Sélection bayésienne de biomarqueurs : application à un problème de protéomique

Noura DRIDI[1], Audrey GIREMUS[1], Jean-Francois GIOVANNELLI[1],

[1]Univ. Bordeaux, IMS, UMR 5218, F-33405 Talence
Mail:{noura.dridi, audrey.giremus, giova}@ims-bordeaux.fr

Résumé – Cet article présente une méthode de sélection de variables en protéomique. L'objectif est de détecter parmi un ensemble des protéines celles qui sont discriminantes (dites biomarqueurs) i.e qui permettent de distinguer entre deux cohortes d'individus sains et malades. Notre approche est fondée sur un modèle bayésien hiérarchique reliant variables biologiques et observations issues d'un spectromètre de Masse. La fonction de sélection optimale est construite de façon à minimiser le risque bayésien d'erreur, ce qui revient à maximiser la probabilité a posteriori. Nous accordons une attention particulière au cas univarié qui nous permet de mettre en évidence l'impact des moyennes ou des variances empiriques des deux cohortes sur la fonction de sélection. D'autre part, en cas d'égalité des moyennes ou des variances entre elles, nous montrons que notre méthode équivaut au calcul des statistiques des tests de Student ou Fisher.

Abstract – In this paper, we propose a variable selection method for proteomics. Our goal is to detect, from a set of proteins, those which enable to distinguish between two subset healthy and pathological individuals (named biomarkers). The approach is based on a hierarchical Bayesian model which relates biological variables and observations from the mass spectrometer. The optimal selection function minimises the bayesian risk, or equivalently maximises the posterior probability. We focus on univariate case, in order to study the impact of the empirical means or variances of the subsets of healthy and pathological individuals on the selection function. Moreover, in case the means and variances are equal for the two subsets, the proposed method is similar to standard statistical tests of Student and Fisher.

1 Introduction

La protéomique est un domaine en pleine expansion qui étudie les profils des protéines, i.e leur localisation, concentration etc [7]. Les biomarqueurs sont les protéines différemment exprimées selon l'état biologique de l'individu (sain ou malade). La sélection de ces biomarqueurs est une étape cruciale, qui permet le diagnostic précoce des maladies comme le cancer [6]. Afin de garantir la fiabilité de l'étude, des techniques de mesures très précises sont nécessaires telles que les mesures de Chromatographie Liquide et de Spectrométrie de Masse (LC-MS). Celles-ci fournissent des spectres avec des pics relatifs à la nature et la concentration des protéines [1, 4]. La protéomique se fonde soit directement sur ces spectres [2], soit sur une estimation des concentrations à partir de ces spectres [4, 5]. Dans ce travail, nous utilisons les concentrations des protéines dans le but de sélectionner des biomarqueurs. Il existe principalement deux classes de méthodes : la première utilise un modèle explicatif de type régression logistique où les variables explicatives sont les concentrations des protéines. La découverte des biomarqueurs revient alors à sélectionner les variables minimisant un critère tel que le Critère d'Information Bayésien (BIC), ou le Critère d'Information d'Akaike (AIC). Cependant, ces méthodes présentent une complexité calculatoire élevée car il faut comparer 2^P modèles pour P protéines. Pour réduire cette complexité, [11] propose d'utiliser la méthode d'échantillonnage de Gibbs pour une présélection. Une autre famille de méthodes telles que le LASSO ou l'algorithme de régulation en filet élastique [3, 12], permet de résoudre ce problème. La deuxième classe de méthodes repose sur l'analyse différentielle. Il s'agit d'appliquer un test univarié sur chaque protéine tel que le test de Student. Une difficulté de ces méthodes est la nécessité de contrôler le *False Discovery Rate(FDR)* ou le *Family Wise Error Rate(FWER)* [10]. De plus, elles traitent séparément chaque protéine, et ne permettent pas de prendre en compte une éventuelle corrélation entre les variables. Nous proposons une méthode alternative qui permet de pallier cette limitation tout en n'imposant pas de relation ad-hoc entre les variables explicatives et les variables expliquées. Elle est fondée sur un modèle bayésien hiérarchique reliant état biologique et concentrations. La fonction de sélection optimale est construite de façon à minimiser le risque bayésien d'erreur, ce qui revient à sélectionner la combinaison de protéines la plus probable a posteriori. Une propriété intéressante d'un point de vue calculatoire est qu'un choix judicieux de lois a priori permet d'obtenir une expression explicite de la loi a posteriori. Dans cette communication, nous détaillons le cas univarié et nous montrons que dans certains cas particuliers, notre statistique de test présente des connections avec des statistiques classiques comme celles de Student ou Fisher.

