Prédicteur Empirique de Bayes basé sur des copules Archimédiennes pour des proportions régionales

Fodé Tounkara* et Louis-Paul Rivest**

*Université du Québec À Montréal (UQÀM) **Université Laval (U. L.)

Vendredi 14 octobre 2016

Introduction

- On s'intéresse aux estimations produites à partir de sous-populations, appelées **Petits domaines**, en particulier lorsque la taille de l'échantillon est petite.

- Ces estimations sont utilisées, par exemple pour
  1. les allocations des fonds du gouvernement
  2. les stratégies de planification régionales

- Exemples d'estimation dans de petits domaines
  1. Incidence de la pauvreté
  2. Taux de chômage des moins de 25 ans
  3. la proportion d'individus sans assurance maladie

Les différentes approches d'estimation

- **Les Estimateurs Directs**
  1. sont adéquats pour les grands domaines
  2. ont souvent de grandes variances et de grands coefficients de variation (CV)
  3. ne peuvent pas être utilisés si on a pas d'observations
- **Les Estimateurs Synthétiques**
- **Les Estimateurs Empiriques de Bayes (EB)** Basés sur des modèles qui sont utilisés pour connecter différents petits domaines
- Deux types de modèles :
  1. modèles au niveau du domaine
  2. modèles au niveau des unités

Notations

- Considérons m domaines de tailles $N_1 \ldots N_m$
- y_{ij} une variable dichotomique, mesurée sur l'unité j du domaine i
- Soit p_i la probabilité de succès conditionnelle pour le domaine i telle que $E(p_i) = \pi_i$
- s_i un échantillon de taille n_i tiré du domaine i
- $y_i = \sum_{j=1}^{n_i} y_{ij}$ le nombre total de succès dans le domaine i
- Soit ρ le coefficient de corrélation intra-grappe

Objectif et Proposition

- **Objectif :** : Estimer les vraies proportions P_i

$$P_i = N_i^{-1}\{\sum_{j \in s_i} y_{ij} + \sum_{j \notin s_i} y_{ij}\}$$

- **Proposition :**

1. Modéliser la variation extra-binomiale par les modèles de copules
2. Appliquer une Inférence Bayésienne pour l'estimation des proportions

Modèle de Base

- Nous considérons des données de **Bernoulli** en grappes telles que
  - les observations à l'**intérieur** de chaque grappe sont **échangeables**

$$F(y_{i1}, \ldots, y_{iN_i}) = F(y_{i\pi(1)}, \ldots, y_{i\pi(N_i)}),$$

$\{\pi(1), \cdots, \pi(n)\}$: une permutation des entiers $\{1, \cdots, n\}$

  - les observations provenant des grappes **différentes** sont **indépendantes**

- But : Modéliser la structure de **dépendance échangeable**

- **Modèle de Base :**

(i) $y_i \mid p_i \sim Bin(n_i, p_i)$

(ii) $p_i \sim G$

Le modèle Beta-Binomial et Le meilleur prédicteur linéaire

- le modèle logit normal : logit $(p_i) \sim N(\mu, \sigma^2)$
- le modèle Beta-binomial (BB) : $p_i \sim \text{Beta}(\alpha, \beta)$,

$$\alpha = \frac{\pi(1-\rho)}{\rho}, \qquad \beta = \frac{(1-\pi)(1-\rho)}{\rho}.$$

- Supposons que $\pi_i = \pi$, pour $i = 1, \ldots, m$
- Sous BB, le prédicteur optimum (B) pour p_i est

$$\hat{p}_i^B(y_i, \varphi) = \gamma \frac{y_i}{n_i} + (1-\gamma)\pi; \qquad \gamma = \frac{n_i \rho}{1 + (n_i - 1)\rho}.$$

- Remarque : $\hat{p}_i^B(y_i, \varphi) = \hat{p}_i^{blup}(y_i, \varphi)$
- L'EQM de $\hat{p}_i^B$ est

$$MSE\left(\hat{p}_i^B(y_i, \varphi)\right) = \frac{\pi(1-\pi)\rho(1-\rho)}{1 + (n_i - 1)\rho}.$$

Modèle de copules pour la corrélation intra-grappe

Approche 1 : Modèles à effet aléatoire

$$p_i = 1 - e^{-a_i \phi_\alpha^{-1}(1-\pi_i)} \tag{2.1}$$

- a_i : variable aléatoire positive avec transformée de Laplace $\phi_\alpha(\cdot)$

Approche 2 : Modèles de copule

$$P(Y_1 \leq y_1, \ldots, Y_{n_i} \leq y_{n_i}) = C_\alpha\{F(y_1), \ldots, F(y_{n_i})\}, \tag{2.2}$$

