Calcul des rangs moyens associés à des individus dans un tableau de données

Doulaye Dembélé

Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC),
CNRS UMR 7104, INSERM U964, Université de Strasbourg,
67404, Illkirch-Graffenstaden, France

Résumé

Dans des applications de la vie courante, des scores sont attribués à des individus en fonction de leurs performances pour des activités. Les résultats de ces expériences sont organisés dans des tableaux où les individus apparaissent dans les lignes et les activités (variables) dans les colonnes. Les individus peuvent être classés (rangs) pour chaque variable en fonction des scores obtenus. On s'intéresse ici au classement moyen de chaque individu afin d'identifier les plus/moins performants. Le classement moyen pour chaque individu est obtenu par la moyenne des rangs de ses différents classements. Cela peut aussi se faire à l'aide d'un calcul matriciel. En utilisant quelques propriétés du calcul de la moyenne, nous réorganisons les rangs de chaque individu pour obtenir un résultat plus maniable. Nous montrons l'avantage du calcul matriciel dans le cas des grandes masses de données avec la présence de valeurs aberrantes. Une applications biomédicale est utilisée en exemple.

1 Introduction

Il existe beaucoup d'applications de la vie courante dans lesquelles on associe des scores à des individus relativement à des variables. Par exemple dans le tableau 1, nous considérons une classe de vingt élèves et leurs rangs en fonction des notes obtenues dans des matières d'apprentissage : espagnol (sp), français (fr), anglais (en), allemand (de) et chinois (cn). La dernière colonne indique le rang moyen pour chaque élève. Nous pouvons aussi considéer des scores (hauteurs) pour différentes plantes relativement à l'utilisation de plusieurs engrais ou des scores de recherche via Internet pour différents mots relativement à des critères définis. Pour toutes ces données, on souhaite connaitre le classement moyen pour chaque individu (élève, plante, mot ...). Cela est obtenu avec une moyenne des rangs. Au lieu d'utiliser ce calcul élémentaire, nous pouvons utiliser un calcul matriciel. *Pourquoi faire simple si on peut faire compliquer ?* Nous utilisons la décomposition en valeur singulière (SVD = singular value decomposition) de la matrice des rangs [2, 3, 4, 6]. Puisque toutes les données (rangs) sont positives, nous pouvons obtenir une bonne approximation de la moyenne à partir de la première composante singulière en raison du théorène de Perron [4]. Remarquons que le calcul de la moyenne pour chaque individu est indépendant de celui des autres. De plus, l'ordre des rangs d'un individu n'a pas d'influence sur la moyenne finale. Ces constations nous conduisent à introduire une idée nouvelle qui consiste à trier les rangs par individu avant de faire la SVD. Dans le cas de données de grandes dimensions, le tri offre la possibilité d'écarter aisémnent des valeurs pour chaque individu avant le calcul du rang moyen.

2 Méthode and résultats

À partir des donnés du tableau 1 nous montrons la démarche adoptée, puis nous montrons comment la méthode proposée se généralise. Notons par X la matrice contenant les rangs de n individus pour m variables (pour le tableau 1, colonnes 3 à 7, n=20 et m=5). La SVD de la matrice

TABLE 1 – Données de rangs des élèves

idx	nom	sp	fr	en	de	cn	moy
01	Lalla	16	16	16	17	17	16.4
02	Issa	14	13	15	12	11	13.0
03	Saly	13	15	17	18	18	16.2
04	Ali	19	20	20	19	20	19.6
05	Fatou	3	4	3	2	3	3.0
06	Bocar	8	6	5	5	5	5.8
07	Assa	11	12	11	11	12	11.4
08	Baba	5	8	4	4	4	5.0
09	Sitan	10	9	6	8	6	7.8
10	Binta	17	18	7	10	16	13.6
11	Adam	2	1	1	3	2	1.8
12	Awa	1	2	2	1	1	1.4
13	Aliou	12	14	14	14	14	13.6
14	Haby	18	17	19	16	15	17.0
15	Sekou	7	11	10	15	10	10.6
16	Nene	15	10	13	13	13	12.8
17	Aiche	4	5	8	6	7	6.0
18	Sira	9	7	9	7	9	8.2
19	Mody	20	19	18	20	19	19.2
20	Fanta	6	3	12	9	8	7.6