Le reste de l'article est organisé comme suit : la deuxième partie est consacrée à la présentation du modèle bayésien hiérarchique et aux notations utilisées, ensuite nous explicitons la fonction de sélection optimale au sens de la minimisation du risque bayésien. Dans la partie 4, nous détaillons le cas univarié. Les résultats sont présentés dans la partie 5 et nous terminons par une conclusion.

$$f_{\boldsymbol{\mathcal{X}},\mathbf{B}|\boldsymbol{\Delta}}(\mathbf{x},\mathbf{b}|\boldsymbol{\delta}) = \int_{\boldsymbol{\Theta}} \prod_{n\in\mathcal{I}_{\mathcal{M}}} \mathcal{N}(\mathbf{x}_n^+;\mathbf{m}_{\mathcal{M}},\boldsymbol{\Gamma}_{\mathcal{M}}) \prod_{n\in\mathcal{I}_{\mathcal{S}}} \mathcal{N}(\mathbf{x}_n^+;\mathbf{m}_{\mathcal{S}},\boldsymbol{\Gamma}_{\mathcal{S}}) \prod_{n\in\mathcal{I}_{\mathcal{C}}} \mathcal{N}(\mathbf{x}_n^-;\mathbf{m}_{\mathcal{C}},\boldsymbol{\Gamma}_{\mathcal{C}})\ p^{N_{\mathcal{M}}}(1-p)^{N_{\mathcal{S}}}\pi_{\boldsymbol{\Theta}}(\boldsymbol{\theta}|\boldsymbol{\delta})\mathrm{d}\boldsymbol{\theta}. \quad (1)$$

2 Notations et position du problème

On dispose de deux cohortes d'individus sains $\mathcal{S}$ et malades $\mathcal{M}$. Les ensembles des sains et des malades sont notés respectivement $\mathcal{I}_{\mathcal{S}}$ et $\mathcal{I}_{\mathcal{M}}$ de tailles $N_{\mathcal{S}}$ et $N_{\mathcal{M}}$, et l'ensemble de tous les individus est noté $\mathcal{I}_{\mathcal{C}} = \mathcal{I}_{\mathcal{S}} \cup \mathcal{I}_{\mathcal{M}}$ de taille $N_{\mathcal{C}}$. Par la suite, les indices $\{\mathcal{M},\mathcal{S},\mathcal{C}\}$ correspondent respectivement aux malades, sains et communs. Pour un individu n on dispose des observations $(\mathbf{x}_n, b_n)$, où $\mathbf{x}_n \in \mathbf{R}^P$ est le vecteur des concentrations de P protéines, et b_n est le statut biologique de l'individu n, $b_n \in \{\mathcal{S},\mathcal{M}\}$. $\boldsymbol{\mathcal{X}}$ et $\boldsymbol{b} = [b_1,...,b_{N_{\mathcal{C}}}]$ sont la matrice des concentrations des P protéines et le vecteur des statuts biologiques pour les $N_{\mathcal{C}}$ individus, respectivement. En désignant par $+$ / $-$ respectivement une protéine discriminante/non discriminante, $\mathbf{x}_n$ s'écrit $\mathbf{x}_n = (\mathbf{x}_n^+, \mathbf{x}_n^-)$, où $\mathbf{x}_n^+$ et $\mathbf{x}_n^-$ désignent respectivement les vecteurs de concentrations des protéines discriminantes et non discriminantes, de tailles P^+ et P^-. b_n est piloté par une distribution de Bernoulli de paramètre p. La distribution des concentrations conditionnellement au statut est supposée normale multivariée, choix classique dans les problématiques de protéomique [2]. $\mathbf{x}_n^+$ est donc distribué selon un mélange de deux gaussiennes multivariées de moyennes et de précision (inverse de covariance) $(\mathbf{m}_{\mathcal{M}},\boldsymbol{\Gamma}_{\mathcal{M}})$ et $(\mathbf{m}_{\mathcal{S}},\boldsymbol{\Gamma}_{\mathcal{S}})$ et de poids respectifs p et $1-p$. Par contre, $\mathbf{x}_n^-$ est modélisé par une seule gaussienne de paramètres $(\mathbf{m}_{\mathcal{C}},\boldsymbol{\Gamma}_{\mathcal{C}})$. De plus les individus sont supposés indépendants et les vecteurs $\mathbf{x}_n^+$ et $\mathbf{x}_n^-$ non corrélés. Pour P protéines, il existe 2^P modèles possibles décrivant $\boldsymbol{\mathcal{X}}$, et qui correspondent à toutes les combinaisons possibles de protéines discriminantes ou non. Ils sont donnés par $\boldsymbol{\delta} \in \{+,-\}^P$. $\boldsymbol{\theta}^{\boldsymbol{\delta}} = [\mathbf{m}_{\mathcal{M}}^{\boldsymbol{\delta}},\boldsymbol{\Gamma}_{\mathcal{M}}^{\boldsymbol{\delta}},\mathbf{m}_{\mathcal{S}}^{\boldsymbol{\delta}},\boldsymbol{\Gamma}_{\mathcal{S}}^{\boldsymbol{\delta}},\mathbf{m}_{\mathcal{C}}^{\boldsymbol{\delta}},\boldsymbol{\Gamma}_{\mathcal{C}}^{\boldsymbol{\delta}},p]$ désigne l'ensemble des paramètres inconnus, qui dépendent du modèle $\boldsymbol{\delta}$. L'indice $\boldsymbol{\delta}$ est omis dans la suite pour alléger les notations.

