Université de Bourgogne
Institut de Mathématiques de Bourgogne
UMR 5584 du CNRS

Mémoire pour l'obtention du diplôme

Habilitation à Diriger des Recherches

Discipline: Mathématiques

Dr. Camelia Goga

Estimation of nonlinear parameters and analysis of large datasets in a finite population context

Soutenue le 23 septembre 2014 devant le jury composé de :

Benoît Cadre	ENS Rennes	Rapporteur
Jean Opsomer	Colorado State University	Rapporteur
Yves Tillé	Université de Neuchâtel	Rapporteur
David Haziza	Université de Montréal	Examinateur
Samuel Herrmann	Université de Bourgogne	Examinateur
Anne Ruiz-Gazen	Université Toulouse 1 Capitole	Directrice de recherche

Dedic aceasta lucrare parintilor mei, tatalui meu care mi-a insuflat pasiunea pentru matematica si mamei mele pentru incurajarile neobosite.
Toate gandurile mele sunt deasemenea pentru H. fara de a carui dragoste n-as fi razbatut.

Si nu in ultimul rand, pentru J. si Y., sperand ca intr-o zi o sa le placa sa se "amuze" cu cifrele...

Acknowledgements

Je souhaite tout d'abord remercier Benoît Cadre, Jean Opsomer et Yves Tillé qui ont (sans hésiter) bien voulu être les rapporteurs de ce mémoire. Je remercie également David Haziza, Samuel Herrmann et Anne Ruiz-Gazen pour leur participation au jury. Je leur suis reconnaissante pour tous leur efforts pour participer à ma soutenance d'HDR en dépit de la distance et de leurs contraintes professionnelles.

Je remercie vivement Anne Ruiz-Gazen pour tout son soutien, ses encouragements et sa confiance en moi. Je la remercie surtout d'avoir eu la volonté et la patience d'écouter il y a dix ans mon travail de thèse (l'histoire des sondages dans le temps sommairement décrite au tableau par "deux patates"), ce qui a été le point de départ de notre collaboration et amitié que, j'espère, seront de longue durée.

Je remercie tous mes collaborateurs et plus particulièrement, Jean-Claude Deville pour m'avoir insufflé d'envie de chercher toujours plus et de ne jamais s'arrêter, Guillaume pour les échanges fructueux (et les blagues) liés à nos projets de recherche, et ce dernier temps à nos habilitations. Et "the last, but not the least", je remercie Hervé, mon co-équipier préféré, toujours à mon écoute et à mes cotés dans mes nouveaux projets. Grace à son intérêt pour la théorie des sondages, il y a maintenant une équipe de recherche en sondages à l'Université de Bourgogne.

Je remercie toute l'équipe de la "salle café" dont la bonne ambiance fait souvent oublier les petits soucis et les discussions (parfois animées) autour des mathématiques donnent des réponses aux questions qu'on se posent et peuvent même en soulever des nouvelles. Plus particulièrement, je remercie Caroline et Catherine pour leur énergie, disponibilité, bonne humeur et je regrette si parfois, j'ai bruité un peu l'ambiance des repas de midi avec mon stress lié à l'HDR. Merci également à Olivier pour son aide précieuse par rapport à la mise en forme du manuscrit. Une pensée spéciale est pour Patrick pour les discussions fructueuses de point de vue mathématique et enrichissantes personnellement.

Enfin, je remercie toute ma famille, roumaine et française, pour leur soutien et je regrette d'avoir été ce dernier temps un peu "sur un nuage" en train de chercher des réponses... Je remercie mes enfants pour leur patience, leurs encouragements et leurs efforts pour comprendre la nature de ce travail en posant des questions simples mais difficiles à répondre, comme par exemple: "Quand tu auras fini de réviser?" ou "Ca sert à quoi une HDR?"...

Contents

Abbreviations

$U = \{1, \ldots, k, \ldots, N\}$	a finite population of size N
$s \subset U$	a sample
n	the sample size
$p(s)$	the probability of drawing the sample s
$\pi_k = \mathbb{P}(k \in s)$	the first-order inclusion probability, for all $k \in U$
$d_k = \frac{1}{\pi_k}$	sampling weight of individual k
$\pi_{kl} = \mathbb{P}(k \,\&\, l \in s)$	the first-order inclusion probability, for all $k, l \in U$, $k \neq l$
$\Delta_{kl} = \pi_{kl} - \pi_k \pi_l$	
$I_k = \mathbf{1}_{\{k \in s\}}$	the sample membership indicator

$\mathcal{Y}$, y_k	study variable and its value for the k-th individual, $k \in U$
t_y	the finite population total of $\mathcal{Y}$
Φ	non-linear parameter
u_k	the linearized variable of Φ

$\mathcal{X}_1, \ldots, \mathcal{X}_p$	auxiliary variables
$\mathbf{x}_k^T = (x_{k1}, \ldots, x_{kp})$	the vector of auxiliary variables for the kth individual
$\mathbf{X} = (\mathbf{x}_k^T)_{k \in U}$	the $N \times p$ matrix of auxiliary information
$\mathbf{X}_s = (\mathbf{x}_k^T)_{k \in s}$	the $n \times p$ sample restriction of $\mathbf{X}$
$\mathbf{b}^T = (B_1, \ldots, B_q)$	B-spline basis functions with K interior knots and of order m; $q = K + m$

Symbols

HT	Horvitz-Thompson
YGS	Yates-Grundy-Sen
ADU	asymptotically design unbiased
BLU	best linear unbiased (estimator)
GREG	generalized regression (estimator)
PC	principal component

$\|\| \cdot \|\|$	the Euclidian norm
$\|\| \cdot \|\|_2$	the spectral norm
$\|\| \cdot \|\|_{HS}$	the Hilbert-Schmidt norm

