Reject Inference, "quantization", interactions, logistic regression trees, and bonuses

Adrien Ehrhardt

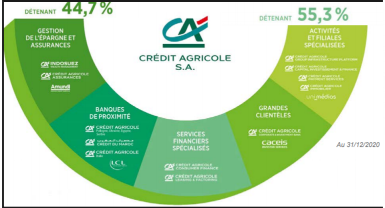
Mission Lane, 08/03/2022

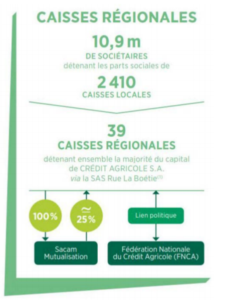
Who am I?

$\approx$ 2016-2019: "CIFRE" PhD student at Inria (consortium of French labs, like CNRS, but specialized in Applied Maths) and Crédit Agricole Consumer Finance (consumer loans).

Who am I?

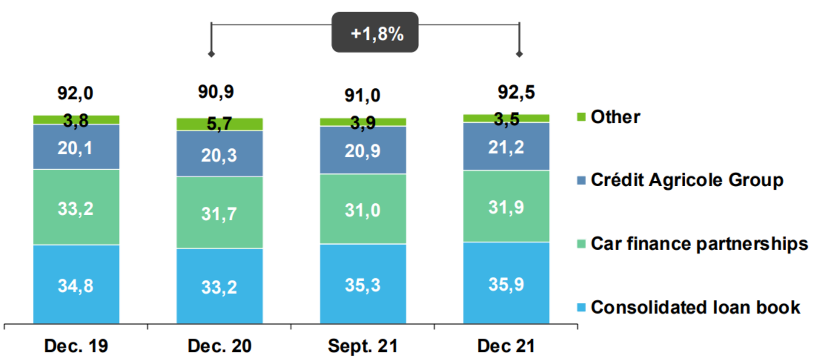
≈ 2016-2019: “CIFRE” PhD student at Inria (consortium of French labs, like CNRS, but specialized in Applied Maths) and Crédit Agricole Consumer Finance (consumer loans).





Who am I?

≈ 2016-2019: "CIFRE" PhD student at Inria (consortium of French labs, like CNRS, but specialized in Applied Maths) and Crédit Agricole Consumer Finance (consumer loans).



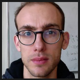
Who am I?

≈ 2016-2019: "CIFRE" PhD student at Inria (consortium of French labs, like CNRS, but specialized in Applied Maths) and Crédit Agricole Consumer Finance (consumer loans).

≈ 2020-now: Machine Learning Engineer at Crédit Agricole S.A. & Associate Professor at École Polytechnique.

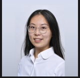
Collaborators

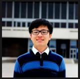

Christophe Biernacki


Vincent Vandewalle


Philippe Heinrich


Elise Bayraktar


Xuwen Liu


Minh Tuan Nguyen

Cléa Laouar

Context and notations: industrial setting

Job	Home	Time in job	Family status	Wages		Repayment
Craftsman	Owner	20	Widower	2000		1
?	Renter	10	Common-law	1700		0
Engineer	Starter	5	Divorced	4000		1
Executive	By work	8	Married	2700		0
Office employee	Renter	12	Married	1400		NA
Worker	By family	2	?	1200		NA

Table: Dataset with outliers and missing values.

Context and notations: industrial setting

Job	Home	Time in job	Family status	Wages		Repayment
Craftsman	Owner	20	Widower	2000		1
?	Renter	10	Common-law	1700		0
Engineer	Starter	5	Divorced	4000		1
Executive	By work	8	Married	2700		0
Office employee	Renter	12	Married	1400		NA
Worker	By family	2	?	1200		NA

Table: Dataset with outliers and missing values.

1. Discarding not financed applicants
2. Feature selection
3. Discretization / grouping
4. Interaction screening
5. Segmentation
6. Logistic regression fitting

Context and notations: industrial setting

Job	Home	Time in job	Family status	Wages		Repayment
Craftsman	Owner	20	Widower	2000		1
?	Renter	10	Common-law	1700		0
Engineer	Starter	5	Divorced	4000		1
Executive	By work	8	Married	2700		0
~~Office employee~~	~~Renter~~	~~12~~	~~Married~~	~~1400~~		NA
~~Worker~~	~~By family~~	~~2~~	~~?~~	~~1200~~		NA

Table: Dataset with outliers and missing values.

1. **Discarding not financed applicants**
2. Feature selection
3. Discretization / grouping
4. Interaction screening
5. Segmentation
6. Logistic regression fitting

Context and notations: industrial setting

Job			Family status	Wages		Repayment
Craftsman			Widower	2000		1
?			Common-law	1700		0
Engineer			Divorced	4000		1
Executive			Married	2700		0
~~Office employee~~	~~Renter~~	~~12~~	~~Married~~	~~1400~~		NA
~~Worker~~	~~By family~~	~~2~~	~~?~~	~~1200~~		NA

Table: Dataset with outliers and missing values.

1. Discarding not financed applicants
2. **Feature selection**
3. Discretization / grouping
4. Interaction screening
5. Segmentation
6. Logistic regression fitting

Context and notations: industrial setting

Job			Family status	Wages		Repayment
Craftsman			Widower	]1500;2000]		1
?			Common-law	]1500;2000]		0
Engineer			Divorced	]2000;∞[		1
Executive			Married	]2000;∞[		0
~~Office employee~~	~~Renter~~	~~12~~	~~Married~~	~~1400~~		NA
~~Worker~~	~~By family~~	~~2~~	~~?~~	~~1200~~		NA

Table: Dataset with outliers and missing values.

1. Discarding not financed applicants
2. Feature selection
3. **Discretization** / grouping
4. Interaction screening
5. Segmentation
6. Logistic regression fitting

Context and notations: industrial setting

Job			**Family status**	Wages		Repayment
?+Low-qualified			**?+Alone**	]1500;2000]		1
?+Low-qualified			**Union**	]1500;2000]		0
High-qualified			**?+Alone**	]2000;∞[		1
High-qualified			**Union**	]2000;∞[		0
~~Office employee~~	~~Renter~~	~~12~~	~~Married~~	~~1400~~		NA
~~Worker~~	~~By family~~	~~2~~	~~?~~	~~1200~~		NA

Table: Dataset with outliers and missing values.

1. Discarding not financed applicants
2. Feature selection
3. Discretization / **grouping**
4. Interaction screening
5. Segmentation
6. Logistic regression fitting

Context and notations: industrial setting

Job			Family status x Wages		Repayment
?+Low-qualified			?+Alone x]1500;2000]		1
?+Low-qualified			Union x]1500;2000]		0
High-qualified			?+Alone x]2000;∞[		1
High-qualified			Union x]2000;∞[		0
~~Office employee~~	~~Renter~~	~~12~~	~~Married~~ ~~1400~~		NA
~~Worker~~	~~By family~~	~~2~~	~~?~~ ~~1200~~		NA

Table: Dataset with outliers and missing values.

1. Discarding not financed applicants
2. Feature selection
3. Discretization / grouping
4. **Interaction screening**
5. Segmentation
6. Logistic regression fitting

Context and notations: industrial setting

Job			Family status x Wages		Repayment
?+Low-qualified			?+Alone x]1500;2000]		1
?+Low-qualified			Union x]1500;2000]		0
High-qualified			?+Alone x]2000;∞[		1
High-qualified			Union x]2000;∞[		0
~~Office employee~~	~~Renter~~	~~12~~	~~Married~~ ~~1400~~		NA
~~Worker~~	~~By family~~	~~2~~	~~?~~ ~~1200~~		NA

Table: Dataset with outliers and missing values.

1. Discarding not financed applicants
2. Feature selection
3. Discretization / grouping
4. Interaction screening
5. **Segmentation**
6. Logistic regression fitting

Context and notations: industrial setting

Job			Family status x Wages	Score	Repayment
?+Low-qualified			?+Alone x]1500;2000]	225	1
?+Low-qualified			Union x]1500;2000]	190	0
High-qualified			?+Alone x]2000;∞[	218	1
High-qualified			Union x]2000;∞[	202	0
~~Office employee~~	~~Renter~~	~~12~~	~~Married~~ ~~1400~~	NA	NA
~~Worker~~	~~By family~~	~~2~~	~~?~~ ~~1200~~	NA	NA

Table: Dataset with outliers and missing values.

1. Discarding not financed applicants
2. Feature selection
3. Discretization / grouping
4. Interaction screening
5. Segmentation
6. **Logistic regression fitting**

Context and notations: available data

Random variables: $\boldsymbol{X}, Y, Z$.

Context and notations: available data

Random variables: $\boldsymbol{X}, Y, Z$.

Observations

$\boldsymbol{x} = (x_1, \ldots, x_d)$	characteristics,
$x_j \in \mathbb{R}$ or $\{1, \ldots, l_j\}$	*e.g.* rent amount, job, . . . ,
$y \in \{0, 1\}$	good or bad,
$z \in \{\text{f}, \text{nf}\}$	financed or not financed.

Context and notations: available data

Random variables: $\boldsymbol{X}, Y, Z$.

Observations

$\boldsymbol{x} = (x_1, \ldots, x_d)$	characteristics,
$x_j \in \mathbb{R}$ or $\{1, \ldots, l_j\}$	*e.g.* rent amount, job, $\ldots$,
$y \in \{0, 1\}$	good or bad,
$z \in \{\mathrm{f}, \mathrm{nf}\}$	financed or not financed.

Samples

$\mathcal{T}_{\mathrm{f}} = (\mathrm{x_f}, \mathrm{y_f}, \mathrm{z_f})$	n-sample of financed clients,
$\mathcal{T}_{\mathrm{nf}} = (\mathrm{x_{nf}}, \mathrm{z_{nf}})$	n'-sample of not-financed clients,
$\mathcal{T} = \mathcal{T}_{\mathrm{f}} \cup \mathcal{T}_{\mathrm{nf}}$	observed sample,
$\mathcal{T}_{\mathrm{c}} = \mathcal{T} \cup \mathrm{y_{nf}}$	complete sample.

The observed data are the following:

$$
\mathcal{T} = \begin{array}{l}
\mathcal{T}_{\mathrm{f}} = \left(\boxed{\mathrm{x}_{\mathrm{f}}} \quad \boxed{\begin{matrix} x_{1,1} & \cdots & x_{1,d} \\ \vdots & \vdots & \vdots \\ x_{n,1} & \cdots & x_{n,d} \end{matrix}} \quad \boxed{\mathrm{y}_{\mathrm{f}}} \quad \boxed{\begin{matrix} y_1 \\ \vdots \\ y_n \end{matrix}} \quad \boxed{\mathrm{z}_{\mathrm{f}}} \quad \boxed{\begin{matrix} \mathrm{f} \\ \vdots \\ \mathrm{f} \end{matrix}} \right). \\
\qquad \cup \\
\mathcal{T}_{\mathrm{nf}} = \left(\boxed{\mathrm{x}_{\mathrm{nf}}} \quad \boxed{\begin{matrix} x_{n+1,1} & \cdots & x_{n+1,d} \\ \vdots & \vdots & \vdots \\ x_{n+n',1} & \cdots & x_{n+n',d} \end{matrix}} \quad \boxed{\mathrm{y}_{\mathrm{nf}}} \quad \boxed{\begin{matrix} \mathrm{NA} \\ \vdots \\ \mathrm{NA} \end{matrix}} \quad \boxed{\mathrm{z}_{\mathrm{nf}}} \quad \boxed{\begin{matrix} \mathrm{nf} \\ \vdots \\ \mathrm{nf} \end{matrix}} \right).
\end{array}
$$

Context and notations: available data

The observed data are the following:

$$\mathcal{T} = \begin{array}{l} \mathcal{T}_{\text{f}} = \left(\boxed{\text{x}_{\text{f}}} \boxed{\begin{matrix} x_{1,1} & \cdots & x_{1,d} \\ \vdots & \vdots & \vdots \\ x_{n,1} & \cdots & x_{n,d} \end{matrix}} \quad \boxed{\text{y}_{\text{f}}} \boxed{\begin{matrix} y_1 \\ \vdots \\ y_n \end{matrix}} \quad \boxed{\text{z}_{\text{f}}} \boxed{\begin{matrix} \text{f} \\ \vdots \\ \text{f} \end{matrix}} \right). \\ \cup \\ \mathcal{T}_{\text{nf}} = \left(\boxed{\text{x}_{\text{nf}}} \boxed{\begin{matrix} x_{n+1,1} & \cdots & x_{n+1,d} \\ \vdots & \vdots & \vdots \\ x_{n+n',1} & \cdots & x_{n+n',d} \end{matrix}} \quad \boxed{\text{y}_{\text{nf}}} \boxed{\begin{matrix} \text{NA} \\ \vdots \\ \text{NA} \end{matrix}} \quad \boxed{\text{z}_{\text{nf}}} \boxed{\begin{matrix} \text{nf} \\ \vdots \\ \text{nf} \end{matrix}} \right). \end{array}$$

