Reject Inference in *Credit Scoring*

Adrien Ehrhardt

Christophe Biernacki, Vincent Vandewalle,
Philippe Heinrich, Sébastien Beben

Crédit Agricole Consumer Finance
Inria Lille - Nord-Europe

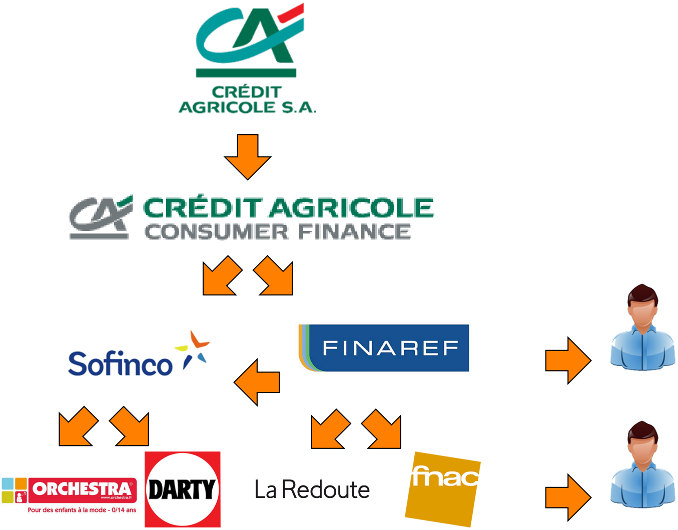
Journées de Statistique 2017 - Avignon

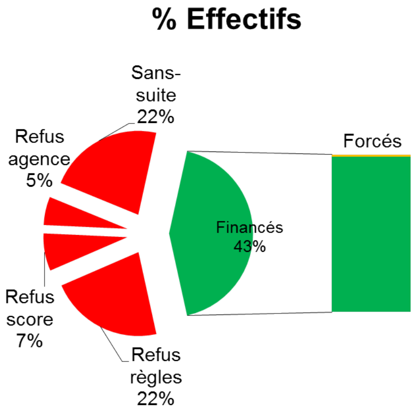
29 mai 2017

Table of contents

Introduction

Introduction: Data generation process





X : random vector of a client's characteristics
$Y \in \{0,1\}$: repayment performance
$Z \in \{\mathrm{f}, \mathrm{nf}\}$: v. a. de financement

n financed clients ($Z = \mathrm{f}$)
m not financed clients ($Z = \mathrm{nf}$)
$\boldsymbol{x}$: observed features of clients
$\boldsymbol{y}$: clients' repayment

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

Introduction: Credit Scoring in practice

"Classical" logistic regression:

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

Parameter estimation:

$$\underbrace{\ell(\theta; \boldsymbol{x}, \boldsymbol{y})}_{\substack{\text{complete}\\ \text{likelihood}}} = \sum_{i=1}^{n} \ln(p_\theta(y_i|x_i)) + \sum_{i=n+1}^{n+m} \ln(p_\theta(y_i|x_i))$$

$$= \underbrace{\ell(\theta; \boldsymbol{x}^{\text{f}}, \boldsymbol{y}^{\text{f}})}_{\substack{\text{observed}\\ \text{likelihood}}} + \underbrace{\ell(\theta; \boldsymbol{x}^{\text{nf}}, \color{red}{\boldsymbol{y}^{\text{nf}}})}_{\color{red}{\textit{unknown}}}$$

Introduction: Credit Scoring in practice

"Classical" logistic regression:

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

Parameter estimation:

$$\underbrace{\ell(\theta; \boldsymbol{x}, \boldsymbol{y})}_{\substack{\text{complete}\\ \text{likelihood}}} = \sum_{i=1}^{n} \ln(p_\theta(y_i|x_i)) + \sum_{i=n+1}^{n+m} \ln(p_\theta(y_i|x_i))$$

$$= \underbrace{\ell(\theta; \boldsymbol{x}^{\text{f}}, \boldsymbol{y}^{\text{f}})}_{\substack{\text{observed}\\ \text{likelihood}}} + \underbrace{\ell(\theta; \boldsymbol{x}^{\text{nf}}, \boldsymbol{y}^{\text{nf}})}_{\textit{unknown}}$$

Sample selection problems

Why and how use $\boldsymbol{x}^{\text{nf}}$?

