Credit Scoring : biais d'échantillon ou *réintégration des refusés*

Adrien Ehrhardt

Christophe Biernacki, Vincent Vandewalle,
Philippe Heinrich, Sébastien Beben

Crédit Agricole Consumer Finance
INRIA Lille - Nord-Europe

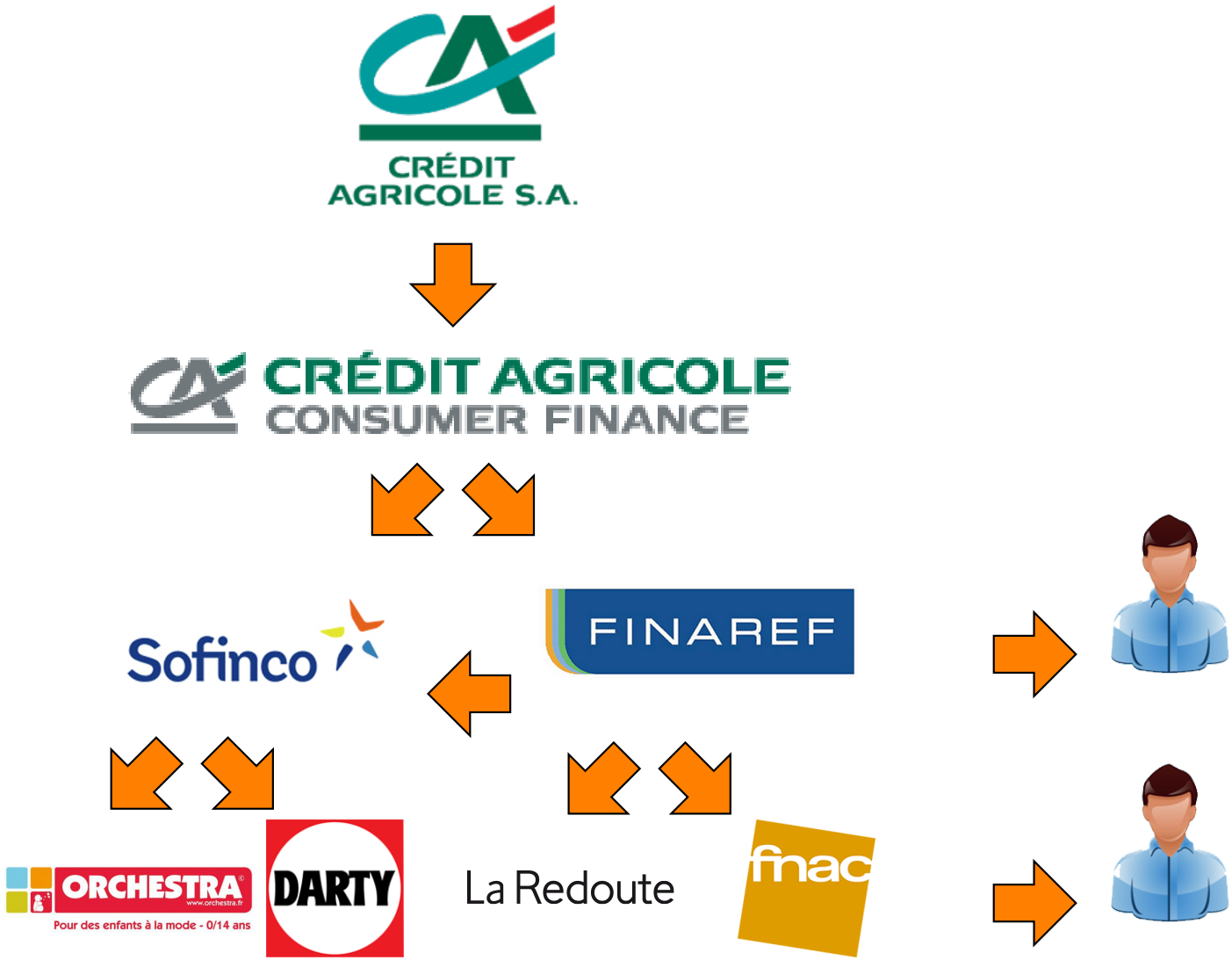
Rencontres des jeunes statisticiens 2017 - Porquerolles

