Modèles basés sur des copules pour l'estimation dans des petits domaines

Louis-Paul Rivest [1], François Verret [2] & Sophie Baillargeon [3]

[1] *Département de mathématiques et de statistique, Université Laval, Québec G1V 0A6, Canada, Louis-Paul.Rivest@mat.ulaval.ca*

[2] *François Verret, Statistics Canada, Ottawa, On K1A 0T6, Canada, Francois.Verret@statcan.gc.ca*

[3] *Département de mathématiques et de statistique, Université Laval, Québec G1V 0A6, Canada, Sophie.Baillargeon@mat.ulaval.ca*

Résumé. Des copules multidimensionnelles sont utilisées pour modéliser la dépendance résiduelle entre les unités échantillonnées dans un petit domaine et pour construire de nouveaux estimateurs de la moyenne de la variable d'intérêt y à l'intérieur d'un petit domaine. Ceci donne des alternatives à l'estimateur EBLUP associé au modèle normal de régression avec ordonnées à l'origine aléatoire de Battese, Harter et Fuller (1988). Le modèle statistique sous-jacent à cette nouvelle classe d'estimateurs fait intervenir (i) le vecteur β des paramètres de la régression de y sur x, un vecteur de variables explicatives dont les totaux sont connus pour chaque petit domaine, (ii), une famille de copules échangeables $\{C_{\alpha,n}\}$ pour la dépendance résiduelle entre les unités d'un même petit domaine et (iii), une fonction de répartition $F_e(\cdot)$ pour la loi marginale des erreurs de la régression de y sur x. Cet exposé met l'emphase sur des modèles où la fonction de répartition F_e n'appartient à aucune famille paramétrique précise; elle est considérée comme un paramètre fonctionnel de dimension infinie. C'est dans ce cadre que des estimateurs EBUP (pour empirical best unbiased predictors) de la moyenne de y dans un petit domaine sont construits.

Mots-clés. Copules, Meilleur prédicteur, ...

1 Introductions aux copules

La notion de copules est attribuée à Sklar (1959) qui a démontré que la fonction de répartion conjointe de plusieurs variable aléatoires peut s'écrire en fonction de leurs lois marginales et d'une copule pour la dépendance entre ces variables. Prenons par exemple un vecteur aléatoire $\mathbf{Y}$ de longueur n de loi $N_n\{\mathbf{0}, \mathbf{\Sigma}(\rho, n)\}$, une loi normale multidimensionnelle de moyenne $\mathbf{0}$ et avec une matrice de variances d'équicorrélation où toutes les variances sont égales à 1 et toutes les corrélations valent $\rho \in [0, 1)$. La fonction de répartition, $F(y_1, \ldots, y_n)$, de $\mathbf{Y}$ est

$$\Pr(Y_1 \leq y_1, \ldots, Y_n \leq y_n) = \int_{-\infty}^{y_1} \cdots \int_{-\infty}^{y_n} \frac{\exp\{-x^\top \mathbf{\Sigma}(\rho, n)^{-1} x/2\}}{(2\pi)^{n/2} |\mathbf{\Sigma}(\rho, n)|^{1/2}} dx_1 \ldots dx_n.$$

La loi marginale des Y_i est $\Phi(z)$, la fonction de répartition normale standardisée, et la copule $C_{\rho,n}$ pour la dépendance entre les Y_i est la distribution conjointe de $\{\Phi(Y_1), \ldots, \Phi(Y_n)\}$,

$$C_{\rho,n}(u_1, \ldots, u_n) = \int_{-\infty}^{\Phi^{-1}(u_1)} \cdots \int_{-\infty}^{\Phi^{-1}(u_n)} \frac{\exp\{-x^\top \mathbf{\Sigma}(\rho, n)^{-1} x/2\}}{(2\pi)^{n/2} |\mathbf{\Sigma}(\rho, n)|^{1/2}} dx_1 \ldots dx_n, \qquad (1)$$

qui est définie sur le carré unité, $u_1, \ldots u_n, \in (0, 1)$. Cette famille de copules est fermée pour la marginalisation, dans le sens où $C_{\rho,n}(u_1, \ldots, u_{n-1}, 1) = C_{\rho,n-1}(u_1, \ldots, u_{n-1})$. De plus, lorsque $\rho = 0$, on obtient la copule d'indépendance, $C_{0,n}(u_1, \ldots, u_n) = u_1 \times \ldots u_n$. Dire que $\mathbf{Y}$ est de loi $N_n\{\mathbf{0}, \mathbf{\Sigma}(\rho, n)\}$ est équivalent à définir la fonction de répartition conjointe des éléments de $\mathbf{Y}$ comme étant

$$F(y_1, \ldots, y_n) = C_{\rho,n}\{\Phi(y_1), \ldots, \Phi(y_n)\}.$$

Le modèle de Battese, Harter & Fuller (1988) pour la prédiction au niveau des unités exprime la variable dépendante pour l'unité j de la petite région $i = 1, \ldots, m$ de la façon suivante,

