# Le calage pénalisé

## Théorie et application

Antoine Rebecq

15 mars 2016

INSEE - DMCSI - Division Sondages

# Sommaire

# Principe

# Équation de calage

L'équation de calage "classique" ([2]) s'écrit:

$$\begin{cases} \min\limits_{w_k} \sum\limits_{k \in s} d_k G(\dfrac{w_k}{d_k}) \\ \text{s.c. } \forall j \in [[1, J]], \sum\limits_{k \in s} w_k x_{jk} = T(X_j) = \sum\limits_{k \in \mathcal{U}} x_{jk} \end{cases} \tag{1}$$

# Équation de calage pénalisé

Idée du calage pénalisé : on relâche la contrainte, et on écrit un programme “ridge”. Pour $\lambda^* \in [0, 1]$ :

$$\min_{w_k} \left[ \lambda^* \sum_{k \in s} d_k G\left(\frac{w_k}{d_k}\right) + (1 - \lambda^*) \left[ \sum_{j=1}^{J} c_j \left( \sum_{k \in s} w_k x_{jk} - T(X_j) \right)^2 \right] \right] \quad (2)$$

C = vecteur de coûts

# Équation de calage pénalisé

Principe : On accepte un calage non-exact sur (certaines) marges, de manière à faciliter la convergence de la procédure.

On peut forcer un calage exact sur certaines marges en fixant un coût infini.

# Choix du $\lambda^*$

Si $\lambda^* \to 1$, le terme de distance aux poids initiaux est prépondérant. Les estimateurs sont peu rapprochés des marges mais les rapports de poids sont proches de 1

Si $\lambda^* \to 0$, le terme de coûts est prépondérant : les contraintes de marges sont satisfaites, mais l'estimateur s'éloigne de l'estimateur initial.

# Choix du $\lambda^*$

Le choix du $\lambda^*$ est finalement conditionné par **l'étendue de la distribution des rapports de poids** que l'on souhaite obtenir ([1])

On parle de **paramètre gap**, très similaire aux paramètres de bornes lorsqu'on utilise les méthodes bornées en calage exact.

# Conséquence sur les méthodes

Les méthodes bornées ne sont pas nécessaires en calage pénalisé.

Si G = distance du khi-deux ($\Leftrightarrow$ méthode linéaire), alors le programme possède une solution analytique.

Il est également possible d'utiliser la méthode du raking ratio (distance de l'entropie), qui assure que tous les poids sont positifs, quel que soit le gap.

# Aspects statistiques

# Propriétés statistiques

| | Calage exact | Calage pénalisé |
|---|---|---|
| Biais | Asymptotiquement nul | Asymptotiquement nul |
| Variance | Variance du total des résidus de la régression linéaire | Variance du total des résidus de la régression ridge |
| Cohérence marges | Oui | Possible avec coût infini |
| Maîtrise rapport des poids | Via distances bornées (logit) | Via paramètre gap |

# Remarque sur la distance aux marges

Remarque :

Il n'est pas possible d'entrer *a priori* une zone "acceptable" pour la distance de l'estimateur aux marges (par exemple si l'on souhaite un calage approché au plus à 10% pour toutes les marges du problème).

On vérifie donc l'écart entre les marges et l'estimateur par calage pénalisé *a posteriori* (voir partie 4).

Il faut ensuite jouer sur les coûts ainsi que le gap pour obtenir une solution satisfaisante pour le statisticien.

# Le package R icarus

# Le package R icarus

Ancien nom : gaston

Reprend l'interface ainsi que nombre des fonctionnalités de Calmar, en R.

Ajout du calage sur bornes minimales ainsi que du calage pénalisé.

# Le package R icarus

Installation :

```
R> install.packages("icarus")
```

Page GitHub : `https://github.com/haroine/icarus`
Wiki : `https://github.com/haroine/icarus/wiki`

# Un exemple de calage simple avec icarus

Un exemple simple de calage sur marges avec *icarus* est disponible sur le wiki :
`https://github.com/haroine/icarus/wiki/Calibration`

Le calage et ses variantes dans *icarus* passent par la fonction **calibration**

```
R> help(calibration)
```

# Un exemple de calage simple avec icarus

```
R> ## Calibration margins
R> mar1 <- c("categ",3,80,90,60)
R> mar2 <- c("sexe",2,140,90,0)
R> mar3 <- c("service",2,100,130,0)
R> mar4 <- c("salaire", 0, 470000,0,0)
R> margins <- rbind(mar1, mar2, mar3, mar4)

R> wCalesLinear <- calibration(data=data,
marginMatrix=margins, colWeights="poids")
```