Credit Scoring aims at estimating $p(y|\boldsymbol{x})$ in the form of a simple **parametric** model $p_{\boldsymbol{\theta}}(y|\boldsymbol{x})$ such as logistic regression:

Context and notations: available data

The observed data are the following:

$$\mathcal{T} = \begin{array}{l} \mathcal{T}_{\mathrm{f}} = \left(\boxed{\mathrm{x_f}} \boxed{\begin{matrix} x_{1,1} & \cdots & x_{1,d} \\ \vdots & \vdots & \vdots \\ x_{n,1} & \cdots & x_{n,d} \end{matrix}} \quad \boxed{\mathrm{y_f}} \boxed{\begin{matrix} y_1 \\ \vdots \\ y_n \end{matrix}} \quad \boxed{\mathrm{z_f}} \boxed{\begin{matrix} \mathrm{f} \\ \vdots \\ \mathrm{f} \end{matrix}} \right). \\ \cup \\ \mathcal{T}_{\mathrm{nf}} = \left(\boxed{\mathrm{x_{nf}}} \boxed{\begin{matrix} x_{n+1,1} & \cdots & x_{n+1,d} \\ \vdots & \vdots & \vdots \\ x_{n+n',1} & \cdots & x_{n+n',d} \end{matrix}} \quad \boxed{\mathrm{y_{nf}}} \boxed{\begin{matrix} \mathrm{NA} \\ \vdots \\ \mathrm{NA} \end{matrix}} \quad \boxed{\mathrm{z_{nf}}} \boxed{\begin{matrix} \mathrm{nf} \\ \vdots \\ \mathrm{nf} \end{matrix}} \right). \end{array}$$

Credit Scoring aims at estimating $p(y|\boldsymbol{x})$ in the form of a simple **parametric** model $p_{\boldsymbol{\theta}}(y|\boldsymbol{x})$ such as logistic regression:

$$\ln \frac{p_{\boldsymbol{\theta}}(1|\boldsymbol{x})}{1 - p_{\boldsymbol{\theta}}(1|\boldsymbol{x})} = (1, \boldsymbol{x})'\boldsymbol{\theta}.$$

Table of Contents

Reject Inference

Feature quantization

Segmentation: logistic regression trees

Missing data imputation

Carbon risk

NLP for extra-financial reports

Conclusion and future work

Reject Inference

Reject Inference: industrial setting

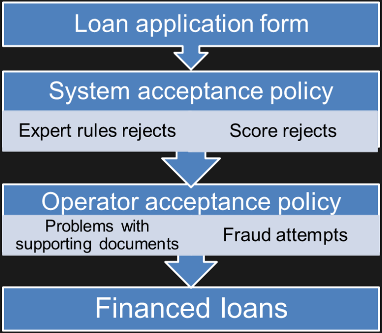


Figure: Simplified financing mechanism at Crédit Agricole Consumer Finance

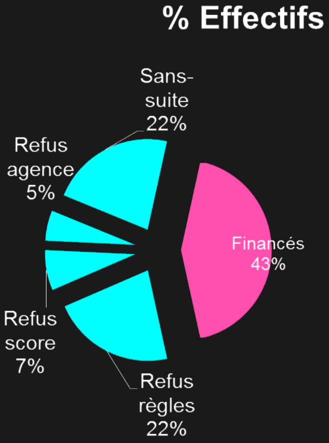


Figure: Proportion of "final" lending decisions for CACF France

Reject Inference: industrial setting

The industry traditionally fits a $\underbrace{\text{logistic regression}}_{\text{modelling constraint}}$ using only $\underbrace{\text{financed clients}}_{\substack{\text{convenience and lack}\\ \text{of better procedure}}}$ (**fixed parameter space** Θ):

GCA $$\hat{\boldsymbol{\theta}}_{\mathrm{f}} = \operatorname*{argmax}_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \mathcal{T}_{\mathrm{f}}) = \sum_{i=1}^{n} \ln p_{\boldsymbol{\theta}}(y_i | \boldsymbol{x}_i),$$

which asymptotically approximates:

$$\boldsymbol{\theta}^{\star}_{\mathrm{f}} = \operatorname*{argmin}_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{X}}[\mathrm{KL}(p || p_{\boldsymbol{\theta}}) | Z = \mathrm{f}].$$

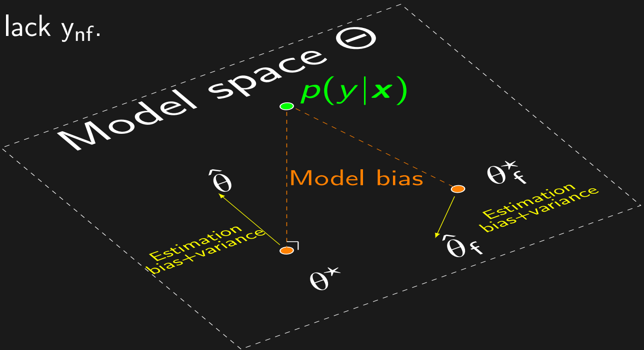
Oracle to be approximated:

$$\begin{aligned}\boldsymbol{\theta}^\star &= \operatorname*{argmin}_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{X}}[\mathrm{KL}(p||p_{\boldsymbol{\theta}})] \\ &= \operatorname*{argmax}_{\boldsymbol{\theta}} \mathbb{E}_{\boldsymbol{x},y\sim p}[\ln p_{\boldsymbol{\theta}}(y|\boldsymbol{x})],\end{aligned}$$

which standard estimator would be:

$$\hat{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \mathcal{T}_c),$$

but we lack y_{nf}.



Reject Inference: Asymptotics

Estimators :

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

[1] Zadrozny, "Learning and evaluating classifiers under sample selection bias".

Reject Inference: Asymptotics

Estimators :

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

[1] Zadrozny, "Learning and evaluating classifiers under sample selection bias".

Reject Inference: Asymptotics

Estimators :

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

[1] Zadrozny, "Learning and evaluating classifiers under sample selection bias".

Estimators :

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

What follows will only hold for "local" model which output depends asymptotically only on $p(y|x)$, such as logistic regression[1].

[1] Zadrozny, "Learning and evaluating classifiers under sample selection bias".

Estimators :

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

What follows will only hold for "local" model which output depends asymptotically only on $p(y|x)$, such as logistic regression[1].

It can be shown that Bayesian classifiers, SVMs, decision trees are "global" learners [1].

[1] Zadrozny, "Learning and evaluating classifiers under sample selection bias".

Reject Inference: modelling the financing mechanism

GCA Due to the financing mechanism, labels y are not MCAR.

Statistician Let $\{p_\phi(z|\boldsymbol{x}, y)\}_{\phi \in \Phi}$ denote this hidden financing mechanism (as a parametrized family).

Reject Inference: modelling the financing mechanism

GCA Due to the financing mechanism, labels y are not MCAR.

Statistician Let $\{p_\phi(z|\boldsymbol{x}, y)\}_{\phi \in \Phi}$ denote this hidden financing mechanism (as a parametrized family).

Combining financing and credit-worthiness probability distributions:

$$p_\gamma(y, z|\boldsymbol{x}) = \underbrace{p_{\theta(\gamma)}(y|\boldsymbol{x})}_{\text{GCA}} \underbrace{p_{\phi(\gamma)}(z|\boldsymbol{x}, y)}_{?}.$$

Reject Inference: modelling the financing mechanism

GCA Due to the financing mechanism, labels y are not MCAR.

Statistician Let $\{p_\phi(z|\boldsymbol{x}, y)\}_{\phi \in \Phi}$ denote this hidden financing mechanism (as a parametrized family).

Combining financing and credit-worthiness probability distributions:

$$p_\gamma(y, z|\boldsymbol{x}) = \underbrace{p_{\theta(\gamma)}(y|\boldsymbol{x})}_{\text{GCA}} \underbrace{p_{\phi(\gamma)}(z|\boldsymbol{x}, y)}_{?}.$$

To estimate γ, we could rely on Maximum Likelihood theory:

Reject Inference: modelling the financing mechanism

GCA Due to the financing mechanism, labels y are not MCAR.

Statistician Let $\{p_\phi(z|\boldsymbol{x}, y)\}_{\phi \in \Phi}$ denote this hidden financing mechanism (as a parametrized family).

Combining financing and credit-worthiness probability distributions:

$$p_\gamma(y, z|\boldsymbol{x}) = \underbrace{p_{\theta(\gamma)}(y|\boldsymbol{x})}_{\text{GCA}} \underbrace{p_{\phi(\gamma)}(z|\boldsymbol{x}, y)}_{?}.$$

To estimate γ, we could rely on Maximum Likelihood theory:

$$\ell(\gamma; \mathcal{T}) = \sum_{i=1}^{n} \ln p_\gamma(y_i, \mathrm{f}|\boldsymbol{x}_i) + \sum_{i=n+1}^{n+n'} \ln \sum_{y \in \{0,1\}} p_\gamma(y, \mathrm{nf}|\boldsymbol{x}_i).$$

Reject Inference: flawed model selection

No free lunch: financial or statistical investment to make.

GCA Because no test-sample $\mathcal{T}^{\text{test}}$ is available $\underbrace{\text{from } p(\boldsymbol{x}, y)}_{\text{funding bad clients at a loss}}$,
we cannot resort to error-rate criteria:

$$\text{Error}(\mathcal{T}^{\text{test}}) = \frac{1}{|\mathcal{T}^{\text{test}}|} \sum_{i \in \mathcal{T}^{\text{test}}} \mathbb{I}(\hat{y}_i \neq y_i).$$

Reject Inference: flawed model selection

No free lunch: financial or statistical investment to make.

GCA Because no test-sample $\mathcal{T}^{\text{test}}$ is available $\underbrace{\text{from } p(\boldsymbol{x}, y)}_{\cancel{\text{funding bad clients at a loss}}}$,
we cannot resort to error-rate criteria:

$$\text{Error}(\mathcal{T}^{\text{test}}) = \frac{1}{|\mathcal{T}^{\text{test}}|} \sum_{i \in \mathcal{T}^{\text{test}}} \mathbb{I}(\hat{y}_i \neq y_i).$$

We should use information criteria on the observed data $\mathcal{T}$ such as:

$$\text{BIC}(\hat{\gamma}; \mathcal{T}) = -2\ell(\hat{\gamma}; \mathcal{T}) + \dim(\mathbf{\Gamma}) \ln n,$$

where $\hat{\gamma} = \text{argmax}_{\gamma}\, \ell(\gamma; \mathcal{T})$, to compare models.

Reject Inference: flawed model selection

No free lunch: financial or statistical investment to make.

GCA Because no test-sample $\mathcal{T}^{\text{test}}$ is available $\underbrace{\text{from } p(\boldsymbol{x}, y)}_{\cancel{\text{funding bad clients at a loss}}}$,

we cannot resort to error-rate criteria:

$$\text{Error}(\mathcal{T}^{\text{test}}) = \frac{1}{|\mathcal{T}^{\text{test}}|} \sum_{i \in \mathcal{T}^{\text{test}}} \mathbb{I}(\hat{y}_i \neq y_i).$$

We should use information criteria on the observed data $\mathcal{T}$ such as:

$$\text{BIC}(\hat{\gamma}; \mathcal{T}) = -2\ell(\hat{\gamma}; \mathcal{T}) + \dim(\boldsymbol{\Gamma}) \ln n,$$

where $\hat{\gamma} = \text{argmax}_\gamma\, \ell(\gamma; \mathcal{T})$, to compare models.

It requires to precisely state the models $\{p_\gamma(y, z|\boldsymbol{x})\}_{\boldsymbol{\Gamma}}$ that compete and their underlying assumptions.

Reject Inference: strategies

We gathered 6 so-called Reject Inference methods from the literature that aim at re-injecting x_{nf} into the estimation procedure of $\boldsymbol{\theta}$.

They usually resemble EM-like algorithms:

$$\mathcal{T}_{\mathrm{c}}^{(1)} = \left(\begin{pmatrix} x_{1,1} & \cdots & x_{1,d} \\ \vdots & \vdots & \vdots \\ x_{n,1} & \cdots & x_{n,d} \\ x_{n+1,1} & \cdots & x_{n+1,d} \\ \vdots & \vdots & \vdots \\ x_{n+n',1} & \cdots & x_{n+n',d} \end{pmatrix}, \begin{pmatrix} y_1 \\ \vdots \\ y_n \\ \hat{y}_{n+1}^{(1)} \\ \vdots \\ \hat{y}_{n+n'}^{(1)} \end{pmatrix}, \begin{pmatrix} \mathrm{f} \\ \vdots \\ \mathrm{f} \\ \mathrm{nf} \\ \vdots \\ \mathrm{nf} \end{pmatrix} \right)$$

Can we reinterpret these empirical methods in the missing data and information criterion frameworks and / or expose their implicit modelling steps?

