Equilibrer un échantillon presque parfaitement

Michaël Leuenberger, Esther Eustache, Raphaël Jauslin & Yves Tillé

Rue de Bellevaux 51, 2000 Neuchâtel, Suisse
esther.eustache@unine.ch

Résumé. Au moyen d'une modification simple, nous proposons une amélioration de la qualité de l'équilibrage d'un échantillon en utilisant la méthode du cube. Les variables auxiliaires pouvant décrire un espace métrique, les unités peuvent être ordonnées dans l'ordre décroissant de leur distance au centre du nuage de points que compose la population. Le centre et la distance peuvent être définis de plusieurs manières différentes. Les premières unités sont décrites comme des unités atypiques, quant aux dernières, elles sont considérées comme centrales ou similaires. La méthode du cube est ensuite appliquée sur les unités ordonnées. Le problème d'arrondi de l'équilibrage sur les variables auxiliaires est ainsi considérablement réduit.

Mots-clés. méthode du cube, profondeur, distance de Mahalanobis.

Abstract. With a simple modification, we propose a strong improvement of the quality of balancing a sample with the cube method. Since the auxiliary variables can described a metric space, the units can be sorted in descending order based on their distance from the center of the population data cloud. The center and distance can be defined in different ways. The first units are described as atypical, while the latest as central, similar. The cube method is then applied on the ordered units. The rounding problem on the balancing variables is greatly reduced.

Keywords. cube method, depth, Mahalanobis distance.

1 La méthode du cube

Considérons une population U contenant N unités. La méthode du cube (Deville and Tillé, 2004) permet la sélection d'échantillons équilibrés sur p variables auxiliaires, dans une population U, tout en respectant les probabilités d'inclusion $\{\pi_k\}_{k\in U}$, qu'elles soient égales ou inégales. Soit $\mathbf{x}_k \in \mathbb{R}^p$ le vecteur des valeurs prises par les p variables auxiliaires pour un individu $k \in U$. Le vecteur $\mathbf{x}_k$ est supposé connu pour toutes les unités de la population U. Un plan d'échantillonnage est dit équilibré si

$$\sum_{k\in U} \frac{a_k \mathbf{x}_k}{\pi_k} = \sum_{k\in U} \mathbf{x}_k, \tag{1}$$

où $\{a_k\}$ est la réalisation d'une variable aléatoire suivant une loi de Bernoulli, qui prend la valeur 1 si l'unité k est sélectionnée dans l'échantillon et 0 sinon. Le vecteur $\mathbf{a} = (a_1, \ldots, a_N)^\top$ est donc l'échantillon aléatoire.

La méthode du cube modifie aléatoirement et progressivement les valeurs $\{\pi_k\}_{k\in U}$ dans le but d'obtenir l'échantillon final composé des $\{a_k\}_{k\in U}$. L'équation d'équilibrage (1) ne peut pas toujours être exactement satisfaite du fait que les $\{a_k\}_{k\in U}$ sont entiers. Lors de la fin de la méthode du cube, les derniers $\{\pi_k\}$ ne sont pas encore entiers, et si aucune solution n'est possible pour obtenir exactement l'équation (1), on obtient une erreur d'arrondi et l'échantillon est dit comme *approximativement équilibré*. Dans ce cas-là, l'erreur d'équilibrage se porte sur les dernières unités traitées.

2 Optimisation en choisissant le meilleur ordre

Considérons que les variables auxiliaires composent un espace métrique de dimension p et que les valeurs du vecteur $\mathbf{x}_k$ correspondent aux coordonnées de l'unité k dans cet espace. On définit également le vecteur des valeurs dilatées par l'inverse de leurs probabilités d'inclusion $\check{\mathbf{x}}_k = \mathbf{x}_k/\pi_k, k \in U$. Un point central, noté $\overline{\check{\mathbf{x}}}$, peut être défini pour le nuage de points des N unités dans cet espace. Il peut être décrit, par exemple, comme la moyenne des valeurs des variables auxiliaires, i.e.