# Un exemple de calage simple avec icarus

Pour le calage pénalisé utilisant un vecteur de coûts **coutsMarges**, il suffit d'ajouter le paramètre "costs" à la fonction **calibration**. *icarus* comprend alors qu'il faut effectuer le calage pénalisé.

```
R> wCalesPen <- calibration(data=data,
marginMatrix=margins, colWeights="poids",
costs=coutsMarges)
```

# Méthodes pour le calage pénalisé

Méthode “raking” disponible à titre expérimental (pas encore testée en production).

```
R> wCalesPen2 <- calibration(data=data,
marginMatrix=margins, colWeights="poids",
method="raking", costs=coutsMarges)
```

# Application au calage de l'enquête ENRJ

# Présentation de l'enquête ENRJ

Enquête Nationale sur les Ressources des Jeunes.

Champ de l'enquête : personnes âgées de 18 à 24 ans. Redressement par correction de la non-réponse et calage.

# Application au calage de l'enquête ENRJ

On souhaite caler sur les structures par sexe $\times$ age [14 modalités].

Et également contrôler l'erreur d'estimation sur des **variables de cadrage** :

- Nombre de bacheliers par sexe/filière
- Nombre d'inscrits à l'université, en BTS, en école d'ingénieur
- Nombre de boursiers sur critères sociaux
- Nombre de bénéficiaires des APL et de la Paje[1]

[1] Prestation accueil du jeune enfant

# Application au calage de l'enquête ENRJ

On teste un calage exact (méthodes linéaire et logit).

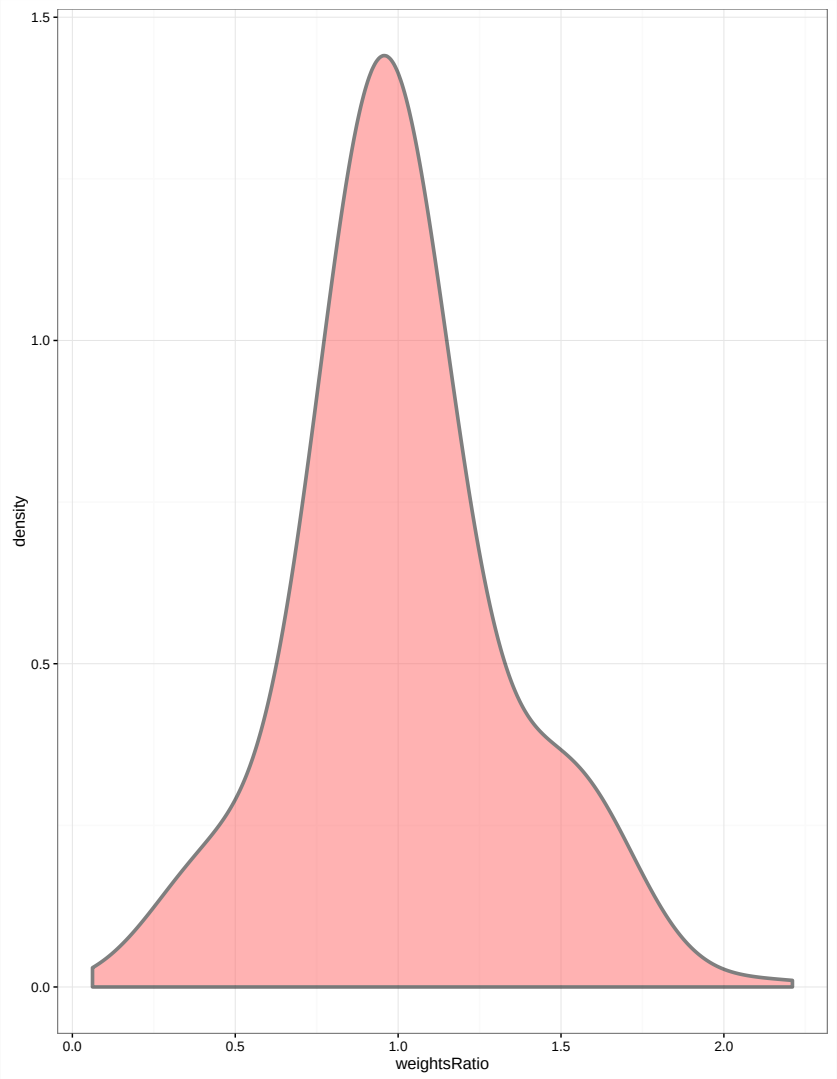
# Application au calage de l'enquête ENRJ