What are the consequences of using "only" $(\boldsymbol{x}^{\text{f}}, \boldsymbol{y}^{\text{f}})$?

Reject Inference methods

Reject Inference methods: Fuzzy Augmentation I

Fuzzy Augmentation can be found, among others, in [Nguyen, 2016].

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

Reject Inference methods: Fuzzy Augmentation II

<u>Step 1:</u> Discard $\boldsymbol{x}^{\text{nf}}$ and estimate $\hat{\theta}^{\text{f}} = \text{argmax}_\theta \, \ell(\theta; \boldsymbol{x}^{\text{f}}, \boldsymbol{y}^{\text{f}})$.

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

Reject Inference methods: Fuzzy Augmentation III

Step 2: Impute $\boldsymbol{y}^{\text{nf}}$ with their estimation given by $\hat{\theta}^{\text{f}}$.

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

Step 3: estimate $\hat{\theta}^{\text{fuzzy}} = \text{argmax}_\theta \, \ell(\theta; \boldsymbol{x}, \boldsymbol{y}^{\text{f}}, \hat{\boldsymbol{y}}^{\text{nf}})$.

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

Reject Inference methods: Maximum likelihood estimation

"Classical" estimation in the *Credit Scoring* field

$$\hat{\theta}^{\text{f}} = \underset{\theta}{\operatorname{argmax}}\, \ell(\theta; \boldsymbol{x}^{\text{f}}, \boldsymbol{y}^{\text{f}}).$$

"Oracle" estimation knowing $\boldsymbol{y}^{\text{nf}}$

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, \ell(\theta; \boldsymbol{x}, \boldsymbol{y}).$$

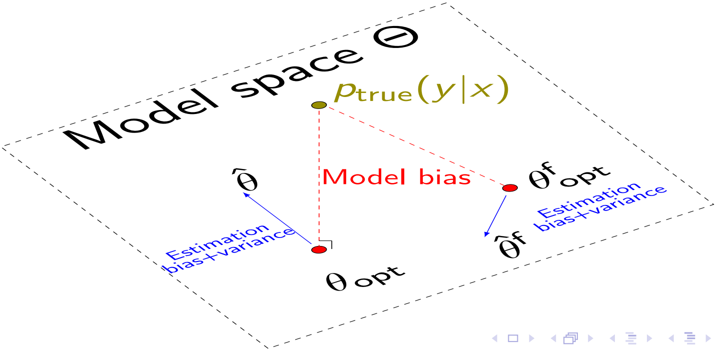
Reject Inference methods: Maximum likelihood estimation

"Classical" estimation in the *Credit Scoring* field

$$\hat{\theta}^{\text{f}} = \underset{\theta}{\text{argmax}}\, \ell(\theta; \boldsymbol{x}^{\text{f}}, \boldsymbol{y}^{\text{f}}).$$

"Oracle" estimation knowing $\boldsymbol{y}^{\text{nf}}$

$$\hat{\theta} = \underset{\theta}{\text{argmax}}\, \ell(\theta; \boldsymbol{x}, \boldsymbol{y}).$$



Reject Inference methods: Maximum likelihood estimation

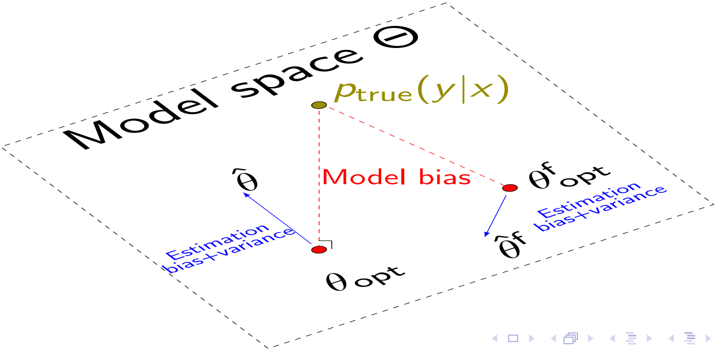
"Classical" estimation in the *Credit Scoring* field

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

"Oracle" estimation knowing $\boldsymbol{y}^{\mathrm{nf}}$

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, \ell(\theta; \boldsymbol{x}, \boldsymbol{y}).$$

Asymptotics ?



