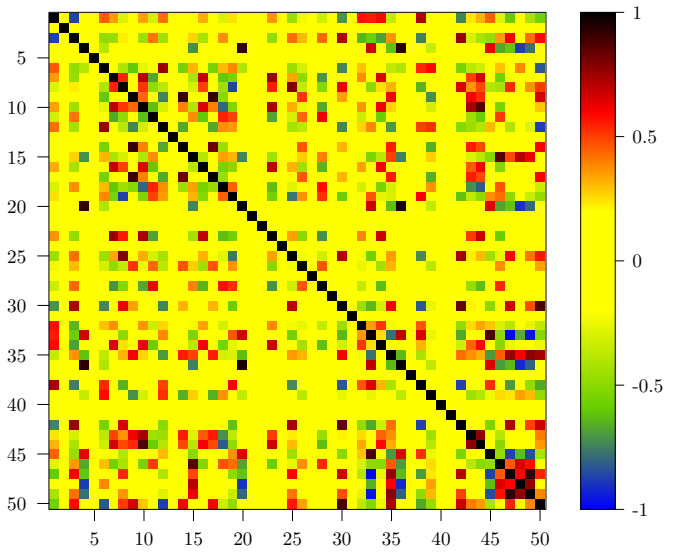

Figure 1: Structure de dépendance $\mathbf{\Sigma}$ des données simulées

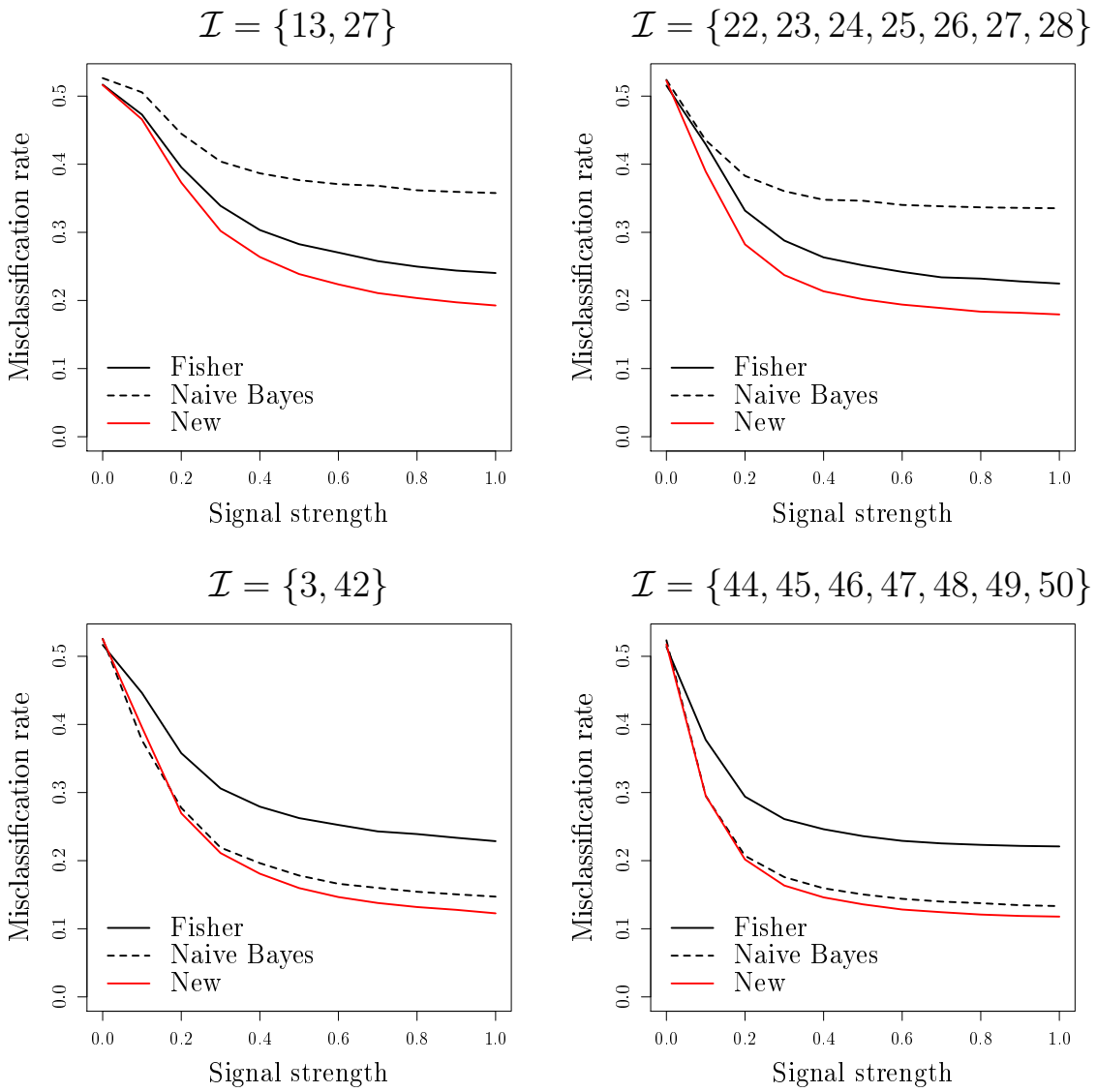

Figure 2: Taux de mauvais classement moyens de chaque méthode dans 4 situations différentes ; en haut, situations favorables à la méthode de Fisher, en bas, situations favorables à la méthode "Naive Bayes"

Les trois méthodes sont également comparées sur des données publiques d'expression de gènes, dont celles concernant le cancer du colon décrites dans [2]. Ces données contiennent les profils d'expression de 2000 gènes pour 62 individus, dont 22 sont atteints du cancer du colon et 40 sont témoins. Les performances des trois méthodes sont évaluées par validation croisée 10 blocs. La procédure de validation croisée est réalisée 10 fois afin de calculer un taux de mauvais classement moyen sur 10 partitions aléatoires différentes

des données. Le taux de mauvais classement moyen obtenu pour la méthode de Fisher est de 26,94%; celui de la méthode "Naïve Bayes" est légèrement meilleur (21,77%). Celui de l'approche proposée est largement inférieur, à 12,26%, ce qui rejoint les résultats obtenus sur les données simulées.

4 Conclusion

Dans le contexte de la classification supervisée où les prédicteurs sont distribués selon une loi normale multivariée, la règle de classification de Fisher est une référence majeure, ce qui s'explique notamment par le fait qu'il s'agit de la contrepartie empirique de la règle optimale de Bayes. Cependant, lorsque la taille d'échantillon est relativement faible au regard du nombre de variables et pour certaines formes de signal d'association entre les prédicteurs et la variable réponse, les performances de la règle de Fisher sont dépassées par celles de la règle "Naïve Bayes". La méthode que nous proposons revient à décorréler de manière adaptative le profil des prédicteurs, afin de minimiser l'erreur de prédiction maximale. L'étude comparative de la présentation sera élargie à d'autres méthodes d'apprentissage statistique en grande dimension.

Bibliographie

[1] Ahdesmäki, M. et Strimmer, K. (2010). Feature selection in omics prediction problems using cat scores and false nondiscovery rate control. *The Annals of Applied Statistics*, 4(1):503–519.

[2] Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., et Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, 96(12):6745–6750.

[3] Bickel, P. J. et Levina, E. (2004). Some theory for Fisher's linear discriminant function,'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli*, 10(6):989–1010.

[4] Dudoit, S., Fridlyand, J., et Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97(457):77–87.