Reject Inference: example of Fuzzy Augmentation[2]

Estimate $\hat{\boldsymbol{\theta}}_{\mathrm{f}} = \mathrm{argmax}_{\boldsymbol{\theta}}\, \ell(\boldsymbol{\theta}; \mathcal{T}_{\mathrm{f}})$, infer for $n+1 \leq i \leq n+n'$:

$$\hat{y}_i = p_{\hat{\boldsymbol{\theta}}_{\mathrm{f}}}(1|\boldsymbol{x}_i),$$

[2] Nguyen, Reject inference in application scorecards.

Reject Inference: example of Fuzzy Augmentation[2]

Estimate $\hat{\boldsymbol{\theta}}_{\mathrm{f}} = \operatorname{argmax}_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \mathcal{T}_{\mathrm{f}})$, infer for $n+1 \leq i \leq n+n'$:

$$\hat{y}_i = p_{\hat{\boldsymbol{\theta}}_{\mathrm{f}}}(1|\boldsymbol{x}_i),$$

and re-estimate $\boldsymbol{\theta}$ using the resulting $\mathcal{T}_{\mathrm{c}}$. For $1 \leq j \leq d$:

$$\frac{\partial \sum_{i=n+1}^{n'+n} \sum_{y_i=0}^{1} p_{\hat{\boldsymbol{\theta}}_{\mathrm{f}}}(y_i|\boldsymbol{x}_i) \ln(p_{\boldsymbol{\theta}}(y_i|\boldsymbol{x}_i))}{\partial \theta_j} = 0 \Leftrightarrow \boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_{\mathrm{f}},$$

[2] Nguyen, Reject inference in application scorecards.

Reject Inference: example of Fuzzy Augmentation[2]

Estimate $\hat{\boldsymbol{\theta}}_{\mathrm{f}} = \mathrm{argmax}_{\boldsymbol{\theta}}\, \ell(\boldsymbol{\theta}; \mathcal{T}_{\mathrm{f}})$, infer for $n+1 \leq i \leq n+n'$:

$$\hat{y}_i = p_{\hat{\boldsymbol{\theta}}_{\mathrm{f}}}(1|\boldsymbol{x}_i),$$

and re-estimate $\boldsymbol{\theta}$ using the resulting $\mathcal{T}_{\mathrm{c}}$. For $1 \leq j \leq d$:

$$\frac{\partial \sum_{i=n+1}^{n'+n} \sum_{y_i=0}^{1} p_{\hat{\boldsymbol{\theta}}_{\mathrm{f}}}(y_i|\boldsymbol{x}_i) \ln(p_{\boldsymbol{\theta}}(y_i|\boldsymbol{x}_i))}{\partial \theta_j} = 0 \Leftrightarrow \boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_{\mathrm{f}},$$

such that:

$$\underset{\boldsymbol{\theta} \in \Theta}{\mathrm{argmax}} \sum_{i=n+1}^{n'+n} \sum_{y_i=0}^{1} p_{\hat{\boldsymbol{\theta}}_{\mathrm{f}}}(y_i|\boldsymbol{x}_i) \ln(p_{\boldsymbol{\theta}}(y_i|\boldsymbol{x}_i)) = \hat{\boldsymbol{\theta}}_{\mathrm{f}}.$$

[2] Nguyen, Reject inference in application scorecards.

Reject Inference: example of Fuzzy Augmentation[2]

Estimate $\hat{\boldsymbol{\theta}}_{\text{f}} = \text{argmax}_{\boldsymbol{\theta}} \, \ell(\boldsymbol{\theta}; \mathcal{T}_{\text{f}})$, infer for $n+1 \leq i \leq n+n'$:

$$\hat{y}_i = p_{\hat{\boldsymbol{\theta}}_{\text{f}}}(1|\boldsymbol{x}_i),$$

and re-estimate $\boldsymbol{\theta}$ using the resulting $\mathcal{T}_{\text{c}}$. For $1 \leq j \leq d$:

$$\frac{\partial \sum_{i=n+1}^{n'+n} \sum_{y_i=0}^{1} p_{\hat{\boldsymbol{\theta}}_{\text{f}}}(y_i|\boldsymbol{x}_i) \ln(p_{\boldsymbol{\theta}}(y_i|\boldsymbol{x}_i))}{\partial \theta_j} = 0 \Leftrightarrow \boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_{\text{f}},$$

such that:

$$\underset{\boldsymbol{\theta} \in \Theta}{\text{argmax}} \sum_{i=n+1}^{n'+n} \sum_{y_i=0}^{1} p_{\hat{\boldsymbol{\theta}}_{\text{f}}}(y_i|\boldsymbol{x}_i) \ln(p_{\boldsymbol{\theta}}(y_i|\boldsymbol{x}_i)) = \hat{\boldsymbol{\theta}}_{\text{f}}.$$

Finally:

$$\underset{\boldsymbol{\theta} \in \Theta}{\text{argmax}} \, \ell(\boldsymbol{\theta}; \mathcal{T}_c) = \underset{\boldsymbol{\theta} \in \Theta}{\text{argmax}} \, \ell(\boldsymbol{\theta}; \mathcal{T}_{\text{f}}) = \hat{\boldsymbol{\theta}}_{\text{f}}.$$

[2] Nguyen, Reject inference in application scorecards.

- **MAR**[3]: $\forall\, \boldsymbol{x}, y, z,\; p(z|\boldsymbol{x}, y) = p(z|\boldsymbol{x})$
  $\rightarrow$ Financing is determined by an old score: $Z = \mathbb{1}_{\{(1,\boldsymbol{x})'\boldsymbol{\theta} > \text{cut}\}}$.

[3] Little and Rubin, Statistical analysis with missing data.
[4] Molenberghs et al., "Every missingness not at random model has a missingness at random counterpart with equal fit".

Reject Inference: missingness mechanism

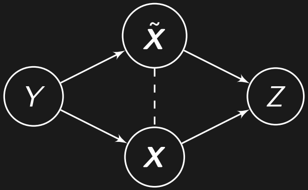
- **MAR**[3]: $\forall\, \boldsymbol{x}, y, z,\ p(z|\boldsymbol{x}, y) = p(z|\boldsymbol{x})$
  $\rightarrow$ Financing is determined by an old score: $Z = \mathbb{1}_{\{(1,\boldsymbol{x})'\boldsymbol{\theta} > \text{cut}\}}$.
- **MNAR**[3]: $\exists\, \boldsymbol{x}, y, z,\ p(z|\boldsymbol{x}, y) \neq p(z|\boldsymbol{x})$
  $\rightarrow$ Operators' hidden "feeling" $\tilde{\boldsymbol{X}}$ influence the financing.
  $\rightarrow$ Expert rules based on both present and hidden features $\boldsymbol{X}$ and $\tilde{\boldsymbol{X}}$ resp. where $\tilde{\boldsymbol{X}}$ cannot be totally explained by $\boldsymbol{X}$.
  $\rightarrow$ Cannot be tested[4].

[3] Little and Rubin, Statistical analysis with missing data.

[4] Molenberghs et al., "Every missingness not at random model has a missingness at random counterpart with equal fit".

Reject Inference: missingness mechanism

- **MAR**[3]: $\forall\, \boldsymbol{x}, y, z,\ p(z|\boldsymbol{x}, y) = p(z|\boldsymbol{x})$
  $\rightarrow$ Financing is determined by an old score: $Z = \mathbb{1}_{\{(1,\boldsymbol{x})'\boldsymbol{\theta} > \text{cut}\}}$.
- **MNAR**[3]: $\exists\, \boldsymbol{x}, y, z,\ p(z|\boldsymbol{x}, y) \neq p(z|\boldsymbol{x})$
  $\rightarrow$ Operators' hidden "feeling" $\tilde{\boldsymbol{X}}$ influence the financing.
  $\rightarrow$ Expert rules based on both present and hidden features $\boldsymbol{X}$ and $\tilde{\boldsymbol{X}}$ resp. where $\tilde{\boldsymbol{X}}$ cannot be totally explained by $\boldsymbol{X}$.
  $\rightarrow$ Cannot be tested[4].



[3] Little and Rubin, Statistical analysis with missing data.

[4] Molenberghs et al., "Every missingness not at random model has a missingness at random counterpart with equal fit".

Reject Inference: research contribution

Fuzzy Augmentation and Twins produce the same coefficient $\hat{\boldsymbol{\theta}}_{\mathrm{f}}$.

Reclassification[5,6,7] is equivalent to a Classification-EM algorithm, thus introducing a bias in the estimation of $\boldsymbol{\theta}$.

	MAR	MNAR
Well-specified model	$\hat{\boldsymbol{\theta}}_{\mathrm{f}}$ is unbiased.	$\hat{\boldsymbol{\theta}}_{\mathrm{f}}$ is biased. Any correction relies on *a priori* unverifiable assumptions about $p_\phi(z \mid \boldsymbol{x}, y)$, *e.g.* the Parcelling[5,6,7] method.
Misspecified model	$\hat{\boldsymbol{\theta}}_{\mathrm{f}}$ is biased: Augmentation[2,5,6,7] could be suitable but introduces a new estimation procedure[8] (which requires $\forall \boldsymbol{x}, p(\mathrm{f} \mid \boldsymbol{x}) > 0$).	

[5] Guizani et al., "Une Comparaison de quatre Techniques d'Inférence des Refusés dans le Processus d'Octroi de Crédit".

[6] Soulié and Viennet, "Le Traitement des Refusés dans le Risque Crédit".

[7] Banasik and Crook, "Reject inference, augmentation, and sample selection".

[8] Zadrozny, "Learning and evaluating classifiers under sample selection bias".

Reject Inference: augmentation

For "local" misspecified models and "global" models:

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{x},y}[\ln[p_{\boldsymbol{\theta}}(y|\boldsymbol{x})]] &= \sum_{y=0}^{1} \int_{\mathcal{X}} \ln p_{\boldsymbol{\theta}}(y|\boldsymbol{x}) p(y|\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x} \\
&= \sum_{y=0}^{1} \int_{\mathcal{X}} p(\mathrm{f}) \ln p_{\boldsymbol{\theta}}(y|\boldsymbol{x}) \frac{p(\boldsymbol{x}|\mathrm{f})}{p(\mathrm{f}|\boldsymbol{x})} p(y|\boldsymbol{x}) d\boldsymbol{x} \\
&= \sum_{y=0}^{1} \int_{\mathcal{X}} p(\mathrm{f}) \frac{\ln p_{\boldsymbol{\theta}}(y|\boldsymbol{x})}{p(\mathrm{f}|\boldsymbol{x})} p(\boldsymbol{x}, y|\mathrm{f}) d\boldsymbol{x} \\
&\approx \frac{1}{n} \sum_{i \in \mathcal{T}_{\mathrm{f}}} \frac{p(\mathrm{f})}{p(\mathrm{f}|\boldsymbol{x}_i)} \ln p_{\boldsymbol{\theta}}(y_i|\boldsymbol{x}_i).
\end{aligned}
$$

Reject Inference: augmentation

For "local" misspecified models and "global" models:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{x},y}[\ln[p_{\boldsymbol{\theta}}(y|\boldsymbol{x})]] &= \sum_{y=0}^{1} \int_{\mathcal{X}} \ln p_{\boldsymbol{\theta}}(y|\boldsymbol{x}) p(y|\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x} \\
&= \sum_{y=0}^{1} \int_{\mathcal{X}} p(\mathrm{f}) \ln p_{\boldsymbol{\theta}}(y|\boldsymbol{x}) \frac{p(\boldsymbol{x}|\mathrm{f})}{p(\mathrm{f}|\boldsymbol{x})} p(y|\boldsymbol{x}) d\boldsymbol{x} \\
&= \sum_{y=0}^{1} \int_{\mathcal{X}} p(\mathrm{f}) \frac{\ln p_{\boldsymbol{\theta}}(y|\boldsymbol{x})}{p(\mathrm{f}|\boldsymbol{x})} p(\boldsymbol{x}, y|\mathrm{f}) d\boldsymbol{x} \\
&\approx \frac{1}{n} \sum_{i \in \mathcal{T}_{\mathrm{f}}} \frac{p(\mathrm{f})}{p(\mathrm{f}|\boldsymbol{x}_i)} \ln p_{\boldsymbol{\theta}}(y_i|\boldsymbol{x}_i).
\end{aligned}$$

This assumes $p(f|\boldsymbol{x}) > 0 \ \forall x$, which is wrong.