$$\overline{\check{\mathbf{x}}} = \frac{1}{N}\sum_{k\in U}\check{\mathbf{x}}_k.$$

Dans un premier temps, les unités sont ordonnées en fonction de leur distance avec le centre $\overline{\check{\mathbf{x}}}$ du nuage de points multivarié. La distance de Mahalanobis peut être utilisée, ou encore une méthode de statistique robuste comme une profondeur (voir parmi Rousseeuw and Struyf, 1998; Liu et al., 1999; Zuo and Serfling, 2000; Mosler, 2013). Les différentes profondeurs sont un ensemble d'outils statistiques robustes qui généralise la notion de quantile aux données multivariées. On peut utiliser les méthodes de "projection" et de "Tukey", implémentées dans le package `R DepthProc` de Kosiorowski and Zawadzki (2020). Il existe plusieurs définitions de la profondeur. Toutes les méthodes permettent d'ordonner les observations de telles sorte que les premières soient celles à la périphérie du nuage de points et les dernières, les plus centrales.

La méthode proposée ici est basée sur les raisonnements heuristiques suivants. A chaque étape du cube, la probabilité d'inclusion d'au moins une unité est arrondie à 0 ou 1 parmi les premières unités avec une probabilité non-entière. Si une unité est considérée au début, elle aura plus de chance, au fil des étapes, d'être arrondie à 0 ou à 1 sans problème, comparé à une unité qui se trouve à la fin,. En ordonnant les unités afin de considérer les unités *centrales* à la fin, nous aurons plus de chance que le problème d'arrondi lié à l'équation (1) se pose pour ces unités centrales. Le problème d'arrondi sera donc moins important. En d'autres termes, en ordonnant les unités en fonction de

leur profondeur ou de l'inverse de leur distance au centre, on impose de traiter d'abord les unités les plus atypiques. A la fin du traitement, les unités restantes seront principalement des unités centrales. Ces unités étant relativement similaires, le plan d'échantillonnage sera mieux équilibré car l'arrondi aura un impact moindre.

Donnons un exemple avec $\overline{\check{\mathbf{x}}}$ pour centre du nuage de points et la distance de Mahalanobis. Une unité est atypique si sa valeur étendue $\check{\mathbf{x}}_k$ est atypique. Soit

$$\mathbf{\Sigma} = \frac{1}{N-1} \sum_{k \in U} (\check{\mathbf{x}}_k - \overline{\check{\mathbf{x}}})(\check{\mathbf{x}}_k - \overline{\check{\mathbf{x}}})^\top.$$

La distance de Mahalanobis $d^2(k, \overline{\check{\mathbf{x}}})$ d'une unité k avec le centre $\overline{\check{\mathbf{x}}}$ est

$$d^2(k, \overline{\check{\mathbf{x}}}) = (\check{\mathbf{x}}_k - \overline{\check{\mathbf{x}}})^\top \mathbf{\Sigma}^{-1} (\check{\mathbf{x}}_k - \overline{\check{\mathbf{x}}}).$$

Les unités $k \in U$ sont ordonnées dans l'ordre décroissant de la distance de Mahalanobis $d^2(k, \overline{\check{\mathbf{x}}})$.

References

Deville, J.-C. and Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika*, 91:893–912.

Kosiorowski, D. and Zawadzki, Z. (2020). *DepthProc An R Package for Robust Exploration of Multidimensional Economic Phenomena*.

Liu, R. Y., Parelius, J. M., and Singh, K. (1999). Multivariate analysis by data depth: descriptive statistics, graphics and inference,(with discussion and a rejoinder by Liu and Singh). *The Annals of Statistics*, 27(3):783–858.

Mosler, K. (2013). Depth statistics. In Becker, C., Fried, R., and Kuhnt, S., editors, *Robustness and complex data structures*, pages 17–34. Springer, New York.

Rousseeuw, P. J. and Struyf, A. (1998). Computing location depth and regression depth in higher dimensions. *Statistics and Computing*, 8(3):193–203.

Zuo, Y. and Serfling, R. (2000). General notions of statistical depth function. *Annals of Statistics*, 28(2):461–482.