Estimation conditionnelle de la proportion d'hypothèses nulles en grande dimension

Chloé Friguet, David Causeur

Estimation conditionnelle de la proportion d'hypothèses nulles en grande dimension

Chloé Friguet & David Causeur

Laboratoire de Mathématiques Appliquées - Agrocampus
65 rue de St Brieuc - CS 84215
35042 Rennes Cedex, France
chloe.friguet@agrocampus-ouest.fr

Résumé L'accessibilité croissante des données biologiques dites à haut-débit a suscité de profondes remises en cause de méthodes statistiques aussi éprouvées que les tests multiples. En effet, le contrôle des taux d'erreurs nécessite des approches adaptées aux données en grande dimension et présentant une structure de corrélation importante. Nous proposons un cadre général permettant la prise en compte de la dépendance dans les Procédures de Tests Multiples, en considérant un modèle d'Analyse en Facteurs (voir Friguet et al. [2009]). Plus particulièrement ici, nous présentons l'intérêt de ce cadre conditionnel pour estimer un paramètre clé des PTM noté π_0, la proportion d'hypothèses nulles, qui intervient dans le contrôle des taux d'erreur. Les méthodes sont illustrés à partir de données issues de biopuces.
Mots clés *Tests multiples, Analyse en Facteur, Dépendance, Proportion d'hypothèses nulles*

Abstract Whole genome microarray experiments has markedly contributed to the development of the statistical methodology for multiple testing in high dimensional data. In this context, the impact of dependence on error control is especially crucial and dedicated methods for large scale, correlated data must be considered. We propose a methodology based on a Factor Analysis model, which shows improvements with respect to existing Multiple Testing Procedures (see Friguet et al. [2009]), and provides a general framework for multiple testing dependence. In this presentation, we focus on the estimation of the number of true null hypotheses, denoted π_0, which is a key parameter to manage with power in MTP. The methods are illustrated thanks to a microarray dataset.
Key words *Multiple testing, Factor analysis model, Dependence, Null hypotheses proportion*

Modélisation de la structure de dépendance

On postule un modèle d'Analyse en Facteurs pour la matrice de covariance [Friguet et al., 2009]. Connue des psychométriciens et des sociologues comme une technique de réduction de l'information, l'Analyse en Facteurs est apparue récemment comme

une technique d'analyse de la dépendance des données en grande dimension issues des expérimentations à "haut-débit" type biopuces [Pournara and Wernisch, 2007, Kustra et al., 2006]. Le modèle d'Analyse en Facteurs décrit la corrélation à travers un petit nombre de variables latentes (facteurs communs) : $\Sigma = B.B' + \Psi$ où B est la matrice $(m \times q)$ des *loadings* associés à la variance commune, Ψ est une matrice diagonale $(m \times m)$ correspondant aux variances spécifiques et q le nombre de facteurs du modèle.

De nombreuses méthodes existent pour estimer les paramètres du modèle d'Analyse en Facteurs : ici, étant donnée la grande dimension des données, et parce qu'on peut voir le modèle d'Analyse en Facteurs comme un modèle à variables latentes, un algorithme de type *EM* peut être implémenté [Rubin and Thayer, 1982].

L'introduction de cette modélisation par l'Analyse en Facteur dans le cadre des tests multiples permet de définir un cadre général pour la prise en compte de la dépendance, non pas au niveau de chaque étape des procédures mais globalement. Une statistique de test ajustée des facteurs communs est définie, améliorant la puissance des procédures, tout en contrôlant le taux d'erreurs de type-I. En effet, ce cadre général permet de réaliser des tests indépendants, conditionnellement à la structure en facteur, et donc de se placer dans le cadre optimal d'utilisation des PTM.

On s'intéresse plus particulièrement à l'étude dans ce cadre de l'estimation d'un paramètre clé des PTM : la proportion d'hypothèses nulles parmi l'ensemble des hypothèses testées, noté π_0. Après un rappel des méthodes d'estimation développées sous l'hypothèse d'indépendance et des problèmes posées par la présence de corrélation, nous présenterons une approche conditionnelle qui permet d'obtenir une estimation plus précise du paramètre.

Estimation de la proportion d'hypothèses nulles

Approche classique Étant donné l'importance de l'estimation de π_0 pour la puissance des tests multiples et sur le contrôle des taux d'erreur [Black, 2004], de nombreux auteurs se sont intéressés à l'estimation de ce paramètre, en particulier depuis l'introduction par Benjamini and Hochberg [1995] du False Dicovery Rate (FDR).