Reject Inference: augmentation

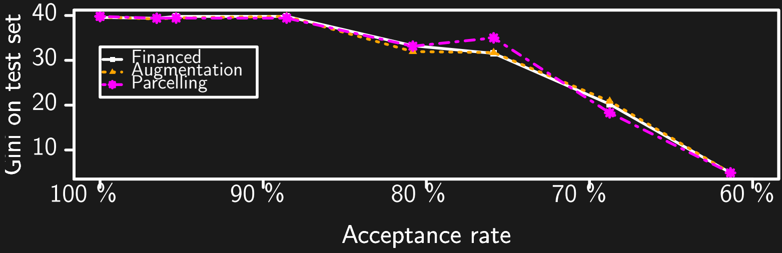
For "local" misspecified models and "global" models:

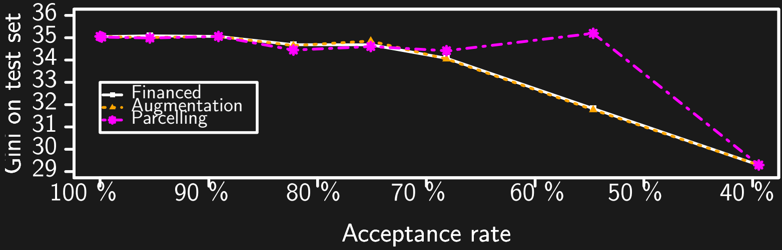
$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{x},y}[\ln[p_{\boldsymbol{\theta}}(y|\boldsymbol{x})]] &= \sum_{y=0}^{1} \int_{\mathcal{X}} \ln p_{\boldsymbol{\theta}}(y|\boldsymbol{x}) p(y|\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x} \\
&= \sum_{y=0}^{1} \int_{\mathcal{X}} p(\mathrm{f}) \ln p_{\boldsymbol{\theta}}(y|\boldsymbol{x}) \frac{p(\boldsymbol{x}|\mathrm{f})}{p(\mathrm{f}|\boldsymbol{x})} p(y|\boldsymbol{x}) d\boldsymbol{x} \\
&= \sum_{y=0}^{1} \int_{\mathcal{X}} p(\mathrm{f}) \frac{\ln p_{\boldsymbol{\theta}}(y|\boldsymbol{x})}{p(\mathrm{f}|\boldsymbol{x})} p(\boldsymbol{x}, y|\mathrm{f}) d\boldsymbol{x} \\
&\approx \frac{1}{n} \sum_{i \in \mathcal{T}_{\mathrm{f}}} \frac{p(\mathrm{f})}{p(\mathrm{f}|\boldsymbol{x}_i)} \ln p_{\boldsymbol{\theta}}(y_i|\boldsymbol{x}_i).
\end{aligned}
$$

This assumes $p(f|\boldsymbol{x}) > 0 \; \forall x$, which is wrong.

Further, one needs to specify / model $p(f|\boldsymbol{x})$.

Reject Inference: industry contribution

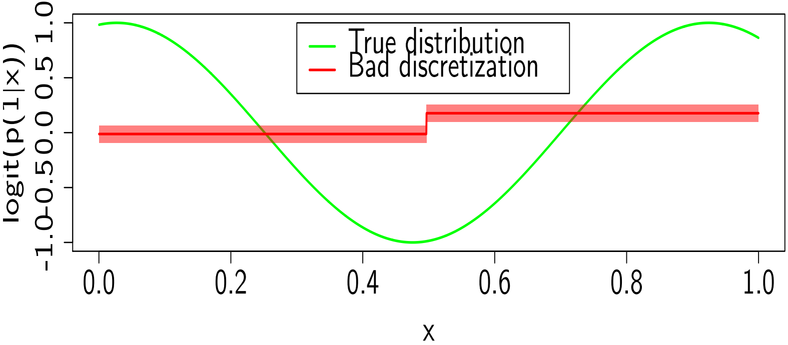




Feature quantization

Feature quantization: by an example

For theoretical reasons: bias-variance tradeoff.



Feature quantization: some more notations I

For practical reasons: interpretability, outliers...
... at the expense of the statistician's time.

Quantized data

$$\boldsymbol{q}(\boldsymbol{x}) = (\boldsymbol{q}_1(x_1), \dots, \boldsymbol{q}_d(x_d))$$

$$\boldsymbol{q}_j(x_j) = (q_{j,h}(x_j))_1^{m_j} \text{ (one-hot encoding)}$$

$$q_{j,h}(\cdot) = \mathbb{1}(x_j \in C_{j,h}), 1 \leq h \leq m_j$$

Feature quantization: some more notations II

Quantization is model selection (illustrated here with BIC).

Oracle

$$\boldsymbol{\theta}^\star, \boldsymbol{q}^\star = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}_{\boldsymbol{q}}, \boldsymbol{q}\in\boldsymbol{Q}}{\operatorname{argmax}} \; \mathbb{E}_{\boldsymbol{x},y}\left[\ln p_{\boldsymbol{\theta}}(y|\boldsymbol{q}(\boldsymbol{x}))\right],$$

Statistician

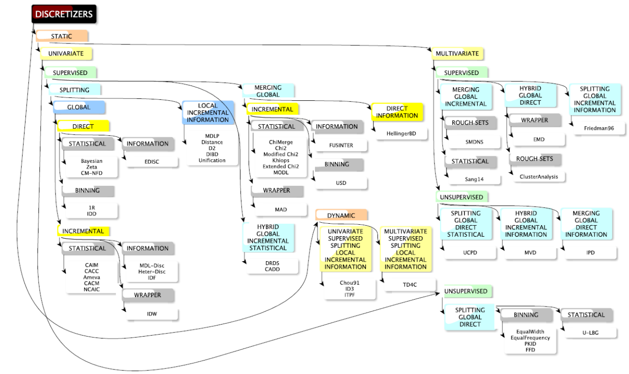
$$\hat{\boldsymbol{\theta}}^{\text{BIC}}, \hat{\boldsymbol{q}}^{\text{BIC}} = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}_{\boldsymbol{q}}, \boldsymbol{q}\in\boldsymbol{Q}}{\operatorname{argmin}} \; \text{BIC}(\hat{\boldsymbol{\theta}}_{\boldsymbol{q}}; \mathrm{y_f}, \boldsymbol{q}(\mathrm{x_f})),$$

$$\text{where } \hat{\boldsymbol{\theta}}_{\boldsymbol{q}} = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}_{\boldsymbol{q}}}{\operatorname{argmax}} \, \ell(\boldsymbol{\theta}; \mathrm{y_f}, \boldsymbol{q}(\mathrm{x_f})).$$

Implicitly assumes quantizations are “well” separated.

Quantization becomes an algorithmic problem.

Feature quantization: existing approaches



These approaches[9] maximize an "intermediary" criterion, *e.g.*:

CACF $$\hat{\boldsymbol{q}}_j^{\chi^2} = \operatorname*{argmax}_{\boldsymbol{q}_j} \chi^2(\boldsymbol{q}_j(\mathrm{x_f}), \mathrm{y_f}) \overset{?}{\approx} \boldsymbol{q}_j^\star,$$

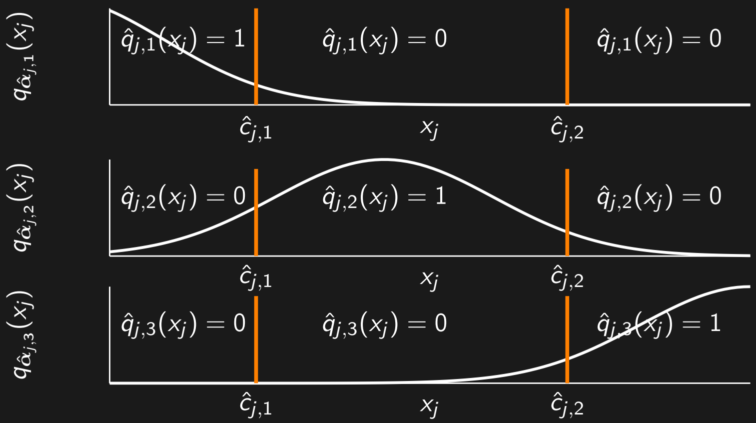
and we **hope** that it's aligned with our original goal s.t.:

CACF $$\hat{\boldsymbol{\theta}}^{\chi^2} = \operatorname*{argmax}_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \mathrm{y_f}, \hat{\boldsymbol{q}}^{\chi^2}(\mathrm{x_f})) \overset{?}{\approx} \boldsymbol{\theta}^\star.$$

[9] Ramirez-Gallego et al., "Data Discretization: Taxonomy and Big Data Challenge".

Feature quantization: MAP estimation

$$\hat{q}_{j,h}(x_j) = 1 \text{ if } h = \underset{1 \leq h' \leq m_j}{\text{argmax}}\, q_{\hat{\alpha}_{j,h'}}, 0 \text{ otherwise}^{10,11}.$$

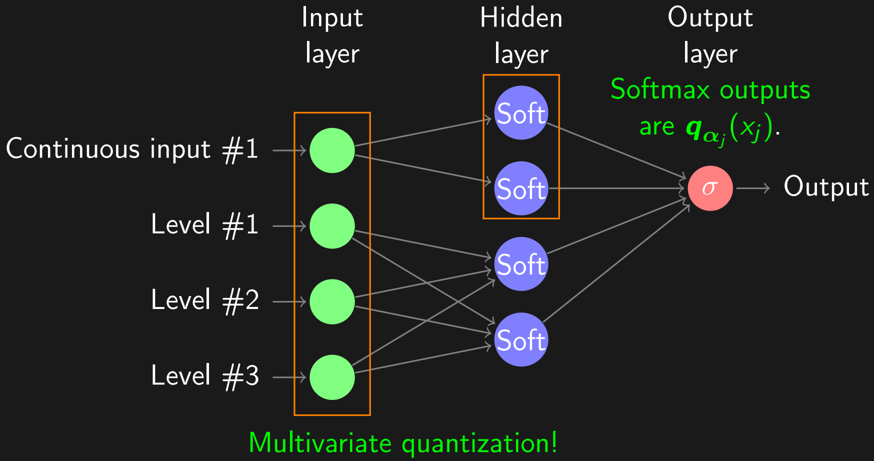


[10] Chamroukhi et al., "A regression model with a hidden logistic process for feature extraction from time series".

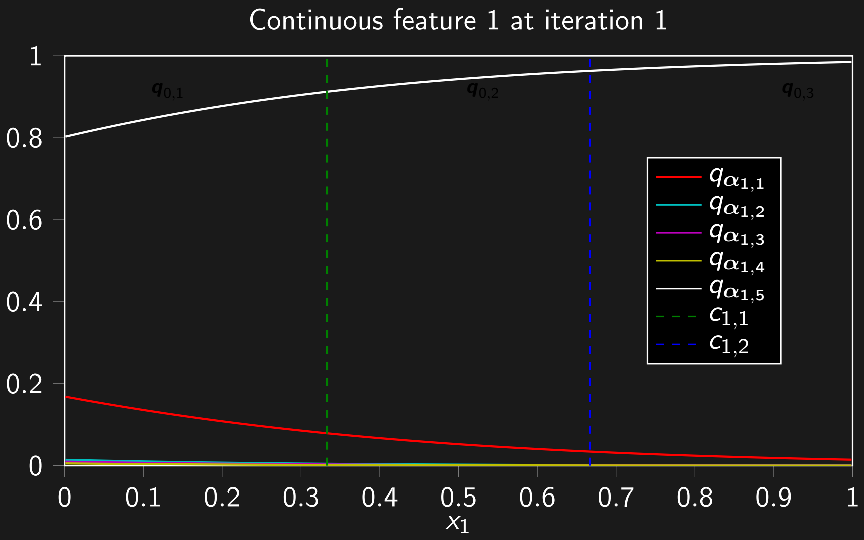
[11] Samé et al., "Model-based clustering and segmentation of time series with changes in regime".

Feature quantization: neural networks

Very simple neural network.
Very fast implementations available, *e.g.* TensorFlow.
No guarantee of global optimum (but works well in practice).



Feature quantization: neural networks



Simulated data

Table: For different sample sizes n, (A) CI of $\hat{c}_{j,2}$ for $c_{j,2} = 2/3$. (B) CI of $\hat{m}$ for $m_1 = 3$. (C) CI of $\hat{m}_3$ for $m_3 = 1$.

n	(A) $\hat{c}_{j,2}$	(B)	$\hat{m}_1$	(C)	$\hat{m}_3$
		1		60	
1,000	[0.656, 0.666]	90		32	
		9		8	
		0		88	
10,000	[0.666, 0.666]	100		12	
		0		0	

Feature quantization: results

CACF data

Table: Gini indices (the greater the value, the better the performance) of our proposed quantization algorithm *glmdisc*, the two baselines and the current scorecard.

