HAL open science

Approche variationnelle pour la fusion de jeux de données d'expression génique

Marie-Christine Roubaud, Bruno Torrésani

To cite this version:

Marie-Christine Roubaud, Bruno Torrésani. Approche variationnelle pour la fusion de jeux de données d'expression génique. 41èmes Journées de Statistique, SFdS, Bordeaux, 2009, Bordeaux, France, France. inria-00386695

HAL Id: inria-00386695

https://inria.hal.science/inria-00386695

Submitted on 22 May 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Approche variationnelle pour la fusion de jeux de données d'expression génique

Marie-Christine Roubaud & Bruno Torrésani

LATP/Université de Provence 39 rue F. Joliot Curie 13453 Marseille Cedex 13

Résumé. L'analyse conjointe de multiples jeux de données de même nature pour en dégager l'information pertinente est un problème complexe. Une approche variationnelle pour calibrer des jeux de données d'expression géniques multiples est proposée, basée sur l'optimisation d'une fonctionnelle de « fonctions de rectification ». L'optimisation est effectuée numériquement par un algorithme itératif. L'approche proposée est illustrée sur une simulation, dans laquelle des jeux artificiels sont constitués à partir d'un jeu de données réel d'expression de *E. Coli*.

Mots clés : données d'expression génique, fusion de données, normalisation, approche variationnelle

Abstract. The joint analysis of multiple datasets of similar nature for extracting relevant common information is a complex problem. A variational approach for calibrating gene expression datasets is proposed, that relies on the optimization of a functional of "rectification functions". The optimization is performed numerically with an iterative algorithm. The proposed approach is illustrated on simulated data, in which articicial datasets are generated from real *E. Coli* expression data.

Key words : gene expression data, data fusion, normalization, variational approach

1 Introduction

La quantité de données d'expression génique disponibles dans de grandes bases de données (voir par exemple *Gene Omnibus* ou encore *ArrayExpress*) connaît actuellement une croissance extrêmement rapide. Pour tirer parti de ces masses de données, un préalable crucial est de s'affranchir « au mieux » de la variabilité entre études, afin d'accéder à la variabilité pertinente pour la problème posé. Schématiquement, on peut distinguer deux types d'approches, que l'on nommera *méta-analyse* et *fusion de données*.

L'approche « méta-analyse », qui fusionne des résultats d'analyse plutôt que les données elles-mêmes, est de loin la plus abordée dans la littérature (voir par exemple Rhodes *et al* (2002), Zintsaras et Ioannidis (2008), et Hong et Breitling (2008) pour une revue). Sa popularité provient notamment du fait que les données sont souvent faciles à obtenir et à traiter : on compare des quantités de même ordre de grandeur (niveau de signification,...) et la comparaison de résultats issus de différentes technologies est possible. Cependant cette approche comporte également de gros inconvénients. Elle simplifie à l'extrême l'information, mais aussi et surtout ne fait pas coopérer les différents jeux de données lors de

l'analyse. Par exemple, elle ne permet pas de résoudre le problème des petits effectifs de certains groupes dans l'analyse différentielle.

Par contre, très peu d'études basées sur la combinaison des jeux de données avant analyse ont été développées, à l'exception d'études portant sur les rangs. Les principales difficultés dans leur mise en oeuvre sont la nécessité de s'affranchir au mieux de l'effet « étude/laboratoire » (Irizarry *et al* (2005)) et la validation de la procédure de fusion. Dans cet article nous nous intéressons à ce type d'approche pour des jeux de données correspondant à une même espèce avec des conditions similaires et issus d'une même technologie. Une méthode de calibration de jeux de données avant leur fusion est proposée et des résultats sont présentées sur une simulation.