L'objectif est de sélectionner le modèle à partir des observations. Pour construire une fonction de sélection, on s'intéresse à ses performances. On définit un coût 0-1 qui affecte respectivement les valeurs $0/1$ à une bonne/mauvaise sélection. Le risque bayésien pour la sélection est défini par le coût moyen, et il s'agit d'une triple moyenne sur : les 2^P modèles concurrents, les observations et les paramètres $\boldsymbol{\theta}$. La fonction de sélection optimale est celle qui minimise le risque bayésien, et consiste à sélectionner le modèle le plus probable a posteriori.

3 Fonction de sélection

Pour chaque modèle candidat $\boldsymbol{\delta}$, en utilisant la formule de Bayes, la probabilité a posteriori $\mathbb{P}_{\boldsymbol{\Delta}|\boldsymbol{\mathcal{X}},\mathbf{B}}(\boldsymbol{\delta}|\mathbf{x},\mathbf{b})$ s'écrit :

$$\mathbb{P}_{\boldsymbol{\Delta}|\boldsymbol{\mathcal{X}},\mathbf{B}}(\boldsymbol{\Delta}=\boldsymbol{\delta}|\mathbf{x},\mathbf{b}) \propto f_{\boldsymbol{\mathcal{X}},\mathbf{B}|\boldsymbol{\Delta}}(\mathbf{x},\mathbf{b}|\boldsymbol{\delta})\mathbb{P}(\boldsymbol{\Delta}=\boldsymbol{\delta}). \quad (2)$$

En tenant compte de la non corrélation entre $\mathbf{x}_n^+$ et $\mathbf{x}_n^-$ et de l'indépendance entre les individus, la vraisemblance est donnée par l'équation (1). Les trois premiers facteurs correspondent à des lois gaussiennes multivariées. Le quatrième facteur est commun pour tous les modèles et n'impacte donc pas le résultat de sélection. Le dernier $\pi_{\boldsymbol{\Theta}}(\boldsymbol{\theta}|\boldsymbol{\delta})$ est la probabilité a priori des paramètres inconnus $\boldsymbol{\theta}$. Ce facteur est particulièrement important. D'une part, il permet de tenir compte des informations a priori. D'autre part, un choix judicieux de lois a priori conjuguées permet d'obtenir une expression analytique de la vraisemblance. Ainsi, nous pouvons noter que les trois premiers facteurs de l'équation (1) peuvent se réécrire sous forme de lois Normale-Wishart. Nous détaillons uniquement la première :

$$\prod_{n\in\mathcal{I}_{\mathcal{M}}} \mathcal{N}(\mathbf{x}_n^+;\mathbf{m}_{\mathcal{M}},\boldsymbol{\Gamma}_{\mathcal{M}}) = (2\pi)^{-PN_{\mathcal{M}}/2}|\boldsymbol{\Gamma}_{\mathcal{M}}|^{N/2}$$
$$\exp\left[-\frac{N_{\mathcal{M}}}{2}\mathrm{tr}\left(\boldsymbol{\Gamma}_{\mathcal{M}}\left[\bar{\mathbf{R}}_{\mathcal{M}} + (\bar{\mathbf{x}}_{\mathcal{M}}-\mathbf{m}_{\mathcal{M}})(\bar{\mathbf{x}}_{\mathcal{M}}-\mathbf{m}_{\mathcal{M}})^t\right]\right)\right].$$