3-7 avril 2017

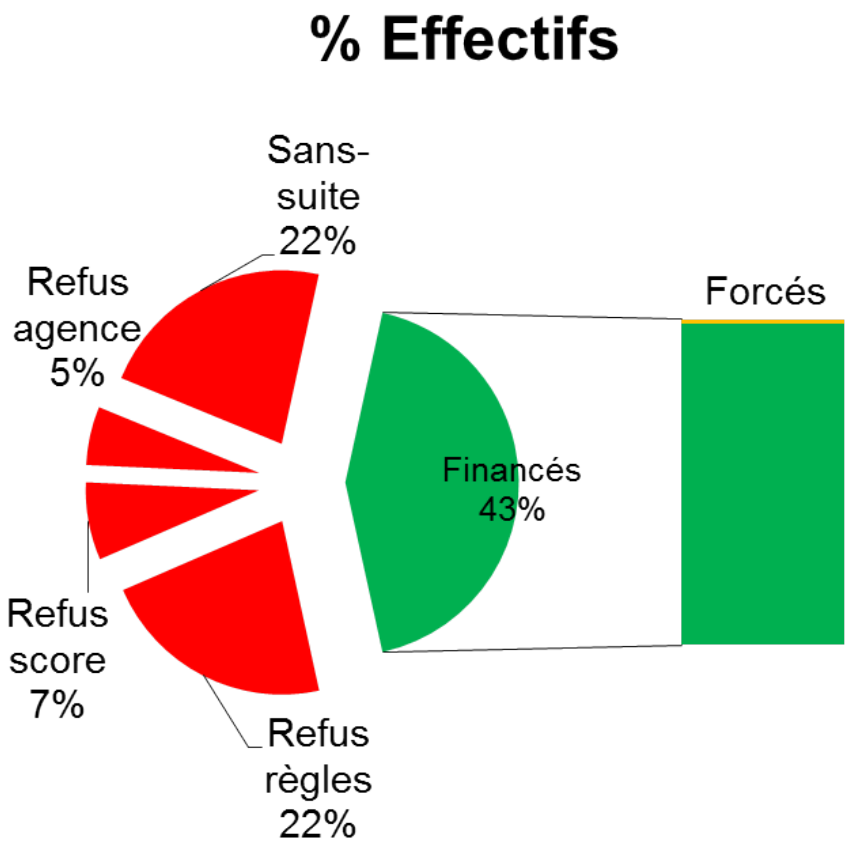
Sommaire

Contexte

Contexte : Entreprise



Contexte : Système d'acceptation



X : vecteur a. des caractéristiques
Y dans $\{0, 1\}$: remboursement
Z dans $\{\mathrm{f}, \mathrm{nf}\}$: v. a. de financement

Contexte : Credit Scoring

$$\exists\, \theta \in \mathbb{R}^{d+1} \text{ s.t. } \forall\, x,\ \ln\left(\frac{P(1|x)}{P(0|x)}\right) = \theta \cdot x$$

n clients financés ($Z = \mathrm{f}$)
m clients non financés ($Z = \mathrm{nf}$)
$\boldsymbol{x}$: *features* observées des clients
$\boldsymbol{y}$: remboursement observé

$$\boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}^{\mathrm{f}} \\ \boldsymbol{x}^{\mathrm{nf}} \end{pmatrix};\ \boldsymbol{y} = \begin{pmatrix} \boldsymbol{y}^{\mathrm{f}} \\ \boldsymbol{y}^{\mathrm{nf}} \end{pmatrix}$$

$$\underbrace{\ell(\theta; \boldsymbol{x}, \boldsymbol{y})}_{\substack{\text{vraisemblance} \\ \text{complète}}} = \left(\sum_{i=1}^{n} + \sum_{i=n+1}^{n+m}\right) \ln(p_\theta(y_i|x_i)) = \underbrace{\ell(\theta; \boldsymbol{x}^{\mathrm{f}}, \boldsymbol{y}^{\mathrm{f}})}_{\substack{\text{vraisemblance} \\ \text{observée}}} + \ell(\theta; \boldsymbol{x}^{\mathrm{nf}}, \boldsymbol{y}^{\mathrm{nf}})$$

Quel intérêt à utiliser $\boldsymbol{x}^{\mathrm{nf}}$?
Quel risque à n'utiliser que $(\boldsymbol{x}^{\mathrm{f}}, \boldsymbol{y}^{\mathrm{f}})$?