2 Fusion de jeux de données d'expression génique

Le cadre de notre étude est celui de données d'expression géniques, et notre objectif est la fusion de plusieurs jeux de données provenant d'études similaires, comportant un nombre suffisant de gènes communs. La variabilité « biologique » est supposée non confondue avec la variabilité « étude ». On se place au niveau de données déjà normalisées, au sein de chaque jeu. Notre objectif est d'effectuer une nouvelle normalisation « entre études », pour les rendre comparables avant de les fusionner et pouvoir ensuite utiliser des méthodes quantitatives sur les données réunies. Une condition nécessaire est que *l'ordre des gènes soit à peu près conservé d'une étude à l'autre.*

Une méthode simple de calibration est fournie par la transformation en données ordinales : la valeur mesurée de l'exression du gène est remplacée par son rang dans la condition considérée. Par exemple dans la librairie `Bioconducor` du langage R, le paquet `Rankprod` de Hong *et al* (2006) effectue une analyse différentielle à partir des rangs. Cette approche engendre cependant une grosse perte d'information, et soulève de nombreuses questions, comme le problème du codage : la transformation doit-elle être faite avant ou après fusion ? En effet le nombre de gènes exploitables peut être différent d'un jeu à un autre et la réunion s'effectuant sur la base des gènes communs, les rangs ne seront pas identiques s'ils sont calculés avant ou après la fusion.

2.1 Approche proposée

Notre approche est fondée sur le constat empirique de la non linéarité de la relation entre les moyennes des différents jeux de données. Notre hypothèse de travail est la suivante : *dans chaque étude, les données mesurées sont les images de niveaux d'expression « vrais » par une fonction (non-linéaire) inconnue.*

Notations et hypothèses. Considérons un ensemble $\{\mathscr{X}_1, \dots \mathscr{X}_K\}$ de jeux de données d'expression, de la forme $\mathscr{X}_k = \{X_{k;gc}\}$, $X_{k;gc}$ représentant le niveau d'expression mesuré du gène g dans la condition c, dans la k-ième étude. On note $\mathbf{X}_{k;g}$ le vecteur des expressions

de g pour toutes les conditions dans l'étude k, G le nombre de gènes communs aux différents jeux, et C_k le nombre de conditions dans l'étude k. On suppose que :

1. Les variables biologiques d'intérêt $\{\Xi_{k;gc}\}$ sont observées dans les différentes études au travers de fonctions d'observation différentes $F_1, \dots F_K$: $X_{k:g,c} = F_k(\Xi_{k;g,c})$
2. Les fonctions d'observation sont non-linéaires, monotones et régulières. Elles dépendent uniquement de l'étude, elles sont en particulier indépendantes par rapport aux gènes et aux conditions dans une même étude.

Approche variationnelle. On recherche des *fonctions de rectification* qui rendent comparables les jeux de données des différentes études. Ces fonctions jouent le rôle de « réciproques régulières » des fonctions d'observation. La méthode proposée est schématisée en Fig 1 dans le cas de trois jeux de données.

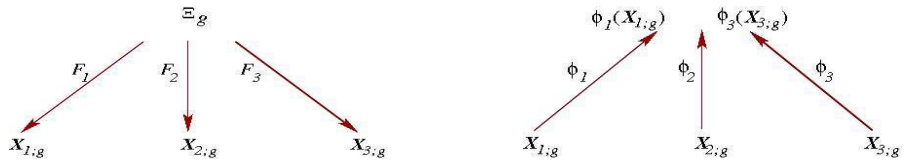

Fig. 1 – Fonctions d'observation et de rectification pour trois jeux de données d'expression

Le problème posé est le suivant : *trouver des fonctions de rectification $\phi_1, \dots \phi_K$ telles que les $\phi_k(\mathbf{X}_{k;g})$, $k = 1 \dots K$ soient les plus proches possible, sous contrainte de régularité.* Les ϕ_k ne sont pas uniques, et sont définies modulo composition avec une fonction inconnue. Pour résoudre ce problème, nous nous basons sur une approche variationnelle, permettant d'imposer de façon simple des hypothèses de régularité sur les fonctions de rectification. La fonctionnelle à minimiser, de type « spline », consiste en une attache aux données quadratique, régularisée par pénalisation des dérivées secondes des fonctions de rectification.

