Variable selection by decorrelated HCT for supervised classification in high dimension.

Emeline Perthame [1] & David Causeur [1]

[1] *Institut de Recherche Mathématique de Rennes (IRMAR) - CNRS UMR 6625 Agrocampus Ouest - Applied Mathematics Department, 65 rue de Saint Brieuc, 35 042 Rennes Cedex, France*

E-mail : emeline.perthame@agrocampus-ouest.fr & david.causeur@agrocampus-ouest.fr

Résumé. Nous considérons le problème de la classification supervisée où Y est une variable aléatoire de Bernoulli et X est un vecteur de covariables distribuées selon une loi normale. Dans ce contexte, l'analyse linéaire discriminante (LDA) atteint de bonnes performances de classification, même en grande dimension où de nombreux algorithmes de sélection de variables peuvent être utilisés pour réduire la dimension. Dans ce cadre, le Higher Criticism Thresholding (HCT) permet d'estimer le support du signal, même en situation de covariables corrélées. Toutefois, certains auteurs suggèrent qu'il peut être amélioré en prenant en compte explicitement cette dépendance.

Dans le contexte des tests multiples, plusieurs auteurs montrent par ailleurs l'impact négatif de la dépendance sur la stabilité de la sélection de variables et suggèrent de travailler sur des données ajustées de la dépendance. Nous proposons une méthode combinant une sélection par HCT suivi d'une LDA, les deux étapes étant fondées sur un même postulat d'indépendance entre les covariables, conditionnellement à un vecteur de facteurs latents.

La méthode HCT s'appuie sur la distribution asymptotique de p-values associées à des statistiques individuelles de sélection, le plus souvent des t-tests. Sous l'hypothèse d'un modèle à facteurs latents, on peut définir des statistiques de sélection décorrélées, par ajustement de l'effet des facteurs, et leurs p-values associées. Un nouveau critère HCT est déduit de l'expression analytique de la fonction de répartition conditionnelle des p-values, qui dépend de la structure de dépendance. L'étape d'estimation du modèle de classification proposée utilise également la structure en facteurs pour gérer la dépendance.

Les propriétés de la méthode sont illustrées sur des simulations et sur des données réelles.

Mots-clés. Apprentissage et classification, données en grande dimension, biostatistique

Abstract. This talk addresses the supervised classification issue in the traditional Linear Discriminant Analysis (LDA) context, where Y is a Bernoulli random variable and X is a vector of covariates normally distributed.

The conceptually simple LDA performs well in many situations, even in the context of high dimensional data, where many variable selection algorithms can be used to lower

dimension. In this setting, Higher Criticism Thresholding (HCT) is known to be effective to estimate support of the signal, even when features are dependent. However, some authors suggest that it can be improved by accounting for dependence.

In similar multiple testing issues, some authors show the negative impact of dependence on the stability of feature selection and suggest to work on decorrelated data, adjusted for latent components of dependence. This talk presents a decorrelated HCT followed by a LDA based on a factor model for the covariates. This model assumes that the features are conditionally independent, given a q-vector Z of latent factors.

We consider subset selectors based on standard t-test vector. Under the above factor model assumption, it is possible to derive decorrelated factor-adjusted test statistics and their corresponding p-values. We give the closed-form expression of the non-null conditional distribution function of p-value, which depends on the dependence structure. A new thresholding strategy is deduced, based on the conditional HC objective function. The proposed classification step also takes advantage of the factor structure to deal with dependence.

The properties of the method are demonstrated through simulation and real data studies.

Keywords. Machine learning and classification, high-dimensional data, biostatistics

1 Introduction

This talk addresses the supervised classification issue in the traditional Linear Discriminant Analysis (LDA) context, where Y is a Bernoulli random variable with $\mathbb{P}(Y = 1) = p_1$ (resp. $\mathbb{P}(Y = 0) = p_0$) and X is a m-vector of covariates such that $X|Y \sim \mathcal{N}_m(\mu_Y, \Sigma)$. In high dimension, variable selection is often used before classification to identify relevant subsets of features.

In this setting, Higher Criticism Thresholding (HCT) is known to be effective to estimate support of the signal (Donoho and Jin (2008)), even when features are dependent. However, some authors (Hall and Jin (2010), Ahdesmäki and Strimmer (2010)) suggest that it can be improved by accounting for dependence.