Reject Inference methods: What is at stake?

Estimators :

1. "Oracle": $\sqrt{n+m}(\hat{\theta} - \theta_{\text{opt}}) \xrightarrow[n,m \to \infty]{\mathcal{L}} \mathcal{N}_{d+1}(0, \Sigma_{\theta_{\text{opt}}})$
2. Current methodology: $\sqrt{n}(\hat{\theta}^{\text{f}} - \theta^{\text{f}}_{\text{opt}}) \xrightarrow[n \to \infty]{\mathcal{L}} \mathcal{N}_{d+1}(0, \Sigma^{\text{f}}_{\theta^{\text{f}}_{\text{opt}}})$

Reject Inference methods: What is at stake?

Estimators :

1. "Oracle": $\sqrt{n+m}(\hat{\theta} - \theta_{\text{opt}}) \xrightarrow[n,m\to\infty]{\mathcal{L}} \mathcal{N}_{d+1}(0, \Sigma_{\theta_{\text{opt}}})$
2. Current methodology: $\sqrt{n}(\hat{\theta}^{\text{f}} - \theta^{\text{f}}_{\text{opt}}) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}_{d+1}(0, \Sigma^{\text{f}}_{\theta^{\text{f}}_{\text{opt}}})$

Reject Inference methods: What is at stake?

Estimators :

1. "Oracle": $\sqrt{n+m}(\hat{\theta} - \theta_{\text{opt}}) \xrightarrow[n,m\to\infty]{\mathcal{L}} \mathcal{N}_{d+1}(0, \Sigma_{\theta_{\text{opt}}})$
2. Current methodology: $\sqrt{n}(\hat{\theta}^{\text{f}} - \theta^{\text{f}}_{\text{opt}}) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}_{d+1}(0, \Sigma^{\text{f}}_{\theta^{\text{f}}_{\text{opt}}})$

Reject Inference methods: What is at stake?

Estimators :

1. "Oracle": $\sqrt{n+m}(\hat{\theta} - \theta_{\text{opt}}) \xrightarrow[n,m \to \infty]{\mathcal{L}} \mathcal{N}_{d+1}(0, \Sigma_{\theta_{\text{opt}}})$
2. Current methodology: $\sqrt{n}(\hat{\theta}^{\text{f}} - \theta^{\text{f}}_{\text{opt}}) \xrightarrow[n \to \infty]{\mathcal{L}} \mathcal{N}_{d+1}(0, \Sigma^{\text{f}}_{\theta^{\text{f}}_{\text{opt}}})$

Reject Inference methods: What is at stake?

Estimators :

1. "Oracle": $\sqrt{n+m}(\hat{\theta} - \theta_{\text{opt}}) \xrightarrow[n,m\to\infty]{\mathcal{L}} \mathcal{N}_{d+1}(0, \Sigma_{\theta_{\text{opt}}})$
2. Current methodology: $\sqrt{n}(\hat{\theta}^{\text{f}} - \theta^{\text{f}}_{\text{opt}}) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}_{d+1}(0, \Sigma^{\text{f}}_{\theta^{\text{f}}_{\text{opt}}})$

Question 1 : asymptotics of the estimators

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

$$\text{(Q2)}\ \Sigma_{\theta_{\text{opt}}} \stackrel{?}{=} \Sigma^{\text{f}}_{\theta^{\text{f}}_{\text{opt}}}$$

Reject Inference methods: Missingness mechanism

- **MAR** : $\forall\, x, y, z,\; p_{\text{true}}(z|x, y) = p_{\text{true}}(z|x)$
  $\rightarrow$ Acceptance is determined by the score : $Z = \mathbb{1}_{\{\theta' X > \text{cut}\}}$.
- **MNAR** : $\exists\, x, y, z,\; p_{\text{true}}(z|x, y) \neq p_{\text{true}}(z|x)$
  $\rightarrow$ Operators' "feeling" X^{c} influence the acceptance.