Calage avec méthode linéaire :

```
calibration(data = enrj, marginMatrix = marges,
            colWeights = "poids_CNR",
            method = "linear", popTotal=5198942)
```

Rapports de poids min / max : 0.06 / 2.2

# Application au calage de l'enquête ENRJ

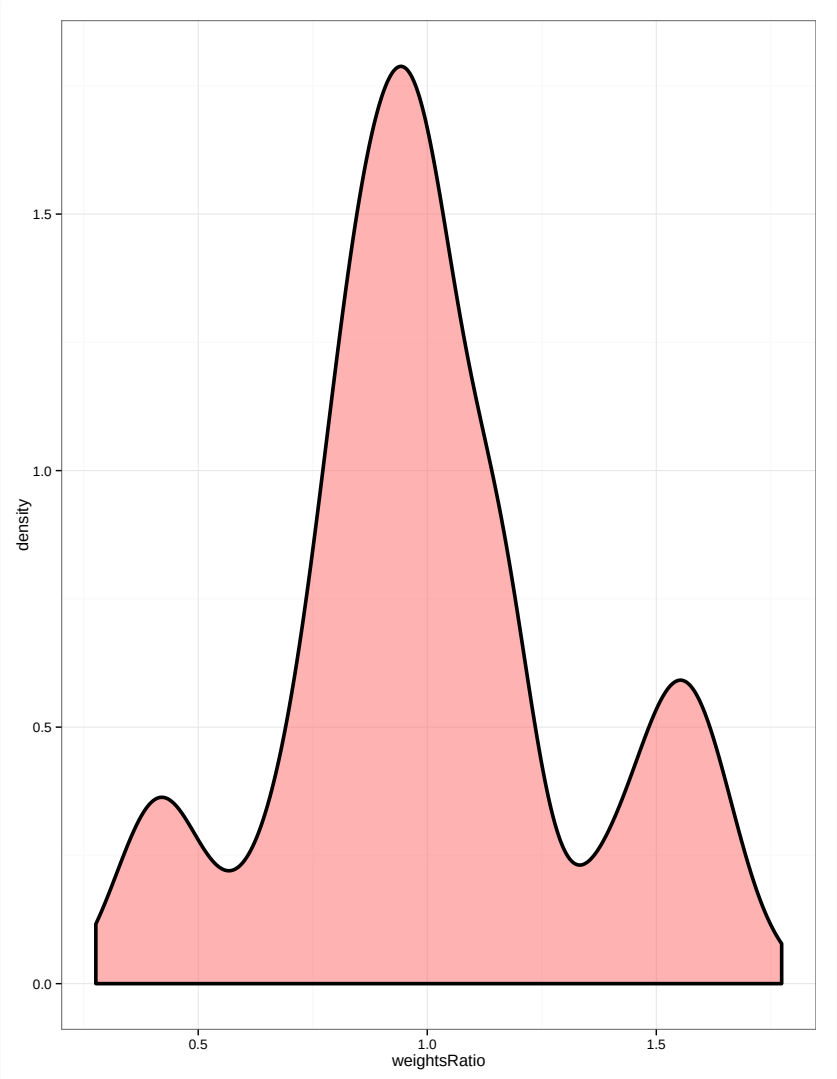


# Application au calage de l'enquête ENRJ

Calage avec méthode logit 0.2 / 1.8 :

```
calibration(data = enrj, marginMatrix = marges,
            colWeights = "poids_CNR",
            method = "logit", bounds=c(0.2,1.8),
            popTotal=5198942)
```

# Application au calage de l'enquête ENRJ



# Application au calage de l'enquête ENRJ

Calage pénalisé avec "gap" de **1.6**.

Les coûts pour les différentes variables du problème sont :

- Structure par sexe et âge : $+\infty$
- Nombre de bacheliers par sexe/filière : **1**
- Nombre d'inscrits à l'université, en BTS, en école d'ingénieur : **700**
- Nombre de boursiers sur critères sociaux : **100**
- Nombre de bénéficiaires des APL et de la Paje[2] : **100**