Portfolio	ALLR	Current performance	*ad hoc* methods	Our proposal: *glmdisc*-NN	Our proposal: *glmdisc*-SEM	*glmdisc*-SEM w. interactions
Automobile	59.3 (3.1)	55.6 (3.4)	59.3 (3.0)	58.9 (2.6)	57.8 (2.9)	**64.8** (2.0)
Renovation	52.3 (5.5)	50.9 (5.6)	54.0 (5.1)	**56.7** (4.8)	55.5 (5.2)	55.5 (5.2)
Standard	39.7 (3.3)	37.1 (3.8)	45.3 (3.1)	43.8 (3.2)	36.7 (3.7)	**47.2** (2.8)
Revolving	62.7 (2.8)	58.5 (3.2)	63.2 (2.8)	62.3 (2.8)	60.7 (2.8)	**67.2** (2.5)
Mass retail	52.8 (5.3)	48.7 (6.0)	61.4 (4.7)	**61.8** (4.6)	61.0 (4.7)	60.3 (4.8)
Electronics	52.9 (11.9)	55.8 (10.8)	56.3 (10.2)	**72.6** (7.4)	62.0 (9.5)	63.7 (9.0)

Segmentation: logistic regression trees

Segmentation: logistic regression trees

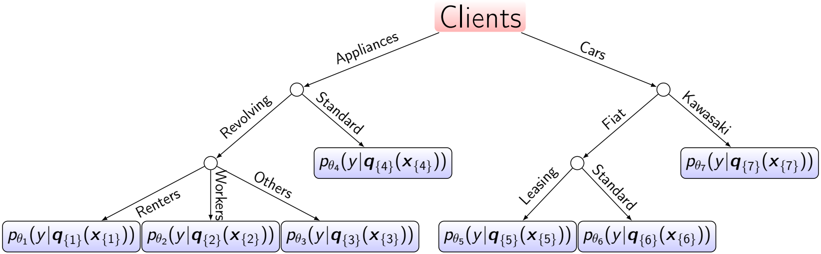


Figure: Scorecards tree structure in acceptance system.

Segmentation: logistic regression trees

Current procedure(s):

Segmentation: logistic regression trees

Current procedure(s):

- Promise a new partner their own score to maximize acceptance;

Segmentation: logistic regression trees

Current procedure(s):

- Promise a new partner their own score to maximize acceptance;
- Merge existing "close" branches that show similar performance;

Segmentation: logistic regression trees

Current procedure(s):

- Promise a new partner their own score to maximize acceptance;
- Merge existing "close" branches that show similar performance;
- **CACF** Try basic "clustering" techniques, *e.g.* visual separation of the data and / or levels on the two first MCA axes.

Segmentation: logistic regression trees

Current procedure(s):

- Promise a new partner their own score to maximize acceptance;
- Merge existing "close" branches that show similar performance;
- **CACF** Try basic "clustering" techniques, *e.g.* visual separation of the data and / or levels on the two first MCA axes.

Problem(s):

Segmentation: logistic regression trees

Current procedure(s):

- Promise a new partner their own score to maximize acceptance;
- Merge existing "close" branches that show similar performance;
- **CACF** Try basic "clustering" techniques, *e.g.* visual separation of the data and / or levels on the two first MCA axes.

Problem(s):

- This structure is not the result of optimization and is probably suboptimal (by how much?);

Segmentation: logistic regression trees

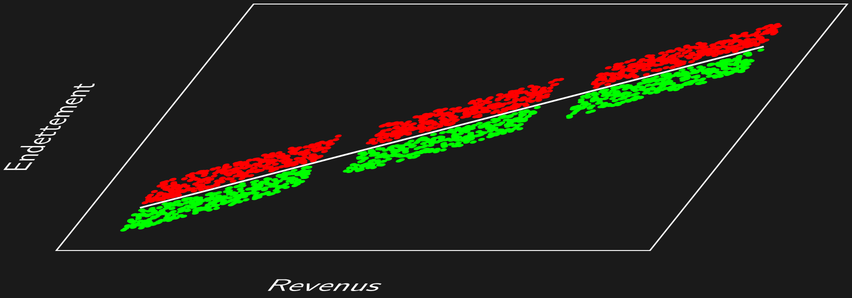
Current procedure(s):

- Promise a new partner their own score to maximize acceptance;
- Merge existing "close" branches that show similar performance;
- **CACF** Try basic "clustering" techniques, *e.g.* visual separation of the data and / or levels on the two first MCA axes.

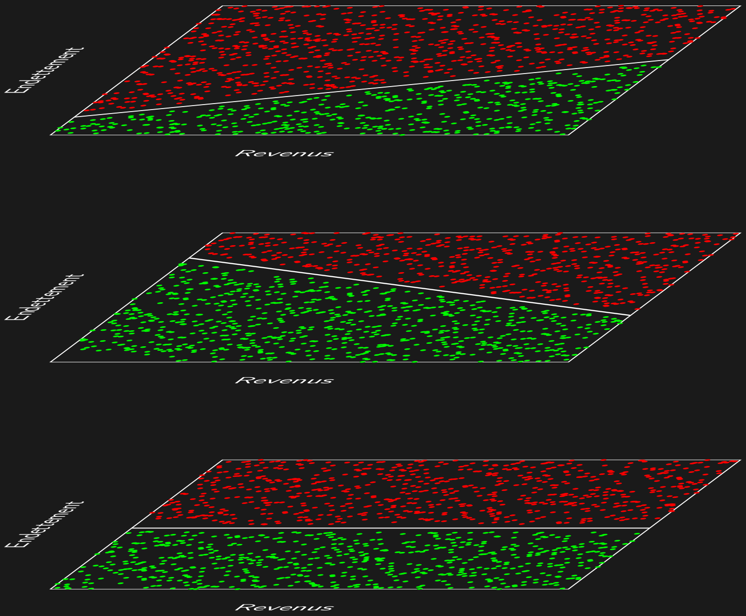
Problem(s):

- This structure is not the result of optimization and is probably suboptimal (by how much?);
- There are situations in which it severely fails.

Segmentation: logistic regression trees



Segmentation: logistic regression trees



Segmentation: logistic regression trees: contribution

Similarly to the quantization proposal: ability to be in several segments at a time.

Segmentation: logistic regression trees: contribution

Similarly to the quantization proposal: ability to be in several segments at a time.

$$p(y|\boldsymbol{x}) = \sum_{c=1}^{K} \underbrace{p_{\boldsymbol{\theta}}(y|\boldsymbol{x};c)}_{\text{“optimized” GCA constraint}} \quad \underbrace{p_{\beta}(c|\boldsymbol{x})}_{\text{“unoptimized” relaxed CACF constraint}},$$

where $p_{\beta}(c|\boldsymbol{x})$ is given by the classification tree as the proportion of training samples in each leaf (**not** majority vote).

Segmentation: logistic regression trees: contribution

Similarly to the quantization proposal: ability to be in several segments at a time.

$$p(y|\boldsymbol{x}) = \sum_{c=1}^{K} \underbrace{p_{\boldsymbol{\theta}}(y|\boldsymbol{x};c)}_{\text{“optimized” GCA constraint}} \quad \underbrace{p_{\beta}(c|\boldsymbol{x})}_{\text{“unoptimized” relaxed CACF constraint}} \quad ,$$

where $p_{\beta}(c|\boldsymbol{x})$ is given by the classification tree as the proportion of training samples in each leaf (**not** majority vote).

$$c_i^{(s)} \sim p_{\boldsymbol{\theta}^{\cdot(s-1)}}(y_i|\boldsymbol{x}_i;\cdot)p_{\beta^{(s-1)}}(\cdot|\boldsymbol{x}_i).$$

Similarly to the quantization proposal: ability to be in several segments at a time.

$$p(y|\boldsymbol{x}) = \sum_{c=1}^{K} \underbrace{p_{\boldsymbol{\theta}}(y|\boldsymbol{x};c)}_{\text{“optimized” GCA constraint}} \underbrace{p_{\beta}(c|\boldsymbol{x})}_{\text{“unoptimized” relaxed CACF constraint}},$$

where $p_\beta(c|\boldsymbol{x})$ is given by the classification tree as the proportion of training samples in each leaf (**not** majority vote).

$$c_i^{(s)} \sim p_{\boldsymbol{\theta}^{\cdot(s-1)}}(y_i|\boldsymbol{x}_i;\cdot) p_{\beta^{(s-1)}}(\cdot|\boldsymbol{x}_i).$$

$$\boldsymbol{\theta}^{c(s)} = \underset{\boldsymbol{\theta}^c}{\operatorname{argmax}} \sum_{i=1}^{n} \mathbb{1}_c(c_i^{(s)}) \ln p_{\boldsymbol{\theta}^c}(y_i|\boldsymbol{x}_i;c_i).$$

Segmentation: logistic regression trees: contribution

Similarly to the quantization proposal: ability to be in several segments at a time.

$$p(y|\boldsymbol{x}) = \sum_{c=1}^{K} \underbrace{p_{\boldsymbol{\theta}}(y|\boldsymbol{x};c)}_{\text{“optimized” GCA constraint}} \quad \underbrace{p_{\beta}(c|\boldsymbol{x})}_{\text{“unoptimized” relaxed CACF constraint}} \quad ,$$

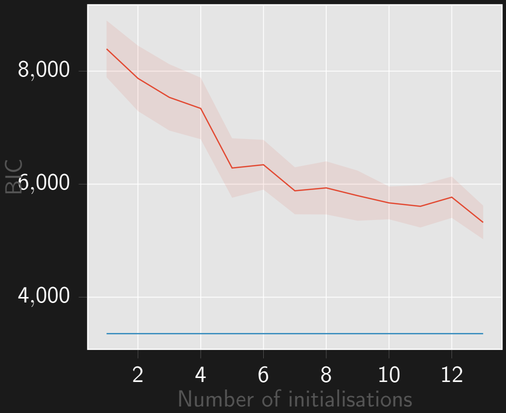
where $p_\beta(c|\boldsymbol{x})$ is given by the classification tree as the proportion of training samples in each leaf (**not** majority vote).

$$c_i^{(s)} \sim p_{\boldsymbol{\theta}^{\cdot(s-1)}}(y_i|\boldsymbol{x}_i;\cdot)p_{\beta^{(s-1)}}(\cdot|\boldsymbol{x}_i).$$

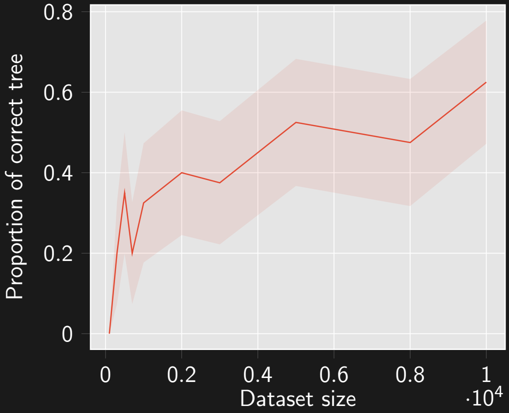
$$\boldsymbol{\theta}^{c(s)} = \underset{\boldsymbol{\theta}^c}{\operatorname{argmax}} \sum_{i=1}^{n} \mathbb{1}_c(c_i^{(s)}) \ln p_{\boldsymbol{\theta}^c}(y_i|\boldsymbol{x}_i;c_i).$$

$$\beta^{(s)} = \mathrm{C4.5}(\mathrm{c}^{(s)}, \mathrm{x}).$$

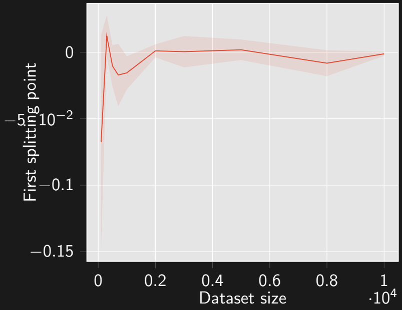
Segmentation: logistic regression trees: some results



Segmentation: logistic regression trees: some results



Segmentation: logistic regression trees: some results



Segmentation: logistic regression trees: some results

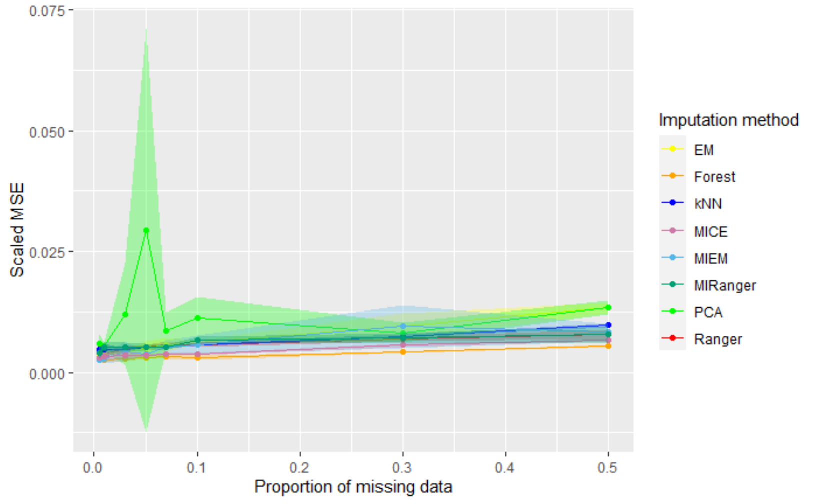
	Logistic regression	Decision Tree	SEM	Gradient Boosting
AUC ($\pm$ vs current method)	-3,02	-2,66	-1,78	-0,17