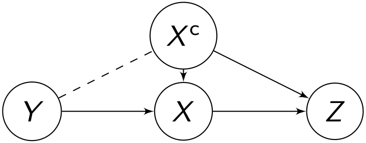


Figure: Dependencies between random variables Y, X^{c}, X and Z

Reject Inference methods: Model specification

- **Well-specified model** : $\exists\, \theta_{\text{true}}, p_{\text{true}}(y|x) = p_{\theta_{\text{true}}}(y|x)$.
  $\rightarrow$ With real data $\Rightarrow$ hypothesis unlikely to be true.
- **Misspecified model** : θ_{opt} is the "best" in the Θ family.
  $\rightarrow$ Logistic regression commonly used for its robustness to misspecification.

$p_\theta(y\|x)$ \ $p_{\text{true}}(z\|x,y)$	MAR	MNAR
Well specified	$\theta^{\text{f}}_{\text{opt}} = \theta_{\text{opt}}$ $\Sigma^{\text{f}}_{\theta^{\text{f}}_{\text{opt}}} \neq \Sigma_{\theta_{\text{opt}}}$	$\theta^{\text{f}}_{\text{opt}} \neq \theta_{\text{opt}}$ $\Sigma^{\text{f}}_{\theta^{\text{f}}_{\text{opt}}} \neq \Sigma_{\theta_{\text{opt}}}$
Misspecified	$\theta^{\text{f}}_{\text{opt}} \neq \theta_{\text{opt}}$ $\Sigma^{\text{f}}_{\theta^{\text{f}}_{\text{opt}}} \neq \Sigma_{\theta_{\text{opt}}}$	

Table: (Q1) and (Q2) w.r.t. model specification and missingness mechanism

Question 2: How to construct a better estimator than $\hat{\theta}^{\text{f}}$?

Reject Inference methods: How to use $\boldsymbol{x}^{\text{nf}}$?

Question 2: How to construct a better estimator than $\hat{\theta}^{\text{f}}$?

Scope for action:

- Change model space Θ,

Reject Inference methods: How to use $\boldsymbol{x}^{\text{nf}}$?

Question 2: How to construct a better estimator than $\hat{\theta}^{\text{f}}$?

Scope for action:

- Change model space Θ,
- Model acceptance/rejection process (i.e. $p_\gamma(z|x, y)$),

Reject Inference methods: How to use $\boldsymbol{x}^{\text{nf}}$?

Question 2: How to construct a better estimator than $\hat{\theta}^{\text{f}}$?

Scope for action:

- Change model space Θ,
- Model acceptance/rejection process (i.e. $p_\gamma(z|x, y)$),
- Use $\boldsymbol{x}^{\text{nf}}$.

Reject Inference methods: How to use $\boldsymbol{x}^{\mathrm{nf}}$?

Question 2: How to construct a better estimator than $\hat{\theta}^{\mathrm{f}}$?

Scope for action:

- Change model space Θ,
- Model acceptance/rejection process (i.e. $p_\gamma(z|x,y)$),
- Use $\boldsymbol{x}^{\mathrm{nf}}$.

Natural way to achieve all three: generative approach

$$p_\alpha(x,y,z) = p_{\beta_\alpha}(x) p_{\theta_\alpha}(y|x) p_{\gamma_\alpha}(z|x,y).$$

$$\left(\boxed{\hat{\theta}_\alpha}, \hat{\beta}_\alpha, \hat{\gamma}_\alpha\right) = \underset{\theta_\alpha,\beta_\alpha,\gamma_\alpha}{\operatorname{argmax}} \ell(\alpha; \boldsymbol{x}, \boldsymbol{y}^{\mathrm{f}}) = \underset{\theta_\alpha,\beta_\alpha,\gamma_\alpha}{\operatorname{argmax}} \sum_{i=1}^{n} \ln(p_{\theta_\alpha}(y_i|x_i)) + \sum_{i=1}^{n+m} \ln(p_{\beta_\alpha}(x_i)) \left(+ \sum_{i=1}^{n} \ln(p_{\gamma_\alpha}(z_i|x_i,y_i)) \right).$$

Reject Inference methods: How to use $\boldsymbol{x}^{\mathrm{nf}}$?