$C[0, \mathcal{T}]$	the space of all continuous real functions on the closed interval $[0, \mathcal{T}]$
$D[0, \mathcal{T}]$	the space of cadlag functions: continue *à droite, limite à gauche*
$L^2[0, \mathcal{T}]$	the space of all square integrable functions on the closed interval $[0, \mathcal{T}]$

Introduction

Ce mémoire est une synthèse de mes travaux de recherche menés ces dix dernières années sur la théorie des sondages. Les sondages sont grandement utilisés pour la statistique publique, lors des analyses économiques et sociales (estimation de l'impact de mesures politiques, calcul d'indices de pauvreté, d'inégalité, ...), pour les études de marketing, mais aussi dans la statistique environnementale et l'écologie (combien y-a-t-il d'arbres dans une forêt, la biodiversité a-t-elle diminuée dans telle région ? ...). Plus récemment, les techniques de sondage ont trouvé des applications nouvelles liées au développement d'instruments de mesure et de récolte automatique des données. On peut citer l'exemple des compteurs intelligents utilisés par EDF qui permettent de mesurer la consommation d'électricité à des pas de temps très fins (toutes les minutes) sur des populations très grandes, plusieurs dizaines de millions de compteurs à terme. L'objectif en théorie des sondages est à la fois d'extraire un sous-ensemble de la population, appelé échantillon, et d'estimer, le mieux possible, à partir de ce sous-ensemble, une fonction, linéaire (total) ou non-linéaire (quantiles, indices de pauvreté, ...), des valeurs d'une ou plusieurs variables d'intérêt calculée sur l'ensemble de la population.

Mes travaux de recherche menés ces dernières années gravitent autours de deux grands axes. J'ai essayé de développer

1. des nouvelles méthodes d'estimation de paramètres non-linéaires et de calcul et d'estimation de leur variance;
2. des nouvelles stratégies d'estimation en présence de grandes bases de données pour des objets de grande dimension.

Ces développements ont nécessité l'emploi d'outils et de résultats provenant de différentes branches de la statistique: théorie des sondages, statistique robuste, estimation non-paramétrique, données fonctionnelles et analyse des données. Ils sont aussi, pour la plupart, issus de collaborations avec des spécialistes de ces domaines et d'encadrement de travaux de thèse : Mohamed Chaouch (2005-2008, co-encadrement avec Jérôme Saracco et Ali Ganoun), Muhammad-Ahmed Shehzad (2009-2012, co-encadrement avec H. Cardot), Pauline Lardin (2010-2012, co-encadrement avec H. Cardot) et Anne De Moliner (2013- , co-encadrement avec H. Cardot).

(1) Etude de paramètres non-linéaires

L'étude de paramètres d'intérêt non-linéaires comme par exemple, les mesures d'inégalités (indice de Gini, Théil) ou les quantiles, est devenu crucial ces dernières années dans beaucoup d'enquêtes françaises et européennes. On cherche dans un premier temps à construire un estimateur qui soit précis et efficace. Ensuite, vient le calcul de la variance et la construction

d'un estimateur de cette variance afin d'avoir un ordre d'idée de la précision de la méthode employée et éventuellement de construire un intervalle de confiance. Je me suis d'abord intéressée à l'estimation de ce type de paramètres dans un cadre temporel, et ensuite à l'amélioration de ces estimateurs par la prise en compte de l'information auxiliaire à l'aide de modèles non-paramétriques.

(*a*) **Sondage temporel.** J'ai commencé par étudier dans ma thèse l'estimation d'une combinaison linéaire de différents totaux lorsque les données proviennent de plusieurs échantillons. Il s'agit par exemple de l'estimation de l'évolution dans le temps du revenu total (voir [18] et [19] dans la liste de mes publications) en autorisant les échantillons aux deux époques à être distincts. Pendant l'année d'ATER passée à l'Université Toulouse 1 Capitole, j'ai commencé à travailler avec A. Ruiz-Gazen sur l'extension de ces travaux au cas d'un paramètre non-linéaire tel que l'indice de Gini. En s'appuyant sur des outils et résultats de la statistique robuste, nous donnons dans [11], [15] une nouvelle classe d'estimateurs ainsi qu'une méthode générale de linéarisation qui permet, dans le cas des enquêtes sur deux échantillons (l'évolution de l'indice de Gini entre deux périodes de temps), d'estimer la variance.

Les méthodes de **réechantillonnage** (jacknife, bootstrap) sont souvent préférées aux techniques de linéarisation car elles évitent les difficultés techniques engendrées par le calcul d'un estimateur de la variance. Leur domaine d'application est pourtant plus restrictif. Avec A. Ruiz-Gazen et G. Chauvet, nous réalisons dans [13], une comparaison bootstrap-linéarisation et une application à l'estimation de l'indice de Gini (voir [22]) est actuellement en révision.

(*b*) **Modèle non-paramétrique.** C'est un fait bien connu en sondage que l'usage de l'information auxiliaire permet très souvent d'améliorer la qualité des estimateurs. J'ai proposé dans ma thèse et publié ensuite dans [16] une nouvelle approche nonparamétrique basée sur les splines de régression pour estimer le total d'une variable d'intérêt sur une population finie. Cette méthode est étendue ensuite lors de l'estimation de l'évolution dans le temps d'un paramètre linéaire de totaux ([17]) et lors de l'estimation des quantiles en population finie ([14]) en collaboration avec Y. Aragon et A. Ruiz-Gazen. En collaboration avec A. Ruiz-Gazen, nous suggérons dans [3] une approche générale pour prendre en compte l'information auxiliaire par des modèles non-paramétriques pour l'estimation de paramètres non-linéaires. Cette méthode est présentée de manière détaillée dans [5] pour le cas particulier de l'estimation d'un odds-ratio, indice très souvent utilisé dans les enquêtes épidémiologiques.

