

Apprentissage supervisé robuste de caractéristiques de classes. Application en protéomique

Pascal SZACHERSKI^{1,2}, Jean-François GIOVANNELLI¹, Laurent GERFAULT², Pierre GRANGEAT²

¹IMS (Université Bordeaux 1 – IPB – CNRS), 351 rue de la Liberté, 33405 Talence Cedex, France

²CEA-LETI, MINATEC Campus, DTBS, 17 rue des Martyrs, 38054 Grenoble Cedex, France

¹prénom.nom@ims-bordeaux.fr, ²prénom.nom@cea.fr

Thème – 3.5 Problèmes inverses et déconvolution ; 4.4 Apprentissage ; 6.3 Bio-ingénierie et sciences de la vie

Problème traité – La protéomique est un domaine en pleine expansion, et son utilisation est notamment envisagée dans le diagnostic de maladies comme le cancer. Un diagnostic basé sur une classification nécessite la connaissance des distributions des biomarqueurs dans chaque classe. L'apprentissage des distributions est réalisé à partir de données acquises suivant un modèle direct hiérarchique. L'acquisition des données fait intervenir des variabilités biologique et technologique. Pour un apprentissage robuste aux fluctuations instrumentales, il est nécessaire d'en tenir compte en les intégrant dans l'estimation.

Originalité – Pour obtenir un apprentissage robuste, nous proposons d'utiliser le cadre des problèmes inverses pour intégrer les variabilités technologique (instrumentale) et biologique qui impactent les données et de tirer profit de la nature hiérarchique du modèle.

Résultats – La distribution des biomarqueurs apprise en intégrant les variabilités technologique et biologique est proche de la vraie distribution. La robustesse est montrée grâce à la proximité des distributions estimées par apprentissage global et par apprentissage idéal avec les variables instrumentales connues, quantifiée par la divergence de Kullback-Leibler.

1 Introduction

Les protéines d'un organisme peuvent être sur- ou sous-exprimées suivant le statut clinique. Le profil protéique permet de procéder à un diagnostic, une détection de maladie précoce, un suivi thérapeutique, de décrire l'efficacité d'un médicament, etc. Cependant, les biomarqueurs, i.e. les protéines différenciellement exprimées spécifiques à une maladie, sont souvent difficilement détectables du fait de leur faible concentration. L'intérêt de l'utilisation d'un cadre bayésien à partir de mesures d'un couplage de chromatographie liquide (LC) et de spectrométrie de masse (MS) pour la quantification robuste a été démontré [1]. Le modèle physique de l'instrument a été réalisé dans [2] et nous sert de base de travail.

Les tâches de classification, telles que la détection ou le diagnostic [3], nécessitent de connaître les caractéristiques des états biologiques d'une manière robuste par rapport aux variabilités mentionnées. Elles doivent être apprises sur les mesures étiquetées provenant de cohortes identifiées. Un apprentissage supervisé directement sur les données brutes LC-MS ou sur des peptides est une méthode incertaine car elle n'intègre pas les variabilités technologique et biologique sous-jacentes. Un autre moyen est d'estimer les concentrations moléculaires des biomarqueurs, et d'apprendre les caractéristiques à partir de ces grandeurs. Cela intègre certes la variabilité biologique, mais la variabilité technologique est omise.

Dans l'approche présentée ci-après, nous extrayons d'un jeu de données d'apprentissage les paramètres de la distribution des biomarqueurs tout en intégrant les variabilités technologique et biologique. Nous disposons de mesures indirectes de la grandeur d'intérêt, d'où le recours au cadre méthodologique des problèmes inverses. L'acquisition se fait dans une cascade de processus séquentiels : un modèle direct hiérarchique explique alors les données observées. Ces processus étant aléatoires, nous intégrons les variabilités à l'aide d'une modélisation probabiliste. L'estimation globale dans un cadre bayésien est réputée plus robuste que l'estimation séparée « paramètre par paramètre » [4]. De plus, cela nous permet d'utiliser la nature hiérarchique du modèle [5].

2 Apprentissage des caractéristiques

Les mesures LC-MS sont étiquetées et acquises indépendamment les unes des autres ce qui permet de regrouper les expériences par classe. L'apprentissage global se sépare alors naturellement en autant de processus d'apprentissages qu'il y a de classes. C'est la raison pour laquelle nous pouvons nous restreindre à un apprentissage des caractéristiques d'une classe.

L'apprentissage des paramètres caractéristiques de la distribution, à savoir la moyenne m et la précision Γ dans le cas présenté, est fait à partir d'une cohorte de N mesures.