X s'écrit :

$$X = UDV^T \tag{1}$$

où U et V sont des matrices carrées orthogonales de tailles n et m, respectivement, et vérifiant $U^TU = I_n$, $V^TV = I_m$. U^T désigne la transposée de la matrice U, I_m est une matrice carrée de taille m dont les composantes sont nulles sauf les éléments de la diagonale qui sont tous égaux à 1. D est une matrice de taille (n, m) dont les composantes sont nulles exceptées les m premières valeurs de la diagonale. Les composantes de la diagonales de D sont appélées valeurs singulières et apparaisssent par ordre décroissant $d_1 \geq d_2 \geq \ldots \geq d_m \geq 0$.

En raison des propriétés des matrices U et V, la matrice X peut s'écrire comme une somme de matrices de rang un :

$$X = \sum_{j=1}^{m} d_j \mathbf{u}_j \mathbf{v}_j^T \tag{2}$$

où $\mathbf{u}_j$ est le vecteur formé par les éléments de la colonne j de la matrice U, $\mathbf{v}_j$ est défini de façon similaire. Une approximation de la matrice X, notée $\hat{X}$, est obtenue en utilisant un nombre de termes q inférieur à m dans la relation (2). Ainsi, nous pouvons écrire :

$$\hat{X} = \sum_{j=1}^{q} d_j \begin{bmatrix} u_{1j}v_{1j} & u_{1j}v_{2j} & \ldots & u_{1j}v_{mj} \\ u_{2j}v_{1j} & u_{2j}v_{2j} & \ldots & u_{2j}v_{mj} \\ \vdots & \vdots & \ddots & \vdots \\ u_{nj}v_{1j} & u_{nj}v_{2j} & \ldots & u_{nj}v_{mj} \end{bmatrix} \tag{3}$$

La valeur moyenne de la ligne i de $\hat{X}$ est :

$$\bar{\hat{x}}_i = \sum_{j=1}^{q} d_j u_{ij} \left(\frac{1}{m} \sum_{k=1}^{m} v_{kj} \right) = \sum_{j=1}^{q} d_j u_{ij} \mathcal{V}_j \tag{4}$$

où l'expression de $\mathcal{V}_j$ se déduit facilement, c'est la valeur moyenne de la colonne j de la matrice V. $\bar{\hat{x}}_i$ est une approximation de la moyenne exacte $\bar{x}_i$ obtenue si $q = m$.

TABLE 2 – Rangs moyens estimés des élèves

		observés		triés	
idx	nom	q=1	q=2	q=1	q=2
01	Lalla	16.401	16.401	16.276	16.400
02	Issa	12.995	12.995	13.010	13.000
03	Saly	16.203	16.204	16.225	16.202
04	Ali	19.599	19.599	19.447	19.598
05	Fatou	2.999	2.999	3.027	2.999
06	Bocar	5.799	5.798	5.835	5.799
07	Assa	11.401	11.401	11.326	11.401
08	Baba	4.998	4.998	5.069	4.998
09	Sitan	7.799	7.798	7.875	7.799
10	Binta	13.606	13.603	13.880	13.600
11	Adam	1.802	1.802	1.851	1.798
12	Awa	1.398	1.399	1.426	1.402
13	Aliou	13.601	13.601	13.525	13.601
14	Haby	16.995	16.995	16.970	16.999
15	Sekou	10.602	10.603	10.731	10.599
16	Nene	12.800	12.801	12.816	12.799
17	Aiche	5.999	6.001	6.080	6.001
18	Sira	8.199	8.199	8.204	8.198
19	Mody	19.201	19.201	19.079	19.201
20	Fanta	7.598	7.600	7.821	7.600

En utilisant q=1 puis q=2, nous avons utilisé l'expression (4) pour estimer les moyennes des rangs des données du tableau 1 en considérant 2 situations : a) utilisation des rangs observés, et b) les rangs de chaque élève sont triés. Les résultats sont montrés dans le tableau 2. Nous observons une bonne approximation pour q=1, elle est encore meilleure avec q=2. Pour les 2 situations, l'utilisation de toutes les composantes (q=m) conduit aux valeurs moyennes exactes, i.e. les valeurs de la dernière colonne du tableau 1. Pour évaluer l'erreur qui résulte du calcul des

moyennes en utilisant la relation (4) avec q=1 ou q=2, nous avons calculé la moyenne des carrés des écarts entre les valeurs exactes et leurs estimés (MSE = mean squared error) :