(2) Données massives et sondages

Les statisticiens sont de plus en plus souvent confrontés à des données qui sont à la fois de grandes dimension et qui proviennent potentiellement de bases de données gigantesques. Nous sommes à l'ère des "Big Data". Dans ces conditions, les techniques de sondages qui offrent un bon compromis entre taille de données à analyser et précision de l'estimation sont des méthodes qui reviennent sur le devant de la scène. Les travaux menés autour de ce thème concernent les

sondages pour des données fonctionnelles, motivés par l'estimation de courbes de charge, et la prise en compte de l'information auxiliaire lorsque les variables auxiliaires sont très nombreuses.

(a) **Sondages pour des données fonctionnelles**. Lorsque les variables statistiques sont mesurées à des pas très fins (seconde ou minute), celles-ci peuvent être considérées comme des courbes ou des fonctions du temps. Lorsque nous avons rencontré Alain Dessertaine en 2006 et qu'il nous a présenté les problématiques de recherche qui se posaient au département R & D d'EDF, liées à l'estimation par sondage de courbes de charge moyenne, les travaux combinant sondage et analyse des données fonctionnelles étaient quasiment inexistants. Les deux premiers travaux ([10], [12]) que nous avons menés sur ce thème portaient sur l'analyse en composantes principales fonctionnelle pour des données obtenues avec des plans de sondages complexes. Ils ont été réalisés en collaboration avec H. Cardot, M. Chaouch et C. Labruère. Suite à ces travaux, EDF a financé la thèse CIFRE de P. Lardin. L'objectif principal était d'améliorer l'estimation de la courbe moyenne d'électricité en considérant plusieurs stratégies ([4], [6], [7]) qui permettaient de prendre en compte de l'information auxiliaire. Un article de synthèse concernant données fonctionnelles et sondages va bientôt paraitre dans un numéro spécial du *Journal de la Société Française de Statistique* ([1]).

Parallèlement j'ai étudié l'estimation du profil médian dans le cas des plans de sondages complexes ([8]) avec M. Chaouch (qui a été recruté comme ingénieur de recherche à EDF après l'obtention de sa thèse).

Les travaux avec EDF se poursuivent actuellement avec une nouvelle thèse, commencée en octobre 2013 par Anne De Moliner (actuellement ingénieure chercheuse à EDF). Ce nouveau projet de recherche porte sur la prise en compte des courbes influentes par des techniques robustes, sur le traitement des valeurs manquantes dans les courbes par différentes méthodes d'imputation et l'estimation dans des petits domaines.

(b) **Information auxiliaire en grande dimension.** Les enquêtes pour lesquelles les variables auxiliaires sont très nombreuses sont de plus en plus fréquentes. Par exemple, les mesures concernant l'audience sur internet pratiquées par Médiamétrie, ont été enrichies ces dernières années par l'insertion de marqueurs ou tags qui permettent une analyse exhaustive du trafic et de la fréquentation d'un site. Dans ces conditions, on peut se poser la question: faut-il considérer toute cette information? En collaboration avec G. Chauvet ([25]), nous avons montré qu'un nombre trop important de variables auxiliaires peut altérer la qualité des estimateurs et nous proposons une méthode pas à pas de sélection des variables les plus pertinentes. De façon alternative, des méthodes comme la régression ridge ou la régression sur composantes principales peuvent être utilisées. La thèse de M.-A. Shehzad a porté sur le développement et l'application de ces méthodes dans le cadre d'estimation d'un total pour des données issues d'enquêtes ([2], [21] et [23]). Une illustration sur des données issues des enquêtes menées par Médiamétrie est réalisée dans [27].

Le manuscript est organisé de la façon suivante. Le chapitre 1 donne une très brève présentation de principaux résultats sur l'estimation de totaux en théorie des sondages. Le chapitre 2 présente une synthèse des travaux concernant l'estimation des paramètres non-linéaires et la prise en compte de l'information auxiliaire par des modèles nonparamétriques. Cependant, la présentation abordée ici est différente de celle de Goga et al. [2009] et Goga and Ruiz-Gazen [2014a]. Cette manière de présenter les résultats permet en particulier de justifier l'usage de la fonction d'influence telle qu'elle a été définie par Deville [1999a]. Il s'agit du résultat de nombreuses discussions avec A. Ruiz-Gazen et je pense que ce mémoire de HDR est la meilleure opportunité pour présenter cette approche. Pour finir, le chapitre 3 présente une synthèse des travaux réalisés autour du thème "données massives et sondages."

Liste des travaux:

1. Lardin, P., Cardot, H. and Goga, C. (2014). Analysing large datasets of functional data: a survey sampling point of view (à paraître dans un numéro spécial dans le *Journal de la Société Française de Statistique*).

2. Goga, C. and Shehzad, M.-A. (2014). A note on partially penalized calibration, *Pakistan Journal of Statistics*, 30, 429-438.

3. Goga, C. and Ruiz-Gazen, A. (2014). Efficient estimation of non-linear finite population parameters using non-parametrics, *Journal of Royal Statistical Society, Series B*, 76, 113-140.

4. Cardot, H., Goga, C. and Lardin, P. (2014). Variance estimation and asymptotic confidence bands for the mean estimator of sampled functional data with high entropy unequal probability sampling designs, *Scandinavian Journal of Statistics*, 41, 516-534.

5. Goga, C. and Ruiz-Gazen, A. (2014). Estimating the odds-ratio using auxiliary information, à paraître dans *Mathematical Population Studies*, numéro spécial.

6. Cardot, H., Dessertaine, A., Goga, C., Josserand, E. et Lardin, P. (2013). Comparison of different sample designs and construction of confidence bands to estimate the mean of functional data: An illustration on electricity consumption, *Survey Methodology*, 39, 283-301.