où $\bar{\mathbf{x}}_{\mathcal{M}}$ et $\bar{\mathbf{R}}_{\mathcal{M}}$ désignent respectivement les moyennes et covariances empiriques des protéines discriminantes pour les individus malades. Pour respecter le principe de conjugaison, nous choisissons donc un a priori séparable de la forme :

$$\pi_{\boldsymbol{\Theta}}(\boldsymbol{\theta}|\boldsymbol{\Delta}) = \pi_{\mathcal{M}}(\mathbf{m}_{\mathcal{M}},\boldsymbol{\Gamma}_{\mathcal{M}})\pi_{\mathcal{S}}(\mathbf{m}_{\mathcal{S}},\boldsymbol{\Gamma}_{\mathcal{S}})\pi_{\mathcal{C}}(\mathbf{m}_{\mathcal{C}},\boldsymbol{\Gamma}_{\mathcal{C}})\pi_p(p),$$

où $\pi_\times$ est une loi Normale-Wishart de paramètres $(\nu_\times,\eta_\times,\boldsymbol{\mu}_\times,\boldsymbol{\Lambda}_\times)$ avec $\times \in \{\mathcal{M},\mathcal{S},\mathcal{C}\}$, et π_p est une loi Bêta de paramètres (α,β).

En remplaçant $\pi_{\boldsymbol{\Theta}}(\boldsymbol{\theta}|\boldsymbol{\Delta})$ dans l'équation (1), et en intégrant par rapport à $\boldsymbol{\theta}$, nous obtenons finalement :

$$f_{\boldsymbol{\mathcal{X}},\mathbf{B}|\boldsymbol{\Delta}}(\mathbf{x},\mathbf{b}|\boldsymbol{\delta}) \propto \frac{K^{\text{pst}}(\mathcal{M})}{K^{\text{pri}}(\mathcal{M})}\frac{K^{\text{pst}}(\mathcal{S})}{K^{\text{pri}}(\mathcal{S})}\frac{K^{\text{pst}}(\mathcal{C})}{K^{\text{pri}}(\mathcal{C})}. \quad (3)$$

où $K^{\text{pri}}(\times)$ et $K^{\text{pst}}(\times)$ sont les constantes de normalisation de la loi a priori et de la loi a posteriori des vecteurs de moyenne et des matrices de précision, respectivement. Elles sont toutes les deux Normale-Wishart.
$K(\times) = (2\pi)^{P^\star/2}2^{\nu_\times P^\star/2}\eta^{-P^\star/2}\det(\boldsymbol{\Lambda}_\times)^{\nu_\times/2}\Gamma_{P^\star}(\nu_\times/2)$, avec $\star \in \{+,-\}$. Les paramètres de la loi a posteriori sont liés aux paramètres a priori comme suit :

$$\begin{aligned}
\nu_\times^{\text{pst}} &= \nu_\times + N_\times, \qquad \eta_\times^{\text{pst}} = \eta_\times + N_\times,\\
\boldsymbol{\mu}_\times^{\text{pst}} &= (N_\times\bar{\mathbf{x}}_\times + \eta_\times\boldsymbol{\mu}_\times)/(N_\times+\eta_\times),\\
(\boldsymbol{\Lambda}_\times^{\text{pst}})^{-1} &= (\boldsymbol{\Lambda}_\times)^{-1} + N_\times\bar{\mathbf{R}}_\times +\\
&\quad N_\times\eta_\times(\boldsymbol{\mu}_\times-\bar{\mathbf{x}}_\times)(\boldsymbol{\mu}_\times-\bar{\mathbf{x}}_\times)^t/(N_\times+\eta_\times).
\end{aligned} \quad (4)$$

