Available online at www.sciencedirect.com
ScienceDirect
C. R. Acad. Sci. Paris, Ser. I 347 (2009) 195–200
COMPTES RENDUS MATHEMATIQUE

Statistique

Ajustement polynomial local de la fonction d'égalisation équipercentile : convergence uniforme presque sûre

Kaouthar El Fassi [a], Belkacem Abdous [b], Mounir Mesbah [a]

[a] *L.S.T.A. – Université Pierre-et-Marie-Curie – Paris 6, 175, rue du Chevaleret, boîte 158, 75013 Paris, France*
[b] *Département de médecine sociale et préventive, Université Laval, pavillon de l'est, local 1138A, Québec, Qc, Canada, G1K 7P4*

Reçu le 22 mai 2008 ; accepté après révision le 15 novembre 2008

Disponible sur Internet le 31 janvier 2009

Présenté par Paul Deheuvels

Résumé

Soient X et Y deux variables aléatoires de fonctions de répartition F et G respectivement. Deux réalisations x et y sont dites équivalentes si et seulement si $F(x) = G(y)$. Cette équation est connue sous le nom « *équation équipercentile* ». Sa résolution, pour x fixé, permet d'exprimer l'équivalent équipercentile de x comme suit : $y(x) = G^{-1} \circ F(x)$, où G^{-1} désigne la fonction inverse de G. Nous proposons dans cette Note divers scénarios d'estimation de la « *fonction d'égalisation equipercentile* » $G^{-1} \circ F$. Ces estimateurs reposent sur la méthode des polynômes locaux. Des résultats de convergence uniforme presque sûre pour chaque scénario sont établis. ***Pour citer cet article : K. El Fassi et al., C. R. Acad. Sci. Paris, Ser. I 347 (2009).***

Abstract

Local polynomial fitting of the equipercentile equating function: strong uniform consistency. Let X and Y be two random variables with cumulative distribution functions F and G respectively. Two given realizations x and y are said to be equivalent if and only if $F(x) = G(y)$. This last equation is known as "*equipercentile equation*". For instance, for a given x, its equipercentile equivalent $y(x)$ is given by $y(x) = G^{-1} \circ F(x)$, where G^{-1} is the inverse of G. In this Note, we propose various nonparametric estimators of the equipercentile equating function $G^{-1} \circ F$. The proposed estimators are based on local polynomial fitting approach. Their uniform strong consistency will be investigated as well. ***To cite this article: K. El Fassi et al., C. R. Acad. Sci. Paris, Ser. I 347 (2009).***

Abridged English version

Equipercentile equating is a common technique that identifies scores on two tests or two health related quality of life instruments that have the same percentile rank. It allows one to use the scores in the two scales interchangeably. Let X and Y be two random variables that might be either two scores on two tests or two forms of the same test or outcomes of two health assessment instruments. Assume that these scores belong to the intervals $[x_1, x_N]$ and $[y_1, y_M]$

Adresse e-mail : kaouthar.el_fassi@upmc.fr (K. El Fassi).

1631-073X/$ – see front matter
doi:10.1016/j.crma.2008.12.019

where N and M are two known integers. Let $F(\cdot)$ and $G(\cdot)$ stand for the associated cumulative distribution functions (cdf). More precisely, two scores x and y are said to be equivalent if and only if

$$F(x) = G(y). \tag{1}$$

In practice such cumulative distributions are unknown and one has to rely on their empirical estimates. Henceforth, we will assume that we have to our disposal two samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ from X-scale and Y-scale respectively. Thus, we will have to solve the empirical version of (1), i.e.

$$F_n(x) = G_m(y), \tag{2}$$

where F_n and G_m are the usual empirical cdf. Besides, since these empirical cdf are discontinuous, it might be difficult to solve this equipercentile equation. Traditionally, this problem is solved by "continuizing" the involved functions by means of smoothing techniques such as kernel and splines methods. The smoothing step could be performed either before solving (2) or after. These two steps are known as "presmoothing" and "postsmoothing" methods. For more details, we refer to von Davier et al. [10] and the references therein.