Deux approches sont envisagées dans la littérature, et reposent sur la condition d'uniformité des probabilités critiques sous H_0, ce qui garantit l'identifiabilité de π_0 [Celisse and Robin, 2008] :

$$\exists t \in]0;1]/\forall k \in \{1;\ldots;m\} \quad p_k \in [t;1] \Rightarrow p_k \sim \mathcal{U}[t;1]$$

Les 2 approches envisagées sont les suivantes :

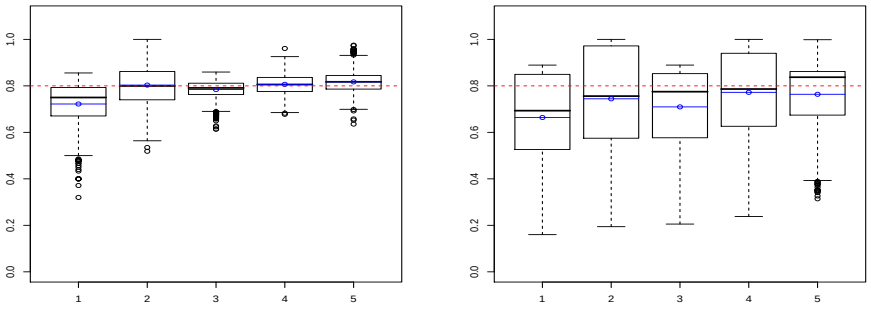
- Schweder and Spjotvoll [1982] ont proposé l'estimateur suivant pour π_0 :

$$\hat{\pi}_0(t) = \frac{W_t}{m(1-t)}$$

où W_t est le nombre de probabilités critiques supérieures à t. En effet, pour t "bien choisi", $\mathbb{E}(W_t) \approx m_0(1-t)$.
Le choix de t s'apparente donc à un compromis biais-variance. De nombreuses méthodes ont été proposées dans la littérature : les plus fréquemment utilisées sont une méthode par *bootstrap* [Storey, 2002] et une méthode par lissage à l'aide de splines cubiques de la courbe des $\hat{\pi}_0(t)$ pour $t \in \{0.01; ...; 0.99\}$ [Storey and Tibsirani, 2003].

- Une autre idée développée par de nombreux auteurs est d'estimer la densité f des probabilités critiques, prenant ainsi en compte à la fois les distributions sous les hypothèses nulle et alternative. Langaas et al. [2005] ont proposé différents estimateurs de f, s'appuyant sur des hypothèses spécifiques à la modélisation de probabilités critiques.

Cependant, ces méthodes sont toutes basées sur une hypothèse d'indépendance. La présence de corrélation induit une grande variation dans l'estimation de π_0 (voir figure 1 qui compare l'étude de l'estimation de π_0 à partir de probabilités critiques issues de tests de Student sur des données indépendantes (1(a)) ou très corrélées (1(b)) avec différentes méthodes).



(a) Variabilité commune : 0% (b) Variabilité commune : 67.56%

FIG. 1 – Estimations de π_0 à partir de probabilités critiques issues de tests de Student sur des données indépendantes (1(a)) ou très corrélées (1(b)) avec différentes méthodes - 1000 jeux de données simulées (m=500 variables, n=2 × 30 individus) pour chacun des deux scénarios, avec une proportion réelle d'hypothèses sous H_0 π_0 de 80%
Méthode 1 : estimateur de Schweder (1982) avec choix de t par *bootstrap* [Storey, 2002]
Méthode 2 : estimateur de Schweder (1982) avec choix de t par ajustement splines [Storey and Tibsirani, 2003]
Méthode 3 : estimation de la densité par une fonction convexe décroissante sur $[0; 1]$ [Langaas et al., 2005]
Méthode 4 : estimation de la densité par une méthode à noyau [Langaas et al., 2005]
Méthode 5 : estimation à partir du plus long intervalle constant [Langaas et al., 2005]