	SEM	LMT	MOB
# segment (current: 9)	2	11	1
AUC ($\pm$ vs current method)	-1,52	-7,70	-5,21

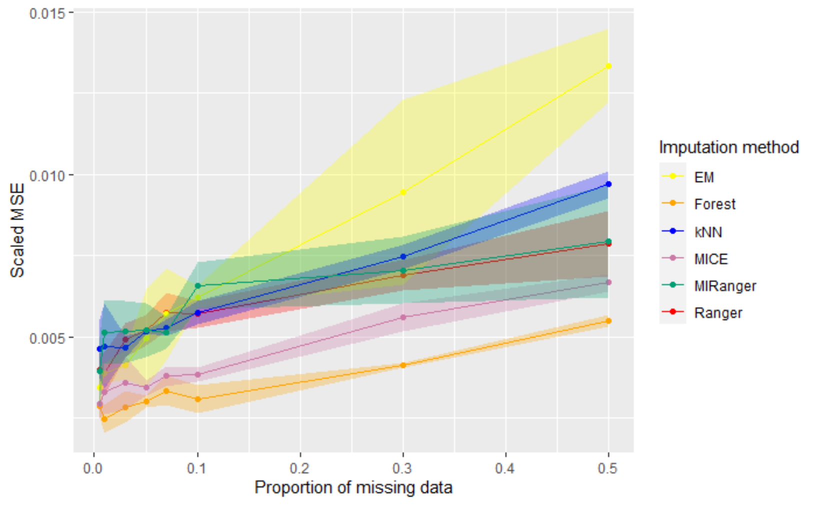
Missing data imputation

Missing data imputation: some results I

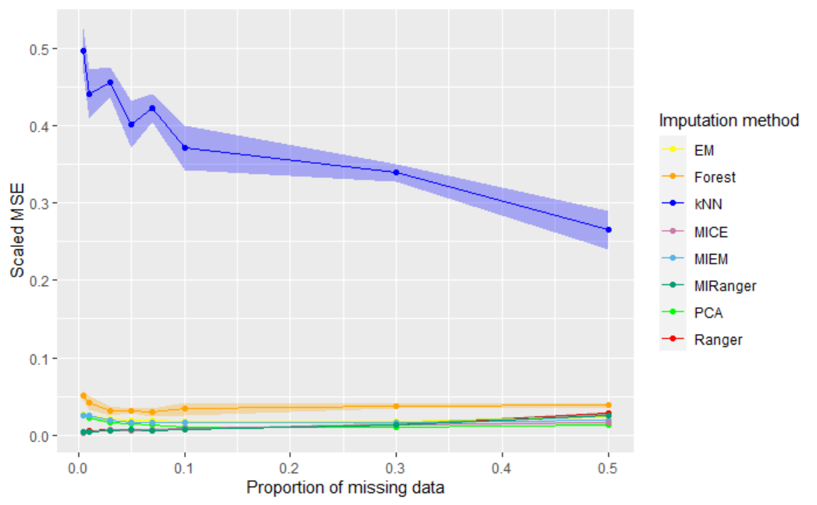
Research internship: comparing missing data imputation methods, mostly in MAR situations.



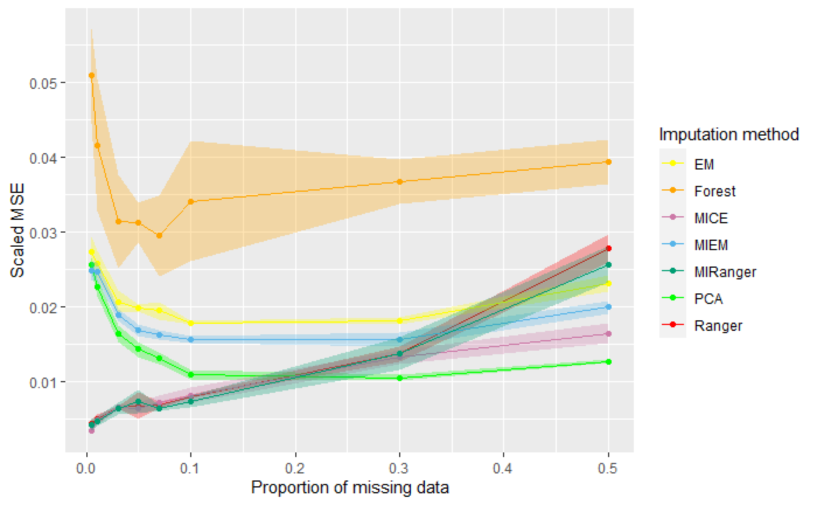
Missing data imputation: some results II



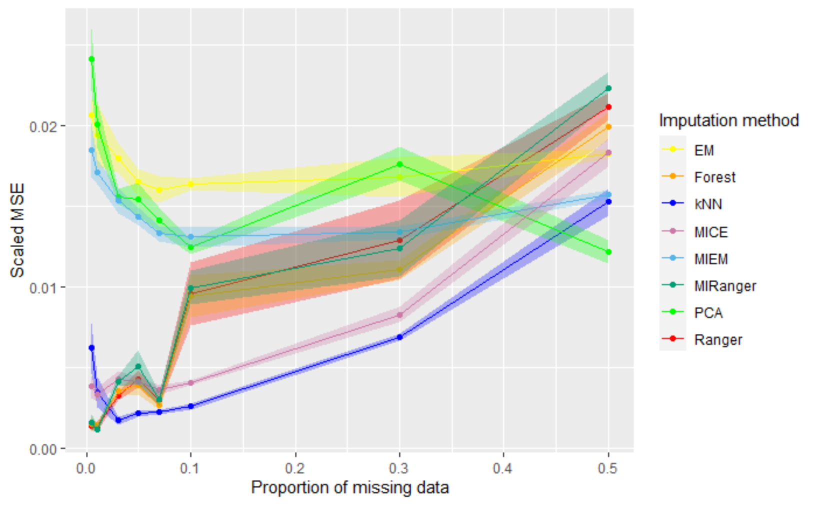
Missing data imputation: some results III



Missing data imputation: some results IV



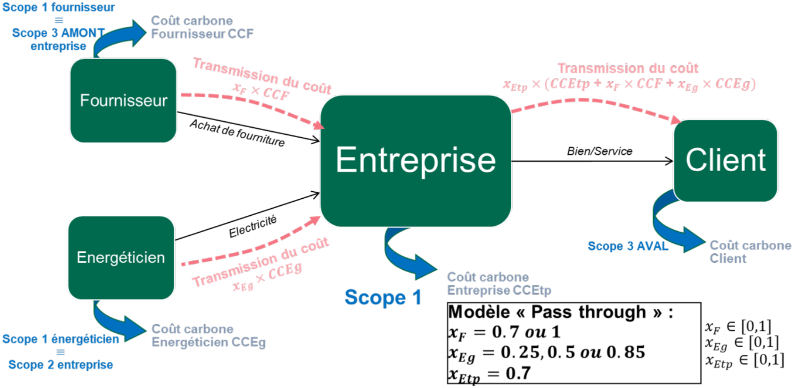
Missing data imputation: some results V



Carbon risk

Carbon risk: some results

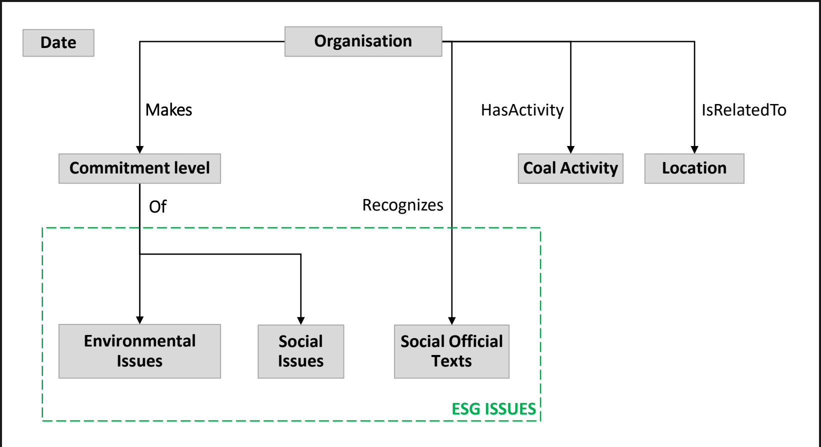
Research internship: use carbon price scenarios to impact the earnings of big corporations and adjust their default probability accordingly.



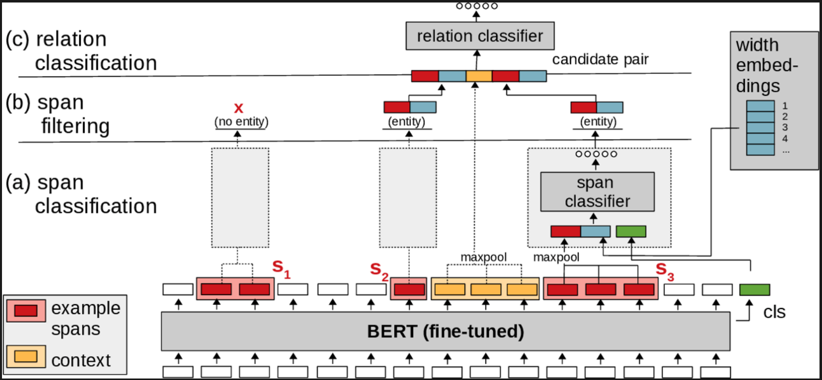
NLP for extra-financial reports

NLP for extra-financial reports: some results I

Research internship: build joint NER and RE models to automatically read through extra-financial reports.



NLP for extra-financial reports: some results II



Conclusion and future work

Conclusions from my PhD

This PhD tackled three main issues of "traditional" Credit Scoring:

1. Reject inference: impact of tossing away not-financed clients,

Conclusions from my PhD

This PhD tackled three main issues of "traditional" Credit Scoring:

1. Reject inference: impact of tossing away not-financed clients,

   Conclusion: sound problem reformulation, no method recommended, scoringTools R package.

Conclusions from my PhD

This PhD tackled three main issues of "traditional" Credit Scoring:

1. Reject inference: impact of tossing away not-financed clients,
2. "Constrained" representation learning: discretization, grouping, interaction screening,

Conclusions from my PhD

This PhD tackled three main issues of "traditional" Credit Scoring:

1. Reject inference: impact of tossing away not-financed clients,
2. "Constrained" representation learning: discretization, grouping, interaction screening,

   Conclusion: better performance, less time-consuming, glmdisc R and Python packages.

Conclusions from my PhD

This PhD tackled three main issues of “traditional” Credit Scoring:

1. Reject inference: impact of tossing away not-financed clients,
2. “Constrained” representation learning: discretization, grouping, interaction screening,
3. Predictive segmentation: logistic regression trees,

Conclusions from my PhD

This PhD tackled three main issues of "traditional" Credit Scoring:

1. Reject inference: impact of tossing away not-financed clients,
2. "Constrained" representation learning: discretization, grouping, interaction screening,
3. Predictive segmentation: logistic regression trees,

   Conclusion: first experiments on simulated and real data are encouraging, glmtree R package.

Future work as presented for my PhD - might be helpful?

There remains a lot of open questions:

1. Credit Scoring for profit: swap "p(2 unpaid instalments)" for $p(\text{profit} > 0)$ or $\mathbb{E}[\text{profit}]$,

Future work as presented for my PhD - might be helpful?

There remains a lot of open questions:

1. Credit Scoring for profit: swap "p(2 unpaid instalments)" for $p(\text{profit} > 0)$ or $\mathbb{E}[\text{profit}]$,

   Perspective: experiment observation-wise misclassification costs.

Future work as presented for my PhD - might be helpful?

There remains a lot of open questions:

1. Credit Scoring for profit: swap "p(2 unpaid instalments)" for $p(\text{profit} > 0)$ or $\mathbb{E}[\text{profit}]$,
2. Representation learning for fine-grained unstructured data,

Future work as presented for my PhD - might be helpful?

There remains a lot of open questions:

1. Credit Scoring for profit: swap "p(2 unpaid instalments)" for $p(\text{profit} > 0)$ or $\mathbb{E}[\text{profit}]$,
2. Representation learning for fine-grained unstructured data,

   Perspective: provide statistically sound methods to aggregate "behavioural" data, *e.g.* web visitation patterns.

Thanks!

References I

[1] John Banasik and Jonathan Crook. "Reject inference, augmentation, and sample selection". In: European Journal of Operational Research 183.3 (2007), pp. 1582–1594. url: http://www.sciencedirect.com/science/article/pii/S0377221706011969 (visited on 08/25/2016).

[2] Faicel Chamroukhi et al. "A regression model with a hidden logistic process for feature extraction from time series". In: International Joint Conference on Neural Networks, 2009. IJCNN 2009. IEEE. 2009, pp. 489–496.