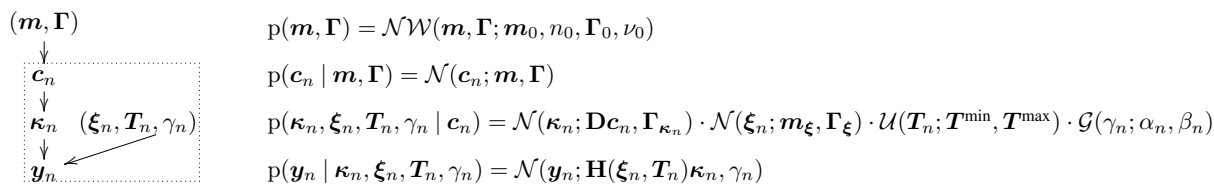


Fig. 1 – Schéma du modèle hiérarchique direct et lois *a priori* associées.

2.1 Problème direct

Suivant la décomposition des molécules analysées, une chaîne d'analyse est décrite par un modèle hiérarchique schématisé sur la Fig. 1.

Les caractéristiques des classes \mathbf{m} et Γ déterminent la concentration des protéines \mathbf{c}_n de la n^{e} mesure, $n = 1, \dots, N$. Les protéines sont découpées par un enzyme en peptides de concentration $\boldsymbol{\kappa}_n = \mathbf{D}\mathbf{c}_n + \boldsymbol{\varepsilon}_n$, perturbé par un bruit de digestion. Le gain de digestion est décrit par la matrice \mathbf{D} . Les peptides sont ensuite séparés par chromatographie, caractérisé par leurs temps de rétention \mathbf{T}_n , puis ionisés en vue de l'introduction dans le spectromètre de masse. Lors de l'ionisation, les peptides subissent un gain noté $\boldsymbol{\xi}_n$. Puis, les données $\mathbf{y}_n = \mathbf{H}(\boldsymbol{\xi}_n, \mathbf{T}_n)\boldsymbol{\kappa}_n + \mathbf{b}_n$, perturbées par un bruit blanc gaussien centré \mathbf{b}_n d'inverse-variance γ_n , sont acquises. La matrice de système \mathbf{H} de toutes les expériences traduit la relation linéaire entre les peptides et la sortie.

2.2 Loi jointe

Du fait des incertitudes et variabilités lors de la préparation et à l'intérieur de l'instrument, les valeurs des paramètres à chaque étape ne sont connues qu'à travers des probabilités. Elles sont explicitées dans la Fig. 1 et ont été choisies pour décrire au mieux les processus tout en restant simples à utiliser. Le temps de rétention prenant sa valeur dans $[\mathbf{T}^{\min}, \mathbf{T}^{\max}]$ de manière équiprobable, sa loi *a priori* est donnée par une loi uniforme $\mathcal{U}(\mathbf{T}^{\min}, \mathbf{T}^{\max})$. La loi pour le couple (\mathbf{m}, Γ) est la loi Normale-Wishart qui est le produit d'une loi normale de moyenne \mathbf{m}_0 et de précision $n_0\Gamma$ et d'une loi Wishart de matrice d'échelle Γ_0 avec ν_0 degrés de liberté. La loi jointe s'exprime alors, compte tenu de l'indépendance des expériences et des indépendances conditionnelles, par

$$p(\mathbf{m}, \Gamma, \mathbf{c}_{1:N}, \boldsymbol{\kappa}_{1:N}, \gamma_{1:N}, \boldsymbol{\xi}_{1:N}, \mathbf{T}_{1:N}, \mathbf{y}_{1:N}) = p(\mathbf{m}, \Gamma) \cdot \prod_{n=1}^N p(\mathbf{y}_n | \boldsymbol{\kappa}_n, \gamma_n) p(\gamma_n) p(\boldsymbol{\xi}_n) p(\mathbf{T}_n) p(\boldsymbol{\kappa}_n | \mathbf{c}_n) p(\mathbf{c}_n | \mathbf{m}, \Gamma) \quad (1)$$

où l'indice $1 : N$ regroupe tous les indices $n = 1, \dots, N$.

2.3 Estimateur des caractéristiques

L'objectif est d'estimer les caractéristiques de l'état biologique. Pour obtenir une estimation robuste, nous choisissons l'estimateur moyenne *a posteriori* (EAP),

$$[\bar{\mathbf{m}}, \bar{\Gamma}] = \int [\mathbf{m}, \Gamma] p(\mathbf{m}, \Gamma | \mathbf{y}_{1:N}) d(\mathbf{m}, \Gamma) \quad (2)$$

qui a la propriété de minimiser l'erreur quadratique moyenne. Ce calcul nécessite de marginaliser les paramètres intermédiaires du processus :

$$[\bar{\mathbf{m}}, \bar{\Gamma}] = \int [\mathbf{m}, \Gamma] p(\mathbf{m}, \Gamma, \mathbf{c}_{1:N}, \boldsymbol{\kappa}_{1:N}, \boldsymbol{\xi}_{1:N}, \mathbf{T}_{1:N}, \gamma_{1:N} | \mathbf{y}_{1:N}) d(\mathbf{m}, \Gamma) d(\mathbf{c}_{1:N}, \boldsymbol{\kappa}_{1:N}, \boldsymbol{\xi}_{1:N}, \mathbf{T}_{1:N}, \gamma_{1:N}). \quad (3)$$