In this work, we will adopt the local polynomial fitting approach as the smoothing tool. Its flexibility together with its ability to avoid boundary value effects are the main advantages over the classical kernel approach. The proposed estimation procedure will evolve around the following equipercentile equating function

$$y_{m,n}(x) = G_m^{-1}\big(F_n(x)\big) \tag{3}$$

where the generalized inverse function H^{-1} of any cdf H is defined by

$$H^{-1}(p) = \inf\big\{u\colon\ H(u) \geqslant p\big\}, \quad 0 < p < 1.$$

Note that the quantity $[G_m^{-1}(F_n(x)) - x]$ has been known for a long time as a Q-Q process or a Q-Q plot. Its first use goes back to Lorenz [9]. Several authors have investigated its finite and asymptotic properties, see e.g., Doksum [7], Aly [2], Beirlant and Deheuvels [3]. All these previous works used Q-Q methods to compare two distributions F and G. In addition, let us point out a completely different approach proposed by Bolshev [4,5] who uses asymptotically normal transformations or orthogonal polynomials expansions to approximate $y(x) = G^{-1}(F(x))$.

Next, since the equipercentile equating function (3) involves two empirical estimates F_n and G_m^{-1}, we might consider several smoothing scenarios. Indeed, the local polynomial fitting framework could be applied to the equipercentile equation $y_{m,n}(x) = G_m^{-1}(F_n(x))$ by smoothing $F_n(\cdot)$ only, or $G_m^{-1}(\cdot)$ only, or both $F_n(\cdot)$ and $G_m^{-1}(\cdot)$ or the Q-Q process $G_m^{-1} \circ F_n(\cdot)$. These scenarios are closely related to presmoothing and postsmoothing considerations. Ultimately, we will investigate the strong uniform consistency of the proposed estimators given in the end of Section 2.

1. Introduction et motivation

Soient X et Y deux variables aléatoires représentant deux scores à deux tests ou deux versions d'un même test ou encore les résultats de deux instruments de mesure de la qualité de vie. Supposons que ces scores appartiennent aux intervalles $[x_1, x_N]$ et $[y_1, y_M]$ respectivement. Les deux entiers N et M sont fixés et connus. Nous désignerons par $F(\cdot)$ et $G(\cdot)$ les fonctions de répartition associées à X et Y respectivement. Nous dirons que deux scores x et y sont équivalents si et seulement si

$$F(x) = G(y). \tag{4}$$

Le but principal de « *l'égalisation équipercentile* » est de trouver pour un score x donné d'une échelle X, le score « équivalent » y dans une autre échelle Y. Par exemple, dans le contexte de la qualité de vie, x représente le score mesuré par un questionnaire donné et y serait le score « équivalent » obtenu par un autre questionnaire. Souvent, les épidémiologistes et cliniciens disposent d'un instrument X et aimeraient traduire (ou interpréter) le score mesuré par cet instrument en un score équivalent qui aurait été obtenu par un autre instrument de référence Y.

En pratique les fonctions de répartition F et G sont inconnues et doivent être remplacées par leurs versions empiriques. Pour ce faire, nous supposerons que nous disposons de deux échantillons $X_1, \ldots, X_n$ et $Y_1, \ldots, Y_m$ issus respectivement de F et G et utiliserons la version empirique de (4), i.e.

$$F_n(x) = G_m(y), \tag{5}$$

où F_n et G_m désignent les fonctions de répartition empiriques usuelles. Étant donné le caractère empirique de (5), sa résolution peut à l'occasion présenter quelques difficultés. Traditionnellement, ce problème est contourné en lissant les fonctions de répartition empiriques par la méthode du noyau ou les splines. Le lissage s'effectue soit avant la résolution de l'équation équipercentile (5), soit après. Ces deux étapes sont appelées « pré-lissage » et « post-lissage ». Pour de plus amples détails et références, nous renvoyons au livre de von Davier et al. [10].