Réintégration des refusés

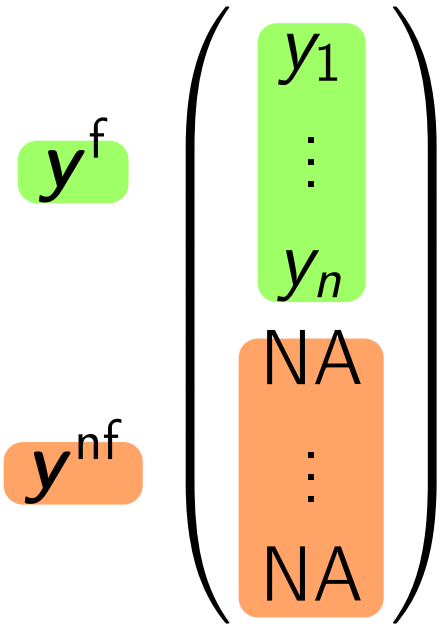
Réintégration des refusés : Exemple Fuzzy Augmentation I

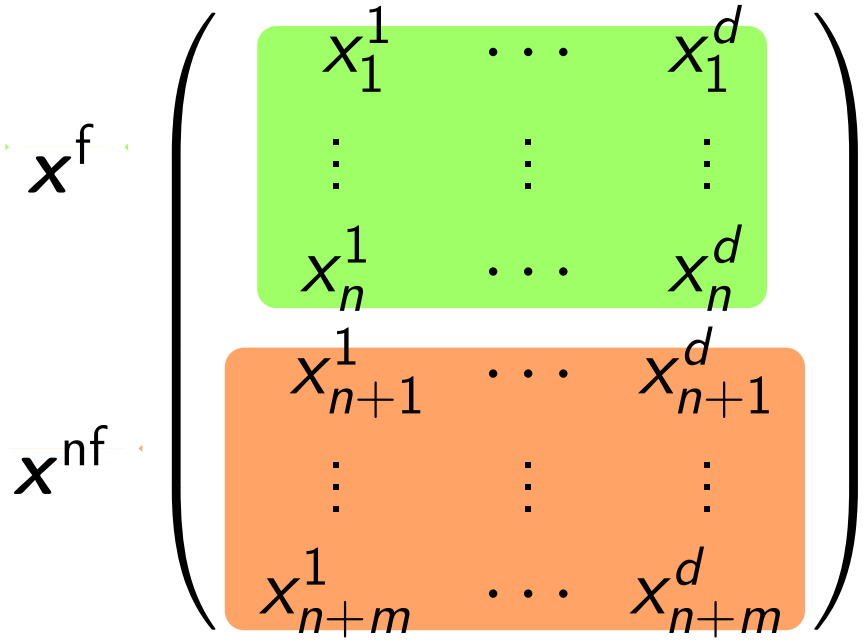
Décrit par [Nguyen, 2016].

$$\begin{array}{c} \boldsymbol{y}^{\mathrm{f}} \\ \\ \\ \\ \boldsymbol{y}^{\mathrm{nf}} \end{array} \begin{pmatrix} y_1 \\ \vdots \\ y_n \\ \mathrm{NA} \\ \vdots \\ \mathrm{NA} \end{pmatrix} \qquad \begin{array}{c} \boldsymbol{x}^{\mathrm{f}} \\ \\ \\ \\ \boldsymbol{x}^{\mathrm{nf}} \end{array} \begin{pmatrix} x_1^1 & \cdots & x_1^d \\ \vdots & \vdots & \vdots \\ x_n^1 & \cdots & x_n^d \\ x_{n+1}^1 & \cdots & x_{n+1}^d \\ \vdots & \vdots & \vdots \\ x_{n+m}^1 & \cdots & x_{n+m}^d \end{pmatrix}$$

Réintégration des refusés : Exemple Fuzzy Augmentation II

Abandon de $\boldsymbol{x}^{\mathrm{nf}}$ et construction de $\hat{\theta}^{\mathrm{f}}$.





Réintégration des refusés : Exemple Fuzzy Augmentation III

Remplacement de $\boldsymbol{y}^{\text{nf}}$ par les proba données par $\hat{\theta}^{\text{f}}$.

$$\begin{array}{c} \boldsymbol{y}^{\text{f}} \\ \\ \boldsymbol{y}^{\text{nf}} \end{array} \begin{pmatrix} y_1 \\ \vdots \\ y_n \\ p_{\hat{\theta}^{\text{f}}}(y_{n+1} = 1 | x_{n+1}) \\ \vdots \\ p_{\hat{\theta}^{\text{f}}}(y_{n+m} = 1 | x_{n+m}) \end{pmatrix} \quad \begin{array}{c} \boldsymbol{x}^{\text{f}} \\ \\ \boldsymbol{x}^{\text{nf}} \end{array} \begin{pmatrix} x_1^1 & \cdots & x_1^d \\ \vdots & \vdots & \vdots \\ x_n^1 & \cdots & x_n^d \\ x_{n+1}^1 & \cdots & x_{n+1}^d \\ \vdots & \vdots & \vdots \\ x_{n+m}^1 & \cdots & x_{n+m}^d \end{pmatrix}$$

Apprendre $\hat{\theta}^{\text{fuzzy}}$ sur le dataset résultant.