- $F(y_j) = (1-\pi_i)^{1-y_j}, \quad y_j = 0, 1, \quad j = 1, \ldots, n_i$
- C_α : $C_\alpha(u_1, \ldots, u_n) = \psi_\alpha\{\sum_{j=1}^{n_i} \psi_\alpha^{-1}(u_j)\}$
- $\psi_\alpha()$: générateur de la copule Archimédiene

Exemples de modèles

Familles	$\phi_\alpha(t) \sim \psi_\alpha(t)$	F_{a_i}
Clayton (C)	$(\alpha t + 1)^{-\frac{1}{\alpha}}$	Gamma$(\alpha, \frac{1}{\alpha})$
Gumbel (G)	$e^{-t^{\frac{1}{\alpha+1}}}$	Loi Stable$(\frac{1}{\alpha+1})$
Frank (F)	$-\frac{1}{\alpha}\log\left\{1 - \frac{(1-e^{-\alpha})}{e^t}\right\}$	Logarithmique$(1-e^{-\alpha})$
Joe (J)	$1 - (1-e^{-t})^{\frac{1}{\alpha+1}}$	Sibuya$(\frac{1}{1+\alpha})$

En permutant les succès (y_i) et les échecs $(n_i - y_i)$, on obtient une nouvelle classe de modèles de copules (dC, dG, dF, dJ) pour la variation extra-binomiale.

Mesures de dépendance

- Le **tau de Kendall** (τ)
- $\tau = 0 \quad \Rightarrow \quad$ l'indépendance, et $\tau \approx 1 \quad \Rightarrow \quad$ une dépendance forte
- Dans le cas des copules Archimédiennes de générateur ψ

$$\tau = 1 + 4 \int_0^1 \frac{\psi(t)}{\psi'(t)} dt.$$

- Clayton : $\tau = \frac{\alpha}{\alpha+2}$, Gumbel : $\tau = \frac{\alpha}{\alpha+1}$
- Pour $\pi_i = \pi$, l'expression du coefficient de corrélation est

$$\rho = \frac{\psi_\alpha\{2\psi_\alpha^{-1}(\pi)\} - \pi^2}{\pi(1-\pi)}.$$

Le meilleur prédicteur optimum (BP)

- Sachant $Y = y$, la T. L. de a_i est donnée par

$$\psi_{\alpha,y}(t) = E\left(e^{-ta}|y,\alpha,\pi\right)$$
$$= \frac{\sum_{\ell=0}^{n-y}\binom{n-y}{\ell}(-1)^{\ell}\psi_{\alpha}\{t+(y+\ell)\psi_{\alpha}^{-1}(\pi)\}}{\sum_{\ell=0}^{n-y}\binom{n-y}{\ell}(-1)^{\ell}\psi_{\alpha}\{(y+\ell)\psi_{\alpha}^{-1}(\pi)\}}.$$

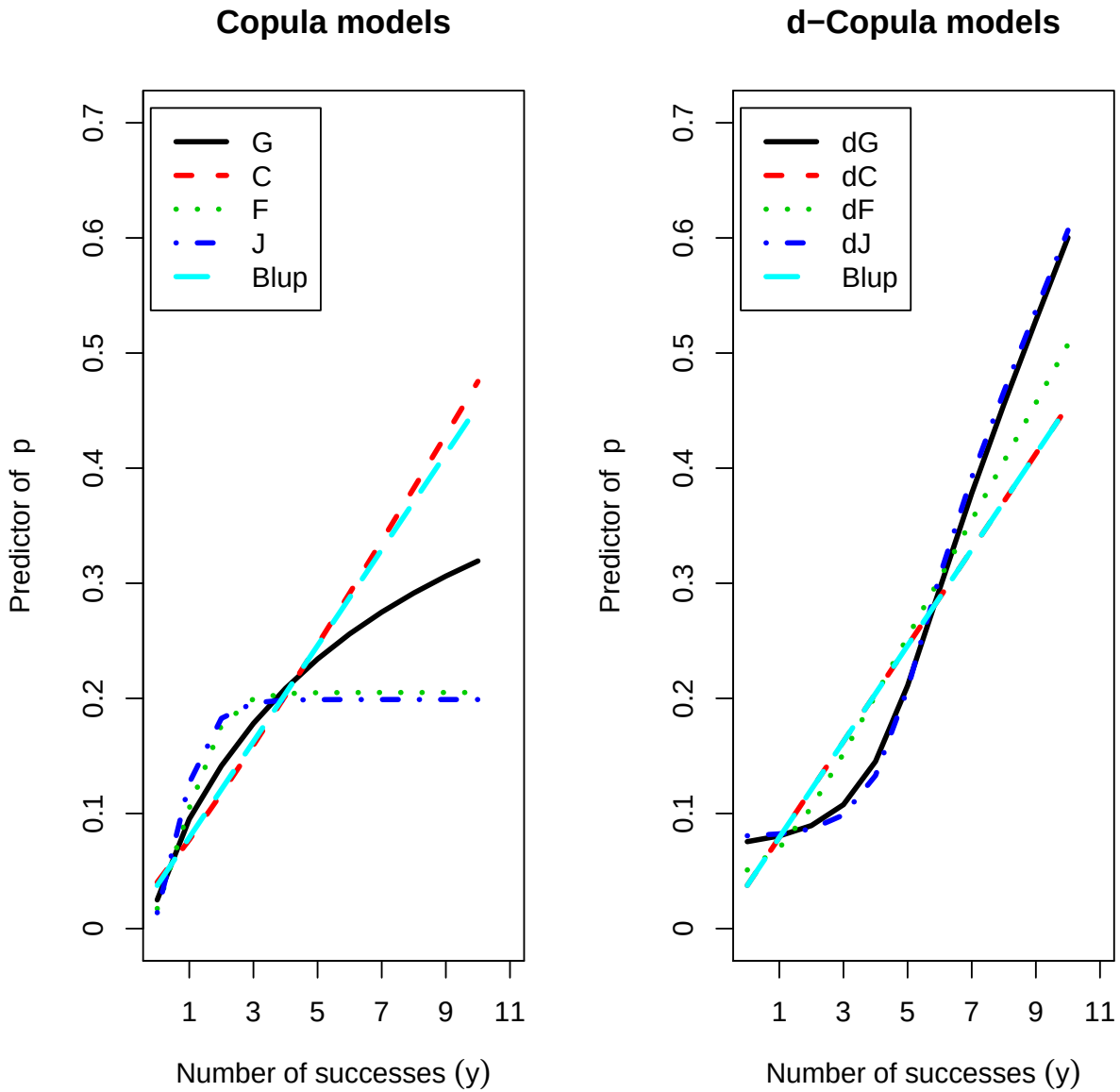
- Si $\varphi = (\alpha,\pi)$ est connu, alors le meilleur prédicteur de p_i est donné par