[3] Asma Guizani et al. "Une Comparaison de quatre Techniques d'Inférence des Refusés dans le Processus d'Octroi de Crédit". In: 45 èmes Journées de statistique. 2013. url: http://cedric.cnam.fr/fichiers/art_2753.pdf (visited on 08/25/2016).

[4] Roderick JA Little and Donald B Rubin. Statistical analysis with missing data. John Wiley & Sons, 2014.

References II

[5] Geert Molenberghs et al. "Every missingness not at random model has a missingness at random counterpart with equal fit". In: Journal of the Royal Statistical Society: Series B 7.2 (2008), pp. 371–388.

[6] Ha Thu Nguyen. Reject inference in application scorecards: evidence from France. Tech. rep. University of Paris West-Nanterre la Défense, EconomiX, 2016. url: http://economix.fr/pdf/dt/2016/WP_EcoX_2016-10.pdf (visited on 08/25/2016).

[7] Sergio Ramirez-Gallego et al. "Data Discretization: Taxonomy and Big Data Challenge". In: Wiley Int. Rev. Data Min. and Knowl. Disc. 6.1 (Jan. 2016), pp. 5–21. issn: 1942-4787. doi: 10.1002/widm.1173. url: http://dx.doi.org/10.1002/widm.1173.

References III

[8] Allou Samé et al. "Model-based clustering and segmentation of time series with changes in regime". In: Advances in Data Analysis and Classification 5.4 (2011), pp. 301–321.

[9] Françoise Fogelman Soulié and Emmanuel Viennet. "Le Traitement des Refusés dans le Risque Crédit". In: Revue des Nouvelles Technologies de l'Information Data Mining et Apprentissage Statistique : application en assurance, banque et marketing, RNTI-A-1 (2007), pp. 22–44.

[10] Bianca Zadrozny. "Learning and evaluating classifiers under sample selection bias". In: Proceedings of the twenty-first ICML. ACM. 2004, p. 114.

Quantization

Quantization: research contribution

"Soft" approximation:

$$\boldsymbol{q}_{\boldsymbol{\alpha}_j}(\cdot) = \left(q_{\boldsymbol{\alpha}_{j,h}}(\cdot)\right)_{h=1}^{m_j} \text{ with } \begin{cases} \sum_{h=1}^{m_j} q_{\boldsymbol{\alpha}_{j,h}}(\cdot) = 1, \\ 0 \leq q_{\boldsymbol{\alpha}_{j,h}}(\cdot) \leq 1, \end{cases}$$

Quantization: research contribution

"Soft" approximation:

$$\boldsymbol{q}_{\boldsymbol{\alpha}_j}(\cdot) = \left(q_{\boldsymbol{\alpha}_{j,h}}(\cdot)\right)_{h=1}^{m_j} \text{ with } \begin{cases} \sum_{h=1}^{m_j} q_{\boldsymbol{\alpha}_{j,h}}(\cdot) = 1, \\ 0 \leq q_{\boldsymbol{\alpha}_{j,h}}(\cdot) \leq 1, \end{cases}$$

For continuous features, we set for $\boldsymbol{\alpha}_{j,h} = (\alpha^0_{j,h}, \alpha^1_{j,h}) \in \mathbb{R}^2$

$$q_{\boldsymbol{\alpha}_{j,h}}(\cdot) = \frac{\exp(\alpha^0_{j,h} + \alpha^1_{j,h} \cdot)}{\sum_{g=1}^{m_j} \exp(\alpha^0_{j,g} + \alpha^1_{j,g} \cdot)}.$$

Quantization: research contribution

"Soft" approximation:

$$\boldsymbol{q}_{\boldsymbol{\alpha}_j}(\cdot) = \big(q_{\boldsymbol{\alpha}_{j,h}}(\cdot)\big)_{h=1}^{m_j} \ \text{ with } \begin{cases} \sum_{h=1}^{m_j} q_{\boldsymbol{\alpha}_{j,h}}(\cdot) = 1, \\ 0 \le q_{\boldsymbol{\alpha}_{j,h}}(\cdot) \le 1, \end{cases}$$

For continuous features, we set for $\boldsymbol{\alpha}_{j,h} = (\alpha^0_{j,h}, \alpha^1_{j,h}) \in \mathbb{R}^2$

$$q_{\boldsymbol{\alpha}_{j,h}}(\cdot) = \frac{\exp(\alpha^0_{j,h} + \alpha^1_{j,h}\cdot)}{\sum_{g=1}^{m_j} \exp(\alpha^0_{j,g} + \alpha^1_{j,g}\cdot)}.$$

For categorical features, we set for $\boldsymbol{\alpha}_{j,h} = (\alpha_{j,h}(1), \dots, \alpha_{j,h}(l_j)) \in \mathbb{R}^{l_j}$

$$q_{\boldsymbol{\alpha}_{j,h}}(\cdot) = \frac{\exp\left(\alpha_{j,h}(\cdot)\right)}{\sum_{g=1}^{m_j} \exp\left(\alpha_{j,g}(\cdot)\right)}.$$

Quantization: research contribution

We wish to maximize the following likelihood:

$$(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\alpha}}) = \underset{\boldsymbol{\theta},\boldsymbol{\alpha}}{\operatorname{argmax}}\, \ell(\boldsymbol{\theta}, \boldsymbol{\alpha}; \mathrm{x_f}, \mathrm{y_f}) = \underset{\boldsymbol{\theta},\boldsymbol{\alpha}}{\operatorname{argmax}} \sum_{i=1}^{n} \ln p_{\boldsymbol{\theta}}(y_i | \boldsymbol{q}_{\alpha}(\boldsymbol{x}_i)).$$

"$\boldsymbol{\alpha}^\star = \lim_{n\to\infty} \hat{\boldsymbol{\alpha}}$" should be such that $\boldsymbol{q}_{\boldsymbol{\alpha}^\star} = \boldsymbol{q}^\star$.

Quantization: research contribution

We wish to maximize the following likelihood:

$$(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\alpha}}) = \operatorname*{argmax}_{\boldsymbol{\theta},\boldsymbol{\alpha}} \ell(\boldsymbol{\theta}, \boldsymbol{\alpha}; \mathrm{x_f}, \mathrm{y_f}) = \operatorname*{argmax}_{\boldsymbol{\theta},\boldsymbol{\alpha}} \sum_{i=1}^{n} \ln p_{\boldsymbol{\theta}}(y_i | \boldsymbol{q}_{\alpha}(\boldsymbol{x}_i)).$$

"$\boldsymbol{\alpha}^\star = \lim_{n\to\infty} \hat{\boldsymbol{\alpha}}$" should be such that $\boldsymbol{q}_{\boldsymbol{\alpha}^\star} = \boldsymbol{q}^\star$.

Problem: $\hat{\boldsymbol{\alpha}}$ has to diverge, the MLE is at the border of the parameter space which could hinder its properties.

Quantization: research contribution

We wish to maximize the following likelihood:

$$(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\alpha}}) = \operatorname*{argmax}_{\boldsymbol{\theta},\boldsymbol{\alpha}} \ell(\boldsymbol{\theta}, \boldsymbol{\alpha}; \mathrm{x_f}, \mathrm{y_f}) = \operatorname*{argmax}_{\boldsymbol{\theta},\boldsymbol{\alpha}} \sum_{i=1}^{n} \ln p_{\boldsymbol{\theta}}(y_i | \boldsymbol{q}_{\alpha}(\boldsymbol{x}_i)).$$

"$\boldsymbol{\alpha}^\star = \lim_{n\to\infty} \hat{\boldsymbol{\alpha}}$" should be such that $\boldsymbol{q}_{\boldsymbol{\alpha}^\star} = \boldsymbol{q}^\star$.

Problem: $\hat{\boldsymbol{\alpha}}$ has to diverge, the MLE is at the border of the parameter space which could hinder its properties.

Anyway, or more generally if there is no true quantization $\boldsymbol{q}^\star$, $\hat{\boldsymbol{q}}$ is used instead as a quantization candidate.

Quantization: research contribution

We wish to maximize the following likelihood:

$$(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\alpha}}) = \underset{\boldsymbol{\theta},\boldsymbol{\alpha}}{\operatorname{argmax}}\, \ell(\boldsymbol{\theta}, \boldsymbol{\alpha}; \mathrm{x_f}, \mathrm{y_f}) = \underset{\boldsymbol{\theta},\boldsymbol{\alpha}}{\operatorname{argmax}} \sum_{i=1}^{n} \ln p_{\boldsymbol{\theta}}(y_i | \boldsymbol{q}_{\alpha}(\boldsymbol{x}_i)).$$

"$\boldsymbol{\alpha}^\star = \lim_{n\to\infty} \hat{\boldsymbol{\alpha}}$" should be such that $\boldsymbol{q}_{\boldsymbol{\alpha}^\star} = \boldsymbol{q}^\star$.

Problem: $\hat{\boldsymbol{\alpha}}$ has to diverge, the MLE is at the border of the parameter space which could hinder its properties.

Anyway, or more generally if there is no true quantization $\boldsymbol{q}^\star$, $\hat{\boldsymbol{q}}$ is used instead as a quantization candidate.

Problem: $\ell(\boldsymbol{\theta}, \boldsymbol{\alpha}; \mathrm{x_f}, \mathrm{y_f})$ cannot be directly maximized.

Solution: Resort to (stochastic) gradient descent which each step (s) will yield $\hat{\boldsymbol{\alpha}}^{(s)}$ and quantization candidate $\hat{\boldsymbol{q}}^{(s)}$.

Quantization: model = quantization selection

Quantization provider to original selection criterion

We have drastically restricted the search space to *iter* well-chosen candidates resulting from the the gradient descent steps.

$$s^\star = \underset{s=1,\ldots,iter}{\operatorname{argmin}} \ \mathrm{BIC}(\hat{\boldsymbol{\theta}}_{\hat{\boldsymbol{q}}^{(s)}})$$

Quantization: model = quantization selection

Quantization provider to original selection criterion

We have drastically restricted the search space to *iter* well-chosen candidates resulting from the the gradient descent steps.

$$s^\star = \underset{s=1,\dots,iter}{\operatorname{argmin}} \operatorname{BIC}(\hat{\boldsymbol{\theta}}_{\hat{\boldsymbol{q}}^{(s)}})$$

We would still need to loop over candidates $\boldsymbol{m}$!

Quantization: model = quantization selection

Quantization provider to original selection criterion

We have drastically restricted the search space to *iter* well-chosen candidates resulting from the the gradient descent steps.

$$s^\star = \underset{s=1,\dots,iter}{\text{argmin}} \; \text{BIC}(\hat{\boldsymbol{\theta}}_{\hat{\boldsymbol{q}}^{(s)}})$$

We would still need to loop over candidates $\boldsymbol{m}$!

In practice if $\forall i, \; q_{\alpha_{j,h}}(x_j) \ll 1$, then level h disappears while performing the argmax.

Quantization: model = quantization selection

Quantization provider to original selection criterion

We have drastically restricted the search space to *iter* well-chosen candidates resulting from the the gradient descent steps.

$$s^\star = \underset{s=1,\ldots,iter}{\operatorname{argmin}} \ \mathrm{BIC}(\hat{\boldsymbol{\theta}}_{\hat{\boldsymbol{q}}^{(s)}})$$

We would still need to loop over candidates $\boldsymbol{m}$!

In practice if $\forall i, \ q_{\alpha_{j,h}}(x_j) \ll 1$, then level h disappears while performing the argmax.

Start with $\boldsymbol{m} = (m_{\max})_1^d$ and "wait" ...

