Fission de données pour l'inférence post-classification : de la théorie à la pratique

Benjamin Hivert[1,2], Denis Agniel[3], Rodolphe Thiébaut[1,2,4] & Boris Hejblum[1,2]

[1] *Univ. Bordeaux, INSERM, INRIA, SISTM team, Bordeaux Population Health, U1219, F-33000 Bordeaux, France*
[2] *Vaccine Research Institute, VRI, Hôpital Henri Mondor, Créteil F-94000, France*
[3] *Rand Corporation, Santa Monica, CA 90401, USA*
[4] *CHU Pellegrin, Groupe Hospitalier Pellegrin, Bordeaux F-33076, France*

benjamin.hivert@u-bordeaux.fr, dagniel@rand.org,
rodolphe.thiebaut@u-bordeaux.fr, boris.hejblum@u-bordeaux.fr

Résumé. Dans divers domaines, tels qu'en génomique, la classification non supervisée pose des défis en raison de son utilisation pour formuler des hypothèses testées sur les mêmes ensembles de données. Cette pratique, appelée inférence post-classification, compromet les propriétés statistiques des tests, en particulier le contrôle de l'erreur de Type I. La fission de données (Leiner et al., 2023) permet d'obtenir deux jeux de données indépendants à partir d'un seul échantillon, en bruitant l'information contenue dans chaque observation en deux parties de manière précise. Ainsi, chaque partie est utilisable indépendamment (pour la classification non supervisée d'une part, et pour les tests d'hypothèses qui suivent d'autre part) sans impacter les propriétés habituelles des méthodes statistiques. Cependant, ses limitations, notamment en termes d'hypothèses distributionnelles et d'adaptabilité à des situations avec de véritables classes inconnues, restreignent son utilisation. L'application rigoureuse de la fission de données pour l'inférence post-classification exige une connaissance préalable des véritables classes et des variances intra-classes associées. Or ces informations sont inconnues en pratique et doivent alors être estimées. Nous démontrons que l'indépendance promise par la fission de données n'est garantie qu'à condition de posséder des estimations non-biaisés des variances, et que par conséquent, elle ne peut théoriquement assurer le contrôle de l'erreur de Type I des tests associés. Nous proposons une nouvelle approche consistant à modéliser chaque observation comme une réalisation d'un processus distinct, avec des paramètres individuels, que nous estimons alors de manière non-paramétrique. Les performances de cette nouvelle approche ont été évaluées au moyen de simulations numérique, révélant l'absolue nécessité d'une très bonne séparation entre les classes afin de garantir des estimations non-biaisées des variances locales, et donc le contrôle effectif de l'erreur de Type I associée. En conclusion, bien que la fission de données ait été initialement envisagée comme une solution aux problèmes d'inférence post-classification, sa mise en pratique est rendue extrêmement délicate par le lien entre l'estimation de la structure des vraies classes et celle de leur variance. Le bon comportement de cette approche pour l'inférence post-classification nécessite indirectement de connaître les vraies classes – cependant inconnues – que l'on cherche aussi à estimer. Notre nouvelle approche de modélisation résout cette difficulté dans certains cas favorables, mais elle souffre des difficultés inhérentes à l'estimation non paramétrique de la variance locale.

Mots-clés. Inférence post-classification, Fission de données, Estimation non-paramétrique, Variance locale

Abstract. In various fields, such as genomics, clustering poses challenges due to its use in formulating hypotheses tested on the same datasets. This practice, known as post-clustering inference, compromises the statistical properties of tests, particularly the Type I error control. To address this, data fission provides an innovative approach by decomposing the information contained in each observation into two parts, independently usable for clustering and subsequent hypothesis testing. However, its limitations, especially in terms of distributional assumptions and adaptability to situations with true unknown classes, restrict its application. In this context, the rigorous application of data fission requires prior knowledge of the true classes and associated intra-class variances. However, in real applications, this information is unknown and must be estimated from the data. We demonstrate that the independence guaranteed by data fission is only valid for unbiased estimators of variance. Therefore, it cannot theoretically ensure control of the Type I error of associated tests due to the complexity of unbiased estimation of unknwow intra-class variances required for its application. Facing these challenges, we propose an alternative approach, modeling each observation as a realization of its own distribution with specific parameters estimated in a non-parametric manner. The performance of this new approach was evaluated through simulations, revealing the need for a clear separation between classes to ensure unbiased estimations and thus effective control of Type I error. In conclusion, although data fission was initially proposed as a solution to post-clustering inference problems, its applicability is compromised by the need to know the true classes. Indeed, the proper behavior of this approach for post-clustering inference depends directly on the true unknown classes to be estimated. The new modeling approach we propose allows overcoming this need for prior knowledge but presents challenges related to non-parametric variance estimation.