$$\hat{p}^{BP}(y,\varphi) = E(p \mid y,\varphi) = \psi_{\alpha,y}\{\psi_{\alpha}^{-1}(\pi)\},$$

et son EQM conditionnelle est

$$V(p|y,\varphi) = \psi_{\alpha,y}\{2\psi_{\alpha}^{-1}(\pi)\} - \left[\psi_{\alpha,y}\{\psi_{\alpha}^{-1}(\pi)\}\right]^2.$$

Comparaison : BP VS BLUP : $\pi = 0.1$ and $\rho = 0.1$



Efficacité Relative

- EQM inconditionnelle sous le modèle de copule

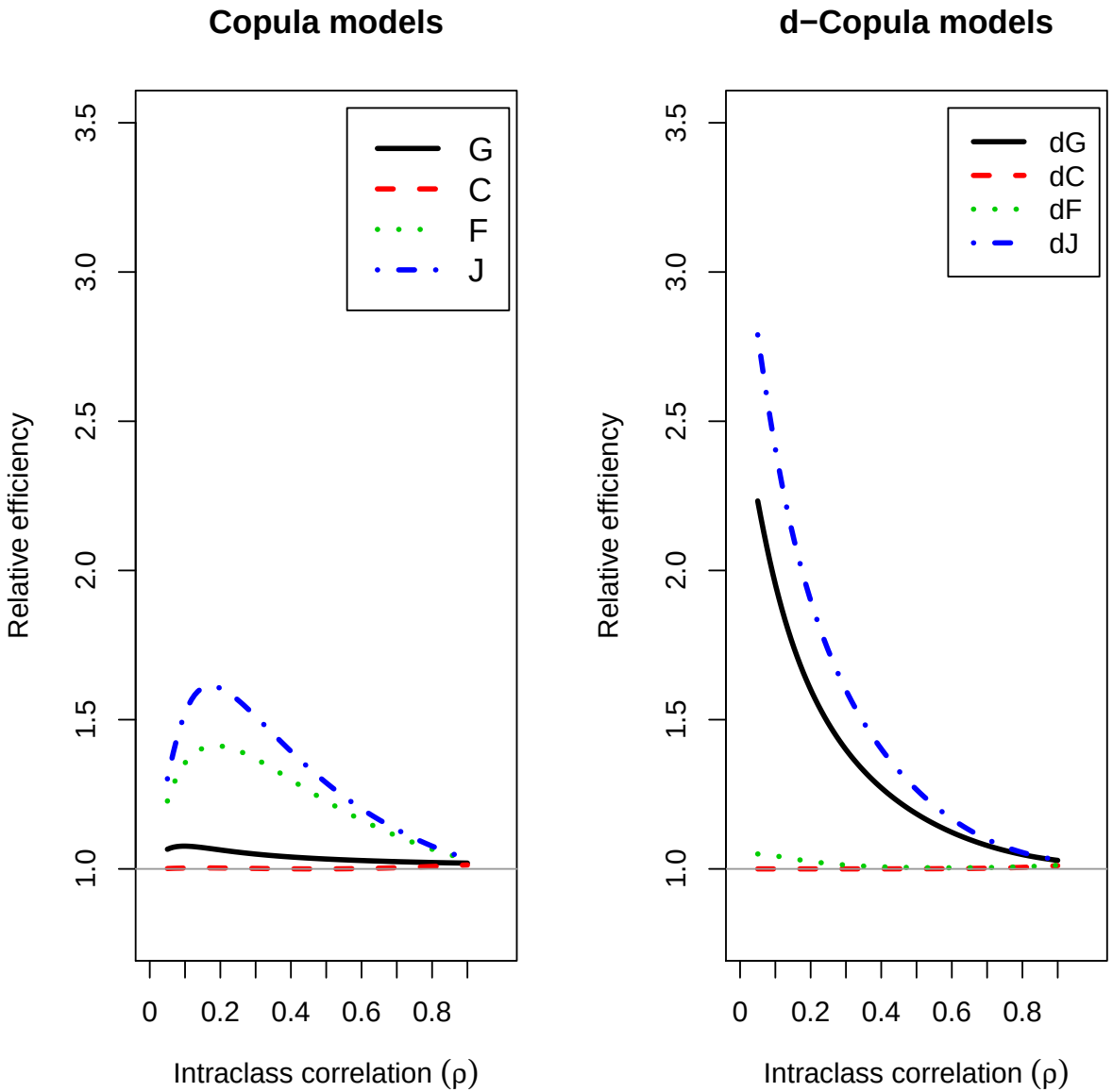
$$MSE\{\hat{p}^{BP}(y,\varphi)\} = E\{V(p|y,\varphi)\}$$
$$= \psi_{\alpha,y}\{2\psi_{\alpha}^{-1}(\pi)\} - \sum_{\ell=0}^{n}\{\hat{p}^{BP}(\ell,\varphi)\}^2 P(Y_i=\ell)$$