Problème : $\hat{\theta}^{\text{fuzzy}} = \hat{\theta}^{\text{f}}$.

Réintégration des refusés : Formalisation du problème

Objet d'intérêt : $p_{\text{vrai}}(y|x)$

Proposition d'un modèle : $p_\theta(y|x)$

Problème (sans surprise. . .) : estimer θ

Données :

1. Cas idéal : $\mathbf{x}^{\text{f}}, \mathbf{x}^{\text{nf}}$ et $\mathbf{y}^{\text{f}}, \mathbf{y}^{\text{nf}}$
2. Cas CACF : $\mathbf{x}^{\text{f}}$ et $\mathbf{y}^{\text{f}}$

Réintégration des refusés : Formalisation du problème

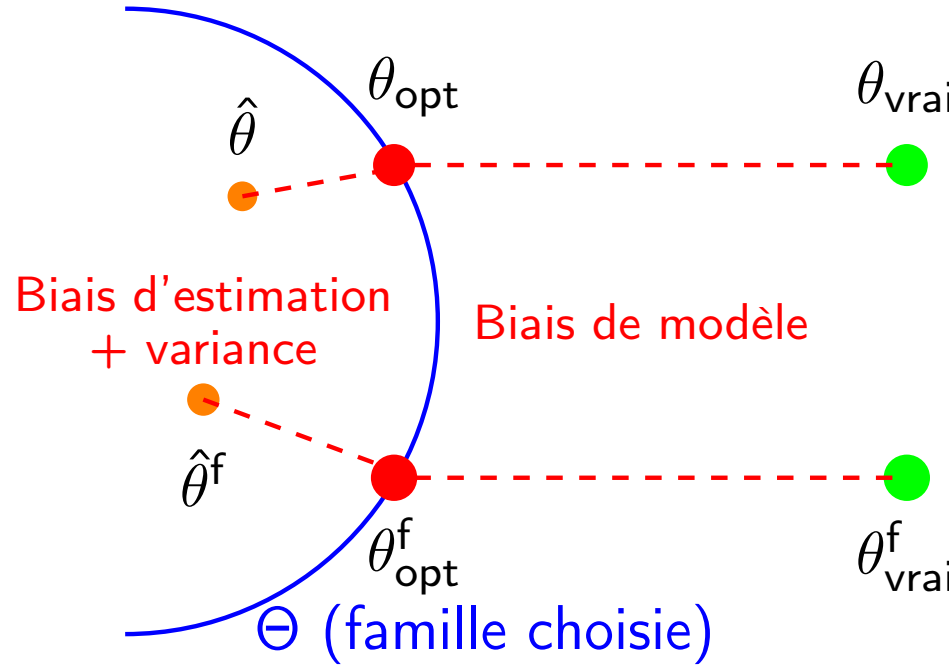
Objet d'intérêt : $p_{\text{vrai}}(y|x)$

Proposition d'un modèle : $p_\theta(y|x)$

Problème (sans surprise...) : estimer θ

Données :

1. Cas idéal : $\mathbf{x}^{\text{f}}, \mathbf{x}^{\text{nf}}$ et $\mathbf{y}^{\text{f}}, \mathbf{y}^{\text{nf}}$
2. Cas CACF : $\mathbf{x}^{\text{f}}$ et $\mathbf{y}^{\text{f}}$



Réintégration des refusés : Formalisation du problème

Estimateurs :

1. Cas idéal : avec toutes les données on obtient $\hat{\theta} = \text{argmax}_\theta \, \ell(\theta; \boldsymbol{x}, \boldsymbol{y})$
   $\sqrt{n+m}(\hat{\theta} - \theta_{\text{opt}}) \xrightarrow[n,m\to\infty]{\mathcal{L}} \mathcal{N}_{d+1}(0, \Sigma_{\theta_{\text{opt}}})$
2. Cas réaliste : avec les financés on obtient $\hat{\theta}^{\text{f}} = \text{argmax}_\theta \, \ell(\theta; \boldsymbol{x}^{\text{f}}, \boldsymbol{y}^{\text{f}})$
   $\sqrt{n}(\hat{\theta}^{\text{f}} - \theta^{\text{f}}_{\text{opt}}) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}_{d+1}(0, \Sigma^{\text{f}}_{\theta^{\text{f}}_{\text{opt}}})$

Réintégration des refusés : Formalisation du problème

Estimateurs :