Bivariate interactions

Bivariate interactions: notations

Upper triangular matrix with $\delta_{k,\ell} = 1$ if $k < \ell$ and features k and ℓ "interact" in the logistic regression.

$$\text{logit}(p_{\boldsymbol{\theta}}(1|\boldsymbol{q}(\boldsymbol{x}))) = \theta_0 + \sum_{j=1}^{d} \theta_j^{\boldsymbol{q}_j(x_j)} + \sum_{1 \leq k < \ell \leq d} \delta_{k,\ell} \theta_{k,\ell}^{\boldsymbol{q}_k(x_k)\boldsymbol{q}_\ell(x_\ell)}.$$

Bivariate interactions: notations

Upper triangular matrix with $\delta_{k,\ell} = 1$ if $k < \ell$ and features k and ℓ "interact" in the logistic regression.

$$\text{logit}(p_{\boldsymbol{\theta}}(1|\boldsymbol{q}(\boldsymbol{x}))) = \theta_0 + \sum_{j=1}^{d} \theta_j^{\boldsymbol{q}_j(x_j)} + \sum_{1 \leq k < \ell \leq d} \delta_{k,\ell} \theta_{k,\ell}^{\boldsymbol{q}_k(x_k)\boldsymbol{q}_\ell(x_\ell)}.$$

Imagine for now that the discretization $\boldsymbol{q}(\boldsymbol{x})$ is fixed. The criterion becomes:

$$(\boldsymbol{\theta}^\star, \boldsymbol{\delta}^\star) = \underset{\boldsymbol{\theta}, \boldsymbol{\delta} \in \{0,1\}^{\frac{d(d-1)}{2}}}{\text{argmin}} \text{BIC}(\hat{\boldsymbol{\theta}}_{\boldsymbol{\delta}}; \mathcal{T}_{\text{f}}).$$

Bivariate interactions: notations

Upper triangular matrix with $\delta_{k,\ell} = 1$ if $k < \ell$ and features k and ℓ "interact" in the logistic regression.

$$\mathrm{logit}(p_{\boldsymbol{\theta}}(1|\boldsymbol{q}(\boldsymbol{x}))) = \theta_0 + \sum_{j=1}^{d} \theta_j^{\boldsymbol{q}_j(x_j)} + \sum_{1 \le k < \ell \le d} \delta_{k,\ell} \theta_{k,\ell}^{\boldsymbol{q}_k(x_k)\boldsymbol{q}_\ell(x_\ell)}.$$

Imagine for now that the discretization $\boldsymbol{q}(\boldsymbol{x})$ is fixed. The criterion becomes:

$$(\boldsymbol{\theta}^\star, \boldsymbol{\delta}^\star) = \underset{\boldsymbol{\theta}, \boldsymbol{\delta} \in \{0,1\}^{\frac{d(d-1)}{2}}}{\mathrm{argmin}} \mathrm{BIC}(\hat{\boldsymbol{\theta}}_{\boldsymbol{\delta}}; \mathcal{T}_{\mathrm{f}}).$$

Analogous to previous problem: $2^{\frac{d(d-1)}{2}}$ models.

Bivariate interactions: model proposal

δ is latent and hard to optimize over: use a stochastic algorithm!

Bivariate interactions: model proposal

δ is latent and hard to optimize over: use a stochastic algorithm!

Strategy used here: Metropolis-Hastings sampling algorithm.
Idea: propose well-chosen interactions and accept / reject them based on the BIC criterion of the resulting logistic regression.

Bivariate interactions: model proposal

δ is latent and hard to optimize over: use a stochastic algorithm!

Strategy used here: Metropolis-Hastings sampling algorithm.
Idea: propose well-chosen interactions and accept / reject them based on the BIC criterion of the resulting logistic regression.

$$p(y|\mathfrak{q}) = \sum_{\boldsymbol{\delta} \in \{0,1\}^{\frac{d(d-1)}{2}}} p(y|\mathfrak{q}, \boldsymbol{\delta}) p(\boldsymbol{\delta})$$

$$p(\boldsymbol{\delta}|\mathfrak{q}, y) \propto \exp(-\text{BIC}[\boldsymbol{\delta}]/2) p(\boldsymbol{\delta})$$

Bivariate interactions: model proposal

δ is latent and hard to optimize over: use a stochastic algorithm!

Strategy used here: Metropolis-Hastings sampling algorithm.
Idea: propose well-chosen interactions and accept / reject them based on the BIC criterion of the resulting logistic regression.

$$p(y|\mathfrak{q}) = \sum_{\boldsymbol{\delta} \in \{0,1\}^{\frac{d(d-1)}{2}}} p(y|\mathfrak{q}, \boldsymbol{\delta}) p(\boldsymbol{\delta})$$

$$p(\boldsymbol{\delta}|\mathfrak{q}, y) \propto \exp(-\mathrm{BIC}[\boldsymbol{\delta}]/2) \cancel{p(\boldsymbol{\delta})} \qquad p(\delta_{p,q}) = \frac{1}{2}$$

Bivariate interactions: model proposal

δ is latent and hard to optimize over: use a stochastic algorithm!

Strategy used here: Metropolis-Hastings sampling algorithm.
Idea: propose well-chosen interactions and accept / reject them based on the BIC criterion of the resulting logistic regression.

$$p(y|\mathbf{q}) = \sum_{\boldsymbol{\delta} \in \{0,1\}^{\frac{d(d-1)}{2}}} p(y|\mathbf{q}, \boldsymbol{\delta}) p(\boldsymbol{\delta})$$

$$p(\boldsymbol{\delta}|\mathbf{q}, y) \propto \exp(-\text{BIC}[\boldsymbol{\delta}]/2) \cancel{p(\boldsymbol{\delta})} \qquad p(\delta_{p,q}) = \frac{1}{2}$$

Which transition proposal $T : (\{0,1\}^{\frac{d(d-1)}{2}}, \{0,1\}^{\frac{d(d-1)}{2}}) \mapsto [0;1]$?

Bivariate interactions: model proposal

$2^{d(d-1)}$ probabilities to calculate...

Bivariate interactions: model proposal

$2^{d(d-1)}$ probabilities to calculate...

We restrict changes to only one entry $\delta_{k,\ell}$.

Bivariate interactions: model proposal

$2^{d(d-1)}$ probabilities to calculate...

We restrict changes to only one entry $\delta_{k,\ell}$.

Proposal: gain/loss in BIC between bivariate models with / without the interaction.

Bivariate interactions: model proposal

$2^{d(d-1)}$ probabilities to calculate. . .

We restrict changes to only one entry $\delta_{k,\ell}$.

Proposal: gain/loss in BIC between bivariate models with / without the interaction.

If the interaction between two features is meaningful when only these two features are considered, there is a good chance that it will be in the full multivariate model.

Bivariate interactions: model proposal

$2^{d(d-1)}$ probabilities to calculate...

We restrict changes to only one entry $\delta_{k,\ell}$.

Proposal: gain/loss in BIC between bivariate models with / without the interaction.

If the interaction between two features is meaningful when only these two features are considered, there is a good chance that it will be in the full multivariate model.

Trick: alternate one discretization / grouping step and one "interaction" step.

SEM-Gibbs quantization

SEM-Gibbs quantization

Originally (and as implemented in the R package glmdisc), the optimization was a bit different:

SEM-Gibbs quantization

Originally (and as implemented in the R package glmdisc), the optimization was a bit different:

- $\boldsymbol{q}$ is considered a latent (unobserved) feature $\mathfrak{q}$;

SEM-Gibbs quantization

Originally (and as implemented in the R package glmdisc), the optimization was a bit different:

- $\boldsymbol{q}$ is considered a latent (unobserved) feature $\mathfrak{q}$;
- A classical EM algorithm is intractable since it requires an Expectation step over all possible quantizations;

SEM-Gibbs quantization

Originally (and as implemented in the R package glmdisc), the optimization was a bit different:

- $\boldsymbol{q}$ is considered a latent (unobserved) feature $\mathfrak{q}$;
- A classical EM algorithm is intractable since it requires an Expectation step over all possible quantizations;
- Solution: random draw $\approx$ Bayesian statistics.

SEM-Gibbs quantization: estimation

“Classical” estimation strategy with latent variables: EM algorithm.

SEM-Gibbs quantization: estimation

"Classical" estimation strategy with latent variables: EM algorithm.

There would still be a sum over $\boldsymbol{\mathcal{Q}_m}$:
$p(y|\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\alpha}) = \sum_{\mathfrak{q} \in \boldsymbol{\mathcal{Q}_m}} p_{\boldsymbol{\theta}}(y|\mathfrak{q}) \prod_{j=1}^{d} p_{\alpha_j}(\mathfrak{q}_j|x_j)$

SEM-Gibbs quantization: estimation

"Classical" estimation strategy with latent variables: EM algorithm.

There would still be a sum over $\boldsymbol{\mathcal{Q}_m}$:
$p(y|\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\alpha}) = \sum_{\mathfrak{q} \in \boldsymbol{\mathcal{Q}_m}} p_{\boldsymbol{\theta}}(y|\mathfrak{q}) \prod_{j=1}^{d} p_{\alpha_j}(\mathfrak{q}_j|x_j)$

Use a Stochastic-EM! Draw $\mathfrak{q}$ knowing that:

SEM-Gibbs quantization: estimation

"Classical" estimation strategy with latent variables: EM algorithm.

There would still be a sum over $\boldsymbol{\mathcal{Q}_m}$:
$p(y|\boldsymbol{x},\boldsymbol{\theta},\boldsymbol{\alpha}) = \sum_{\mathfrak{q}\in\boldsymbol{\mathcal{Q}_m}} p_{\boldsymbol{\theta}}(y|\mathfrak{q}) \prod_{j=1}^{d} p_{\alpha_j}(\mathfrak{q}_j|x_j)$

Use a Stochastic-EM! Draw $\mathfrak{q}$ knowing that:

$$p(\mathfrak{q}|\boldsymbol{x},y) = \underbrace{\frac{p_{\boldsymbol{\theta}}(y|\mathfrak{q}) \prod_{j=1}^{d} p_{\alpha_j}(\mathfrak{q}_j|x_j)}{\sum_{\mathfrak{q}\in\boldsymbol{\mathcal{Q}_m}} p_{\boldsymbol{\theta}}(y|\mathfrak{q}) \prod_{j=1}^{d} p_{\alpha_j}(\mathfrak{q}_j|x_j)}}_{\text{still difficult to calculate}}$$

SEM-Gibbs quantization: estimation

"Classical" estimation strategy with latent variables: EM algorithm.

There would still be a sum over $\boldsymbol{\mathcal{Q}_m}$:
$p(y|\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\alpha}) = \sum_{\mathfrak{q} \in \boldsymbol{\mathcal{Q}_m}} p_{\boldsymbol{\theta}}(y|\mathfrak{q}) \prod_{j=1}^{d} p_{\alpha_j}(\mathfrak{q}_j|x_j)$

Use a Stochastic-EM! Draw $\mathfrak{q}$ knowing that:

$$p(\mathfrak{q}|\boldsymbol{x}, y) = \frac{p_{\boldsymbol{\theta}}(y|\mathfrak{q}) \prod_{j=1}^{d} p_{\alpha_j}(\mathfrak{q}_j|x_j)}{\underbrace{\sum_{\mathfrak{q} \in \boldsymbol{\mathcal{Q}_m}} p_{\boldsymbol{\theta}}(y|\mathfrak{q}) \prod_{j=1}^{d} p_{\alpha_j}(\mathfrak{q}_j|x_j)}_{\text{still difficult to calculate}}}$$

Gibbs-sampling step:

$$p(\mathfrak{q}_j|\boldsymbol{x}, y, \mathfrak{q}_{\{-j\}}) \propto p_{\boldsymbol{\theta}}(y|\mathfrak{q}) p_{\alpha_j}(\mathfrak{q}_j|x_j)$$

SEM-Gibbs quantization: algorithm

Initialization

$$\begin{pmatrix} x_{1,1} & \cdots & x_{1,d} \\ \vdots & \vdots & \vdots \\ x_{n,1} & \cdots & x_{n,d} \end{pmatrix} \overset{\text{at random}}{\Rightarrow} \begin{pmatrix} \mathfrak{q}_{1,1} & \cdots & \mathfrak{q}_{1,d} \\ \vdots & \vdots & \vdots \\ \mathfrak{q}_{n,1} & \cdots & \mathfrak{q}_{n,d} \end{pmatrix}$$

Loop

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \overset{\substack{\text{logistic} \\ \text{regression}}}{\Rightarrow} \begin{pmatrix} \mathfrak{q}_{1,1} & \cdots & \mathfrak{q}_{1,d} \\ \vdots & \vdots & \vdots \\ \mathfrak{q}_{n,1} & \cdots & \mathfrak{q}_{n,d} \end{pmatrix} \overset{\substack{\text{polytomous} \\ \text{regression}}}{\Rightarrow} \begin{pmatrix} x_{1,1} & \cdots & x_{1,d} \\ \vdots & \vdots & \vdots \\ x_{n,1} & \cdots & x_{n,d} \end{pmatrix}$$

Updating $\mathfrak{q}$

$$\begin{pmatrix} p(y_1, \mathfrak{q}_{1,j} = k | \boldsymbol{x}_i) \\ \vdots \\ p(y_n, \mathfrak{q}_{n,j} = k | \boldsymbol{x}_i) \end{pmatrix} \overset{\substack{\text{random} \\ \text{sampling}}}{\Rightarrow} \begin{pmatrix} \mathfrak{q}_{1,j} \\ \vdots \\ \mathfrak{q}_{n,j} \end{pmatrix}$$

Calculating $\mathfrak{q}^{\text{MAP}}$

$$\begin{pmatrix} \mathfrak{q}^{\text{MAP},1,j} \\ \vdots \\ \mathfrak{q}^{\text{MAP},n,j} \end{pmatrix} \overset{\substack{\text{MAP} \\ \text{estimate}}}{=} \begin{pmatrix} \operatorname{argmax}_{\mathfrak{q}_j} p_{\boldsymbol{\alpha_j}}(\mathfrak{q}_j | x_{1,j}) \\ \vdots \\ \operatorname{argmax}_{\mathfrak{q}_j} p_{\boldsymbol{\alpha_j}}(\mathfrak{q}_j | x_{n,j}) \end{pmatrix}$$

SEM-Gibbs quantization: simulations