- Efficacité Relative (E. R.) de BP par rapport au BLUP

$$ER(\varphi) = \frac{MSE\{\hat{p}^{Blup}(y,\varphi)\}}{MSE\{\hat{p}^{BP}(y,\varphi)\}}$$

- $ER(\varphi) > 1 \quad \Rightarrow \quad$ BP plus éfficace que BLUP

E. R. : $\pi = 0.1$


Copula models
d−Copula models
Relative efficiency
Intraclass correlation (ρ)
G
C
F
J
dG
dC
dF
dJ
1.0
1.5
2.0
2.5
3.0
3.5
0
0.2
0.4
0.6
0.8

Estimations et selection d'un modèle de copules

- Les paramètres $\varphi = (\alpha, \pi)$ sont estimés par la méthode de maximun de vraisemblance à partir de $\{(y_i, n_i), i = 1, \ldots, m\}$, un échantillon tiré des m domaines.

- Pour la sélection d'un modèle de copule, on utilise une estimation du critère AIC, défini par

$$AIC = -2 \log L(\hat{\alpha}, \hat{\pi}) + 2k,$$

- $\hat{\varphi} = (\hat{\alpha}, \hat{\pi})$ est l'estimateur de φ et $L(\cdot, \cdot)$ la vraisemblance et $k = 2$

- On choisit le modèle de copule avec la petite valeur AIC

Le meilleur prédicteur empirique (EBP)

- Soit $\hat{\varphi} = (\hat{\alpha}, \hat{\pi})$ est l'estimateur du M. V. pour $\varphi = (\alpha, \pi)$
- Le meilleur prédicteur empirique de Bayes est obtenu, en remplaçant $\varphi = (\alpha, \pi)$ par son estimateur $\hat{\varphi} = (\hat{\alpha}, \hat{\pi})$ dans $\hat{p}^{BP}(y, \varphi)$

$$\hat{p}^{EBP} = \hat{p}^{BP}(y_i, \hat{\varphi}),$$

- $EQM(\hat{p}^{EBP}) = E(\hat{p}^{EBP} - p)^2$
- Une estimation de EQM peut être $V(p_i|y_i, \hat{\varphi})$, où $\mathrm{E}V(p_i|y_i, \hat{\varphi})$
- **Problème** : Ces variances ignorent la variabilité associée à l'estimation des paramètres, donc sous-estiment la vraie EQM.