7. Cardot, H., Goga, C. and Lardin, P. (2013). Uniform convergence and asymptotic confidence bands for model-assisted estimators of the mean of sampled functional data, *Electronic Journal of Statistics*, 7, 562-596.

8. Chaouch, M. and Goga, C. (2012). Using complex surveys to estimate the L1-median of a functional variable: application to electricity load curves, *The International Statistical Review, Special Issue On Energy*, 80, 40-59.

9. Chaouch, M. and Goga, C. (2010). Design-based estimation for geometric quantiles with application to outlier detection, *Computational Statistics and Data Analysis*, 54, 2214-2229.

10. Cardot, H, Chaouch, M., Goga, C. et Labruère, C. (2010). Properties of Design-based Functional Principal Components Analysis. *Journal of Statistical Planning and Inference*, 140, 75-91.

11. Goga, C., Deville, J.C. and Ruiz-Gazen, A. (2009). Use of functionals in linearization and composite estimation with application to two-survey data, *Biometrika*, 96, 691-709.

12. Cardot, H., Chaouch, M., Goga, C. and Labruère, C. (2008). Functional Principal Components Analysis with Survey Data. In *Functional and Operatorial Statistics*, Sophie Dabo-Niang and Frédéric Férraty (eds.), Springer-Verlag Heidelberg.

13. Chauvet, G., Goga, C. et Ruiz-Gazen, A. (2008). Estimation de variance en présence de deux échantillons : linéarisation et bootstrap. In *Méthodes de sondages*, Guibert, P., Haziza, D., Ruiz-Gazen, A. et Tillé, Y. (eds.), Dunod, Sciences Sup, Paris 2008, pp. 369-374.

14. Aragon, Y., Goga, C. et Ruiz-Gazen, A. (2006). Estimation non paramétrique de quantiles en présence d'information auxiliaire. In *Méthodes d'enquêtes et sondages. Pratiques européenne et nord-américaine*, Pierre Lavallée et Louis-Paul Rivest (eds.), Dunod, Sciences Sup, Paris 2006, pp. 377-382.

15. Goga, C., Deville, J.-C. and Ruiz-Gazen, A. (2006). Linéarisation par la fonction d'influence pour des données issues de deux échantillons. In *Méthodes d'enquêtes et sondages. Pratiques européenne et nord-américaine* Pierre Lavallée et Louis-Paul Rivest (eds.), Dunod, Sciences Sup, Paris 2006, pp. 382-388.

16. Goga, C. (2005). Réduction de la variance dans les sondages en présence d'information auxiliaire : une approche nonparamétrique par splines de régression. *Canadian Journal of Statistics/Revue Canadienne de Statistique*, 33, 2, 163-180.

17. Goga, C. (2005). Estimation de l'évolution d'un total en présence d'information auxiliaire. Une approche par splines de régression. *Comptes Rendus de l'Academie des Sciences*, Ser. 1, 339 (2005), 441-444.

18. Deville, J.C. et Goga, C. (2004). Estimation par régression par polynômes locaux dans des enquêtes sur plusieurs échantillons". In *Echantillonnage et méthodes d'enquêtes.*, Pascal Ardilly (eds.), Dunod, Sciences Sup, Paris 2004, pp. 156-162.

19. Deville, J.C et Goga, C. (2004). Les enquêtes répétées : une synthèse et quelques nouveautés. In *Echantillonnage et méthodes d'enquêtes*, Pascal Ardilly (eds.), Dunod, Sciences Sup, Paris 2004, pp. 148-156.

20. Goga, C. (1999). O generalizare a modelului statistic al lui Franklin. *Mathematical Reports* (en roumain), 1(51), 211-226.

Travaux soumis et en révision

21. Goga, C., Cardot, C. and Shehzad, M. A. (2014). *Calibration and Partial Calibration on Principal Components when the Number of Auxiliary Variables is large* (soumis).

22. Chauvet, G. and Goga, C. (2013). *Gini coefficient and Gini coefficient change : linearization versus bootstrap to estimate the variance* (en révision pour *Survey Methodology*).

Travaux en cours

23. Goga, C. and Shehzad, M. A. (2014). *Penalization in survey sampling: a unified point of view.*

24. Goga, C. and Ruiz-Gazen, A. (2013). *Nonparametric B-spline Calibration.*

25. Chauvet, G. and Goga, C. (2013). *Bootstrap variance estimation and variable selection.*

26. Cardot, H., De Moliner, A. and Goga, C. (2014). *Estimating with kernel smoothers the mean of functional data in a finite population when some data are missing.*

Actes de congrès avec comité de lecture

27. Goga, C., Shehzad, M. A. and Vanheuverzwyn, A. (2011). Principal component regression with survey data. Application on the French Media Audiance. Int. Statistical Inst., Proceedings of the 58th ISI World Statistics Congress - Dublin 2011 (Session CPS002), 3847-3852.

Chapter 1

A brief overview on the estimation of finite population totals

This chapter gives a brief presentation of the main estimators as well as some well known asymptotic results in survey sampling theory, especially for the estimation of finite population totals. The results presented here are not new (except result 1.4) and the stress is put on some results needed in the next chapters. The required assumptions are introduced and discussed. Section 1.1 fixes the notations and introduces the Horvitz-Thompson (HT) estimator of a finite population total as well as its variance and variance estimator. The asymptotic properties of the HT estimator are recalled in Section 1.2 and Section 1.3 presents the Hájek approximation of the variance for πps sampling designs. Finally, Section 1.5 deals with different approaches for improving the HT estimator.