Keywords. Post-clustering inference, Data fission, Non-parametric estimation, Local variance

1 Introduction

Lors d'analyses exploratoires, on applique couramment des méthodes de classification non supervisée pour identifier la structure des données en regroupant les observations en sous-groupes (appelés « classes ») homogènes et séparés. Ces classes peuvent ensuite être utilisées pour générer des hypothèses, qui seront testées par des tests statistiques. Par exemple, il est courant en transcriptomique d'effectuer des tests univariés pour identifier les variables (i.e. les gènes) dont l'expression est significativement associée à une classe, dans le but de décrire et d'interpréter cette dernière. Cependant, étant donné que les classes sont obtenues à partir des mêmes données que celles utilisées pour les tests, ces analyses en deux étapes ne respectent plus le cadre théorique des tests statistiques, où les hypothèses de tests doivent être posées avant les analyses et ne peuvent dépendre des données. Ce problème connu sous le nom de « double-dipping »(Kriegeskorte et al., 2009) en anglais, compromet les bonnes

propriétés statistiques des tests utilisés. En particulier dans le cas de tests post-classification, on observe alors un mauvais contrôle de l'erreur de Type I (Gao et al., 2022). Récemment, des tests d'hypothèses spécialement conçus pour résoudre les problèmes d'inférence post-classification ont été proposés (Zhang et al., 2019; Chen and Gao, 2023; Hivert et al., 2024). Ils assurent un contrôle effectif de l'erreur de Type I dans ce contexte de double utilisation des données, en se basant notamment sur les concepts d'inférences sélectives. Cependant, leur application pratique est entravée par leur temps de calcul conséquent, et leur restriction à des méthodes de classification non-supervisée particulières, ou la distribution et la dimension des données.

Leiner et al. (2023) ont récemment introduit une approche attractive permettant l'utilisation répétée d'observations issues d'un même échantillon : la fission de données. Comparable aux divisions classiques en échantillons respectivement d'apprentissage et de test fréquemment utilisées en apprentissage automatique, cette méthode vise à décomposer l'information contenue dans chaque observation en deux parties indépendantes. Il est possible d'utiliser la première pour construire les classes, et la seconde pour tester des différences entre elles. Cependant, cette méthode repose sur des hypothèses distributionnelles fortes, et seules deux distributions (la distribution de Poisson et la distribution gaussienne) bénéficient d'une procédure de fission directement exploitable dans le contexte de l'inférence post-classification. Plus récemment, Neufeld et al. (2023) ont élargi la versatilité des distributions pouvant être décomposées en généralisant la fission de données grâce au « data thinnig ». Bien que théoriquement la fission de données et sa généralisation aient été proposées pour répondre aux défis du double-dipping dans le cadre de la classification non-supervisée, nous démontrons ici que ces méthodes sont extrêmement difficiles à appliquer en pratique. Elles reposent sur des hypothèses distributionnelles exigeant une absence de classe, et adapter ces méthodes aux situations impliquant de véritables classes inconnues requiert l'estimation d'hyperparamètres spécifiques à chacune de ces classes inconnues. Nous mettons alors en lumière l'impact d'une estimation biaisée de ces hyperparamètres sur l'erreur de Type I des tests suivant la classification non-supervisée, malgré l'utilisation de la fission de données. Nous proposons également une stratégie d'estimation non-paramétrique de la variance locale pour le cas gaussien, offrant ainsi la possibilité de s'affranchir de la nécessité de connaître les vraies classes dans certains cas favorables.