L'estimation de la variabilité pour EBP

- Soit $\hat{p}_i^{EBP}$ le prédicteur empirique de p_i
- Objectif : estimation de $= E(\hat{p}_i^{EBP} - p_i)^2$, et de $E\left[(\hat{p}_i^{EBP} - p_i)^2 \mid y_i\right]$
- Pour $E(\hat{p}_i^{EBP} - p_i)^2$, nous avons utilisé la méthode du **Jackknife** proposée par Jiang and Lahiri (2002) et la méthode du **Boostrap** de Butar and Lahiri (2003)
- Pour $E\left[(\hat{p}_i^{EBP} - p_i)^2 \mid y_i\right]$, nous avons utilisé la méthode du **Jackknife** proposée par Lohr and Rao (2007)

Extension au niveau des unités : prédiction de p_{ij}

Modèle

$$p_{ij} = e^{-a_i \psi_\alpha^{-1}(\pi_{ij})}, \quad g(\pi_{ij}) = x_{ij}^t \beta,$$

La vrais. marginale pour $\phi = (\alpha, \beta)$:

$$l_i(\mathbf{y}_i|\mathbf{x}_i, \phi) = \sum_v (-1)^{\sum_{j=1}^{n_i} v_j} \psi_\alpha \left\{ \sum_{j=1}^{n_i} \psi_\alpha^{-1}\{\bar{F}_{ij}(y_{ij} + v_j)\} \right\}.$$

La T.L. conditionnelle :

$$\psi_{\phi,\mathbf{y}_i,\mathbf{x}_i}(t) = \frac{\sum_v (-1)^{\sum_{j=1}^{n_i} v_j} \psi_\alpha \left[t + \sum_{j=1}^{n_i} \psi_\alpha^{-1}\{\bar{F}_{ij}(y_{ij} + v_j)\} \right]}{l_i(\mathbf{y}_i|\mathbf{x}_i, \phi)}.$$

Estimateur de Bayes p_{ij}

$$\hat{p}_{ij,\mathbf{y}_i,\mathbf{x}_i}^{BP}(\phi) = \psi_{\phi,\mathbf{y}_i,\mathbf{x}_i}\{\psi_\alpha^{-1}(\pi_{ij})\}.$$

Variance a posteriori :

$$MSE^c\{\hat{p}_{ij,\mathbf{y}_i,\mathbf{x}_i}^{BP}(\phi)\} = \psi_{\phi,\mathbf{y}_i,\mathbf{x}_i}\{2\psi_\alpha^{-1}(\pi_{ij})\} - [\psi_{\phi,\mathbf{y}_i,\mathbf{x}_i}\{\psi_\alpha^{-1}(\pi_{ij})\}]^2$$