Dans ce travail, nous optons pour la méthode d'ajustement polynomial local. Cette approche a l'avantage d'être fléxible et de ne pas être soumise aux effets de bord. La procédure de lissage sera basée sur la fonction d'égalisation équipercentile suivante

$$y_{m,n}(x) = G_m^{-1}\big(F_n(x)\big) \tag{6}$$

où l'inverse généralisé de toute fonction de répartition H est défini par

$$H^{-1}(p) = \inf\big\{u\colon\ H(u) \geqslant p\big\}, \quad 0 < p < 1.$$

Notons que la quantité $[G_m^{-1}(F_n(x)) - x]$ est bien connue dans la littérature. Elle est appelée processus Quantile-Quantile et sa première utilisation remonte à Lorenz [9]. Ses propriétés asymptotiques ont été étudiées par plusieurs auteurs, par exemple, Doksum [7], Aly [2], Beirlant et Deheuvels [3]. Ces travaux utilisent les processus Q-Q pour comparer deux distributions F et G. Par ailleurs, signalons les travaux de Bolshev [4,5] qui propose une approche complètement différente pour approximer $y(x)$. Cette approche repose sur la notion de transformations asymptotiquement normales et les approximations par polynômes orthogonaux.

Nous présenterons dans la section suivante l'approche par ajustement poynomial local dans un cadre général. Ensuite, nous appliquerons cette technique à l'estimation des fonctions F_n, G_m^{-1} et $G_m^{-1} \circ F_n$. La dernière Section sera dédiée à la convergence unifrome presque sûre des divers estimateurs proposés.

2. Ajustement par polynômes locaux

Soit $H(\cdot)$ une fonction de distribution arbitraire et inconnue. Désignons par $\Phi(x, H)$ une fonctionnelle indexée par $x \in (a, b)$ avec $a < b$ deux constantes connues et finies. Supposons qu'on dispose d'un échantillon $Z_1, \ldots, Z_n$ issu de H et de $\Phi_n(x)$ un estimateur empirique de $\Phi(x, H)$. Le principe d'ajustement polynomial local repose sur le fait que toute fonction $\Phi(\cdot, H)$, suffisamment régulière aux voisinages d'un point x, peut être approximée localement par un polynôme. Plus formellement, cette idée se traduit par la résolution du problème des moindres carrés pondérés suivant

$$\min_{a_0,\ldots,a_r} \int_a^b \left\{ \Phi(z, H) - \sum_{k=0}^{r} a_k (z-x)^k \right\}^2 \frac{1}{h} K\left(\frac{z-x}{h}\right) \mathrm{d}z, \tag{7}$$

où $K(\cdot)$ est une densité de probabilité arbitraire tandis que $h = h(n)$ désigne une suite de paramètres de lissage. En pratique, la fonction $\Phi(z, H)$ est inconnue et doit être remplacée par son estimateur empirique $\Phi_n(z)$. Dans ce travail, nous nous limiterons aux ajustements linéaires et quadratiques i.e. $r = 1, 2$. En effet, si nous fixons $r = 1, 2$ dans le critère (7) et utilisons la valeur optimale obtenue de a_0 comme estimateur de $\Phi(u)$ alors nous obtenons les estimateurs linéaires et quadratiques suivants

$$\Phi_{nr}(u) = \int_a^b \frac{1}{h} K_r\left(\frac{z-u}{h}\right) \Phi_n(z)\, \mathrm{d}z, \quad r = 1, 2 \tag{8}$$

où

$$K_r(y) = \begin{cases} \dfrac{\mu_2 - \mu_1 y}{\mu_0\mu_2 - \mu_1^2} K(y), & \text{pour } r = 1; \\ \dfrac{(\mu_2\mu_4 - \mu_3^2) - (\mu_1\mu_4 - \mu_2\mu_3)y + (\mu_1\mu_3 - \mu_2^2)y^2}{(\mu_2\mu_4 - \mu_3^2)\mu_0 - (\mu_1\mu_4 - \mu_2\mu_3)\mu_1 + (\mu_1\mu_3 - \mu_2^2)\mu_2} K(y), & \text{pour } r = 2 \end{cases} \tag{9}$$

avec

$$\mu_l = \frac{1}{h}\int_a^b \left(\frac{z-u}{h}\right)^l K\left(\frac{z-u}{h}\right) \mathrm{d}z = \int_{(a-u)/h}^{(b-u)/h} z^l K(z)\,\mathrm{d}z, \quad \text{pour } l = 0, 1, \ldots, 4.$$