Le calcul de cette intégrale est analytiquement impossible. C'est la raison pour laquelle nous allons approcher la loi *a posteriori* jointe $p(\mathbf{m}, \Gamma, \mathbf{c}_{1:N}, \boldsymbol{\kappa}_{1:N}, \boldsymbol{\xi}_{1:N}, \mathbf{T}_{1:N}, \gamma_{1:N} | \mathbf{y}_{1:N})$ en utilisant une méthode MCMC. Ceci permettra de marginaliser les paramètres dans l'équation (3) et ensuite d'approcher l'estimateur de (2).

3 Mise en œuvre algorithmique

À l'intérieur de l'approche MCMC, nous adaptons une structure de Gibbs qui permet de transformer un problème global en une succession de plusieurs sous-problèmes plus simples. Pour cela, on calcule les lois *a posteriori* conditionnelles. Pour les concentrations (protéines et peptides), le gain et la précision du bruit, les lois *a posteriori* conditionnelles sont de la même famille que leur lois *a priori* grâce à leur conjugaison par la fonction de vraisemblance associée. Ceci permet d'échantillonner facilement sous ces lois puisqu'elles sont connues complètement. Quant au temps de rétention où il n'y a pas de conjugaison des lois, une étape de *Metropolis-Hasting Marche Aléatoire* est intégrée dans la boucle de Gibbs [5, sect. 11.5]. Ainsi, on approche ainsi un tirage sous la loi *a posteriori* jointe. La marginalisation des paramètres intermédiaires est obtenue en ne gardant que les tirages des paramètres d'intérêt. Enfin, l'approximation du EAP du couple (\mathbf{m}, Γ) revient à moyenner les tirages aléatoires marginalisés. Les premiers tirages aléatoires ne se font pas sous la loi *a posteriori* jointe [5, sect. 11.6]. Pour ne pas fausser l'approximation de l'estimation, nous ne prenons en compte les tirages qu'après un temps de chauffe de K_0 itérations. Les étapes de l'algorithme sont résumées dans Alg. 1.

Alg. 1 – Résumé de l’algorithme d’estimation des caractéristiques.

<ol style="list-style-type: none"> 1. Initialisation de l’algorithme. 2. Boucle de Gibbs <ul style="list-style-type: none"> – pour $k = 1$ à $K_0 + K$ <ol style="list-style-type: none"> (a) pour $n = 1$ à N, échantillonner : <ol style="list-style-type: none"> i. $\gamma_n^{(k+1)} \sim p(\gamma_n \mathbf{y}_n, \boldsymbol{\kappa}_n^{(k)})$ ii. $\boldsymbol{\xi}_n^{(k+1)} \sim p(\boldsymbol{\xi}_n \mathbf{y}_n, \boldsymbol{\kappa}_n^{(k)}, \mathbf{T}_n^{(k)}, \gamma_n^{(k+1)})$ 	<ol style="list-style-type: none"> iii. $\mathbf{T}_n^{(k+1)} \sim p(\mathbf{T}_n \mathbf{y}_n, \boldsymbol{\kappa}_n^{(k)}, \boldsymbol{\xi}_n^{(k+1)}, \gamma_n^{(k+1)})$ iv. $\boldsymbol{\kappa}_n^{(k+1)} \sim p(\boldsymbol{\kappa}_n \mathbf{y}_n, \gamma_n^{k+1}, \mathbf{T}_n^{(k+1)}, \boldsymbol{\xi}_n^{(k+1)}, \mathbf{c}_n^{(k)})$ v. $\mathbf{c}_n^{(k+1)} \sim p(\mathbf{c}_n \boldsymbol{\kappa}_n^{(k+1)}, \mathbf{m}^{(k)}, \boldsymbol{\Gamma}^{(k)})$ <p>(b) échantillonner $(\mathbf{m}^{(k+1)}, \boldsymbol{\Gamma}^{(k+1)}) \sim p(\mathbf{m}, \boldsymbol{\Gamma} \mathbf{c}_{1:N}^{(k+1)})$</p> <ol style="list-style-type: none"> 3. Approximation du EAP : $[\hat{\mathbf{m}}, \hat{\boldsymbol{\Gamma}}] = \frac{1}{K} \sum_{k=K_0+1}^{K_0+K} [\mathbf{m}^{(k)}, \boldsymbol{\Gamma}^{(k)}]$
---	--