Simulations

- Nous examinons, à partir des échantillons Monte Carlo, la performance

  (i) des Prédicteurs empiriques EBP et linéaires EBLUP

  (ii) du critère de sélection AIC

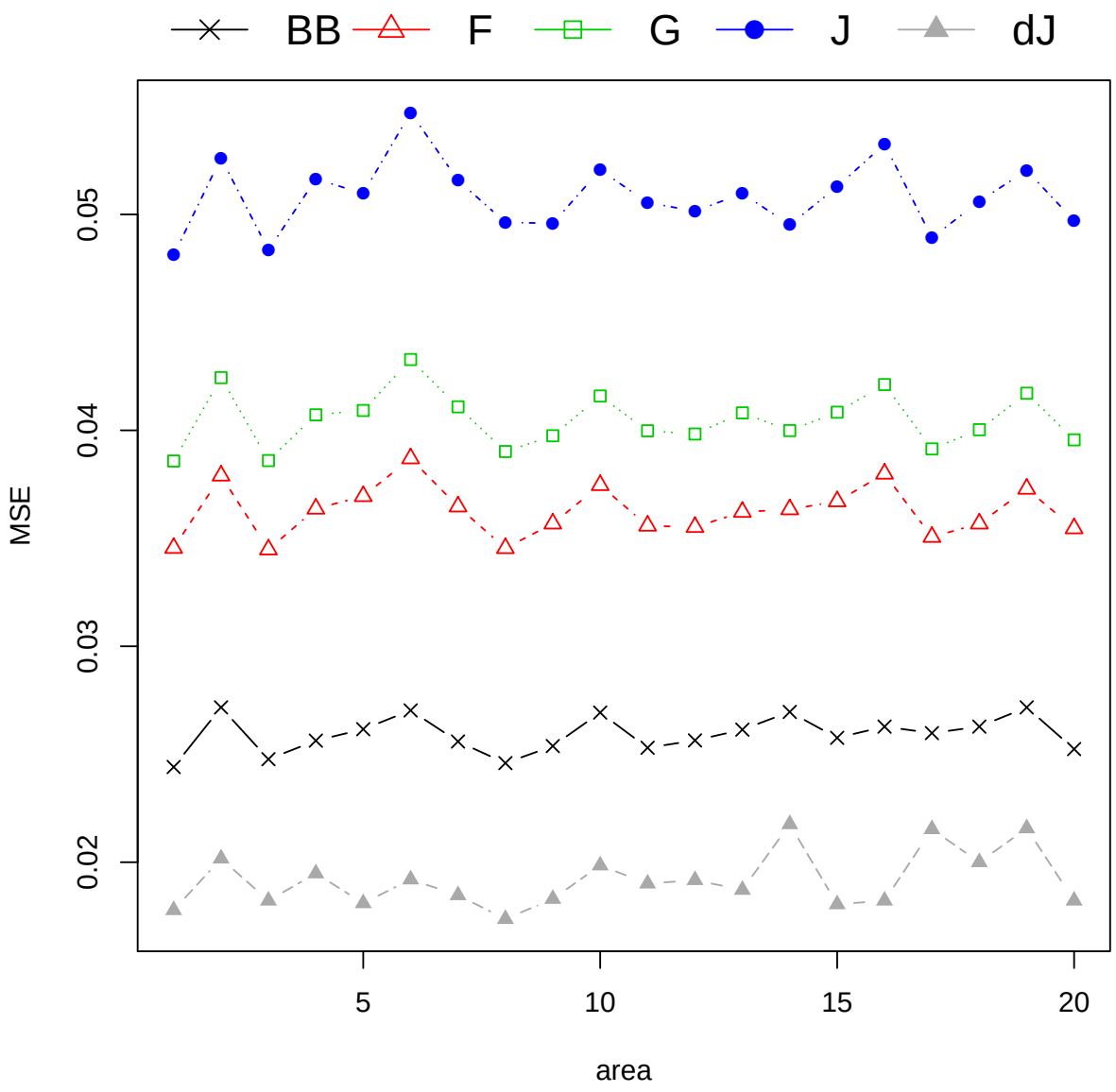
  (iii) des Estimateurs des EQMs

- **Modèle de Joe** : $p_i = 1 - e^{-a_i \phi_\alpha^{-1}(1-\pi)}$, où $a_i \sim$ **Sibuya** $(\frac{1}{1+\alpha})$

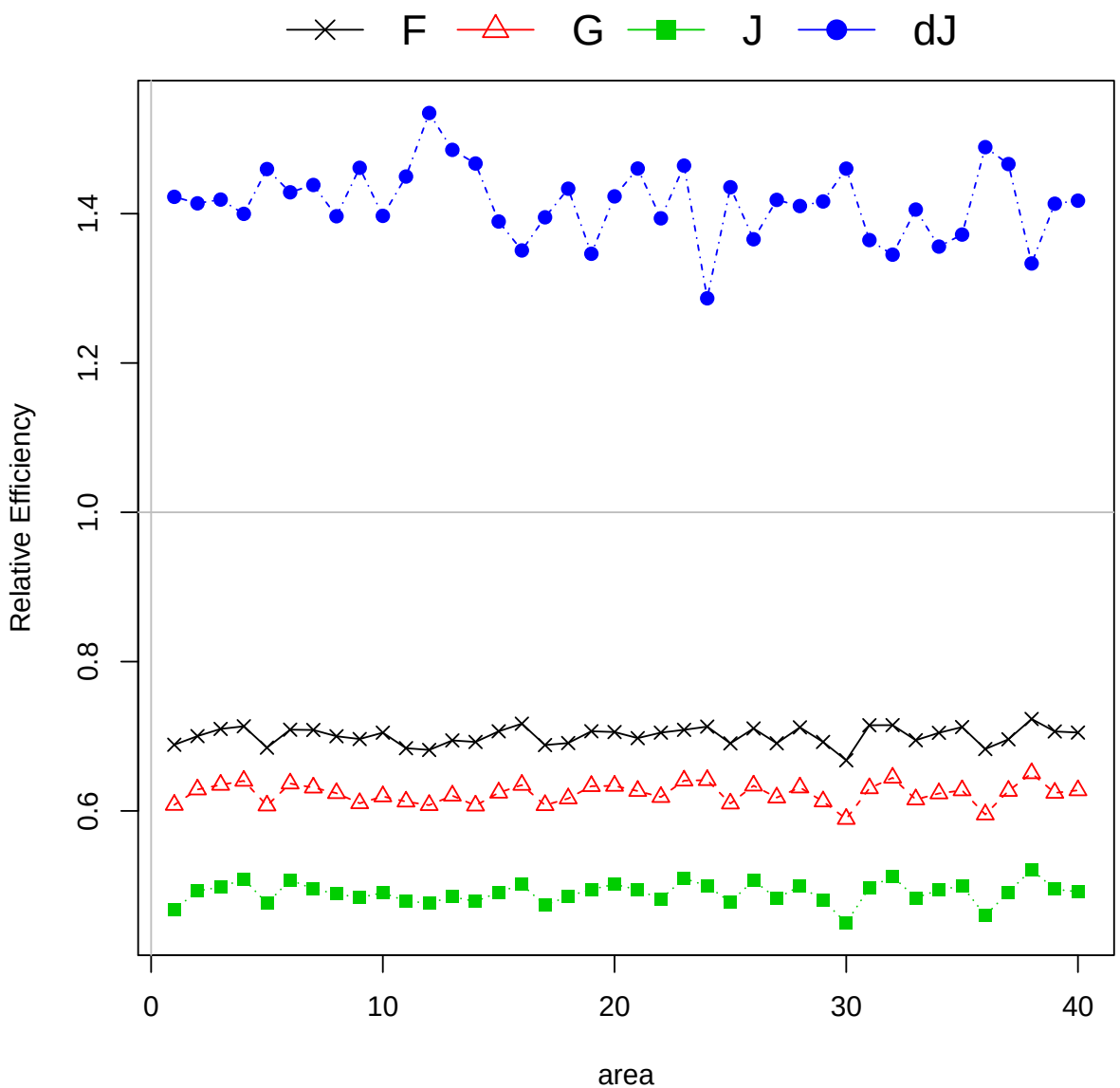
- Vrais paramètres : π=0.5 et ρ= 0.1, 0.3