Question 2: How to construct a better estimator than $\hat{\theta}^{\mathrm{f}}$?

Scope for action:

- ~~Change model space Θ~~ logistic regression,
- Model acceptance/rejection process (i.e. $p_\gamma(z|x,y)$),
- Use $\boldsymbol{x}^{\mathrm{nf}}$.

~~Natural way to achieve all three: generative approach~~

$$p_\alpha(x,y,z) = p_{\beta_\alpha}(x) p_{\theta_\alpha}(y|x) p_{\gamma_\alpha}(z|x,y).$$

$$(\boxed{\hat{\theta}_\alpha}, \hat{\beta}_\alpha, \hat{\gamma}_\alpha) = \underset{\alpha}{\operatorname{argmax}}\, \ell(\alpha; \boldsymbol{x}, \boldsymbol{y}^{\mathrm{f}}) = \underset{\theta_\alpha,\beta_\alpha,\gamma_\alpha}{\operatorname{argmax}} \sum_{i=1}^{n} \ln(p_{\theta_\alpha}(y_i|x_i)) + \sum_{i=1}^{n+m} \ln(p_{\beta_\alpha}(x_i)) \left(+ \sum_{i=1}^{n} \ln(p_{\gamma_\alpha}(z_i|x_i,y_i)) \right).$$

Question 2: How to construct a better estimator than $\hat{\theta}^{\text{f}}$?

Scope for action:

- ~~Change model space Θ~~ logistic regression,
- ~~Model acceptance/rejection process (i.e. $p_\gamma(z|x,y)$)~~ γ cannot be estimated,
- Use $\boldsymbol{x}^{\text{nf}}$.

~~Natural way to achieve all three: generative approach~~

$$p_\alpha(x,y,z) = p_{\beta_\alpha}(x) p_{\theta_\alpha}(y|x) p_{\gamma_\alpha}(z|x,y).$$

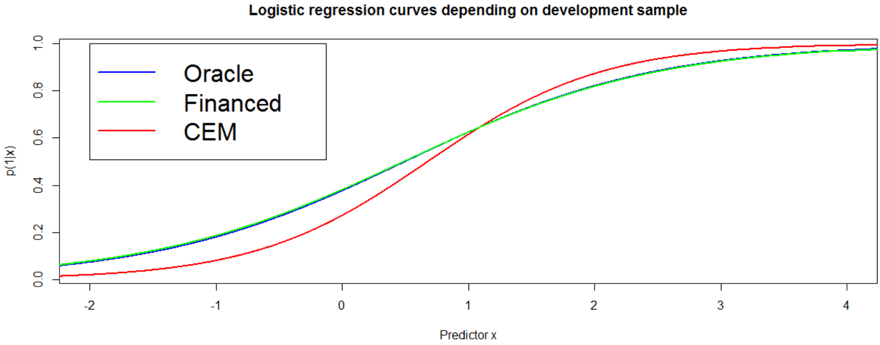
$$(\boxed{\hat{\theta}_\alpha}, \hat{\beta}_\alpha, \hat{\gamma}_\alpha) = \underset{\alpha}{\operatorname{argmax}}\, \ell(\alpha; \boldsymbol{x}, \boldsymbol{y}^{\text{f}}) = \underset{\theta_\alpha,\beta_\alpha,\gamma_\alpha}{\operatorname{argmax}} \sum_{i=1}^{n} \ln(p_{\theta_\alpha}(y_i|x_i)) + \sum_{i=1}^{n+m} \ln(p_{\beta_\alpha}(x_i)) \left(+ \sum_{i=1}^{n} \ln(p_{\gamma_\alpha}(z_i|x_i,y_i)) \right).$$

Reject Inference methods: A possible reinterpretation

Reclassification[1] :

$$(\hat{\theta}^{\text{CEM}}, \hat{\boldsymbol{y}}^{\text{nf}}) = \underset{\theta, \boldsymbol{y}^{\text{nf}}}{\text{argmax}}\, \ell(\theta; \boldsymbol{x}, \boldsymbol{y}^{\text{f}}, \boldsymbol{y}^{\text{nf}}) \text{ where } \hat{y}_i = \underset{y_i}{\text{argmax}}\, p_{\hat{\theta}^{\text{f}}}(y_i|x_i).$$

Problem: inconsistent estimator.