1.1 The Horvitz-Thompson (HT) estimator of finite population totals

Consider the finite population $U = \{1, \ldots, k, \ldots, N\}$ and suppose we wish to estimate the total t_y of a study variable $\mathcal{Y}$ over the population U :

$$t_y = \sum_{k \in U} y_k,$$

where y_k is the value of $\mathcal{Y}$ for the kth unit. Let $s \subset U$ be a probability sample selected from U according to a sampling design $p(\cdot)$. More exactly, $p(\cdot)$ is a probability distribution defined on the set $\mathcal{S}$ of all possible subsets of U and $p(s)$ is the probability of selecting the sample s. Given $p(\cdot)$, each unit k from the population is assigned a known inclusion probability $\pi_k = \Pr(\mathrm{k} \in \mathrm{s}) > 0 = \sum_{\mathrm{k} \ni \mathrm{s}} \mathrm{p(s)}$, and a corresponding sampling design weight $d_k = 1/\pi_k$. We suppose that y_k is known for all $k \in s$ (complete response).

Without auxiliary information, the total t_y can be estimated by the well-known HT estimator (Horvitz and Thompson [1952]):

$$\hat{t}_{yd} = \sum_{k \in s} d_k y_k = \sum_{k \in s} \frac{y_k}{\pi_k}. \tag{1.1}$$

If we denote by $I_k = \mathbf{1}_{\{k \in s\}}$ the sample membership indicator, then the HT estimator may be written as

$$\hat{t}_{yd} = \sum_{k \in U} d_k y_k I_k.$$

If all the first-order inclusion probabilities are positive, $\pi_k > 0$, then the HT estimator is unbiased for t_y with respect to the sampling design $p(\cdot)$, namely

$$\mathbb{E}_p(\hat{t}_{yd}) = t_y,$$

where $\mathbb{E}_p$ is the expectation with respect to the sampling design. Its variance with respect to $p(\cdot)$ is given by

$$\mathbb{V}_p(\hat{t}_{yd}) = \sum_{k \in U} \sum_{k \in U} (\pi_{kl} - \pi_k \pi_l) d_k d_l y_k y_l, \tag{1.2}$$

where $\pi_{kl} = \Pr(\mathrm{k,l} \in \mathrm{s})$ is the second-order inclusion probability. If all $\pi_{kl} > 0$, then $\mathbb{V}_p(\hat{t}_{yd})$ can be estimated unbiasedly by the HT variance estimator:

$$\hat{V}(\hat{t}_{yd}) = \sum_{k \in s} \sum_{k \in s} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} d_k d_l y_k y_l. \tag{1.3}$$

Proportional-to-size designs

For fixed-size sampling designs, the variance $\mathbb{V}_p$ given in (1.2) may be written as

$$\mathbb{V}_{YGS}(\hat{t}_{yd}) = -\frac{1}{2} \sum_{k \in U} \sum_{l \in U} (\pi_{kl} - \pi_k \pi_l) \left(d_k y_k - d_l y_l\right)^2. \tag{1.4}$$

This formula is due to Yates and Grundy [1953] and Sen [1953]. This variance can be estimated by

$$\hat{V}_{YGS}(\hat{t}_{yd}) = -\frac{1}{2} \sum_{k \in s} \sum_{l \in s} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} \left(d_k y_k - d_l y_l\right)^2, \tag{1.5}$$

which may be different from the HT variance estimator given in (1.3).

We can remark from (1.4), that $\mathbb{V}_{YGS}(\hat{t}_{yd}) = 0$ for a sampling design such that π_k is proportional to y_k, for all $k \in U$. In practice, we can take π_k to be proportional to a real auxiliary variable

$\mathcal{X}$ which is nearly proportional to the variable of interest and whose values x_k, supposed to be positive, are known for all $k \in U$. The inclusion probabilities are then given by:

$$\pi_k = n\frac{x_k}{\sum_{k\in U} x_k}. \tag{1.6}$$

If some x_k values are very large, it may happen that the above $\pi_k > 1$ for some elements. In this situation, we could set $\pi_k = 1$ for all k such that $nx_k > \sum_{k\in U} x_k$ and let π_k be proportional to X for the remaining elements k. Without replacement designs satisfying (1.6) are called *πps designs*. For given first-order inclusion probabilities π_k, there are many such sampling designs (see *e.g* Brewer and Hanif [1983] and Tillé [2006]). In particular, the balanced sampling (Deville and Tillé [2004]) balanced on π_k leads to a πps sampling design.

1.2 Asymptotic properties of the HT estimator

We are interested in investigating the asymptotic properties of certain estimators, among which the HT estimator receives a special attention, as the sample size and population size become large. We consider the asymptotic framework introduced by Isaki and Fuller [1982] and a sequence of growing and nested populations U_N of size N tending to infinity. A sample s_N of size n_N growing to infinity is drawn from U_N according to the sampling design $p_N(s_N)$. Note that while the sequence of populations is nested, the sequence of samples is not. The first and second order inclusion probabilities are respectively denoted by π_{kN} and π_{klN}. For simplicity of notations and when there is no ambiguity, the subscript N is dropped. Usually, we are interested in establishing the following asymptotic properties.

Definition 1.1. An estimator $\hat{\theta}$ is consistent for θ if

$$\text{for all } \varepsilon > 0, \quad \lim_{N\to\infty} \mathbb{P}(|\hat{\theta} - \theta| > \varepsilon) = 0.$$

Definition 1.2. An estimator $\hat{\theta}$ is asymptotically design-unbiased (ADU) for θ if

$$\lim_{N\to\infty} \mathbb{E}_p(\hat{\theta}) = \theta.$$

We are often interested in estimating nonlinear estimators such as the calibration or the GREG-type estimators which can be shown to be asymptotically equivalent to linear combination of HT-type estimators. So, the consistency and the ADU-ness of the HT estimator $\hat{t}_{yd}$ are briefly recalled. In order to obtain these properties, the following assumptions on the sampling designs and on the study variable $\mathcal{Y}$ are generally supposed.

Assumption S1. Assume that $\lim_{N\to\infty} \frac{n}{N} = \pi \in (0, 1)$.