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\bar{x}_i - \bar{\hat{x}}_i)^2 \qquad (5)$$

Avec q=1, nous avons un plus faible valeur MSE pour les rangs observés (5.43E-6) qu'avec les rangs triés (1.16E-2). La situation est inverse avec q=2, MSE=4.77E-6 et 1.29E-6 pour les rangs observés et triés respectivement. Nous ne pouvons pas conclure que cela sera le cas pour toutes données.

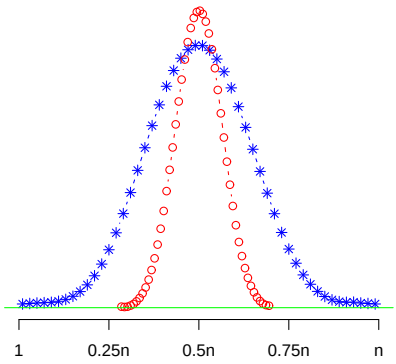

FIGURE 1 – Histogramme des rangs moyens

Les relations (1) - (4) sont indépendantes des données de l'exemple académique du tableau 1. Dans la pratique, n et m peuvent être de l'ordre de milliers. Pour des raisons techniques, il est possible d'observer des scores absents pour des individus relativement à des variables. Un individu peut aussi se voir attribuer un score douteux (aberrant) comparé à ses autres scores. Pour n élevé, nous pouvons tracer l'histogramme de rangs moyens comme montré dans la figure 1. Si certains individus des données se classement fréquemment en début ou fin pour les variables, nous obtenons le tracé en bleu ($\star-$). Dans les cas ou chaque individu peut se classer en début et fin (rangs aléatoires) nous obtenons le tracé rouge ($o-$). Pour les données avec un grand nombre de variables incluant des rangs aberrants la courbe de l'histogramme se situera entre les deux courbes de la figure 1. Pour ne pas prendre en compte des valeurs liées aux valeurs aberrantes, il suffit de supprimer des colonnes en début et fin du tableau de rangs triés. Ceci permet d'obtenir des rangs moyens robustes et une courbe d'histogramme plus proche de la courbe bleue.

Nous présentons dans le paragraphe suivant une application de la méthode présentée dans le domaine biomédicale de la recherche sur le cancer.

2.1 Application

Dans la recherche médicale en oncologie, la détermination de gènes marqueurs peut se faire à l'aide d'une analyse cytogénétique moléculaire en utilisant une approche d'hybridation génomique comparative sur lame (aCGH : array comparative genomic hybridization) [7] ou par séquençage [8]. Avec la technologie aCGH, des sondes sont utilisées pour interroger toutes les régions du génome. Une sonde est une séquence de nucléotides consécutifs qui sont un sous-ensemble de ceux du génome. Généralement, chaque gène est représenté par une ou plusieurs sondes. Pour caractériser la maladie étudiée, l'échantillon du patient est comparé à un échantillon de référence ou à un pool d'échantillons normaux. Une mesure de (rapport) est associée à chaque sonde interogée. Pour une sonde, un rapport égal à 1 indique qu'elle ne réagit pas à la maladie ; un rapport plus grand (plus petit) que 1 indique est obtenue pour les sondes qui réagissent positivement (négativement) à la maladie. Les sondes qui réagissent permettent de déterminer les gènes marqueurs et les régions du genome qui les contiennent. Nous considérons une étude aCGH dans laquelle plusieurs patients sont utilisés et nous sommes intéressés par les régions anormales récurrentes (communes) pour la plupart des patients. Des méthodes ont été décrites dans la littérature pour résoudre ce problème [9, 10]. Dans ces méthodes, la recherche est effectuée au niveau des chromosomes et une segmentation des résultats de chaque échantillon est souvent nécessaire. Nous proposons une approche pour détecter des régions anormales récurrentes avec un traitement simultané des données tous les chromosomes sans une étape initiale de segmentation. Soit n le nombre de sondes dans l'étude avec k patients. Les rapports obtenus avec chaque patient peuvent être classés et les rangs issus de ce classement sont affectés aux sondes. Ceci permet