Remarquons que les paramètres a posteriori dépendent, en plus des hyperparamètres $(\nu_\times,\eta_\times,\boldsymbol{\mu}_\times,\boldsymbol{\Lambda}_\times)$, de $\bar{\mathbf{x}}_\times$, $\bar{\mathbf{R}}_\times$ et $\bar{p}$. $\bar{\mathbf{x}}_\times$ et $\bar{\mathbf{R}}_\times$ désignent respectivement les moyennes et covariances empiriques calculées à partir des concentrations des protéines concernées pour les $N_\times$ individus. On définit également $\bar{p} = N_{\mathcal{M}}/N_{\mathcal{C}}$ qui est la proportion des malades. Ce dernier terme apparaît en écrivant les paramètres empiriques de la cohorte globale en fonction de ceux des cohortes $\mathcal{M}$ et $\mathcal{S}$:

$$\begin{aligned}
\bar{\mathbf{x}}_{\mathcal{C}} &= \bar{p}\bar{\mathbf{x}}_{\mathcal{M}} + (1-\bar{p})\bar{\mathbf{x}}_{\mathcal{S}}\\
\bar{\mathbf{R}}_{\mathcal{C}} &= \bar{p}\bar{\mathbf{R}}_{\mathcal{M}} + (1-\bar{p})\bar{\mathbf{R}}_{\mathcal{S}} + \bar{p}(1-\bar{p})(\bar{\mathbf{x}}_{\mathcal{S}}-\bar{\mathbf{x}}_{\mathcal{M}})(\bar{\mathbf{x}}_{\mathcal{S}}-\bar{\mathbf{x}}_{\mathcal{M}})^t.
\end{aligned} \quad (5)$$

Ces relations sont importantes pour la formulation et l'interprétation des résultats dans la suite. Nous étudions en particulier,

l'impact de $\bar{\mathbf{x}}_\times$, $\bar{\mathbf{R}}_\times$ et $\bar{p}$ sur la probabilité a posteriori.

Remarques :

-Le nombre de modèles possibles augmente exponentiellement en fonction de P, ce qui peut paraître coûteux calculatoirement. Cependant du fait qu'on a une formule explicite de la probabilité a posteriori, la complexité calculatoire reste réduite même pour des valeurs élevées de P. D'autre part, nous calculons le vecteur des moyennes et la matrice des covariances des P protéines pour les sous cohortes $\mathcal{M}$ et $\mathcal{S}$. Les moyennes et covariances des communs sont déduites en utilisant les équations (5). Ensuite, nous en extrayons les quantités empiriques nécessaires pour tous les modèles concurrents.

-La vraisemblance (3) dépend des paramètres de la Normale-Wishart $(\nu_\times, \eta_\times, \boldsymbol{\mu}_\times, \boldsymbol{\Lambda}_\times)$. Dans le cas non informatif où ces paramètres tendent vers 0, le coefficient de proportionnalité a une forme indéterminée. Pour ajuster ces paramètres, nous nous plaçons dans un cadre peu informatif où on connait l'ordre de grandeur de $(\mathbf{m}_\times \ \boldsymbol{\Gamma}_\times)$. Le calcul est donné dans [8].Cet ajustement permet de lever l'indétermination sans prépondérance de l'information a priori sur celle apportée par les données.

Les performances de la méthode pour $P > 1$ ont été abordées dans [8], et la contribution du présent papier consiste à approfondir le cas $P = 1$, et établir des connections avec des tests univariés classiques.

4 Etude du cas univarié

Dans le cas où $P = 1$, il s'agit de déterminer si la protéine considérée est un biomarqueur ou non. On a alors seulement deux modèles concurrents $\boldsymbol{\Delta} = \{+, -\}$, supposés équiprobables a priori. De même, les hyperparamètres $(\nu_\times, \eta_\times, \boldsymbol{\mu}_\times, \boldsymbol{\Lambda}_\times)$ où $\times \in \{\mathcal{M}, \mathcal{S}, \mathcal{C}\}$ sont supposés identiques. La probabilité que la protéine soit discriminante $P(\boldsymbol{\Delta} = +|\mathbf{x}, \boldsymbol{b})$, notée Pr^+, dépend dans ce cas uniquement des constantes de la loi a posteriori. Celles-ci sont fonctions des paramètres a posteriori donnés par les équations (4), où il apparaît que seul le paramètre $\boldsymbol{\Lambda}_\times^{\text{pst}}$ dépend des observations. Pour un a priori peu informatif, nous obtenons $\text{Pr}^+ \propto (1+T)^{-1}$ où :

$$T \propto \frac{\left(\bar{p}\bar{\mathbf{R}}_\mathcal{M} + (1-\bar{p})\bar{\mathbf{R}}_\mathcal{S} + \bar{p}(1-\bar{p})(\bar{\mathbf{x}}_\mathcal{M} - \bar{\mathbf{x}}_\mathcal{S})^2\right)^{-N_\mathcal{C}/2}}{\bar{\mathbf{R}}_\mathcal{M}^{-N_\mathcal{M}/2}\bar{\mathbf{R}}_\mathcal{S}^{-N_\mathcal{S}/2}}. \tag{6}$$