Le problème global est donc de minimiser, par rapport à la référence $\mathbf{M} = \{M_g, g = 1, \dots G\}$ et aux fonctions de rectification ϕ_k, $k = 1, \dots K$, la quantité

$$\Phi[\phi_1, \dots \phi_K] = \sum_{k=1}^{K} \left(\sum_g \left[\overline{\phi_k(\mathbf{X}_{k;g})} - M_g \right]^2 + \lambda \int \phi_k''(x)^2 \, dx \right) \tag{1}$$

où λ est un paramètre servant à « régler » l'importance relative des termes d'attache aux données et de régularité, et où on a noté $\overline{\phi_k(\mathbf{X}_{k;g})} = \frac{1}{C_k} \sum_{c=1}^{C_k} \phi_k(X_{k;gc})$ la moyenne des valeurs rectifiées pour le gène g dans l'étude k.

Algorithme. Nous recherchons un minimum de Φ via un algorithme itératif. Partant d'un choix initial pour la référence $\mathbf{M}$, les deux étapes suivantes sont itérées :

1. Estimation des fonctions de rectification ϕ_k, sachant $\mathbf{M}$. Fixer $\mathbf{M}$ découple le problème en K problème indépendants, qui sont des problèmes de régression « spline ».

2. Estimation de $\mathbf{M}$ sachant les ϕ_k ; ceci revient à un simple calcul de moyenne.

Application aux données d'expression. Partant d'une famille de jeux de données d'expression $\{\mathscr{X}_1, \dots \mathscr{X}_K\}$, les opérations suivantes sont effectuées :

- Estimation de la moyenne M et des fonctions de rectification $\phi_1, \dots \phi_K$.
- Rectification des données : $\tilde{X}_{k;gc} = \phi_k(X_{k;gc})$.
- Fusion des jeux rectifiés en un « grand jeu ».

Remarques : - Il n'y a pas unicité de la solution du problème d'inversion des fonctions d'observation. L'algorithme fournit une solution permettant de « recaler » les données, sans pour autant retrouver les « vrais » niveaux d'expression géniques Ξ_{gc}.

- Il n'est pas nécessaire (et parfois pas souhaitable) d'utiliser tous les gènes pour estimer les fonctions de rectification. Le sous-ensemble de gènes utilisé doit comporter des gènes *peu variables* dans chaque jeu, et *échantillonnant* l'ensemble du domaine des valeurs.

2.2 Simulations

L'algorithme a été testé sur deux jeux de données artificiel construits à partir d'un jeu de données d'expression chez *E. coli* obtenu par Covert *et al* (2006). Ce jeu correspond à l'étude de l'expression de 7295 gènes dans deux situations différentes : 20 aérobie (O) et 22 anaérobie (N). Son intérêt est qu'il exhibe une variabilité biologique claire due aux deux situations (voir le premier plan factoriel des données originales, Fig. 2).

Constitution de deux jeux de données artificiels. Nous avons généré deux jeux de données en tirant au hasard 10 (O) et 11 (N) dans le jeu de départ. Puis nous avons appliqué deux transformations non-linéaires différentes F_1 et F_2 à chacun des jeux (ici, deux variantes de l'arcsinus hyperbolique). Chaque transformation simulant la fonction d'observation d'une étude.

Une projection dans le premier plan factoriel de l'ensemble des données déformées fait apparaître clairement l'effet « étude » comme premières sources de variabilité et la variabilité biologique est totalement masquée (voir Fig. 2).