Assumption S2. Assume that $\min_{k\in U}\pi_k \geq \tilde{\lambda}$ and $\min_{k,l\in U}\pi_{kl} \geq \lambda^*$ with $\tilde{\lambda}$ and λ^* some positive constants and

$$\overline{\lim}_{N\to\infty} n \max_{k\neq l\in U} |\pi_{kl} - \pi_k\pi_l| < C_1 < \infty,$$

with C_1 a positive constant.

Assumption V1. Assume that $\lim_{N\to\infty} \frac{1}{N}\sum_{k\in U} y_k^2 < \infty$.

Assumptions (S1) and (S2) deal with first and second order inclusion probabilities and are rather classical in survey sampling theory (see also Robinson and Särndal [1983] and Breidt and Opsomer [2000]) but exclude situations in which the sampling fraction is negligible. They are satisfied for many sampling designs including the simple random sampling without replacement, the stratified sampling and the rejective sampling (Hájek [1981], Boistard et al. [2012]). Assumption (S2) is not satisfied for systematic sampling design. It may be weakened as seen in Breidt and Opsomer [2008] to include the cluster sampling design, but the rates of convergence of the HT estimator are generally slower.

Theorem 1.3. *(Consistency of the HT estimator) Make assumptions (S1)-(S2) and (V1). Then*

$$\lim_{N\to\infty} n\mathbb{E}_p\left[\frac{1}{N}(\hat{t}_{yd} - t_y)\right]^2 < \infty.$$

By the Markov inequality, we can obtain easily the consistency and the ADU-ness of the HT estimator. Isaki and Fuller [1982] have also proved the consistency of the HT estimator but assuming a slightly different assumption of the moments on $\mathcal{Y}$. The book by Fuller [2009], chapter 1, discusses also the consistency of the HT estimator.

1.3 Hájek approximation for the variance

The HT variance given in (1.2) is difficult to compute since it contains double sums and requires the knowledge of $\mathrm{Cov}_p(I_k, I_l) = \pi_{kl} - \pi_k\pi_l$ which involves the second-order inclusion probabilities π_{kl} for all $k \neq l$. These quantities are difficult to compute or even unknown, especially for unequal probability sampling designs such as the πps designs. As suggested by Hájek [1981], attempts should be done in order to find fixed-size designs simple to put into practice in which $\mathrm{Cov}_p(I_k, I_l)$ would be approximated by a simple structure as follows:

$$\mathrm{Cov}_p(I_k, I_l) \simeq c_k c_l, \quad k \neq l, \tag{1.7}$$

allowing in this way to compute easily an approximation of the variance. Among unequal probability sampling designs, the Poisson (PO) sampling design has received a special attention. The PO sampling design is a without replacement design for which the variables I_k are independent

and distributed as

$$P(I_k = 1) = \pi_k, \quad P(I_k = 0) = 1 - \pi_k, \quad k \in U.$$

This means that $\pi_{kl} = \pi_k \pi_l$ for $k \neq l \in U$ and the variance HT formula becomes easy to calculate in this case: the covariance terms are zero and the variance terms are equal to $\pi_k(1 - \pi_k)$. However, an important drawback of the PO design is the fact that its size is random. By conditioning on the event $n_s = n$, we obtain the PO sampling design conditioned to size or the rejective sampling design in which a sample is selected according to a PO design and rejected if it does not have the desired size. The operation is continued until a sample of size n is obtained. Hájek [1964] showed that the second-order inclusion probabilities for the rejective sampling may be approximated as follows

$$\pi_{kl} = \pi_k \pi_l \left\{ 1 - \frac{(1 - \pi_k)(1 - \pi_l)}{D(\pi)} [1 + o(1)] \right\}$$

where $D(\pi) = \sum_{k \in U} \pi_k(1 - \pi_k)$ and if $\lim_{N \to \infty} D(\pi) = \infty$. This means that in the case of the rejective sampling, the covariance terms $\mathrm{Cov}_p(I_k, I_l)$ for $k \neq l$ satisfy relation (1.7) and they are all negative ensuring the positivity of the YGS variance estimator given in (1.5). Further, many sampling designs may be derived by means of conditioning as developed in Tillé [2006] so that the above approximation is valid. Hájek [1981] proves that this approximation is also valid for the Sampford-Durbin sampling.
Taking into account these considerations, the following assumption on the second-order inclusion probabilities will be assumed for the rest of this document.

Assumption S3. Assume that

$$\pi_{kl} = \pi_k \pi_l \left\{ 1 - \frac{(1 - \pi_k)(1 - \pi_l)}{D(\pi)} [1 + o(1)] \right\}, \quad \text{for} \quad k, l \in U$$

where $D(\pi) = \sum_{k \in U} \pi_k(1 - \pi_k)$ and $\lim_{N \to \infty} D(\pi) = \infty$.

If assumption (S3) is satisfied, then the YGS variance may be approximated by

$$\mathbb{V}_H(\hat{t}_{yd}) = \sum_{k \in U} (y_k - R\pi_k)^2 \left(\frac{1}{\pi_k} - 1 \right) \tag{1.8}$$

for $R = \dfrac{\sum_{k \in U} y_k(1 - \pi_k)}{D(\pi)}$. This formula is also called the *Hájek variance approximation*. Moreover, the assumption S3 implies that $\limsup_{N \to \infty} n \max_{k \neq l \in U} |\pi_{kl} - \pi_k \pi_l| < C_1 < \infty$.