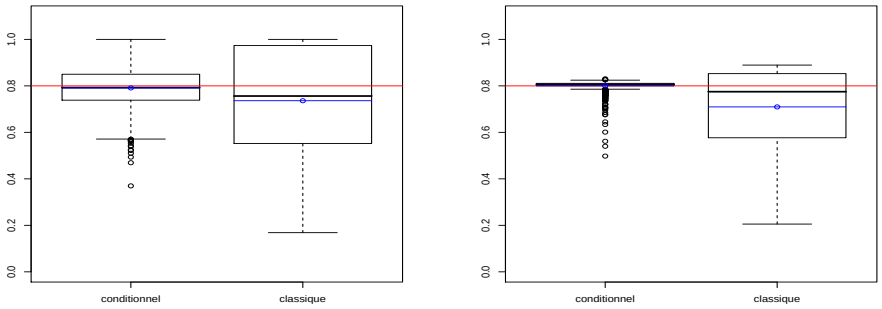
Approche conditionnelle Nous proposons de prendre en compte la dépendance par un modèle d'Analyse en Facteur. Dans cet exposé, on s'intéressera en particulier à l'estimation de π_0. On montre que la variance de l'estimateur de Schweder and Spjotvoll [1982] repose explicitement sur un terme qui dépend de la corrélation (voir aussi Owen [2005] et Efron [2007]). On définit alors un estimateur conditionnel de π_0, qui corrige l'estimateur classique en tenant compte de la dépendance :

$$\pi_0^{(c)}(t) = \frac{W_t}{m(1 - \bar{t_Z}(t))}$$

où $\bar{t_Z}(t) = \frac{1}{m_0} \sum_{k \in \mathcal{M}_0} \mathbb{P}\left(p_k \leq t | Z\right)$.

Par ailleurs, la dépendance entre les données induit une grande variabilité de l'estimation de f autour de la densité réelle, en particulier au voisinage de 1, conduisant à une grande variabilité de l'estimation du paramètre. L'approche conditionnelle est la suivante : nous utilisons les algorithmes d'estimation de la densité, appliqués à la distribution des probabilités critiques associées aux statistiques de test conditionnelles.

On peut alors comparer les approches classiques avec les approches conditionnelles, sur les mêmes données simulées que précédemment. Les résultats sont présentés sur la figure 2, dans le cas d'une forte dépendance. Le cadre conditionnel, tenant compte du modèle d'Analyse en Facteurs, permet de proposer des estimateurs qui améliorent l'estimation de π_0 en terme de variabilité et de biais.



(a) choix de t par ajustement splines [Storey and Tibsirani, 2003] : comparaison de l'estimateur classique [Schweder and Spjotvoll, 1982] et de l'estimateur conditionnel

(b) estimation de la densité des probabilités critiques ajustées (conditionnel) et de la densité des probabilité de Student (classique) - hypothèse de convexité de la densité [Langaas et al., 2005]

FIG. 2 – Estimations de π_0 - Données très corrélées (Variabilité commune : 67.56%) avec différentes méthodes - 1000 jeux de données simulées (m=500 variables, n=2 × 30 individus), avec une proportion réelle d'hypothèses sous H_0 π_0 de 80%.

Références

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate : a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57 :289–300, 1995.

M. A. Black. A note on the adaptative control of false discovery rates. *Journal of the Royal Statistical Society. Series B*, 66 :297–304, 2004.

A. Celisse and S. Robin. Nonparametric density estimation by explicit leave-p-out cross-validation. *Comput. Statist. Data Analysis*, 52 :2350–2368, 2008.

B. Efron. Correlation and large-scale simultaneous testing. *Journal of the American Statistical Association*, 102 :93–103, 2007.

C. Friguet, M. Kloareg, and D. Causeur. A factor model approach to multiple testing under dependence. *to appear*, 2009.

R. Kustra, R. Shioda, and M. Zhu. A factor analysis model for functional genomics. *BMC Bioinformatics*, 7, 2006.

M. Langaas, B. H. Lindqvist, and E. Ferkingstad. Estimating the proportion of true null hypotheses, with application to dna microarray data. *Journal of the Royal Statistical Society. Series B*, 67 :555–572, 2005.

A.B. Owen. Variance of the number of false discoveries. *Journal of the Royal Statistical Society. Series B*, 67 :411–426, 2005.

I. Pournara and L. Wernisch. Factor analysis for gene regulatory networks and transcription factor activity profiles. *BMC Bioinformatics*, 8, 2007.

D. B. Rubin and D. T. Thayer. Em algorithms for ml factor analysis. *Psychometrika*, 47 :69–76, 1982.

T. Schweder and E. Spjotvoll. Plots of p-values to evaluate many tests simultaneously. *Biometrika*, 69 : 493–502, 1982.

J. Storey and R Tibsirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100 :9440–9445, 2003.

J. D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society. Series B*, 64 :479–498, 2002.