Le noyau K_r dépend du point d'intérêt u à travers les moments μ_l. Cette dépendance s'estompe lorsque le point u appartient à l'intérieur du support $[a, b]$ et que le noyau initial K est de support $[-1, 1]$. Il prend des formes différentes selon que le point u est proche de a, est à l'intérieur de $[a, b]$ ou est proche de b. C'est cette propriété qui lui confère sa résistance aux effets de bord dont souffre le noyau classique. En somme, cette approche fournit des estimateurs à noyau qui ont l'avantage de s'adapter à de nombreux problèmes d'estimation fonctionnelle, d'être flexibles et de ne pas être soumis aux problèmes des effets de bord (voir, Abdous et al. [1] pour plus de détails).

L'adaptation de cette technique d'estimation à notre problème d'égalisation equipercentile est immédiate. En effet, nous avons plusieurs scénarios possibles. En reprenant l'estimateur empirique de la fonction equipercentile $y_{m,n}(x) = G_m^{-1}(F_n(x))$, on voit qu'il est possible de lisser $F_n(\cdot)$ uniquement, ou bien lisser $G_m^{-1}(\cdot)$ uniquement, ou bien lisser simultanément et séparément $F_n(\cdot)$ et $G_m^{-1}(\cdot)$ ou encore lisser l'estimateur $G_m^{-1} \circ F_n(\cdot)$. Ces divers scénarios nous conduisent à considérer les 5 estimateurs suivants

(i) $y_{m,n}^{[1]}(x) = G_m^{-1}(F_n(x))$, pour $x \in [x_1, x_N]$, i.e. le processus Q-Q initial sans lissage.

(ii) $y_{m,n}^{[2]}(x) = G_m^{-1}(\widetilde{F}_n(x))$, pour $x \in [x_1, x_N]$. Seule la f.r. empirique F_n est lissée. Elle est remplacée par $\widetilde{F}_n$, obtenue de (8) après avoir remplacé Φ_n par F_n et posé $a = x_1$ et $b = x_N$.

(iii) $y_{m,n}^{[4]}(x) = \widetilde{G_m^{-1}}(\widetilde{F}_n(x)) = \int_0^1 \frac{1}{k} K_r(\frac{z-\widetilde{F}_n(x)}{k}) G_m^{-1}(z)\,\mathrm{d}z$, avec $\widetilde{F}_n(x) = \int_{x_1}^{x_N} \frac{1}{h} L_r(\frac{z-x}{h}) F_n(z)\,\mathrm{d}z$. Ici, nous avons lissé séparément G_m^{-1} et F_n en prenant deux fenêtres différentes k et h et deux noyaux K et L.

(iv) $y_{m,n}^{[5]}(x) = \widetilde{G_m^{-1} \circ F_n(x)}) = \int_{x_1}^{x_N} \frac{1}{h} K_r(\frac{z-x}{h}) G_m^{-1} \circ F_n(z)\,\mathrm{d}z$, pour $x \in [x_1, x_N]$, i.e. on lisse en une seule fois le processus $G_m^{-1} \circ F_n(\cdot)$.

La section suivante traite de la convergence uniforme presque sûre de ces estimateurs.

3. Convergence uniforme presque sûre

Théorème 3.1. *Soient X et Y deux variables aléatoires de fonctions de répartition F et G et de supports $S(F) = [x_1, x_N]$ et $S(G) = [y_1, y_M]$ respectivement. Supposons que G soit continûment dérivable et telle que $G' = g$, $\inf_{0\leqslant u\leqslant 1} g(G^{-1}(u)) > 0$ et $\sup_{0\leqslant u\leqslant 1} |g'(G^{-1}(u))| < \infty$. Alors,*

$$\sup_{x\in S(F)} \left|y_{m,n}^{[i]}(x) - G^{-1}\big(F(x)\big)\right| \xrightarrow{\text{p.s.}} 0, \quad \textit{pour } i = 1, \ldots, 5.$$

Schéma des preuves. La preuve intégrale de ce résultat est fournie dans El Fassi et al. [8]. Nous en donnons les grandes lignes ci-après. En effet, la démonstration de la convergence uniforme presque sûre des estimateurs proposés est basée sur le lemme de Bochner et la convergence uniforme presque sûre des processus Q-Q.