1. Cas idéal : avec toutes les données on obtient $\hat{\theta} = \text{argmax}_\theta \, \ell(\theta; \boldsymbol{x}, \boldsymbol{y})$
   $\sqrt{n+m}(\hat{\theta} - \theta_{\text{opt}}) \xrightarrow[n,m\to\infty]{\mathcal{L}} \mathcal{N}_{d+1}(0, \Sigma_{\theta_{\text{opt}}})$
2. Cas réaliste : avec les financés on obtient $\hat{\theta}^{\text{f}} = \text{argmax}_\theta \, \ell(\theta; \boldsymbol{x}^{\text{f}}, \boldsymbol{y}^{\text{f}})$
   $\sqrt{n}(\hat{\theta}^{\text{f}} - \theta^{\text{f}}_{\text{opt}}) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}_{d+1}(0, \Sigma^{\text{f}}_{\theta^{\text{f}}_{\text{opt}}})$

Question 1 : propriétés asymptotiques des deux estimateurs

$$\boxed{\text{(Q1)}\ \theta_{\text{opt}} \stackrel{?}{=} \theta^{\text{f}}_{\text{opt}}}$$

$$\boxed{\text{(Q2)}\ \text{ARE}(\hat{\theta}^{\text{f}}, \hat{\theta}) = \left(\varphi \frac{|\Sigma_{\theta_{\text{opt}}}|}{|\Sigma^{\text{f}}_{\theta^{\text{f}}_{\text{opt}}}|} \right)^{\frac{1}{d+1}}}$$

(Q2a) $\dfrac{n}{n+m} \xrightarrow[n,m\to\infty]{a.s.} \varphi$ (Q2b) $\Sigma_{\theta_{\text{opt}}} \stackrel{?}{=} \Sigma^{\text{f}}_{\theta^{\text{f}}_{\text{opt}}}$

Réintégration des refusés : Données manquantes

- **MCAR** : $\forall\, x, y, z,\ p_{\text{vrai}}(z|x, y) = p_{\text{vrai}}(z)$
  $\rightarrow$ Inadapté au Credit Scoring.

Réintégration des refusés : Données manquantes

- **MCAR** : $\forall\, x, y, z,\ p_{\text{vrai}}(z|x,y) = p_{\text{vrai}}(z)$
  $\rightarrow$ Inadapté au Credit Scoring.
- **MAR** : $\forall\, x, y, z,\ p_{\text{vrai}}(z|x,y) = p_{\text{vrai}}(z|x)$
  $\rightarrow$ Le score détermine l'acceptation : $Z = \mathbb{1}_{\{\theta' X > \text{cut}\}}$.

Réintégration des refusés : Données manquantes

- **MCAR** : $\forall\, x, y, z,\ p_{\text{vrai}}(z|x,y) = p_{\text{vrai}}(z)$
  $\rightarrow$ Inadapté au Credit Scoring.
- **MAR** : $\forall\, x, y, z,\ p_{\text{vrai}}(z|x,y) = p_{\text{vrai}}(z|x)$
  $\rightarrow$ Le score détermine l'acceptation : $Z = \mathbb{1}_{\{\theta' X > \text{cut}\}}$.
- **MNAR** : $\exists\, x, y, z,\ p_{\text{vrai}}(z|x,y) \neq p_{\text{vrai}}(z|x)$
  $\rightarrow$ Influence du "feeling" des conseillers X^c.

Réintégration des refusés : Données manquantes

- **MCAR** : $\forall\, x, y, z,\ p_{\text{vrai}}(z|x,y) = p_{\text{vrai}}(z)$
  $\rightarrow$ Inadapté au Credit Scoring.
- **MAR** : $\forall\, x, y, z,\ p_{\text{vrai}}(z|x,y) = p_{\text{vrai}}(z|x)$
  $\rightarrow$ Le score détermine l'acceptation : $Z = \mathbb{1}_{\{\theta' X > \text{cut}\}}$.
- **MNAR** : $\exists\, x, y, z,\ p_{\text{vrai}}(z|x,y) \neq p_{\text{vrai}}(z|x)$
  $\rightarrow$ Influence du "feeling" des conseillers X^c.