Among fixed-size sampling designs with fixed first-order inclusion probabilities π_k, the conditional PO sampling design possesses a remarquable property: it is the sampling design with the highest entropy (Hájek [1981], Chen et al. [1994]), where the entropy of a sampling design is

defined as

$$H(p) = -\sum_{s\in\mathcal{S}} p(s)\ln(p(s)), \tag{1.9}$$

with the convention $0\ln 0 = 0$ and $\mathcal{S}$ is the set of all possible samples. So, the entropy is a measure of spread of the sampling design and as Hájek [1981] remarked, "we shall try to spread probabilities as much as is reasonable and compatible with prescribed probabilities of inclusion $\pi_1, \ldots, \pi_N$ excluding eventually those samples which are a priori undesirable." Berger [1998b] proves that the Hájek variance approximation may be used for designs asymptotically maximum entropy. Berger [1998b] used the Kullback-Leibler divergence $\mathcal{K}(p, p_{rej})$ with respect to p_{rej} and defined by:

$$\mathcal{K}(p, p_{rej}) = \sum_{s\in\mathcal{S}} p(s) log\left(\frac{p(s)}{p_{rej}(s)}\right) \tag{1.10}$$

as a measure of "divergence" of a design p with respect to the rejective sampling p_{rej}. A design is then asymptotically maximum entropy if $\lim_{N\to\infty} K(p, p_{rej}) = 0$.

Variants and refinements of the Hájek variance formula as well as variance estimators are proposed in Deville and Tillé [2005]. Matei and Tillé [2005] show on simulations that these approximations to the variance of HT estimators are effective, even for moderate sample sizes, provided that the entropy of the underlying sampling design is high enough. Recently Deville and Tillé [2005] and Fuller [2009] consider balanced, or approximately balanced, sampling algorithms which can be useful to build designs with fixed size and given inclusion probabilities. They relate these sampling designs to the rejective sampling, so that the Hájek variance approximation can be used.

1.4 Variance estimation

In order to obtain the consistency of the HT variance estimator given by (1.3), an additional assumption on the fourth-order inclusion probabilities is needed:

Assumption S4. Assume that $\lim_{N\to\infty} \max_{(k,l,k',l')\in D_{4,n}} |\mathbb{E}_p\{(I_{kl} - \pi_{kl})(I_{k'l'} - \pi_{k'l'})\}| = 0$ where $D_{4,n}$ is the set of all distinct 4-tuples from U,

as well as on the fourth-moment of the study variable $\mathcal{Y}$:

Assumption V2. Assume that $\lim\limits_{N\to\infty} \dfrac{1}{N}\sum\limits_{k\in U} y_k^4 < \infty$.

Assumption (S4) was considered for the first time by Breidt and Opsomer [2000] in a non-parametric framework in order to prove the consistency of the variance estimator for the local

polynomial nonparametric model-assisted estimator. Their proof may be adapted easily to the HT variance estimator $\hat{V}(\hat{t}_{yd})$. Assumption (S4) is true for simple random sampling without replacement, stratified sampling and rejective sampling (Boistard et al. [2012]). More generally, it also holds for unequal probability designs with large entropy (see their definitions in formula 1.9, Section 1.3) as shown in the following proposition:

Proposition 1.4. *(Cardot et al. [2014b]) Let p be a sampling design with the same first order inclusion probabilities as the rejective design p_{rej}. If $\lim_{N\to\infty} D(\pi) = \infty$, for $D(\pi) = \sum_{k\in U} \pi_k(1-\pi_k)$ then*

$$\max_{(k,l,k',l')\in D_{4,n}} |\mathbb{E}_p\{(I_{kl}-\pi_{kl})(I_{k'l'}-\pi_{k'l'})\}| \leq \frac{C}{D(\pi)} + \sqrt{\frac{\mathcal{K}(p,p_{rej})}{2}},$$

for some constant C and $\mathcal{K}(p,p_{rej})$ is the Kullback-Leibler divergence with respect to p_{rej} given in (1.10).

Theorem 1.5. *(Consistency of the variance estimator) Make assumptions (S1), (S2), (S4) and (V2). Then:*

$$\lim_{N\to\infty} n\mathbb{E}_p|\hat{V}(\hat{t}_{yd}) - \mathbb{V}_p(\hat{t}_{yd})| = 0.$$

When the aim is to build confidence intervals, the asymptotic distribution of the HT estimator is required. Asymptotic normality is not easy to show in survey sampling framework and it has been shown in case-by-case studies: Erdös and Rényi [1959] and Hájek [1960] checked it for simple random sampling without replacement, Hájek [1964] for the rejective sampling by assuming a Lindeberg-Feller condition. Vísek [1979] has given a simpler proof of the Hájek's result with application to the rejective, Sampford and successive sampling designs. Berger [1998a] extends the result of Hájek [1964] to sample designs which are asymptotically of maximum entropy. In order to check the Lindeberg condition, the variable $\mathcal{Y}$ is usually assumed to have a finite $(2+\delta)$th moment, $\delta > 0$ (Fuller [2009], Thompson [1997]).

Taking into account these considerations, the following assumption will be assumed as soon as the aim is to build confidence intervals:

Assumption S5. Assume that the sampling design and the study variable is such that the HT estimator is asymptotically normal, namely

$$\frac{\sqrt{n}}{N}(\hat{t}_{yd} - t_y) \to_{\mathcal{D}} \mathcal{N}(0,\sigma^2)$$

as $N\to\infty$ and $\lim_{N\to\infty} \frac{n}{N^2}\mathbb{V}_p(\hat{t}_{yd}) = \sigma^2$.