The framework of HCT is the rare and weak feature model: all entries of μ_0 are zero and a small number of entries of μ_1 are non-zero. In order to detect the non-zero entries of μ_1, we consider subset selectors based on standard t-test vector $T = (T_1, \ldots, T_m)$ and define the corresponding p-values $p = (p_1, \ldots, p_m)$. Standard HCT first compute the HC objective function defined as

$$HC(i, p_{(i)}) \quad = \quad \sqrt{m}\frac{i/m - p_{(i)}}{\sqrt{i/m(1 - i/m)}},$$

where $p_{(i)}$ stands for the i-th order statistics of p. In practice, the HC function is maximized on $\{i, 1 \leq i \leq \alpha_0 m\}$, where $\alpha_0 \in [0; 1]$ ($\alpha_0 = 0.1$ is recommended in the literature

(Donoho and Jin (2008))). The maximum is achieved at index $\hat{\imath}$. A feature is selected if it T-score exceeds a threshold $\hat{t}_{HC}$ in magnitude, defined as $\hat{t}_{HC} = |T|_{(\hat{\imath})}$.

Afterwards, a classification step follows dimension reduction. The conceptually simple Linear Discriminant Analysis (LDA) performs well in many situations, even in high dimension. As LDA needs to invert the covariance matrix, some authors ignore dependence (Bickel and Levina (2004)) but others recommend to infer the dependence structure (Zuber and Strimmer (2009)). In this framework, the proposed classification procedure also deals with dependence by decorrelating data before classification.

2 Proposed method

2.1 A factor model approach

In multiple testing issues, some authors (Friguet et al (2009), Leek and Storey (2008)) show the negative impact of dependence on the stability of feature selection and suggest to work on decorrelated data, adjusted for latent components of dependence. Factor model assumes that the features are conditionally independent, given a $q-$vector Z of latent factors :

$$X|Y,Z \;\sim\; \mathcal{N}_m(\mu_Y + BZ, \Psi),$$

where Z is distributed according to $\mathcal{N}_q(0, \mathbb{I}_q)$, Ψ is a diagonal matrix of specific variance and B is a $m \times q$ matrix of dependence shared by covariates.

Under the above factor model assumption, it is possible to derive decorrelated factor-adjusted test statistics (Friguet et al (2009)):

$$T_i^* \;=\; \frac{T_i - b_i' Z_T}{\sqrt{\Psi_i}}$$

where $Z_T = \frac{\bar{Z}_1 - \bar{Z}_0}{\sqrt{\frac{1}{n_1}+\frac{1}{n_0}}}$, $\bar{Z}_1$ (resp. $\bar{Z}_0$) is the empirical mean of latent factors for observations in group 1 (resp. group 0) and b_i stands for the i-th row of matrix B. The corresponding p-values are $p^* = (p_1^*, \ldots, p_m^*)$. These test statistics $(T_1^*, \ldots, T_m^*)$ inherits conditional independence of covariates given latent factors.

2.2 Decorrelated HCT

In order to decorrelate standard HCT, we give the closed-form expression of the non-null conditional distribution function of p-value p_i^*, which depends on the dependence structure:

$$\begin{aligned} F_i(x,z) &= \mathbb{P}_{Z=z}(p_i^* \leq x) \\ &= 1 - \Phi\left(\Phi^{-1}\left(1-\frac{x}{2}\right) - \frac{\delta_i}{\sqrt{\Psi_i}}\right) + \Phi\left(-\Phi^{-1}\left(1-\frac{x}{2}\right) - \frac{\delta_i}{\sqrt{\Psi_i}}\right) \end{aligned}$$

where δ_i is the standardized means difference $\frac{\mu_1^i - \mu_0^i}{\sqrt{\frac{1}{n_1}+\frac{1}{n_0}}}$ for variable i and Φ is the cumulative distribution function of standard gaussian distribution.

Inspired by the definition of HC objective function proposed by Klaus and Strimmer (2013), a Factor-Adjusted Higher Criticism objective function is deduced:

$$FAHC_i(x,z) \quad = \quad \frac{|F_i(x,z) - x|}{\sqrt{F_i(x,z)(1-F_i(x,z))}}.$$

An EM algorithm, proposed in Perthame et al (2015) gives estimations of factor model parameters Ψ, B, latent factors and group means estimations so that factor adjusted test statistics T^* and corresponding p-values p^* are computed.