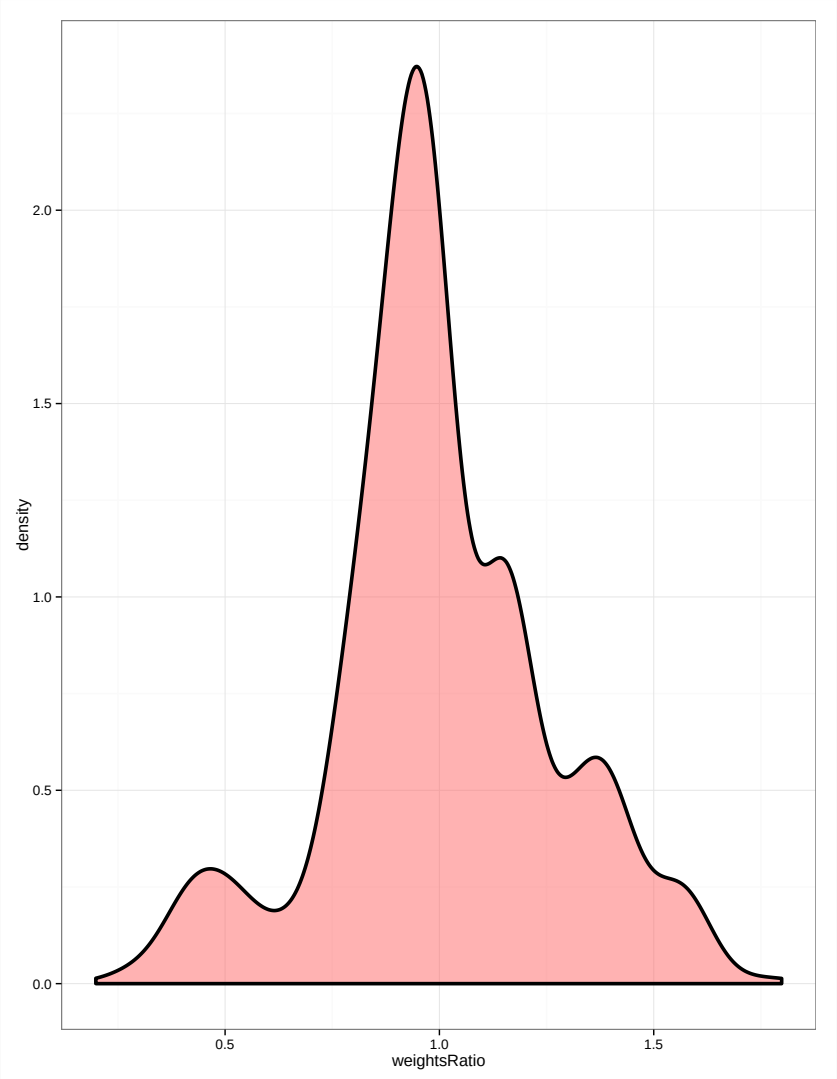
[2] Prestation accueil du jeune enfant

# Application au calage de l'enquête ENRJ

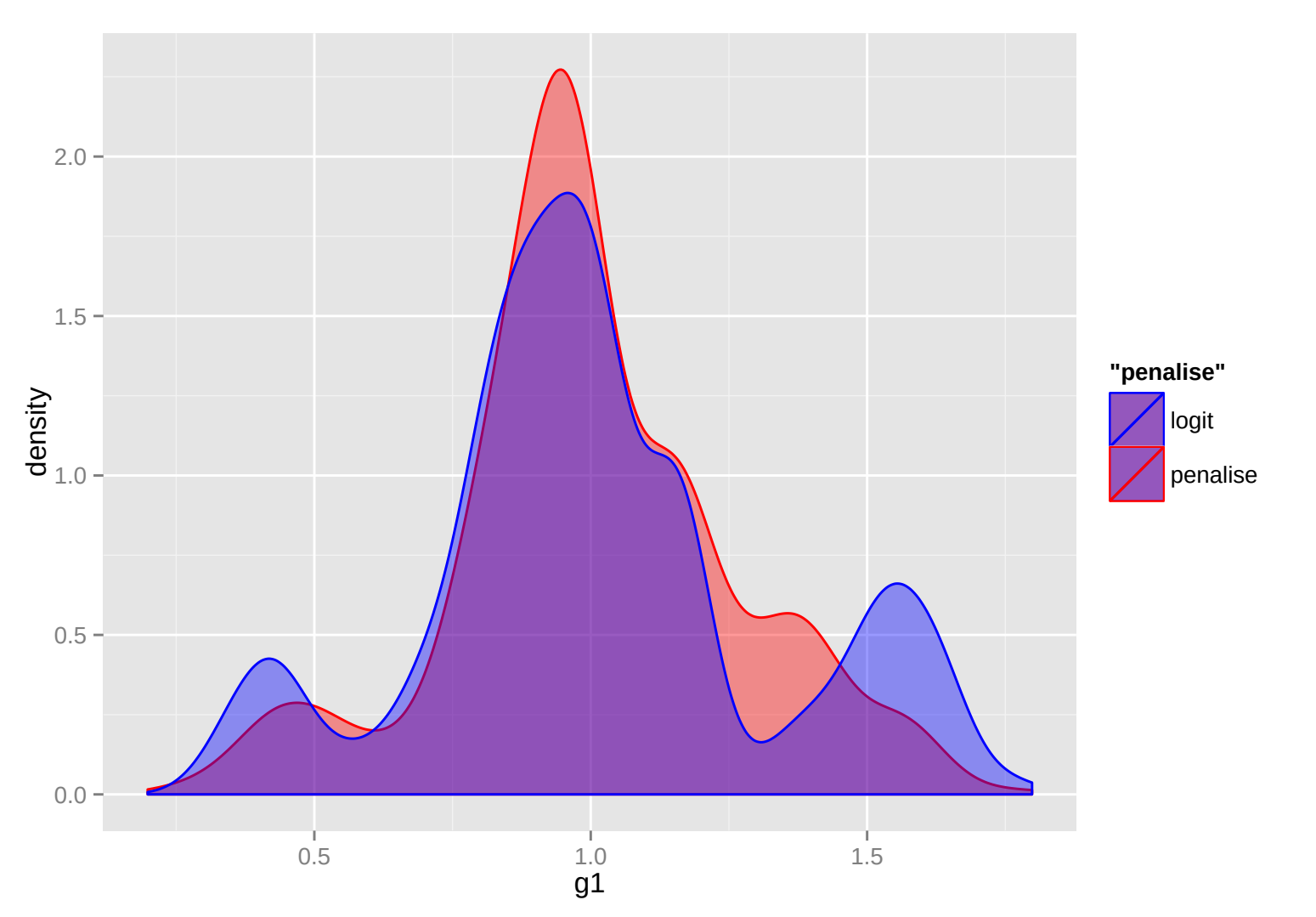
```
coutsENRJ <- c(Inf,700,700,700,700,700,1,1,1,1,1,
               100,100,100,100)

calibration(data = enrj, marginMatrix = marges,
            colWeights = "poids_CNR", gap=1.6,
            costs=coutsENRJ, popTotal=5198942)
```

# Application au calage de l'enquête ENRJ



# Application au calage de l'enquête ENRJ



# Application au calage de l'enquête ENRJ

| Variable | Coût | Écart à la marge, post-stratification | Écart à la marge, calage pénalisé |
|---|---|---|---|
| Bac général | 100 | 3,5% | 3,4% |
| Bac technologique | 100 | 2,2% | -11,4% |
| Bac professionnel | 100 | 7,0% | -7,1% |
| Bac - Hommes | 1 | 8,0% | 0,4% |
| Bac - Femmes | 1 | 0,4% | -2,8% |
| Inscrits à l'université - Hommes | 1 | 2,2% | 8,6% |
| Inscrits à l'université - Femmes | 1 | -4,0% | 3,4% |
| Inscrits en BTS - Hommes | 700 | 97,9% | 10,2% |
| Inscrits en BTS - Femmes | 700 | 90,4% | 10,6% |

# Application au calage de l'enquête ENRJ

| | | | |
|---|---|---|---|