(i) Le comportement de l'estimateur $y_{m,n}^{[1]}(x)$ est identique à celui du processus Q-Q : $[G_m^{-1}(F_n(x)) - x]$. Voir par exemple Doksum [7], Aly [2], Beirlant and Deheuvels [3].

(ii) La convergence du second estimateur $y_{m,n}^{[2]}(x)$ repose sur l'inégalité suivante

$$\begin{aligned}\left|y_{m,n}^{[2]}(x) - G^{-1}\big(F(x)\big)\right| &\leqslant \sup_{x\in S(F)} \left|G_m^{-1}\big(\widetilde{F}_n(x)\big) - G^{-1}\big(\widetilde{F}_n(x)\big)\right| + \sup_{x\in S(F)} \left|G^{-1}\big(\widetilde{F}_n(x)\big) - G^{-1}\big(F(x)\big)\right| \\ &:= \Delta_1 + \Delta_2.\end{aligned}$$

En utilisant le fait que $\sup_{u\in[0,1]} |G_m^{-1}(u) - G^{-1}(u)| \xrightarrow{\text{p.s.}} 0$, $m \to \infty$ (voir, e.g. Csörgő [6]), il est aisé de voir que Δ_1 converge p.s. vers 0. Quant à Δ_2, notons d'abord que

$$\begin{aligned}\left|\widetilde{F}_n(x) - F(x)\right| &\leqslant \sup_{z\in S(F)} \left|F_n(z) - F(z)\right| \int_{(x_1-x)/h}^{(x_N-x)/h} \left|K_r(z)\right| \mathrm{d}z + \left| \int_{x_1}^{x_N} \frac{1}{h} K_r\left(\frac{x-z}{h}\right)\big[F(z) - F(x)\big]\,\mathrm{d}z \right| \\ &:= \Gamma_1 + \Gamma_2.\end{aligned}$$

Le lemme de Bochner (voir El Fassi et al. [8]) permet de montrer que $\Gamma_2 \to 0$, tandis que le résultat classique $\sup_{x\in S(F)} |F_n(x) - F(x)| \xrightarrow{\text{p.s.}} 0$ et le fait que la quantité $|\int_{(x_1-x)/h}^{(x_N-x)/h} |K_r(z)|\,\mathrm{d}z$ soit bornée pour tout x nous assurent la convergence presque sûre de Γ_1. Pour conclure, il suffit d'utiliser la continuité uniforme de $G^{-1}(\cdot)$ sur $[0, 1]$.

(iii) Considérons $y_{m,n}^{[3]}(x)$ et notons que

$$\begin{aligned}\big|y_{m,n}^{[3]}(x) - G^{-1}\circ F(x)\big| &\leqslant \left|\int_0^1 \frac{1}{h}K_r\left(\frac{z-F_n(x)}{h}\right)\big[G_m^{-1}(z) - G^{-1}(z)\big]\,\mathrm{d}z\right| \\ &\quad + \left|\int_0^1 \frac{1}{h}K_r\left(\frac{z-F_n(x)}{h}\right)\big[G^{-1}(z) - G^{-1}\circ F_n(x)\big]\,\mathrm{d}z\right| \\ &\quad + \big|G^{-1}\circ F_n(x) - G^{-1}\circ F(x)\big| \leqslant \Delta_{31} + \Delta_{32} + \Delta_{33}.\end{aligned}$$