In practice, the selection procedure is performed by plugging-in estimators of Ψ, Z and $\delta = (\delta_1, \ldots, \delta_m)$ in the FAHC function. FAHC is evaluated in p^* and maximized so that Factor-Adjusted Higher Criticism Thresholding is defined as:

$$\hat{\imath}_{FAHC} \quad = \quad argmax_{1 \leq i \leq \alpha_0 m} \frac{|\hat{F}_{(i)}(p^*_{(i)}, \hat{Z}) - p^*_{(i)}|}{\sqrt{\hat{F}_{(i)}(p^*_{(i)}, \hat{Z})(1 - \hat{F}_{(i)}(p^*_{(i)}, \hat{Z}))}}.$$

A feature is selected if its T^*-score exceeds the threshold $\hat{t}_{FAHC} = |T^*|_{\hat{\imath}_{FAHC}}$ in magnitude.

In the case of independence, it can be proved that the maximum of the standard HC function is reached close to the minimum of the number of misclassified features (sum of False Positive and False Negative). This property does not hold when variables are correlated. Some empirical results demonstrate that the threshold $\hat{t}_{FAHC}$ achieved by the proposed procedure is close to the ideal threshold which minimizes the number of misclassified features.

2.3 Conditional Bayes classifier

As in the conditional Bayes classifier rule proposed by Perthame et al (2015), the proposed classification step also takes advantage of the factor structure to deal with dependence. Assuming a factor model, the log-ratio of posterior probabilities for an observation x given latent factors z is:

$$LR(x,z) \quad = \quad \log \frac{p_1}{p_0} - 0.5(\mu_1' \Psi^{-1} \mu_1 - \mu_0' \Psi^{-1} \mu_0) + (x - Bz)' \Psi^{-1} (\mu_1 - \mu_0) \qquad (1)$$

where $(x - Bz)$ are defined as factor adjusted data.

This classification rule leads to the optimal classifier under a factor model assumption. In practice, the proposed procedure consists in the following steps:

(1) Variable selection by FA HCT

(2) Computation of factor adjusted data $x - Bz$

(3) Computation of $\widehat{LR}(x, z)$ by plugging-in estimators of parameters in expression (1)

(4) Prediction is 1 if $LR \geq 0$ and 0 otherwise.

3 Conclusion

The properties of the method are demonstrated through simulation and real data studies. The proposed procedure is compared to standard HCT and to other methods which account for dependence both in selection and in classification steps such as Shrinkage Discriminant Analysis (Ahdesmäki and Strimmer (2010)). We show that variable selection by decorrelated HCT leads to more sparse models as the subsets of selected features are smaller. It also appears that our classification step leads to smaller cross-validated misclassification rates.

References

[1] Ahdesmäki, M. and Strimmer, K. (2010), Feature selection in omics prediction problems using cat scores and false non-discovery rate control, *Annals of Applied Statistics*, 4, 503-519.
[2] Bickel, P. and 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.
[3] Donoho, D. and Jin, J. (2008), Higher criticism thresholding: Optimal feature selection when useful features are rare and weak, *Proceedings of The National Academy of Sciences*, 105:39,14790-14795.
[4] Friguet, C. and Kloareg, M. and Causeur, D. (2009), A factor model approach to multiple testing under dependence, *Journal of the American Statistical Association*, 104:488, 1406-1415.
[5] Hall, P. and Jin, J. (2010), Innovated higher criticism for detecting sparse signals in correlated noise, *The Annals of Statistics*, 38:3, 1686-1732.
[6] Klaus, B. and Strimmer, K. (2013), Signal identification for rare and weak features: higher criticism or false discovery rates?, *Biostatistics*, 14:1, 129-143.
[7] Leek, J. T. and Storey, J. (2008), A general framework for multiple testing dependence, *Proceedings of the National Academy of Sciences*, 105, 18718-18723.
[8] Perthame, E. and Friguet, C. and Causeur, D. (2015), Stability of feature selection in classification issues for high-dimensional correlated data, *Under review.*
[9] Zuber, V. and Strimmer, K., (2009), Gene ranking and biomarker discovery under correlation, *Bioinformatics*, 25, 2700-2707.