$$Y_{ij} = x_{ij}^\top \beta + a_i + b_{ij} \quad i = 1, \ldots, m; j = 1, \ldots, N_i, \qquad (2)$$

où N_i est la taille de la petite région i, x_{ij} est le vecteur des variables explicatives pour l'unité (i, j), β est un vecteur de paramètres de régression, $a_i \sim N(0, \sigma_a^2)$ est l'effet aléatoire pour la petite région i et $b_{ij} \sim N(0, \sigma^2)$ représente l'erreur expérimentale. Selon ce modèle, la fonction de répartition conjointe des Y_{ij} dans le petit domaine i est une loi normale multidimensionnelle, $N_{N_i}\{\mathbf{X}_i\beta, (\sigma^2 + \sigma_a^2)\Sigma(\rho, N_i)\}$, où $\mathbf{X}_i$ et la matrice des x_{ij}, $\Sigma(\rho, N_i)$ est la matrice d'équicorrélation définie précédemment, avec $\rho = \sigma_a^2/(\sigma^2 + \sigma_a^2)$, et $\sigma^2 + \sigma_a^2$ est la variance marginale de Y_{ij}. Sous (2), la fonction de répartition conjointe des Y_{ij} peut également s'écrire à l'aide de la copule normale (1) de la façon suivante,

$$C_{\rho,N_i}\left[\Phi\left(\frac{y_{i1} - x_{i1}^\top \beta}{(\sigma^2 + \sigma_a^2)^{1/2}}\right), \ldots, \Phi\left(\frac{y_{iN_i} - x_{iN_i}^\top \beta}{(\sigma^2 + \sigma_a^2)^{1/2}}\right)\right]. \qquad (3)$$

2 Modèles de copules pour petites régions

Une fois que le modèle de Battese, Harter et Fuller (1988) a été écrit sous la forme (3), on peut facilement en construire des généralisations. La famille de copules normales peut être remplacée par une autre famille de copules échangeables telle que la copule t ou une copule Archimédienne multidimensionnelle, voir Mai et Scherer (2012), chapitre 2. Les hypothèses concernant la loi marginale des erreurs peuvent également être assouplies en supposant que cette dernière est une fonction de répartition quelconque $F_e(e)$, de moyenne 0. Le modèle proposé pour la fonction de répartition conjointe des Y_{ij} pour le petit domaine i est donc

$$C_{\alpha,N_i}\left[F_e\left(y_{i1} - x_{i1}^\top \beta\right), \ldots, \Phi\left(y_{iN_i} - x_{iN_i}^\top \beta\right)\right], \quad i = 1, \ldots, m, \qquad (4)$$

où $\beta \in \Re^p$ est un vecteur de paramètres de régression inconnus, F_e est une fonction de répartition quelconque de moyenne 0 et $\{C_{\alpha,n}\}$ est une famille paramétrique de copules échangeables indexées par le paramètre univarié $\alpha > 0$ associé à la force de la dépendance résiduelle.

Supposons maintenant que des échantillons aléatoires simples de tailles $n_i < N_i$ ont été tirés dans les domaines $i = 1, \ldots, m$. Pour ajuster (4), il faut estimer trois paramètres, β, α, et $F_e(\cdot)$. Puisqu'aucune hypothèse n'est faite concernant $F_e(\cdot)$, la méthode du maximum de vraisemblance est difficilement applicable et il faut trouver des estimateurs ad hoc. Pour β nous suggérons d'utiliser l'estimateur $\hat{\beta}$ obtenu en ajustant le modèle (2) aux données. Si les données ne sont pas normales cet estimateur n'est pas optimal cependant il donne tout de même un estimateur convergent au sens où $\sqrt{m}(\hat{\beta}-\beta)$ est borné en probabilité lorsque m tend vers l'infini. Les résidus du modèle s'écrivent $e_{ij} = Y_{ij} - x_{ij}^\top \hat{\beta}$ pour $i = 1, \ldots, m; j = 1, \ldots, N_i$ et un estimateur naïf de la fonction de répartition $F_e(\cdot)$ est la fonction de répartition empirique des résidus,

$$\hat{F}_e(z) = \frac{1}{(\sum n_i) + 1} \sum_{i=1}^{m} \sum_{j \in s_i} 1_{\{\hat{\epsilon}_{ij} \leq z\}}.$$

Il s'agit d'un estimateur convergent de $F_e(z)$ lorsque m tend vers l'infini. Finalement pour estimer le paramètre de dépendance α une version échangeable du tau de Kendall standard est disponible. Elle fait intervenir les $\sum_{i<k} n_i(n_i-1)n_k(n_k-1)$ paires de vecteurs bidimensionnels ordonnés où les deux éléments d'une paire viennent de deux domaines différents. La statistique $\hat{\tau}$ est la proportion de paires concordantes moins la proportion des paires discordantes dans cet ensemble. Cette statistique est étudiée par Romdhani, Lakhal et Rivest (2014). Il s'agit d'un estimation convergente du taux de Kendall caractérisant la dépendance de la copule $C_{\alpha,2}(u_1, u_2)$. Lorsque l'on travaille avec la copule normale, la formule $\tau = 2\arcsin(\rho)/\pi$ donne un lien entre τ et la corrélation ρ qui permet d'estimer ρ à partir de $\hat{\tau}$.