Si on suppose l'égalité des variances $\bar{\mathbf{R}}_\mathcal{S} = \bar{\mathbf{R}}_\mathcal{M} = \bar{\mathbf{R}}$,

$$T \propto \left(1 + \bar{p}(1-\bar{p})(\bar{\mathbf{x}}_\mathcal{M} - \bar{\mathbf{x}}_\mathcal{S})^2/\bar{\mathbf{R}}\right)^{-N_\mathcal{C}/2}. \tag{7}$$

La probabilité a posteriori que la protéine soit discriminante est alors fonction du rapport des écarts des moyennes et de la variance empirique $(\bar{\mathbf{x}}_\mathcal{S} - \bar{\mathbf{x}}_\mathcal{M})^2/\bar{\mathbf{R}}$. Remarquons que ce facteur correspond à la statistique du test de Student [9].

En cas d'égalité des moyennes $\bar{\mathbf{x}}_\mathcal{S} = \bar{\mathbf{x}}_\mathcal{M}$, nous obtenons :

$$T \propto \left(\bar{\mathbf{R}}_\mathcal{M}/\bar{\mathbf{R}}_\mathcal{S}\right)^{N_\mathcal{M}/2}\left(1 - \bar{p} + \bar{p}(\bar{\mathbf{R}}_\mathcal{M}/\bar{\mathbf{R}}_\mathcal{S})\right)^{-N_\mathcal{C}/2}. \tag{8}$$

La probabilité a posteriori que la protéine soit discriminante dépend alors uniquement du rapport des variances empiriques des deux cohortes, comme pour la statistique de Fisher utilisée pour tester l'égalité de deux variances [9].

5 Résultats

Les performances de l'algorithme de sélection proposé sont illustrées par des calculs numériques. Les résultats présentés concernent uniquement le cas univarié $P = 1$. Pour $P > 1$ des résultats sont fournis et analysés dans [8]. Dans un premier temps, l'objectif est d'évaluer l'impact des paramètres empiriques $\bar{\mathbf{R}}_\times$, $\bar{\mathbf{x}}_\times$ et $\bar{p}$ sur la probabilité de détecter un biomarqueur Pr^+. A chaque fois, deux des paramètres sont fixés et nous faisons varier le troisième. Dans un second temps, nous nous focalisons sur les performances de l'algorithme en terme de taux d'erreur de sélection, et le comparons à des tests classiques.

Égalité des variances : nous supposons que les variances empiriques des concentrations des sains et des malades sont égales, nous fixons la moyenne empirique des sains $\bar{\mathbf{x}}_\mathcal{S} = 10$ et nous faisons varier celle des malades $\bar{\mathbf{x}}_\mathcal{M} \in \{0, ..., 20\}$. La figure 1 illustre la variation du logarithme de la probabilité a posteriori que la protéine soit discriminante Pr^+ en fonction des écarts de moyennes $\bar{\mathbf{x}}_\mathcal{M} - \bar{\mathbf{x}}_\mathcal{S}$, pour différentes valeurs de la variance empirique $\bar{\mathbf{R}}$ et de la probabilité empirique d'être malade $\bar{p}$. En observant la figure 1, il est clair que Pr^+ est une

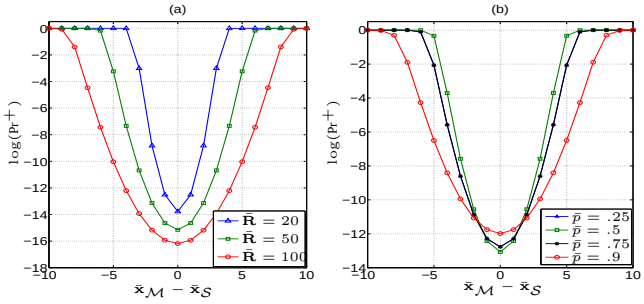

FIGURE 1 – Log de Pr^+ en fonction de $\bar{\mathbf{x}}_\mathcal{M} - \bar{\mathbf{x}}_\mathcal{S}$ et $\bar{p}$, $N_\mathcal{C} = 100$.

fonction symétrique et croissante avec l'écart des moyennes en valeur absolue. De plus, pour une valeur fixe de cet écart, Pr^+ est une fonction décroissante de $\bar{\mathbf{R}}$, résultat cohérent avec l'équation (7). Sur la figure 1-(a), nous constatons que plus la variance empirique est grande plus l'écart de moyennes doit être élevé pour déclarer un biomarqueur.

