Méthodes de réduction de dimension basées sur l'algorithme FINE pour le clustering de patients à partir de données de cytométrie en flux

Walid Laziri, Frédéric Allemand, Anne Gégout-Petit, Sophie Wantz-Mézières

Méthodes de réduction de dimension basées sur l'algorithme FINE pour le clustering de patients à partir de données de cytométrie en flux

Walid Laziri [1], Frédéric Allemand [2],
Anne Gégout-Petit [3] & Sophie Wantz-Mézières [4]

[1] *INRIA et EMOSIS, France, walid.laziri@inria.fr*
[2] *EMOSIS, France, frederic.allemand@emosis-diagnostics.com*
[3] *Université de Lorraine, CNRS, INRIA, IECL, F-54000 France, anne.gegout-petit@univ-lorraine.fr*
[4] *Université de Lorraine, CNRS, INRIA, IECL, F-54000 France, sophie.mezieres@univ-lorraine.fr*

Résumé. La cytométrie en flux est une technique utilisée en médecine pour établir le diagnostic de pathologies complexes à partir de données multidimensionnelles rapides (quelques secondes) d'un grand nombre de cellules individuelles ($< 10^4$ pour un tube d'échantillons sanguins). La classification (« clustering ») et l'analyse de ces sous-populations, et donc la discrimination diagnostique entre « malades » et « non malades », est cependant encore essentiellement faite « à la main », à partir d'histogrammes ou de l'examen visuel de nuages de points. Une approche de type algorithmique computationnelle qui automatiserait la recherche des différences ou des similitudes entre ces sous-populations améliorerait la qualité du diagnostic. L'approche considérée ici relève de la géométrie de l'information, qui va réduire la dimensionalité de l'espace des observations multiparamétriques au sous-espace des densités de probabilités du modèle statistique qui décrit ces observations, et qui possède une structure géométrique particulière (variété). L'objectif du présent travail est d'explorer l'algorithme de géométrie de l'information Fisher Information Non-parametric Embedding (FINE) en l'appliquant à des données de cytomètrie en flux, dans le contexte d'une pathologie particulière, la thrombopénie induite par l'héparine (TIH).

Mots-clés. Cytométrie en flux, Géométrie de l'Information, Réduction de dimension.

Abstract. Flow cytometry is used in medicine to diagnose complex disorders from the multiparametric measure (up to 20 parameters) performed in few seconds on tens of thousands of cells. Clustering and analysis of those data is however still done "by the hand", on histograms and 2D scatter plots, which impedes the quality of diagnostic discrimination between "disease" and "non disease" patients. Computational algorithmic approach that would automate and deepens the search of differences or similarities between cell subpopulations could thus increase the quality of diagnosis. The approach considered here is information geometry, whose principle is to lower the dimensionality of multiparametric observations by considering the subspace of the parameters of the statistical model describing the observation, whose points are probability density functions, and which is equipped with a special geometrical structure (manifold). The objective of the reported study is to explore an algorithm belonging to the field of information geometry, Fisher Information Non-parametric Embedding (FINE), by applying it to flow cytometry data in the context of a specific severe disorder, heparin-induced thrombocytopenia.

Keywords. Flow Cytometry, Information Geometry, Dimension Reduction

1 Introduction

1.1 Données de cytométrie en flux

La cytométrie en flux est une technique utilisée pour analyser des cellules ou des particules individuelles contenues dans un échantillon biologique. L'échantillon passe dans un cytomètre, où les cellules sont irradiées par un laser, ce qui les fait se disperser et émettre une lumière fluorescente. Un certain nombre de détecteurs rassemblent et analysent ensuite la lumière diffusée et émise, produisant une quantité importante d'informations quantitatives sur chaque cellule. Ces données peuvent contenir des informations sur leur taille, leur granularité et l'expression de protéines particulières ou d'autres marqueurs. Chaque cellule est ainsi caractérisée par un ensemble de paramètres (jusqu'à 20), permettant d'identifier dans l'échantillon des sous-populations cellulaires partageant des caractéristiques communes spécifiques d'un diagnostic. La distribution des paramètres mesurés au sein de la population de l'échantillon est souvent représentée dans les données de cytométrie en flux sous forme d'histogrammes ou de graphiques en points. Cependant, analyser manuellement ces mêmes diagrammes de dispersion ou des histogrammes ne prend pas en compte le caractère multidimensionnel des données, d'où l'intêret d'utiliser d'autres méthodes afin d'obtenir de meilleures analyses.