Un raisonnement analogue à celui utilisé au point (ii) ci-haut permet d'établir la convergence uniforme presque sûre des termes Δ_{31} et Δ_{33}. Quant au terme restant Δ_{32}, il est toujours loisible d'écrire la décomposition suivante

$$\begin{aligned}\int_{-F_n(x)/h}^{(1-F_n(x))/h} K_r(z)\Phi(z)\,\mathrm{d}z &= \mathbb{I}\big(F_n(x) = \alpha h\big)\int_{-\alpha}^{1} K_r^L(z)\Phi(z)\,\mathrm{d}z + \mathbb{I}\big(h < F_n(x) < 1-h\big)\int_{-1}^{1} K_r^I(z)\Phi(z)\,\mathrm{d}z \\ &\quad + \mathbb{I}\big(F_n(x) = 1-\alpha h\big)\int_{-1}^{\alpha} K_r^R(z)\Phi(z)\,\mathrm{d}z,\end{aligned}$$

où $\alpha \in [0, 1]$, Φ est une fonction arbitraire, tandis que K_r^L, K_r^I et K_r^R sont les versions du noyau K_r restreint aux régions de gauche, de centre et de droite de l'intervalle $[0, 1]$. Ainsi, nous avons

$$\Delta_{32} \leqslant \int_{-F_n(x)/h}^{(1-F_n(x))/h} \big|K_r(y)\big|\big|G^{-1}\big(F_n(x)+hy\big) - G^{-1}\circ F_n(x)\big|\,\mathrm{d}y \leqslant C \sup_{|u-v|\leqslant h}\big|G^{-1}(u) - G^{-1}(v)\big|$$

avec C une constante positive. La continuité uniforme de G^{-1} permet de conclure.

(iv)–(v) La preuve de ces deux résultats ne pose pas de problèmes particuliers, elle suit les mêmes lignes que les preuves ci-haut.

4. Conclusion

Nous avons présenté un résultat théorique sur la convergence uniforme presque sûre qui valide le bon potentiel des estimateurs polynômiaux locaux de la fonction d'égalisation équipercentile. Quand les fonctions F et G sont remplacées par leurs estimateurs empiriques, la solution $y_{n,m}(x) = G_m^{-1}(F_n(x))$ peut ne pas exister en raison de la discontinuité. Pour contourner ce problème, nous avons fait appel à la méthode d'ajustement local polynomial. Cette méthode a l'avantage, non seulement de lisser une fonction arbitraire, mais également d'être flexible et ne pas être soumise aux problèmes des effets de bord. Ces effets de bord se rencontrent en particulier quand on utilise sans précaution la méthode du noyau qui est un cas particulier de notre méthode. Une validation par des simulations et une application sur des données réelles en qualité de vie, est en cours et complétera les résultats théoriques présentés dans cette Note.

Références

[1] B. Abdous, A. Berlinet, N. Hengartner, A general theory for kernel estimation of smooth functionals of the distribution function and their derivatives, Rev. Roumaine Math. Pures Appl. 48 (3) (2003) 217–232.

[2] E.-E. Aly, Strong approximations of the Q-Q process, J. Multivariate Anal. 20 (1) (1986) 114–128.

[3] J. Beirlant, P. Deheuvels, On the approximation of P-P and Q-Q plot processes by Brownian bridges, Statist. Probab. Lett. 9 (3) (1990) 241–251.

[4] L.N. Bolshev, On transformations of random variables, Theory Probab. Appl. IV (1959) 129–141.
[5] L.N. Bolshev, Asymptotically Pearson transformations, Theory Probab. Appl. VIII (1963) 121–146.
[6] M. Csörgő, Quantile Processes with Statistical Applications, SIAM, Philadelphia, 1983.
[7] K. Doksum, Empirical probability plots and statistical inference for nonlinear models in the two-sample case, Ann. Statist. 2 (1974) 267–277.
[8] K. El Fassi, B. Abdous, M. Mesbah, Égalisation equipercentile et polynômes locaux, Preprint, L.S.T.A., Université Pierre et Marie Curie, 2008.
[9] M.O. Lorenz, Methods on measuring the concentration of wealth, J. Amer. Statist. Assoc. 70 (1905) 209–219.
[10] A.A. von Davier, P.W. Holland, D.T. Thayer, The Kernel Method of Test Equating, Statistics for Social Science and Public Policy, Springer, New York, 2004.