Toujours dans le cas d'égalité des variances, nous représentons sur la figure 1-(b), la variation du logarithme de Pr^+ en fonction des écarts des moyennes, pour différentes valeurs de $\bar{p}$. Nous remarquons que les courbes pour $\bar{p} = .25$ et $\bar{p} = .75$ sont confondues, ce résultat est expliqué par la symétrie de Pr^+ par rapport à $\bar{p}$ et $1 - \bar{p}$ comme confirmé par (7). D'autre part, il est noté que plus la proportion des malades $\bar{p}$ augmente, plus les écarts des moyennes doivent être significatifs pour déclarer un biomarqueur.

Égalité des moyennes : nous supposons maintenant que les moyennes empiriques sont égales $\bar{\mathbf{x}}_\mathcal{S} = \bar{\mathbf{x}}_\mathcal{M} = 20$, nous fixons la variance empirique des sains $\bar{\mathbf{R}}_\mathcal{S} = 20$ et nous faisons varier celle des malades $\bar{\mathbf{R}}_\mathcal{M} \in \{1, ..., 120\}$. La figure 2 illustre la variation du logarithme de Pr^+ en fonction du rapport des variances et de $\bar{p}$. En observant la figure 2, il est clair que plus la proportion des malades est élevée, plus la variance des malades doit être grande par rapport à celle des sains pour sélectionner un biomarqueur, et réciproquement.

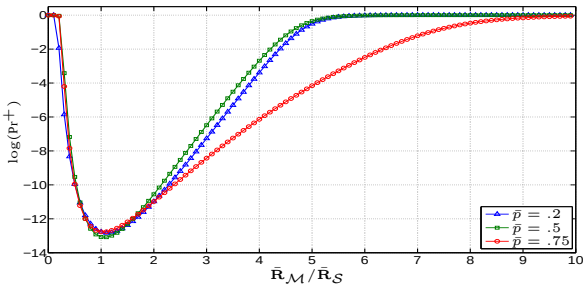

FIGURE 2 – Log de $\Pr^+$ en fonction de $\bar{\mathbf{R}}_{\mathcal{M}}/\bar{\mathbf{R}}_{\mathcal{S}}$ et $\bar{p}$

Performances de l'algorithme de sélection : dans la suite les performances de l'algorithme sont évaluées en terme d'erreur de sélection. Pour $N_{\mathcal{C}} = 100$ individus et une seule protéine, $N_r = 10^6$ réalisations sont générées pour chacune des configurations testées, désignées par $+/-$ selon si la protéine est discriminante ou non. Pour chaque configuration, les statuts des individus et le vecteur des concentrations de la protéine sont générés selon les lois décrites dans la partie 2. Les paramètres de ces lois, p et $(\mathbf{m}_\times\ \mathbf{\Gamma}_\times)$ avec $\times \in \{\mathcal{M}, \mathcal{S}, \mathcal{C}\}$, sont également simulés selon des lois de probabilité définies dans la partie 2. Les hyperparamètres $(\nu_\times, \eta_\times, \boldsymbol{\mu}_\times, \mathbf{\Lambda}_\times)$ sont calculés comme décrit dans [8]. La probabilité a posteriori est donnée par l'équation (3). Parmi les deux modèles concurrents, nous sélectionnons celui qui maximise cette probabilité.
Les résultats sont comparés avec le test de Student. Ce dernier teste l'égalité des moyennes : la protéine est déclarée comme biomarqueur si l'hypothèse d'égalité des moyennes empiriques de deux sous cohortes malade et saine est rejetée. Le risque de première espèce est fixé $\alpha = 0.1\%$. Nous calculons les matrices de confusion : une case du tableau correspond au pourcentage des réalisations pour lesquelles le vrai modèle est M^* et celui estimé est $\hat{M}$. D'après les résultats de la table 1, il

Test de Student			Algorithme bayésien		
M^* \ $\hat{M}$	-	+	M^* \ $\hat{M}$	-	+
-	99.895	0.104	-	99.955	0.044
+	2.681	97.318	+	2.665	97.334

TABLE 1 – Pourcentage de faux positifs et négatifs ainsi que de vrais positifs et négatifs. $+/-$ pour protéine discriminante/non discriminante

est clair que l'algorithme bayésien sélectionne plus souvent le bon modèle. La proportion de bonne sélection, donnée par la moyenne des proportions des vrais positifs et vrais négatifs, est égale à 98.644%, ce qui est équivaut à un taux de fausse déclaration inférieur à 2%. Pour la répartition des erreurs, on remarque qu'il s'agit souvent de faux négatifs.