2 Méthode

2.1 Décompositions en variables aléatoires indépendantes

Soit X une variable aléatoire. La fission de données proposée par Leiner et al. (2023) a pour objectif de décomposer X en deux nouvelles variables aléatoires $X^{(1)}$ et $X^{(2)}$ grâce à l'ajout de deux bruits paramétriques précisément reliés. Ces deux transformations contiennent chacune de l'information sur X, mais incomplète. De plus, la répartition de la quantité d'information issue de X entre $X^{(1)}$ ou $X^{(2)}$ dépend d'un paramètre τ. Formellement, la fission de données décompose X en deux parties $X^{(1)}$ et $X^{(2)}$ telles que soit $\boldsymbol{\mathcal{P}_1}$) : $X^{(1)}$ et $X^{(2)}$ sont

indépendantes de distributions connues ; soit $\boldsymbol{\mathcal{P}_2}$) : $X^{(1)}$ a une distribution (marginale) connue et $X^{(2)}|X^{(1)}$ a une distribution (conditionnelle) connue soient vérifiées. Dans notre cadre de l'inférence post-classification, la propriété $\boldsymbol{\mathcal{P}_1}$ est requise pour garantir l'indépendance entre la classification servant à construire les classes (les hypothèses de tests) sur les réalisations de $X^{(1)}$ et les tests statistiques entre ces classes appliqués sur les réalisations de $X^{(2)}$. Seules la loi Poisson et la loi normale admettent une décomposition vérifiant $\boldsymbol{\mathcal{P}_1}$. La méthode du *data thinning* (Neufeld et al., 2023) généralise la fission de données en proposant une méthode de décomposition garantissant $\boldsymbol{\mathcal{P}_1}$ pour une famille de distributions de probabilité plus large (incluant notamment la loi Poisson, la loi normale et la loi binomiale négative). Nous allons ici d'avantage nous intéresser à la fission de données dans le cas gaussien, mais les résultats présentés plus bas sont généralisables au *data thinning* d'autres distributions de probabilité.

Soit $\boldsymbol{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ une variable aléatoire gaussienne à p dimensions. Alors, la fission de $\boldsymbol{X}$ est donnée, pour $\tau \in]0, +\infty)$, par :

$$\boldsymbol{X}^{(1)} = \boldsymbol{X} + \tau \boldsymbol{Z} \qquad \text{et} \qquad \boldsymbol{X}^{(2)} = \boldsymbol{X} - \frac{1}{\tau}\boldsymbol{Z} \qquad \text{pour } \boldsymbol{Z} \sim \mathcal{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma}) \tag{1}$$

$\boldsymbol{X}^{(1)}$ et $\boldsymbol{X}^{(2)}$ ainsi construites respectent la propriété $\boldsymbol{\mathcal{P}_1}$ (Leiner et al., 2023). Il en découle que $\boldsymbol{X}^{(1)} \sim \mathcal{N}_p\left(\boldsymbol{\mu}, (1+\tau^2)\boldsymbol{\Sigma}\right)$ et $\boldsymbol{X}^{(2)} \sim \mathcal{N}_p\left(\boldsymbol{\mu}, (1+\frac{1}{\tau^2})\boldsymbol{\Sigma}\right)$. Ainsi, la fission de données dans le cas gaussien se traduit par la construction de deux nouvelles variables aléatoires, dont la variance est augmentée par rapport à la variable aléatoire originale $\boldsymbol{X}$.