- **m**=20, 40 et n_i=5

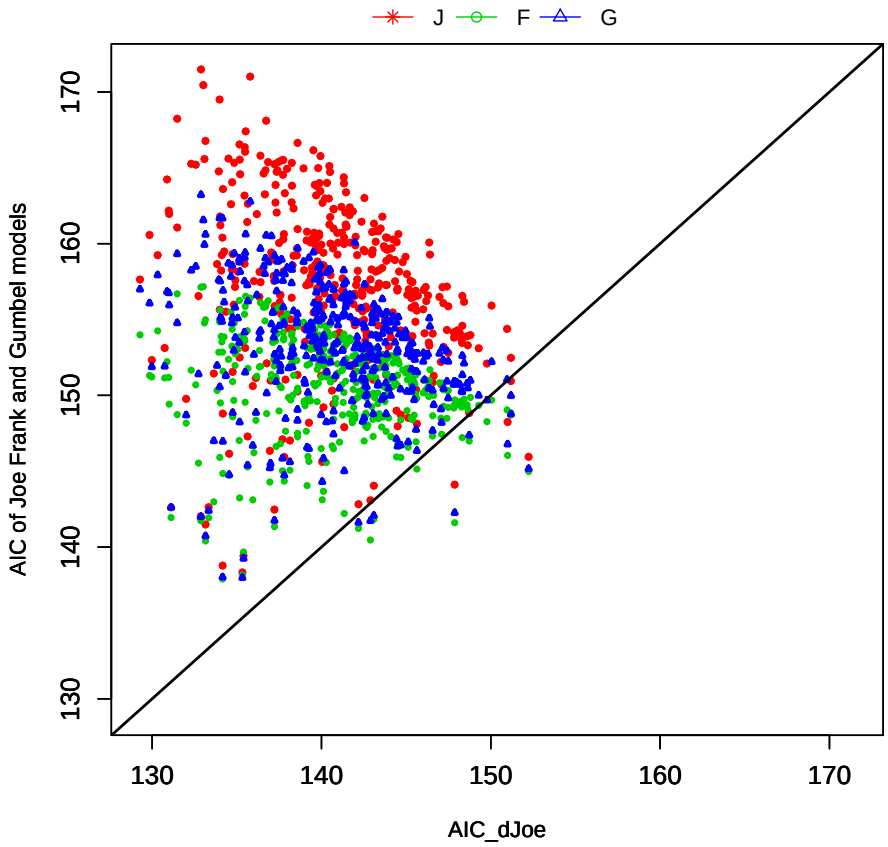
EQM empirique pour EBPs and EBLUP $\rho = 0.3$



Efficité Relative $\rho = 0.3$



Sélection d'un modèle de copule : Données dJoe avec $\rho = 0.3$



Le modèle dJoe est choisi dans plus de 95 % des cas face à ses compétiteurs.

Biais Relatif des estimateurs pour EQMs

- $\hat{MSE}_i = (1/1000) \sum_{r=1}^{1000} \hat{MSE}_i^r$
- Biais relatif : $BR_i = 100 \frac{\hat{MSE}_i - MSE_i}{MSE_i}$
- Biais Relatif moyen pour les estimateurs de EQM