3 Meilleure prédiction non biaisée

Dans cette section on considère un seul petit domaine de taille N duquel un échantillon aléatoire simple s de taille n a été tiré. On suppose que les paramètres (β, α, F_e) sont connus et que les erreurs expérimentales $(\varepsilon_1, \ldots, \varepsilon_n)$ sont observées. On veut prédire ε_r, l'erreur expérimentale d'une unité non échantillonnée. La meilleure prédiction est donnée par l'espérance conditionnelle de ε_r sachant $(\varepsilon_1, \ldots, \varepsilon_n)$. Si $c_{\alpha,n}$ dénote la densité de la copule $C_{\alpha,n}$, alors la distribution conditionnelle de ε_r s'écrit

$$F_e(\varepsilon_r|s) = \int_{-\infty}^{\varepsilon_r} \frac{c_{\alpha,n+1}\{F_e(x), F_e(\varepsilon_1), \ldots, F_e(\varepsilon_n)\}}{c_{\alpha,n}\{F_e(\varepsilon_1), \ldots, F_e(\varepsilon_n)\}} dF_e(x).$$

On peut écrire l'espérance conditionnelle de ε_r à l'aide de la fonction de poids

$$w_1\{F_e(x), F_e(\epsilon_j) : \ j \in s\} = \frac{c_{\alpha,n+1}\{F_e(x), F_e(\varepsilon_1), \ldots, F_e(\varepsilon_n)\}}{c_{\alpha,n}\{F_e(\varepsilon_1), \ldots, F_e(\varepsilon_n)\}},$$

de la façon suivante,

$$E(\epsilon_r|s) = \int_R x \ w_1\{F_e(x), F_e(\varepsilon_j) : \ j \in s\} dF_e(x).$$

Ainsi la meilleure prédiction non biaisée de la moyenne de Y dans le petit domaine est

$$\begin{aligned} \tilde{\bar{y}}_U &= \bar{x}_U^\top \beta + \frac{1}{N}\left\{\sum_{j\in s} \varepsilon_j + (N-n)E(\varepsilon_r|s)\right\} \\ &= \bar{x}_U^\top \beta + \frac{1}{N}\sum_{j\in s} \varepsilon_j + \left(1 - \frac{n}{N}\right) E(\varepsilon_r|s), \end{aligned}$$

où $\bar{x}_U$ est la moyenne de x dans le petit domaine. On évalue facilement la variance de l'erreur de prédiction

$$E\{(\tilde{\bar{y}}_U - \bar{y}_U)^2\} = \frac{N-n}{N^2} E\{\mathrm{Var}(\varepsilon_r|s)\} + \frac{(N-n)(N-n-1)}{N^2} E\{\mathrm{Cov}(\varepsilon_{r1}, \varepsilon_{r2}|s)\}.$$

où $(\varepsilon_{r1}, \varepsilon_{r2})$ représentent les erreurs expérimentales de deux unités non échantillonnées. Lorsque N est grand cette expression se réduit à $E\{\mathrm{Cov}(\varepsilon_{r1}, \varepsilon_{r2}|s)\}$. Elle généralise la variance a posteriori qui mesure la précision d'une prédiction BLUP lorsque tous les paramètres sont connus.

En pratique les paramètres sont inconnus et il faut remplacer (β, α, F_e) par les estimateurs donnés à la section précédente. De plus un simple estimateur "plug-in" de $E\{\mathrm{Cov}(\varepsilon_{r1}, \varepsilon_{r2}|s)\}$ sous-estime la variance de l'erreur de prédiction car il ne tient pas compte de l'estimation des paramètres et des ajustements sont nécessaires. Ces questions seront discutées durant la présentation qui illustrera la méthodologie proposée à l'aide d'une application au jeu de donées sur les municipalités suisses du package R sampling, voir Tillé et Matei (2013).

Bibliographie

[1] Battese, G. E. Harter, R. M. & Fuller, W. A. (1988), An Error-Components Models for Prediction of County Crop Areas Using Survey and Satellite Data, *Journal of the American Statistical Association*, 83, 28-36

[2] Mai, J.-M. & Scherer, M. (2012), *Simulating Copulas; Stochastic Models, Sampling Algorithms and Applications. Series in Quantitative Finance: Volume 4.* World Scientific Publishing Company.

[3] Romdhani, H., Lakhal-Chaïeb, L. & Rivest, L.-P. (2014), An exchangeable Kendall's tau for clustered data, *Canadian Journal of Statistics*, accepté pour publication
[4] Sklar, A. (1959). Fonctions de Répartition à n Dimensions et leurs Marges. *Publications de l'Institut de statistique de l'Université de Paris*, 8, 229–231.
[5] Tillé, Y. et Matei, A (2013). sampling: Survey Sampling. R package version 2.6. http://CRAN.R-project.org/package=sampling