2.2 Classes & variances inconnues : des limitations pratiques de la fission de données

La décomposition formulée dans l'équation (1) révèle deux limitations majeures de la fission de données, mettant en lumière la complexité de la mise en oeuvre pratique de ces approches. Tout d'abord, cette procédure de fission de données est uniquement applicable aux réalisations d'une variable aléatoire gaussienne. Cependant, cette approche de modélisation se heurte à la réalité de la présence de vraies classes inconnues dans les données. En effet, ces données sont plus généralement modélisées comme des réalisations de mélanges de gaussiennes, où chaque composante (elle-même gaussienne) représente sa propre classe. Ainsi, la fission de données, telle qu'elle a été initialement introduite, ne peut être appliquée que dans des situations où l'on présume une unique classe homogène. Dans un contexte de mélange gaussien, elle ne peut uniquement être appliquée qu'à chacune des composantes du mélange de manière indépendante. Or cela n'est faisable que si les composantes du mélange, et donc les classes, sont connues. C'est cette condition préalable de la connaissance des classes qui représente une limitation significative en pratique : la classification non-supervisée visant à identifier ces classes inconnues demeure l'une des motivations principales de l'application de la fission de données. De plus, pour appliquer la fission de données, il est nécessaire de connaître la matrice de covariance des observations $\boldsymbol{\Sigma}$. Étant inconnue, il est possible de l'estimer par la matrice de covariance empirique $\hat{\boldsymbol{\Sigma}}$. Cependant, pour appliquer la fission de données à chaque composante d'un mélange, il est alors nécessaire de considérer les matrices de covariances intra-composantes, qui sont impossibles à estimer sans la connaître la vraie structure en

classes. De surcoît, l'ensemble des propriétés théoriques de la fission de données nécessite de connaître la vraie matrice de variance $\boldsymbol{\Sigma}$, et en particulier pour l'indépendance entre $\boldsymbol{X}^{(1)}$ et $\boldsymbol{X}^{(2)}$. On peut montrer que considérer une estimation de $\boldsymbol{\Sigma}$ par $\hat{\boldsymbol{\Sigma}}$ résulte en une covariance entre ces deux nouvelles variables aléatoires donnée par : $\text{Cov}\left(\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}\right) = \boldsymbol{\Sigma} - \hat{\boldsymbol{\Sigma}}$.

2.3 Modélisation individualisée pour une fission de données généralisée & Estimation non-paramétrique de la variance

Afin de généraliser la fission de données quelque soit le nombre de classe dans les données, nous proposons de modéliser chaque observation comme étant une réalisation de sa propre distribution avec ses propres paramètres individuels, c'est-à-dire :

$$\boldsymbol{X}_i \sim \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i). \tag{2}$$

Cette approche consiste donc à inclure des moyennes et des variances qui ne sont pas spécifiques aux classes (inconnues), mais plutôt spécifiques à chaque observation. La connaissance de la structure du mélange n'est alors plus nécessaire pour l'estimation de la variance. Seule la proximité entre individus importe, puisque cette modélisation suppose que deux individus provenant d'une même composante présentent des valeurs similaires pour ces paramètres.

Cependant, l'estimation de la variance $\boldsymbol{\Sigma}_i$ devient encore plus délicate. Les estimateurs classiques tels que la variance empirique ne sont pas adaptés. Nous proposons d'estimer ces variance de façon non-paramétrique en pondérant les observations dans leurs calculs à l'aide d'un noyau configuré de manière à ce qu'un individu proche de l'individu i (et donc vraisemblablement issu de la même composante du mélange) ait un poids important mais que les individus très éloignés de l'individu i se voient attribuer un poids insignifiant.

3 Résultats

3.1 Impact des Biais dans l'estimation de la variance sur la classification non-supervisée et l'erreur de Type I

Tout d'abord, il est crucial de noter qu'un biais dans l'estimation de la variance induit directement une covariance non-nulle entre $\boldsymbol{X}^{(1)}$ et $\boldsymbol{X}^{(2)}$, et donc une perte de l'indépendance. Le contrôle de l'erreur de Type I pour les tests sur $\boldsymbol{X}^{(2)}$ entre les classes estimées sur $\boldsymbol{X}^{(1)}$ n'est plus garanti. Nous établissons précisément la relation entre ce biais d'estimation sur la variance et l'erreur de Type I résultante des analyses dans le contexte gaussien univarié. Si l'on considère n réalisations d'une variable aléatoire gaussienne $X \sim \mathcal{N}(\mu, \sigma^2)$, l'application de la fission de données sur ces réalisations de X en utilisant un estimateur $\widehat{\sigma^2}$ de σ^2 pour une classification en deux classes entraîne une déviation de la statistique du t-test de Student

sous $\mathcal{H}_0$ égale à :

$$\frac{\sqrt{n}\sqrt{\frac{2}{\pi}\operatorname{Cor}\left(X^{(1)}, X^{(2)}\right)^2}}{\sqrt{1-\frac{2}{\pi}\operatorname{Cor}\left(X^{(1)}, X^{(2)}\right)^2}}. \tag{3}$$