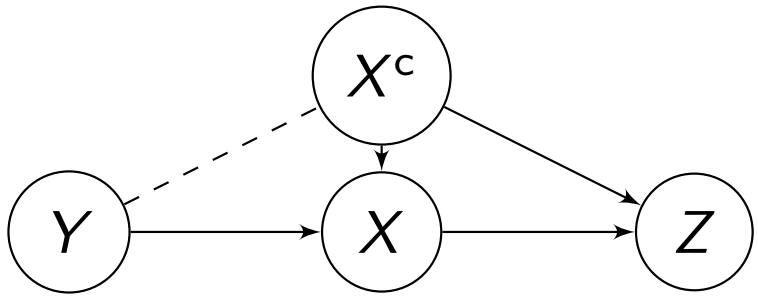


Figure – Schéma des dépendances des variables aléatoires Y, X^c, X et Z

Réintégration des refusés : Bon/mauvais modèle

- **Bon modèle** : $\exists\, \theta_{\text{vrai}}, p_{\text{vrai}}(y|x) = p_{\theta_{\text{vrai}}}(y|x)$.
  $\rightarrow$ Données réelles $\Rightarrow$ hypothèse peu probable.
- **Mauvais modèle** : θ_{opt} minimise l'ignorance sur la vraie loi.
  $\rightarrow$ Utilisation de la régression logistique pour sa robustesse à la misspecification.

Réintégration des refusés : Bon/mauvais modèle

- **Bon modèle** : $\exists\, \theta_{\text{vrai}}, p_{\text{vrai}}(y|x) = p_{\theta_{\text{vrai}}}(y|x)$.
  $\rightarrow$ Données réelles $\Rightarrow$ hypothèse peu probable.
- **Mauvais modèle** : θ_{opt} minimise l'ignorance sur la vraie loi.
  $\rightarrow$ Utilisation de la régression logistique pour sa robustesse à la misspecification.

$p_\theta(y\|x,z)$ \ $p_{\text{vrai}}(z\|x)$	MCAR	MAR	MNAR
Bon	$\theta^{\text{f}}_{\text{opt}} = \theta_{\text{opt}}$ $\Sigma^{\text{f}}_{\theta^{\text{f}}_{\text{opt}}} = \Sigma_{\theta_{\text{opt}}}$	$\theta^{\text{f}}_{\text{opt}} = \theta_{\text{opt}}$ $\Sigma^{\text{f}}_{\theta^{\text{f}}_{\text{opt}}} \neq \Sigma_{\theta_{\text{opt}}}$	$\theta^{\text{f}}_{\text{opt}} \neq \theta_{\text{opt}}$ $\Sigma^{\text{f}}_{\theta^{\text{f}}_{\text{opt}}} \neq \Sigma_{\theta_{\text{opt}}}$
Mauvais		$\theta^{\text{f}}_{\text{opt}} \neq \theta_{\text{opt}}$ $\Sigma^{\text{f}}_{\theta^{\text{f}}_{\text{opt}}} \neq \Sigma_{\theta_{\text{opt}}}$	

Table – Réponses à (Q1) et (Q2b) selon le mécanisme des données manquantes et l'hypothèse du bon modèle

Réintégration des refusés : Améliorer l'estimateur $\hat{\theta}^{\mathrm{f}}$?

Question 2 : Comment améliorer $\hat{\theta}^{\mathrm{f}}$?

Réintégration des refusés : Améliorer l'estimateur $\hat{\theta}^{\mathrm{f}}$?

Question 2 : Comment améliorer $\hat{\theta}^{\mathrm{f}}$?

Leviers :

- Changer le modèle (i.e. l'espace Θ),

Réintégration des refusés : Améliorer l'estimateur $\hat{\theta}^{\mathrm{f}}$?

Question 2 : Comment améliorer $\hat{\theta}^{\mathrm{f}}$?

Leviers :

- Changer le modèle (i.e. l'espace Θ),
- Modéliser la sélection (i.e. $p_\alpha(z|x,y)$),

Réintégration des refusés : Améliorer l'estimateur $\hat{\theta}^{\mathrm{f}}$?

Question 2 : Comment améliorer $\hat{\theta}^{\mathrm{f}}$?

Leviers :

- Changer le modèle (i.e. l'espace Θ),
- Modéliser la sélection (i.e. $p_\alpha(z|x,y)$),
- Utiliser $\boldsymbol{x}^{\mathrm{nf}}$.

Réintégration des refusés : Améliorer l'estimateur $\hat{\theta}^{\mathrm{f}}$?

Question 2 : Comment améliorer $\hat{\theta}^{\mathrm{f}}$?

Leviers :

- Changer le modèle (i.e. l'espace Θ),
- Modéliser la sélection (i.e. $p_\alpha(z|x,y)$),
- Utiliser $\boldsymbol{x}^{\mathrm{nf}}$.

Moyen "naturel" de faire les trois : le modèle génératif $p_\theta(y|x)$, $p_\beta(x)$ et $p_\gamma(z|x,y)$ où θ, β et γ sont fonctionnellement dépendants (en général).

$$\hat{\theta}^{\mathrm{gen}} = \operatorname{argmax} \ell(\theta; \boldsymbol{x}, \boldsymbol{y}^{\mathrm{f}}) = \ell(\theta; \boldsymbol{x}^{\mathrm{f}}, \boldsymbol{y}^{\mathrm{f}}) + \ell(\theta; \boldsymbol{x}^{\mathrm{nf}})$$

généralement obtenu par algorithme EM.