6 Conclusion

La détection de biomarqueurs peut être considérée comme un problème de sélection de variables. Plusieurs approches statistiques peuvent alors être utilisées : univariées ou multivariées. Dans ce travail, nous avons considéré une approche multivariée, fondée sur une modélisation bayésienne hiérarchique où l'ensemble des informations sur les variables sont décrites par des lois de probabilités. La méthode consiste à minimiser le risque bayésien d'erreur ce qui revient à sélectionner le modèle le plus probable a posteriori. La difficulté réside dans le calcul de la probabilité a posteriori qui nécessite l'intégration par rapport aux paramètres inconnus. Nous avons proposé un choix des lois a priori qui nous a permis de calculer analytiquement la probabilité a posteriori. Le modèle sélectionné est celui qui maximise cette probabilité. Dans cet article, nous avons approfondi le cas d'une seule protéine, en particulier l'égalité des moyennes ou variances empiriques. Les simulations numériques, montre une cohérence avec les formules du calcul, et une amélioration des performances comparés au test de Student.

Références

[1] G. Strubel. *Reconstruction de profils molt'eculaires : modélisation et inversion d'une chaîne de mesure protómique.* PhD thesis, Ecole Polytechnique de Grenoble. France, 2008

[2] P. Szacherski et J.-F. Giovannelli et P. Grangeat. *Joint Bayesian hierarchical inversion-classification and application in proteomics* IEEE Statistical Signal Processing Workshop. pp : 121-124, 2011

[3] H. Zou, T. Hastie. *Regularization and variable selection via the elastic net*. Journal of the Royal Society : Series B(Statistical Methodology). vol 67, pp :301-320, 2005.

[4] P. Grangeat, L. Gerfault, C. Paulus, V. Kritsotakis, M.N. Tsiknakis, F. Lisacek, P.A. Binz, M. Perez, M. Trauchessec, V. Brun. *First demonstration on NSE biomarker of a computational environment dedicated to lab-on-chip based cancer diagnosis.* Poster at 58th ASMS Conference, Salt Lake City. 2010

[5] G. Strubel, J.P. Giovannelli, C. Paulus, L. Gerfault, P. Grangeat. *Bayesian estimation for molecular profile reconstruction in proteomics based on liquid chromatography and mass spectrometry*, IEE International Conference on Engineering in Medicine and Biology Society. pp : 5979-5982, 2007.

[6] Martin D.B, Nelson P.S. *From genomics to proteomics : techniques and applications in cancer research* Trends in Cell Biology. vol :11, pp :S60-S65, 2001.

[7] R.E. Banks, M.J. Dunn, D.F. Hochstrasser, J.C. Sanchezc, W. Blackstock, D.J. Pappin, P.J. Selby. *Proteomics : new perspectives, new biomedical opportunities*. Journal The Lancet. vol 356, pp :1749-56 , 2000.

[8] F. Adjed, J.-F. Giovannelli, A. Giremus, N. Dridi, P. Szacherski *Variable selection for a mixed population applied in proteomics*. International Conference on Acoustics, Speech, and Signal Processing Mai 2012.

[9] G. Saporta *Probabilités, analyse des données et statistique*. Technip, Paris, 1990.

[10] Y. Benjamin, Y. Hochberg *Controlling the false discovery rate : a practical and powerful approach to multiple testing*, Journal of the Royal Statistical Society, Series B (Methodological) vol :57(1), pp :289-300, 1995.

[11] George E, McCulloch R. *Variable selection via the Gibbs sampling*. J. Amer. Statist. Assoc. vol :88, pp :881-889,1993.

[12] P. Bühlmann, T. Hothorn. *Boosting algorithms : regularization, prediction and model fitting (with discussion)* Statistical Science. vol :22(4), pp 477-505, 2007