| Inscrits en écoles de commerce ou ingénieur | 1 | -20,8% | -8,0% |
| Boursiers - Hommes | 100 | 15,3% | 7,6% |
| Boursiers - Femmes | 10 | 10,0% | 10,2% |
| Bénéficiaires des APL non en couple | 100 | -30,9% | -7,2% |
| Bénéficiaires de la Paje | 100 | -31,7% | 18,3% |

# Conclusion

# Conclusion

Le calage pénalisé :

- Permet d'intégrer un grand nombre de marges sans déformer excessivement les poids
- Peut-être mis en œuvre en pratique en utilisant le package R *icarus*

Merci pour votre attention !

# References I



J. Bocci and C. Beaumont.
**Another look at ridge calibration.**
*Metron*, 66(1):5–20, 2008.


J.-C. Deville and C.-E. Särndal.
**Calibration estimators in survey sampling.**
*Journal of the American statistical Association*, 87(418):376–382, 1992.


C. Goga and M. A. Shehzad.
**Overview of ridge regression estimators in survey sampling.**
2011.

# References II


F. Guggemos and Y. Tillé.
**Penalized calibration in survey sampling: Design-based estimation assisted by mixed models.**
*Journal of statistical planning and inference*, 140(11):3199–3212, 2010.

A. Rebecq.
**Icarus: an r package for calibration in survey sampling.**
R package version 0.2.0, 2016.