1.2 Algorithme FINE

L'algorithme FINE (voir Carter et al, (2009)) est un algorithme de réduction de dimension provenant de la géométrie de l'information. En effet les données d'un tube seront vues comme

une densité multiparamétrique qui appartient à une variété statistique.

Soit notre jeu de données patients $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n\}$ tel que $\mathbf{X}_i$ soit le tube d'échantillon du ième patient de taille Nxd (avec N le nombre de cellules $N \sim 10000$ et d le nombre de paramètres observés sur chaque cellule $d \sim 20$). La première étape de l'algorithme FINE consiste à estimer la densité multivariée de chacun des tubes patients en utilisant une estimation par noyau. On pose alors $P = \{\hat{p}_1, \hat{p}_2, ..., \hat{p}_n\}$ les densités estimées.

La seconde étape est de construire une matrice contenant les distances entre les densités estimées. Sur une variété statistique, on mesure la distance comme le chemin le plus court entre deux points de la variété appelée géodésique; celle-ci diffère de la distance Euclidienne classique. Les auteurs de FINE proposent d'utiliser une distance basée sur l'information de Fisher. Cette distance se calcule difficilement dans le cas général, une solution est d'approximer cette métrique d'information de Fisher avec une distance basée sur la divergence de Kullback-Leibler (voir Kass.R et Vos.R, (2001)). Cette distance entre deux densités p_1 et p_2 de notre variété est donnée par :

$$D_{KL}(p_1, p_2) = KL(p_1||p_2) + KL(p_2||p_1).$$

Cependant l'approximation ne fonctionne que lorsque les deux points de la variété sont proches. Ainsi pour mesurer la taille de la géodésique, les auteurs proposent de construire une chaîne de proches voisins; l'approximation de la distance basée sur l'information de Fisher devient alors :

$$\tilde{G}(p_1, p_2; P) = \min_{m,P} \sum_{i=1}^{m-1} D_{KL}(p_{(i)}, p_{(i+1)})$$

avec $p_{(i)} \to p_{(i+1)}$ formant une chaîne de plus proches voisins de P entre p_1 et p_2.

Nous pouvons alors construire une matrice de dissimilarité G tel que $G(i,j) = \tilde{G}(p_i, p_j; P)$.

La dernière étape de l'algorithme FINE consiste à réduire la dimension à partir de la matrice de dissimilarité G précédemment calculée, en utilisant une cMDS (classical Multi-Dimensional Scaling, voir Saeed et al (2018)). La cMDS nous donne un nuage de points dans un espace de dimension réduite (très souvent égale à deux).

En résumé l'algorithme FINE fonctionne ainsi :

Data: Notre jeu de données $\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n\}$ et d' la taille de l'espace réduit
Résultat: Le nuage de points $\mathbf{Y}$ de dimension réduite
Pour *i allant de 1 à n* **calculer**
  $\hat{p}_i$ la densité estimée de X_i;
fin
Calculer la matrice de dissimilarité G avec $G(i,j) = \tilde{G}(p_i, p_j; P)$
Construire l'espace $\mathbf{Y}$ de dimension Nxd' en utilisant la cMDS sur G.

Algorithme 1: Fisher information non-parametric embedding

2 Adaptations de l'algorithme FINE

On se propose désormais dans ce travail d'explorer différentes méthodes. Nous nous concentrerons particulièrement sur des alternatives à la seconde étape i.e la construction de la matrice de distance G. De plus, nous nous intéresserons à deux approches distinctes pour cette dernière : quantifier la distance comme la somme du chemin de plus proches voisins (comme dans l'algorithme FINE classique) ou calculer directement la distance entre les deux points d'intérêt. Nous évaluerons ces alternatives sur leur capacité à isoler des clusters bien différents et sur le temps de calcul.

2.1 Divergence de KL avec hypothèse de distribution gaussienne multivariée

Une première option consiste à faire l'hypothèse que nos densités sont distribuées suivant des gaussiennes multivariées. L'étape d'estimation des densités n'est plus requise car il existe alors une formule explicite de la distance de Kullback-Leibler à partir du vecteur des moyennes $\mu_i = \{\mu_i^1, \mu_i^2, .., \mu_i^d\}$ contenant les moyennes de chacun des d marqueurs du jeu de données issu de l'échantillon X_i et de la matrice de variance/covariance Σ_i correspondante.