Ce résultat démontre que seul un biais très faible dans l'estimation de $\widehat{\sigma^2}$ est autorisé pour garantir un contrôle de l'erreur de Type I. La Figure 1 **A**, représentant l'évolution de l'erreur de Type I en fonction du biais relatif sur l'estimation de σ^2, illustre la concordance de ce résultat théorique sur une étude de simulations et souligne ainsi la nécessité d'une méthode d'estimation précise et très efficace de la variance locale.

3.2 Performances de l'estimation non-paramétrique de la variance et de son application dans la fission de données

Nous avons réalisé une étude de simulations afin d'évaluer deux aspects cruciaux portant sur notre estimateur de la variance : *i*) ses performances directes en termes de qualité de l'estimation et *ii*) son application dans la fission de données pour des problèmes d'inférence post-classification. Pour ce faire, nous avons généré $n = 100$ réalisations d'un modèle de mélange de gaussiennes univarié à deux composantes ayant des variances égales. Nous avons exploré différentes valeurs de la différence de moyenne δ entre les deux composantes, allant de $\delta = 0$ (mélange à une seule classe) à $\delta = 100$, ainsi que diverses valeurs de la variance intra-composantes partagée σ^2. La fission de données a été appliquée à chaque observation en utilisant son estimation non-paramétrique de la variance locale associée, comme décrit en section 2.3. Par la suite, $X^{(1)}$ a été utilisé pour une classification en 3 classes via l'algorithme des k-means, induisant ainsi une composante faussement séparée en deux classes. Ensuite, une différence de moyenne entre ces deux classes artificielles a été testée à l'aide du t-test de Student sur $X^{(2)}$.

La Figure 1 **B** représente l'évolution du biais relatif sur l'estimation de σ^2 en fonction de la séparabilité entre les deux composantes du mélange, décrite par le ratio $\delta \,/\, \sigma$. Il est clair d'après cette figure qu'une bonne séparabilité entre les deux classes (une grande valeur du ratio $\delta \,/\, \sigma$) est nécessaire pour que le biais d'estimation soit suffisament faible. La Figure 1 **C** illustre les performances de notre approche en termes de contrôle de l'erreur de Type I avec la fission de données appliquée à l'inférence post-classification. Comme attendu, l'absence de biais dans l'estimation de la variance est cruciale pour garantir un contrôle effectif de l'erreur de Type I. Pour notre méthode d'estimation, cela se traduit donc par la nécessité d'une très bonne séparabilité entre les classes.