Réintégration des refusés : Améliorer l'estimateur $\hat{\theta}^{\mathsf{f}}$?

Question 2 : Comment améliorer $\hat{\theta}^{\mathsf{f}}$?

Leviers :

- ~~Changer le modèle (i.e. l'espace Θ)~~ régression logistique,
- Modéliser la sélection (i.e. $p_\alpha(z|x,y)$),
- Utiliser $\boldsymbol{x}^{\mathsf{nf}}$.

Moyen "naturel" de faire les trois : le modèle génératif $p_\theta(y|x)$, $p_\beta(x)$ et $p_\gamma(z|x,y)$ où θ, β et γ sont fonctionnellement dépendants (en général).

$$\hat{\theta}^{\mathsf{gen}} = \operatorname{argmax} \ell(\theta; \boldsymbol{x}, \boldsymbol{y}^{\mathsf{f}}) = \ell(\theta; \boldsymbol{x}^{\mathsf{f}}, \boldsymbol{y}^{\mathsf{f}}) + \ell(\theta; \boldsymbol{x}^{\mathsf{nf}})$$

généralement obtenu par algorithme EM.

Réintégration des refusés : Améliorer l'estimateur $\hat{\theta}^{\mathsf{f}}$?

Question 2 : Comment améliorer $\hat{\theta}^{\mathsf{f}}$?

Leviers :

- ~~Changer le modèle (i.e. l'espace Θ)~~ régression logistique,
- ~~Modéliser la sélection (i.e. $p_\alpha(z|x,y)$)~~ relève de la croyance,
- Utiliser $\boldsymbol{x}^{\mathsf{nf}}$.

Moyen "naturel" de faire les trois : le modèle génératif $p_\theta(y|x)$, $p_\beta(x)$ et $p_\gamma(z|x,y)$ où θ, β et γ sont fonctionnellement dépendants (en général).

$$\hat{\theta}^{\mathsf{gen}} = \operatorname{argmax} \ell(\theta; \boldsymbol{x}, \boldsymbol{y}^{\mathsf{f}}) = \ell(\theta; \boldsymbol{x}^{\mathsf{f}}, \boldsymbol{y}^{\mathsf{f}}) + \ell(\theta; \boldsymbol{x}^{\mathsf{nf}})$$

généralement obtenu par algorithme EM.

Réintégration des refusés : Réinterprétation des méthodes

Reclassification
([Viennet et al., 2006, Banasik and Crook, 2007, Guizani et al., 2013]) :

$$(\hat{\theta}^{\text{CEM}}, \hat{\boldsymbol{y}}^{\text{nf}}) = \underset{\theta, \boldsymbol{y}^{\text{nf}}}{\text{argmax}}\, \ell(\theta; \boldsymbol{x}, \boldsymbol{y}^{\text{f}}, \boldsymbol{y}^{\text{nf}}) = \ell(\theta; \boldsymbol{x}^{\text{f}}, \boldsymbol{y}^{\text{f}}) + \sum_{i=n+1}^{m+n} \ln(p_\theta(\hat{y}_i | x_i)).$$

où $\hat{y}_i = \text{argmax}_{y_i}\, p_{\hat{\theta}^{\text{f}}}(y_i | x_i)$.

C'est une tentative de "mimer" le modèle génératif par l'utilisation d'un algorithme Classification-EM.

Problème : produit un estimateur inconsistent, aucune garantie d'amélioration.