[1] [Soulié and Viennet, 2007, Banasik and Crook, 2007, Guizani et al., 2013]

Reject Inference methods: A possible reinterpretation

Augmentation[2]: MAR / misspecified model.

$$\ell_{\text{Aug}}(\theta; \boldsymbol{x}^{\text{f}}, \boldsymbol{y}^{\text{f}}) = \sum_{i=1}^{n} \frac{1}{p_{\text{true}}(\text{f}|x_i)} \ln(p_\theta(y_i|x_i)).$$

Problem: estimation of $p_{\text{true}}(\text{f}|x_i)$.

Parcelling [3]:

$$\ell(\theta; \boldsymbol{x}, \boldsymbol{y}^{\text{f}}, \hat{\boldsymbol{y}}^{\text{nf}}) \text{ where } \hat{y}_i = \begin{cases} 1 \text{ w.p. } \alpha_i p_{\hat{\theta}^{\text{f}}}(1|x_i, \text{f}) \\ 0 \text{ w.p. } 1 - \alpha_i p_{\hat{\theta}^{\text{f}}}(1|x_i, \text{f}) \end{cases}.$$

Problem: MNAR assumptions hidden in $\hat{\boldsymbol{y}}^{\text{nf}}$ (α_i) impossible to test.

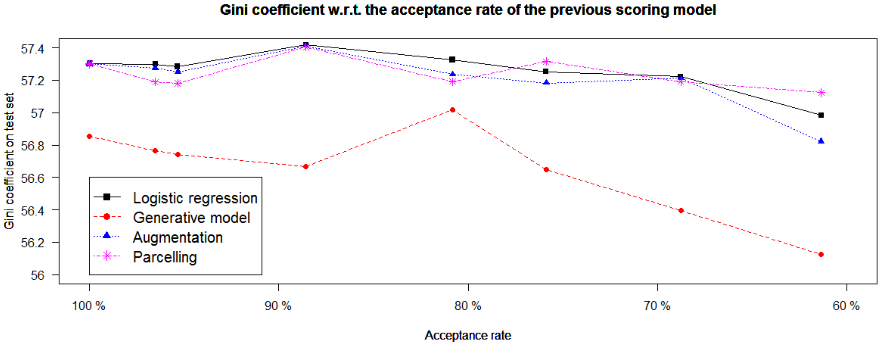
[2] [Soulié and Viennet, 2007, Banasik and Crook, 2007, Guizani et al., 2013, Nguyen, 2016]

[3] [Soulié and Viennet, 2007, Banasik and Crook, 2007, Guizani et al., 2013]

Reject Inference methods: Experimental results

Portfolio from a consumer electronics distributor partner:

- Approximately 250,000 applications
- Approximately 10 discrete (or discretized) features (with interactions):
  - Socio-professional category(-ies)
  - Amount of rent
  - Number of children
  - Years in current job position
- Approximately 3 % of default



Conclusion

Conclusion

- **Fuzzy Augmentation**, **Reclassification** (and **Twins**) were proved useless.
- **Augmentation** and **Parcelling** seem legitimate depending on model specification and missingness mechanism but difficult/impossible to set forth in practice.
- **Recommendation**: do not practice Reject Inference since high risk/low return of all tested Reject Inference techniques.
- For the most part, Reject Inference had previously been tackled only experimentally.
- Research paper in progress.

Thank you for your attention!

Any 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 èmes Journées de statistique.

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

Soulié, F. F. and Viennet, E. (2007).
Le traitement des refusés dans le risque crédit.
Revue des Nouvelles Technologies de l'Information, Data Mining et Apprentissage Statistique : application en assurance, banque et marketing, RNTI-A-1:22–44.