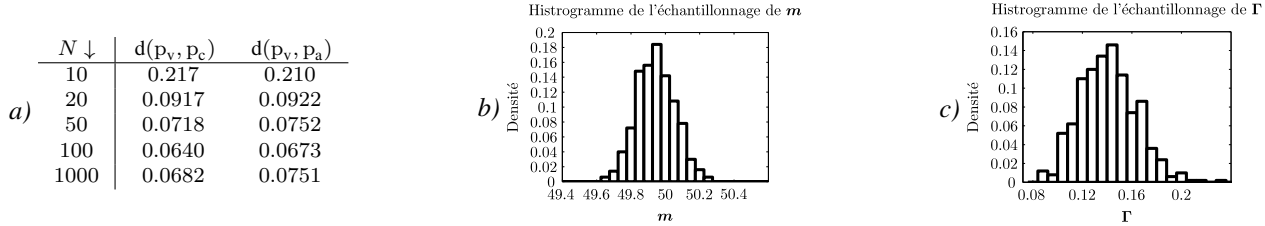


Fig. 2 – a) Résultats moyens sur 55 expériences du test de performance de l’apprentissage des distributions avec les variables technologiques connues (p_c) et estimées p_a . $d(p_v, \cdot)$ désigne la divergence de Kullback-Leibler avec la distribution vraie p_v . b) Histogramme des tirages aléatoires pour m . c) *Idem* pour Γ .

4 Résultats

Nous présentons les résultats obtenus sur une protéine composée d’un peptide. La concentration vraie utilisée dans la cohorte d’apprentissage de taille N a été modélisée selon une loi gaussienne $p_v(c) = \mathcal{N}_{\text{vrai}}(c; m_{\text{vrai}}, \gamma_{\text{vrai}})$ autour de la valeur $m_{\text{vrai}} = 50$ (unité arbitraire) de précision $\gamma_{\text{vrai}} = 0.2$. les autres paramètres ont une variabilité de 2% pour le temps de rétention et de 10% pour le gain de système. Le bruit de mesure est simulé avec $\gamma_n \sim \gamma_{\text{min}} + \mathcal{G}(0.495, 2)$ et une précision minimale $\gamma_{\text{min}} = 0.01$.

La divergence de Kullback-Leibler entre les distributions vraie et estimée est utilisée pour évaluer les performances de l’approche présentée [5, Annexe B]. Ceci est fait en comparant deux stratégies. Premièrement, lors de l’inversion, les paramètres instruments sont fixés aux valeurs vraies pour connaître l’approche idéale de la méthode. Deuxièmement, tous les paramètres instruments sont estimés lors de l’inversion. Les distributions estimées sont notées respectivement p_c et p_a . Les résultats de ces évaluations pour différentes tailles de cohorte sont présentés dans la table 2a).

On constate que les performances de la dernière évaluation, grâce à l’apprentissage robuste, sont proches du cas idéal pour lequel uniquement les paramètres biologiques sont incertains. Les figures 2b) et c) présentent les histogrammes marginaux des tirages aléatoires pour les caractéristiques des classes. La robustesse est également soulignée par les faibles coefficients de variation *a posteriori* pour m et γ , respectivement $CV_m^{\text{post}} = 2.2 \cdot 10^{-3}$, $CV_\gamma^{\text{post}} = 0.16$.

5 Conclusion

Ce document présente une méthode d’apprentissage robuste aux variabilités qui surviennent lors de l’acquisition des données, grâce à l’utilisation du contexte des problèmes inverses. Les distributions apprises par l’apprentissage interviennent lors de l’approche d’inversion-classification bayésienne que nous avons proposée pour analyser les données LC-MS en associant quantification et analyse différentielle [3]. La méthode est transposable à l’apprentissage des paramètres caractéristiques des lois de tous les paramètres intervenant, nous assurant une connaissance des lois *a priori* dans de futures expériences.

Références

- [1] P. Grangeat et al. 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, USA, 2010.
- [2] G. Strubel. *Reconstruction de profils moléculaires : modélisation et inversion d’une chaîne de mesure protéomique*. PhD thesis, École Polytechnique de Grenoble, France, 2008.
- [3] P. Szacherski, J.-F. Giovannelli, and P. Grangeat. Joint bayesian hierarchical inversion-classification and application in proteomics. In *submitted to IEEE SSP*, Nice, France, June 2011.
- [4] L. Gerfault, G. Strubel, C. Paulus, J.-F. Giovannelli, and P. Grangeat. Évaluation statistique d’un algorithme bayésien pour la reconstruction de profils moléculaires par spectrométrie de masse. In *XXII^{ème} Colloque GRETSI*, Dijon, France, 2009.
- [5] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis, Second Edition (Texts in Statistical Science)*. Chapman & Hall/CRC, 2 edition, July 2003.