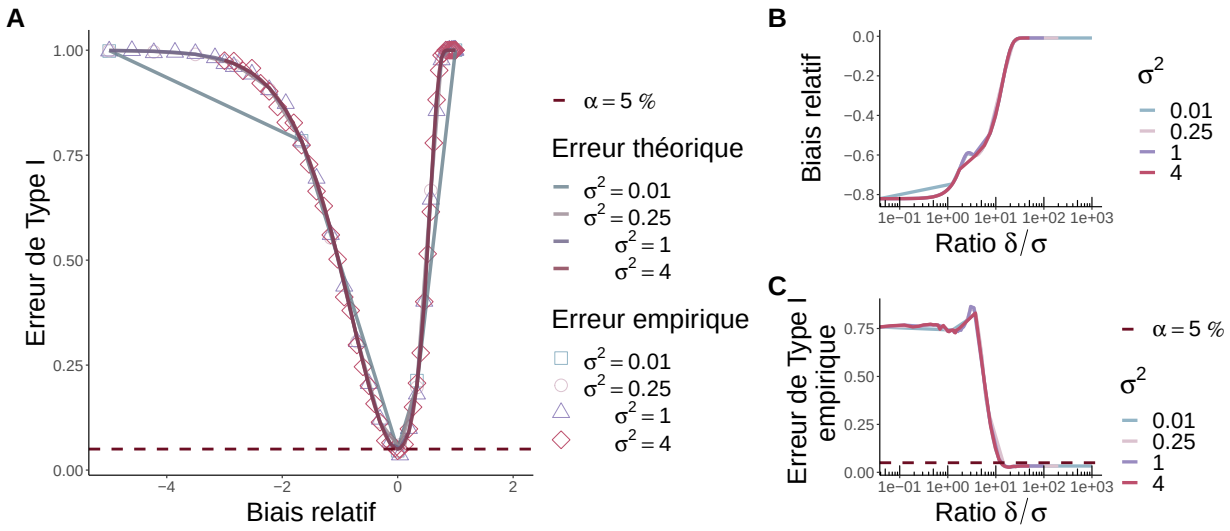

FIGURE 1 : Importance de l'estimation de la variance pour la fission de donnée. **A.** Comparaison de l'erreur de Type I théorique décrite en (3) et de l'erreur de Type I empirique associées à la fission de données pour de l'inférence post-classification en fonction du biais relatif sur l'estimation de σ^2. **B.** Évolution du biais relatif sur l'estimation de σ^2 en fonction de la séparabilité entre les classes donnée par le ratio $\delta \,/\, \sigma$. **C.** Évolution de l'erreur du Type I associée à la fission de données en fonction de la séparabilité entre les classes donnée par le ratio $\delta \,/\, \sigma$.

4 Discussion

Bien que la fission de données et sa généralisation, le *data thinning*, aient été initialement proposées comme des solutions aux problèmes d'inférence post-classification, ces méthodes se heurtent à des défis majeurs, liés aux hypothèses paramétriques qu'elles exigent. En particulier, l'absence de mélange dans les données et la connaissance préalable des hyperparamètres. Ces conditions restrictives rendent leur application extrêmement difficile en pratique, car la structure sous-jacente des données, essentielle à leur bon fonctionnement, demeure inconnue dans le contexte de la classification non-supervisée.

Nous soulignons ces limites en nous concentrant sur le cas gaussien, où l'estimation de la variance (indispensable à l'utilisation pratique de ces méthodes) devient particulièrement délicate en raison de sa dépendance aux composantes inconnues et à estimer. Nos analyses théoriques ont démontré l'impact significatif d'une mauvaise estimation de la variance sur l'inflation de l'erreur de Type I lors de l'inférence post-classification, mettant en évidence les risques inhérents à l'application aveugle de la fission de données et du *data thining*. Dans le but de surmonter ces limites, nous avons proposé une nouvelle approche reposant sur une estimation non-paramétrique de la variance locale, éliminant ainsi le besoin de connaissance préalable des composantes. Cependant, cette méthode alternative nécessite notamment une très bonne séparabilité entre les vraies composantes afin de garantir un contrôle effectif de l'erreur de Type I. À noter que, bien que nos résultats et analyses soient spécifiques au cas

gaussien, les limites discutées demeurent vraies pour d'autres distributions ayant un processus de fission ou de *data thinning* reposant sur la connaissance d'hyperparamètres. Ainsi, la nécessité de comprendre et de traiter ces limitations devient cruciale lors de l'application de telles méthodes à des données réelles.

Références

Chen, Y. T. and Gao, L. L. (2023). Testing for a difference in means of a single feature after clustering. *arXiv preprint arXiv:2311.16375*.

Gao, L. L., Bien, J., and Witten, D. (2022). Selective inference for hierarchical clustering. *Journal of the American Statistical Association*.

Hivert, B., Agniel, D., Thiébaut, R., and Hejblum, B. P. (2024). Post-clustering difference testing: valid inference and practical considerations with applications to ecological and biological data. *Computational Statistics & Data Analysis*, 107916.

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S., and Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nature neuroscience*, 12(5):535–540.

Leiner, J., Duan, B., Wasserman, L., and Ramdas, A. (2023). Data fission: splitting a single data point. *Journal of the American Statistical Association*.

Neufeld, A., Dharamshi, A., Gao, L. L., and Witten, D. (2023). Data thinning for convolution-closed distributions. *arXiv preprint arXiv:2301.07276*.

Zhang, J. M., Kamath, G. M., and David, N. T. (2019). Valid post-clustering differential analysis for single-cell RNA-Seq. *Cell systems*, 9(4):383–392.