Réintégration des refusés : Réinterprétation des méthodes

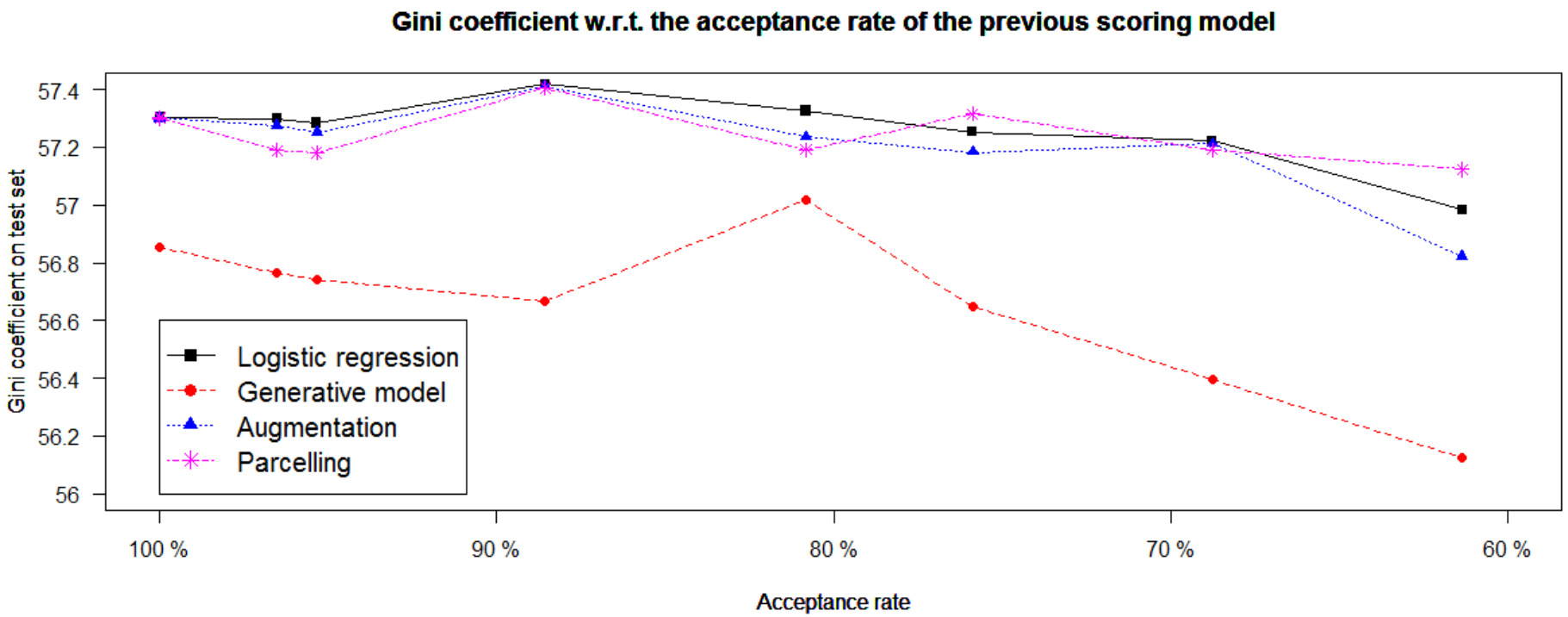
Augmentation ([Viennet et al., 2006, Banasik and Crook, 2007, Guizani et al., 2013, Nguyen, 2016]) : critère issu de l'*Importance Sampling* permettant de corriger le biais de modèle **mais** algorithme insatisfaisant.

Twins (méthode interne CACF) : le critère conduit à $\hat{\theta}^{\text{twins}} = \hat{\theta}^{\text{f}}$.

Parcelling
([Viennet et al., 2006, Banasik and Crook, 2007, Guizani et al., 2013]) : hypothèses MNAR invérifiables.

Génératif : biais de modèle important.

Réintégration des refusés : Résultats expérimentaux



Conclusion

Conclusion

- **Fuzzy Augmentation**, **Reclassification** et **Twins** ont été écartées.
- Le choix d'une (autre) méthode relève de la croyance.
- La formalisation a permis de clotûrer un débat resté empirique jusqu'alors.
- Article en préparation.

Conclusion

- **Fuzzy Augmentation**, **Reclassification** et **Twins** ont été écartées.
- Le choix d'une (autre) méthode relève de la croyance.
- La formalisation a permis de clotûrer un débat resté empirique jusqu'alors.
- Article en préparation.
- **Sujet en cours** permettant de limiter le biais de modèle : discrétisation de variables continues, regroupement de modalités pour variables qualitatives et croisements automatiques de variables.
- **Sujet à venir** : augmentation du nombre de prédicteurs permettant d'abaisser l'erreur de Bayes.

Merci pour votre attention !

Questions ?


Banasik, J. and Crook, J. (2007).
Reject inference, augmentation, and sample selection.
European Journal of Operational Research, 183(3) :1582–1594.


Guizani, A., Souissi, B., Ammou, S. B., and Saporta, G. (2013).
Une comparaison de quatre techniques d'inférence des refusés dans le processus d'octroi de crédit.
In 45 emes Journ'ees de statistique, page pp.


Nguyen, H. T. (2016).
Reject inference in application scorecards : evidence from France.
Technical report, University of Paris West-Nanterre la Défense, EconomiX.


Viennet, E., Soulié, F. F., and Rognier, B. (2006).
Evaluation de techniques de traitement des refusés pour l'octroi de crédit.
arXiv preprint cs/0607048.