d'obtenir un tableau similaire au tableau 1 dans lequel les individus sont les sondes alors que les variables sont les différents patients. Les sondes les plus (moins) fréquemment classées au début (à la fin) auront les rangs moyens plus faibles (grands). Après l'identification des sondes d'intérêt, nous pouvons identifier les chromosomes et les gènes auxquels elles appartiennent. Une dernière étape va consister à agréger les sondes significatives pour obtenir les régions anormales récurrentes du génome pour la maladie. Les détails de la méthode sont décrits dans [1].

3 Conclusions

La méthode présentée est assez générale et adaptable à plusieurs situations réelles. Elle s'apparente à l'analyse en composante principale (ACP) utilisée dans la visualisation des données. Nous nous sommes intéressés à la détermination des individus en tête ou queue d'un tableau relativement à leurs scores pour des variables. Ce problème peut être une étape d'un problème plus général comme l'exemple de notre application. L'idée de trier les rangs avant le calcul des moyennes via la SVD est nouvelle. La méthode proposée permet d'obtenir des rangs moyens robustes, i.e. insensibles aux valeurs abérrantes couramment rencontrées pour des données de grandes dimensions.

A Annexe

Il y a un lien entre notre démarche et le théorème de Perron-Frobenius (PF) qui s'applique à des matrices carrées dont les composantes ont des valeurs non négatives. Le théorème PF stipule que pour les matrices carrées à valeurs non-négatives, la première valeur propre de sa décomposition en valeur propre (ED : eigenvalue decomposition) est positive et est plus importante en module comparée aux autres valeurs propres [5, 4, 6]. Nous pouvons obtenir une matrice carrée symétrique A à partir de la matrice X (n,m) :

$$A = X^T X \tag{6}$$

La ED de la matrice A s'écrit [3] :

$$A = W\Gamma W^T \tag{7}$$

où W est une matrice orthogonale et Γ est une matrice diagonale avec les valeurs propres qui apparaissent par ordre décroissant. En utilisant l'expression de (1) de X dans (6) combinée avec (7), nous pouvons écrire :

$$A = VDU^T UDV^T = VD^2V^T = W\Gamma W^T \tag{8}$$

Cette dernière relation montre que la matrice V de la SVD de X est la même que la matrice W de la ED de A alors que les valeurs singulières dans la matrice D sont les racines des valeurs propres dans la matrice Γ. Ainsi, la première valeur singulière est la plus importante. Le trie de rangs conduit à augmenter le poids des 2 premières composantes

Références

[1] D. Dembélé. Analysis of high throughput biological data using their rank values. *Stat Methods in Med Res*, aa(bb) :ccc–ddd, 2018.

[2] F. R. Gantmacher. *Théories des matrices*, volume 2. Dunod, Paris, 1966.

[3] G. H. Golub and C. F. V. Loan. *Matrix computations*. The Johns Hopkins Univ Press, Baltimore, 3 edition, 1996.

[4] R. R. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge Univ Press, 1996.

[5] M. Marcus and H. Ming. *A Survey of Matrix Theory and Matrix Inequalities*. Allyn and Bacon, Inc., Boston, 1964.

[6] C. Meyer. *Matrix Analysis and Applied Linear Algebra*. SIAM, Philadelphia, 2000.

[7] D. Pinkel and al. High Resolution Analysis of DNA Copy Number Variation Using Comparative Genomic Hybridization to Microarrays. *Nature Genet*, 20 :207–211, 1998.

[8] E. A. Pleasance and al. A Comprehensive Catalogue of Somatic Mutations from a Human Cancer Genome. *Nature*, 463(14) :191–196, 2010.

[9] D. E. Stange and al. Expression of an ASCL2 related stem cell signature and IGF2 in colorectal cancer liver metastases with 11p15.5 gain. *Gut*, 59 :1236–1244, 2010.

[10] E. van Dyk and al. A scale-space method for detecting recurrent DNA copy number changes with analytical false discovery rate control. *Nucleic Acids Res*, 41(9) :e100, 2013.