Estimation et validation : La méthode décrite ci-dessus a été appliquée à ces jeux simulés, les fonctions de rectification ϕ_1, ϕ_2 étant estimées à partir d'un sous-jeu de gènes « invariants ». Puis la rectification a été effectuée sur les jeux complets. Les résultats (Fig. 2) sont analysés par comparaison des expressions moyennes de chaque gène (groupe de 4 figures en haut à gauche), des *boxplots* (groupe de 4 figures en haut à droite) et des projections sur le premier plan factoriel d'une ACP (groupe de 4 figures en bas).

La rectification permet de supprimer la distorsion et de linéariser la dépendance entre les moyennes des expressions de gènes, comme le montre les tracés des moyennes de la Fig. 2 (représentant dans l'ordre le jeu 1 et le jeu 2 contre leur moyenne, le jeu 1 contre le jeu 2, et le jeu 1 rectifié contre le jeu 2 rectifié).

La projection sur le premier plan factoriel montre que la variabilité induite par la distorsion est réduite par la rectification, les données rectifiées retrouvant la variabilité

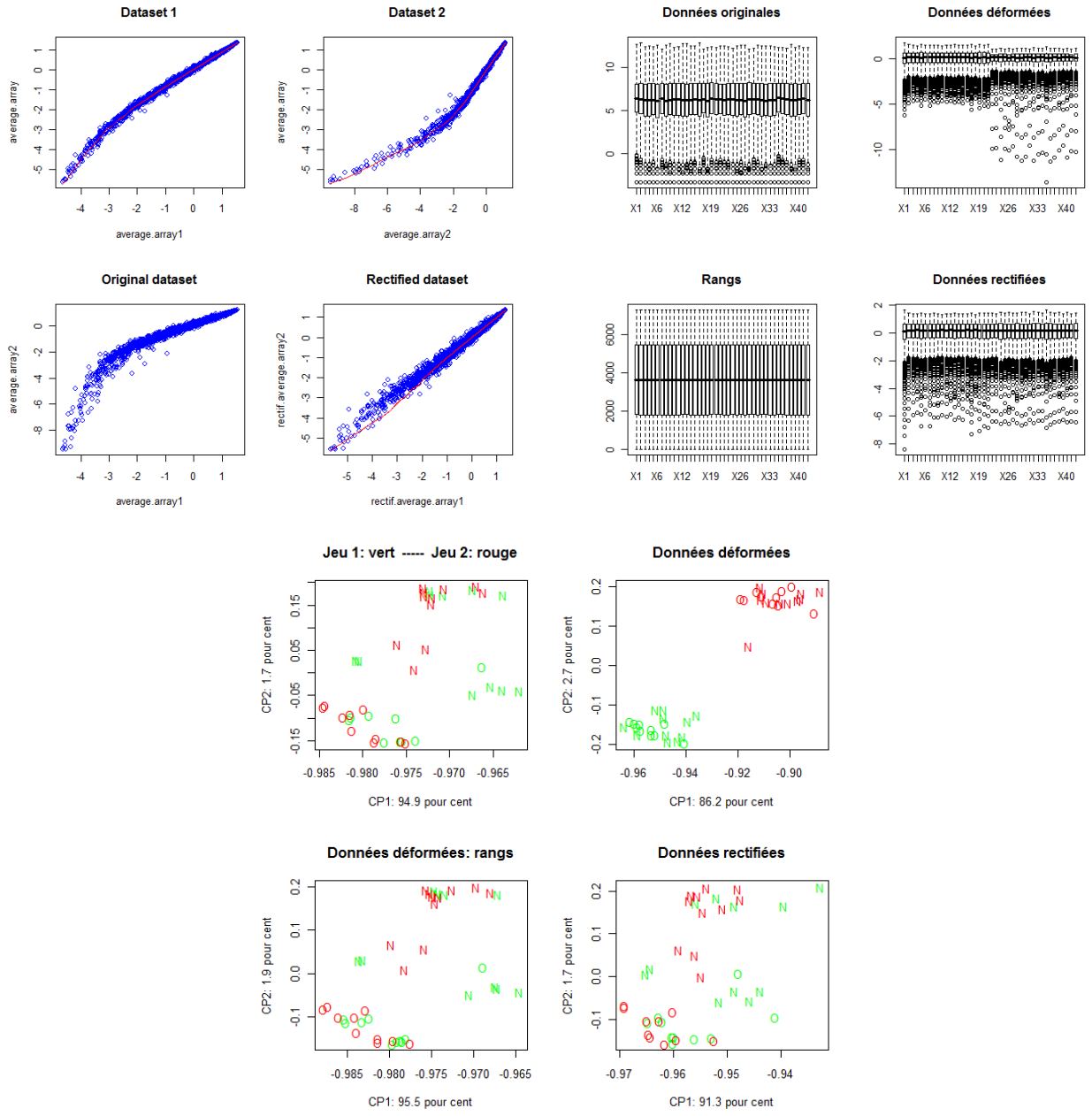

FIG. 2 – SIMULATION. Groupe en haut à gauche : expressions moyennes de chaque gène, avant rectification (3 premières courbes) et après rectification. Groupe en haut à droite : boxplots des données initiales, déformées, ordinales et rectifiées. Groupe en bas : projections sur le premier plan factoriel des données initiales, déformées, ordinales et rectifiées.

biologique de départ. Ce résultat est à peu près comparable au résultat obtenu à partir des données ordinales. La rectification évite toutefois la perte d'information inhérente à la transformation en rangs.

Les boxplots illustrent le fait que la rectification, comme prévu, ne permet pas d'inverser la transformation non-linéaire, mais atteint son but en fournissant des distributions comparables, mais pas identiques comme le fait le passage aux rangs.

3 Conclusion