inconditionnelle $E\left[(\hat{p}_i^{EBP} - p_i)^2\right]$

m	Naive	Jackknife	Bootstrap
20	-38.60	12.22	3.69
40	-24.55	4.38	2.44

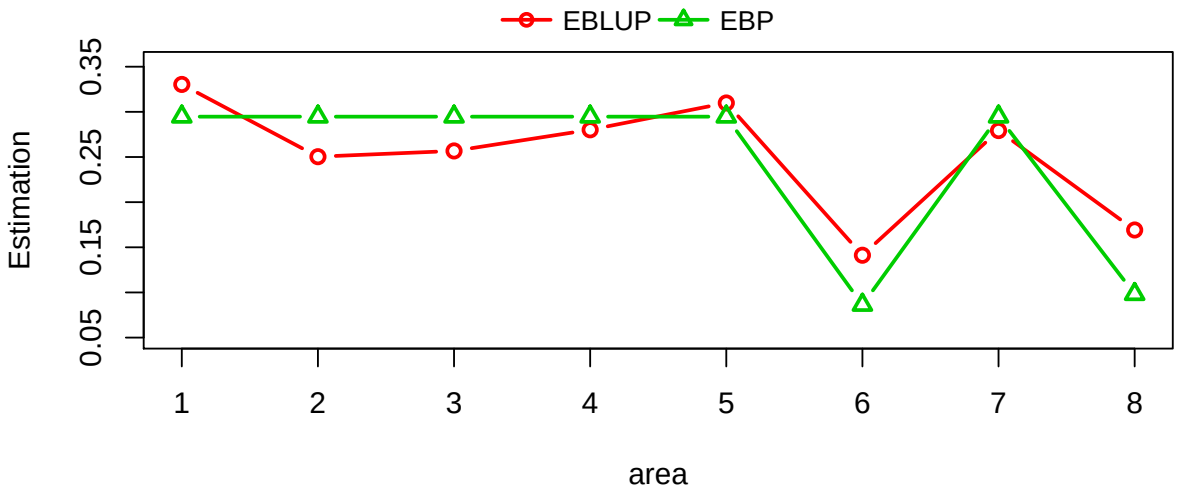
conditionnelle $E\left[(\hat{p}_i^{EBP} - p_i)^2 \mid y_i\right]$

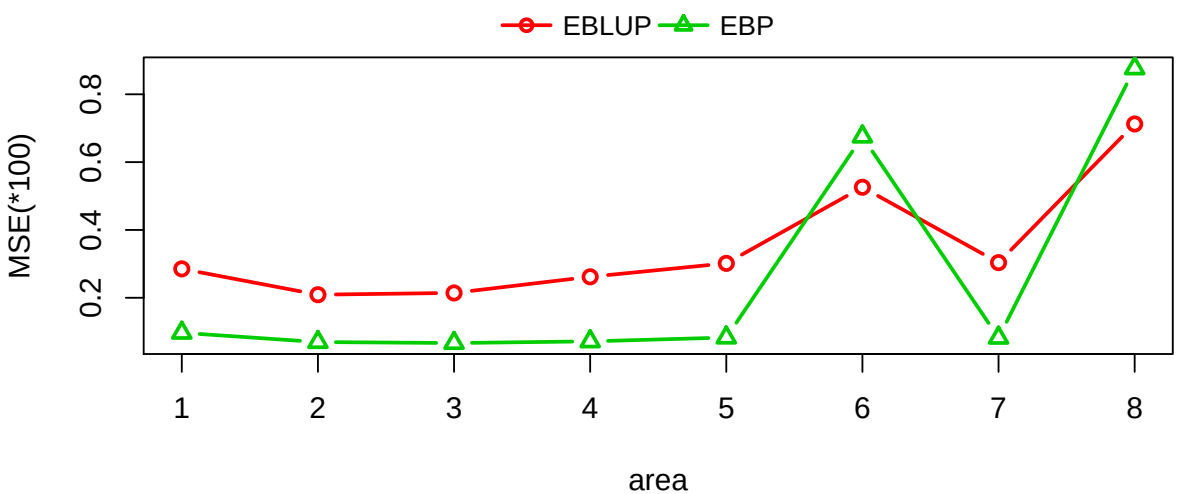
m	Naive	Jackknife
20	-39.29	-11.49
40	-24.57	-0.59

Exemple numérique : Estimation de la pauvreté

- Jeux de données (incomedata) sur le revenu des individus en Espagne.
- La variable binaire indique si le revenu d'un individu est en dessous du seuil de pauvreté (6557.143).
- L'objectif est d'estimer l'incidence de la pauvreté pour 8 provainces d'Espagne.
- L'analyse des données montre qu'une estimation de l'AIC est de 56.59 pour le modèle BB, et de 52.69, le meilleur résultat parmi les familles de copules, obtenu avec la copule de Joe
- Estimation des paramètres est : $\hat{\pi} = 0.246$, $\hat{\rho} = 0.056$ et $\hat{\tau} = 0.12$.

Estimation (En haut) et MSE (En bas)





Conclusion

1. Une classe de modèles est proposée pour analyser des données de Bernoulli échangeables.

2. Plusieurs modèles de copules sont disponible.

3. Les modèles sont utilisés pour produire des EBPs pour des proportions régionales.

4. On a des expressions explicites pour les prédicteurs empiriques.

5. Une estimation de EQM pour les prédicteurs empiriques est proposée.

6. Une comparaison entre EBP et EBLUP à partir des données simulées et réelles est aussi présentée, et a montré de belles propriés pour l'EBP.

7. L'etape de choix d'un modèle de copule est importante.

References I


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


Rao, J. N. (2003). *Small Area Estimation*. New Jersey : Wiley

FIN