Pour deux distributions normales multidimensionnelles $f_1 = \mathcal{N}_d(\mu_1, \Sigma_1)$ et $f_2 = \mathcal{N}_d(\mu_2, \Sigma_2)$ (voir Kullback.S et Leibler.R, (1951)) elle est donnée par :

$$I_{KL}(f_1||f_2) = \frac{1}{2}\{(tr(\Omega_2\Sigma_1) + (\mu_2 - \mu_1)^T\Omega_1(\mu_2 - \mu_1) - d - ln|\Sigma_1| + ln|\Sigma_2|\}$$

avec $\Omega = \Sigma^{-1}$.

Pour cette alternative nous utiliserons le package {rags2ridges} de R.

2.2 Divergence KL directement depuis les données

Une seconde approche consiste à estimer la divergence de Kullback-Leibler directement sur les données de cytométrie (voir Pérez-Cruz.F, (2008)) sans l'hypothèse sur la distribution. Cette méthode a pour objectif de se passer de l'estimation des densités et de calculer la divergence de Kullback-Leibler en utilisant les k-nearest-neighbour.

Pour cette approche le package {RANN} de R sera utilisé.

2.3 Distance de Wasserstein avec hypothèse gaussienne

Dans cette section nous souhaitons utiliser la distance de Wasserstein pour calculer la matrice de dissimilarité, nous calculons cette distance sous l'hypothèse gaussienne. La distance de Wasserstein est souvent utilisée pour résoudre des problèmes de transports optimaux, particulièrement pour mesurer la distance entre deux distributions de probabilités.

Pour deux distributions normales multidimensionnelles $f_1 = \mathcal{N}_d(\mu_1, \Sigma_1)$ et $f_2 = \mathcal{N}_d(\mu_2, \Sigma_2)$ (voir Perterson.A et al, (2016) Dowson.D et al, (1982)) la formule de la distance de Wasserstein est :

$$W_2(f_1, f_2)^2 = ||\mu_1 - \mu_2||_2^2 + Trace((\Sigma_1 + \Sigma_2) - 2 \times (\Sigma_2^{1/2}\Sigma_1\Sigma_2^{1/2})^{1/2}).$$

Pour calculer la distance de Wasserstein nous utiliserons le package {dad} de R.

2.4 Chemins

Pour chaque alternative présentée ci-dessus, nous allons tester deux approches pour quantifier la distance entre deux densités. Dans un premier temps nous calculerons la distance telle qu'elle est présentée dans l'algorithme FINE classique, i.e en construisant un chemin de plus proches voisins puis en sommant ces distances. Ensuite nous mesurerons la distance directement entre nos densités, i.e sans construire un chemin de plus proches voisins.

3 Application aux données

Les données utilisées ici proviennent d'une étude effectuée sur des patients potentiellement atteints de la TIH. La TIH (pour Thrombopénie Induite par Héparine) est une pathologie rare et potentiellement mortelle qui survient lors de l'utilisation de l'héparine. Les seules variables nécessaires pour l'analyse seront : FSC (forward scatter) pour la taille de la cellule analysée, SSC (side scatter) pour la granulosité de la cellule, la variable correspondant au marqueur CD-41 permettant de détecter si la cellule analysée est une plaquette ou non, et le marqueur CD-62 mesurant l'activation de la plaquette au contact de l'héparine. L'analyse se fait principalement sur ces quatre variables sur un échantillon biologique composé de plaquettes sanguines d'un patient. Dans ce cas l'hypothèse gaussienne que nous ferons plus tard est naturellement induite par la distribution des variables observées.

Notre jeu de données est constitué de 48 tubes : 12 tubes contrôles positifs à la TIH, 12 tubes contrôles négatifs à la TIH, 12 tubes patients sains (appelés QCB) et 12 tubes patients qui sont à priori atteints de TIH (appelés QCM). Nous chercherons donc au travers de ce travail à séparer ces groupes en clusters bien définis.

Nous pouvons constater à la figure 1, que globalement les projections en dimension réduite de nos tubes se rapprochent du résultat attendu sur l'ensemble des méthodes. Les différences sont constatées principalement dans un premier temps pour les clusters « NEG » et « QCB » qui peuvent être confondus au vu de leur caractère négatif à la TIH. Par ailleurs on observe que le temps de calcul varie entre les approches, en particulier, il augmente considérablement lors de la recherche du chemin de plus proches voisins. Des résultats plus détaillés seront donnés dans la présentation.