Nous proposons une méthode de normalisation de jeux de données multiples, basée sur une hypothèse de relation non-linéaire entre différentes études et une approche variationnelle. Les résultats obtenus sur des jeux de données simulées sont encourageants. La rectification proposée est dans ce cas suffisante pour réduire l'effet « étude » dans les jeux fusionnés. Cependant plusieurs points restent à approfondir : les pré-traitements (log, stabilisation de la variance,...), le choix de la référence M dans l'algorithme de rectification, la sélection de gènes invariants ainsi que la prise en compte de distorsions sur la variance. La prochaine étape est l'application de cette méthode sur des jeux de données réels. *Rendre comparables plusieurs jeux de données similaires* est un véritable enjeu. Ceci permet d'envisager des études sur de très gros jeux de données avec un gain important en robustesse.

Bibliographie

[1] Covert, M. W., Knight, E.M., Reed, J.L., Herrgard, M.J. et Palsson, B.O. (2004) Integrating high-throughput and computational data elucidates bacterial networks, *Nature*, 429 (2004), 92-96.

[2] Hong, F et Breitling, R. (2008) A comparison of meta-analysis methods for detecting differentially expresed genes in microarray experiments, *Bioinformatics*, 24 :3, 374-387.

[3] Hong, F., Breitling, R., McEntee, C.W., Wittner, B.S., Nemhauser, J.L. et Chory, J. (2006) RankProd : a Bioconductor package for detecting differentially expressed genes in meta-analysis, *Bioinformatics Application Note*, 22 :22, 2825-2827.

[4] Irizarry, R. A. *et al* (2005) Multiple-laboratory comparison of microarray platforms, *Nature Methods*, 2 :6, 477.

[5] Rhodes, D. R., Barrette, T.R., Rubin, M.A., Ghosh, D. et Chinnaiyan, A.M. (2002) Meta-analysis of microarrays : interstudy validation of gene expression profiles reveals pathway disregulation in prostate cancer, *Cancer Research*, 62, 4427-4433.

[6] Zintsaras, E. et Ioannidis, J. P. A. (2008) Meta-analysis for ranked discovery datasets : Theoretical framework and empirical demonstration for microarrays, *Comp. Biol. and Chem.*, 32, 39-47.