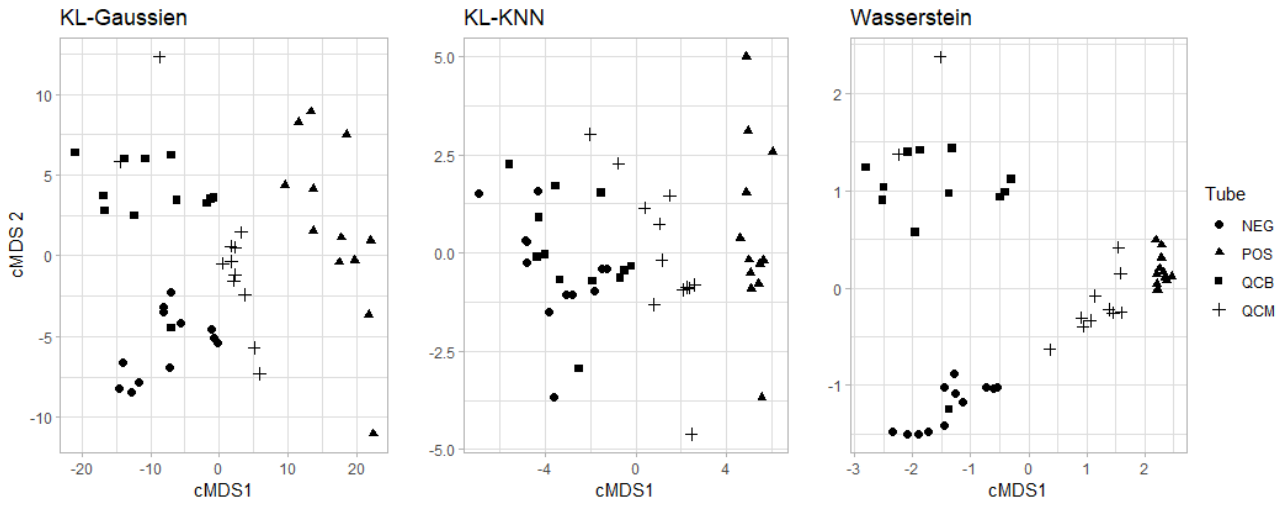

Figure 1: Visualisation des clusters obtenus selon les différentes approches dont les distances ont été quantifiées sans calcul de chemin. Nuage de points projeté après cMDS.

4 Conclusion

Nous avons introduit des alternatives à l'algorithme FINE permettant de réduire la dimension sur des objets multidimensionnels, ici un jeu de tubes analysés par cytométrie en flux contenant des contrôles et sujets sains ou non. Dans la communication nous présenterons l'ensemble des résultats ainsi que les temps de calcul correspondants. Généralement les résultats restent très satisfaisants pour la séparation en différents clusters. À l'avenir on pourra se pencher sur d'autres méthodes de réduction de dimensions (étape 3 de FINE) telles que d'autres versions du Multi-Dimensionnal Scaling ou ISOMAP. Un travail prospectif sera aussi de s'intéresser à une façon d'ajouter de nouveaux tubes patients afin de les placer à posteriori sur le nuage de points déjà construit, dans un objectif diagnostic de la TIH.

Bibliographie

Pérez-Cruz, F., (2008), July. Kullback-Leibler divergence estimation of continuous distributions. In *2008 IEEE International Symposium on Information Theory* 1666-1670.

Kullback, S., Leibler, R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1), 79-86.

Carter, K.M., Raich, R., Finn, W.G. and Hero III, A.O., (2009). Fine: Fisher information nonparametric embedding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11), 2093-2098.

Saeed, N., Nam, H., Haq, M. I. U., Muhammad Saqib, D. B. (2018). A survey on Multidimensional Scaling. *ACM Computing Surveys (CSUR)*, 51(3), 1-25.

Kass, R. E., Vos, P. W. (2011). *Geometrical Foundations of Asymptotic Inference. John Wiley Sons.*

Peterson, A., Mueller, H.G (2016). Functional Data Analysis for Density Functions by Transformation to a Hilbert Space. *The Annals of Statistics*, 44 (1), 183-218.

Dowson, D.C., Ladau, B.V. (1982). The Fréchet Distance between Multivariate Normal Distributions. *Journal of Multivariate Analysis*, 12, 450-455.