Moreover, if the variance $\mathbb{V}_p(\hat{t}_{yd})$ is estimated by a consistent estimator $\hat{V}(\hat{t}_{yd})$ (result 1.5), we have with Slutsky's theorem:

$$\frac{\hat{t}_{yd} - t_y}{\sqrt{\hat{V}(\hat{t}_{yd})}} \to_{\mathcal{D}} \mathcal{N}(0,1).$$

1.5 Improving the HT estimator by taking into account auxiliary information

Let $\mathcal{X}_1, \ldots, \mathcal{X}_p$ be p auxiliary variables and let $\mathbf{x}_k^T = (x_{k1}, \ldots, x_{kp})$ be the vector of the values of the auxiliary variables for the k-th individual. Usually, we know only the finite population total of $\mathbf{x}_k$, denoted by $t_{\mathbf{x}}$. When the vector $\mathbf{x}_k$ is known for all the population units, we have a *complete information.*
It is of interest to improve the HT estimator $\hat{t}_{yd} = \sum_{k \in s} d_k y_k$ of t_y by using this auxiliary information. With multipurpose surveys, the main goal is to derive a unique weight w_{ks} for each unit k in the sample with w_{ks} containing the auxiliary information but not depending on the study variables. In this way, it is possible to estimate any linear combination of totals and other finite population parameters (Särndal [2007]).

Then, estimators of finite population totals are weighted sums such that :

$$\hat{t}_{yw} = \sum_{k \in s} w_{ks} y_k. \tag{1.11}$$

Mainly, there are two ways to incorporate auxiliary information depending on whether or not a model is fitted to the data. If a model relates $\mathbf{x}_k$ to y_k, then there are at least two ways to construct estimators: model-assisted estimators and model-based estimators. If no model is explicitly stated, we have *the calibration approach* as suggested by Deville and Särndal [1992].

Calibration approach
Deville and Särndal [1992] suggested the calibration method to use effectively the known population totals of $\mathcal{X}_j, j = 1, \ldots, p$ at the estimation stage. It is a very popular method extensively used in practice due to its simple formulation. Särndal [2007] gives a very comprehensive overview of the original method as well as its various applications and extensions.

The calibration estimator of t_y is a weighted estimator $\hat{t}_{yw}^{\text{cal}} = \sum_{k \in s} w_{ks}^{\text{cal}} y_k$ where the calibration weights $\mathbf{w}_s^{\text{cal}} = (w_{ks}^{\text{cal}})_{k \in s}$ minimize a distance measure Υ_s to the sampling weights d_k and subject to some calibration constraints. More exactly,

$$\mathbf{w}_s^{\text{cal}} = \text{argmin}_{\mathbf{w}} \Upsilon_s(\mathbf{w}) \tag{1.12}$$

subject to

$$\sum_{k\in s} w_{ks}^{\text{cal}} \mathbf{x}_k = t_{\mathbf{x}}, \tag{1.13}$$

where $t_{\mathbf{x}} = \sum_{k\in U} \mathbf{x}_k$ is the vector of known totals of $\mathcal{X}_j, j = 1, \ldots, p$ and $\mathbf{w} = (w_k)_{k\in s}$. The calibration constraints (1.13) may be interpreted as a property of consistency with known totals of the weight system. This property is mostly looked for in national statistical agencies.

Several distance functions Υ_s have been suggested in Deville and Särndal [1992] and all resulting estimators are asymptotically equivalent to the one obtained by minimizing the chi-squared distance function

$$\Upsilon_s(\mathbf{w}) = \sum_{k\in s} q_k^{-1} \frac{(w_k - d_k)^2}{d_k}, \tag{1.14}$$

where the q_k's are known positive constants used to control the variability of the observations and are unrelated to d_k. With the chi-square distance Υ_s given in (1.14), the resulting calibration weights are

$$w_{ks}^{\text{cal}} = d_k - q_k d_k \mathbf{x}_k^T \left(\sum_{k\in s} q_k d_k \mathbf{x}_k \mathbf{x}_k^T \right)^{-1} (\hat{t}_{\mathbf{x}d} - t_{\mathbf{x}}), \quad k \in s \tag{1.15}$$

where $\hat{t}_{\mathbf{x}d} = \sum_{k\in s} d_k \mathbf{x}_k$ is the HT estimator of $t_{\mathbf{x}}$. The calibration estimator $\hat{t}_{yw}^{\text{cal}}$ is

$$\hat{t}_{yw}^{\text{cal}} = \hat{t}_{yd} - \left(\hat{t}_{\mathbf{x}d} - t_{\mathbf{x}}\right)^T \hat{\boldsymbol{\beta}}_{\mathbf{x}}(\mathbf{d}, \mathbf{q}^{-1}), \tag{1.16}$$

where $\hat{\boldsymbol{\beta}}_{\mathbf{x}}(\mathbf{d}, \mathbf{q}^{-1}) = (\sum_{k\in s} d_k q_k \mathbf{x}_k \mathbf{x}_k^T)^{-1} \sum_{k\in s} d_k q_k \mathbf{x}_k y_k$[1] with $\mathbf{d} = (d_k)_{k\in s}$ and $\mathbf{q}^{-1} = (q_k^{-1})_{k\in s}$. Using the distance

$$\Upsilon_s(\mathbf{w}) = -\sum_{k\in s} d_k \log\left(\frac{w_k}{d_k}\right),$$

results in the empirical likelihood studied by Chen and Qin [1993] in a finite population context. Note also that the well-known poststratified estimator is also a particular case of calibration estimator.

Under mild regularity assumptions, Deville and Särndal [1992] prove that the calibration estimator $\hat{t}_{yw}^{\text{cal}}$ is asymptotically equivalent to the difference estimator

$$\tilde{t}_{y,\mathbf{x}}^{\text{diff}} = \hat{t}_{yd} - \left(\hat{t}_{\mathbf{x}d} - t_{\mathbf{x}}\right)^T \tilde{\boldsymbol{\beta}}_{\mathbf{x}}(\mathbf{q}^{-1}), \tag{1.17}$$

[1] This general notation for the estimator of the regression coefficient will be useful in Section 3.2

with $\tilde{\boldsymbol{\beta}}_{\mathbf{x}}(\mathbf{q}^{-1}) = (\sum_{k\in U} q_k \mathbf{x}_k \mathbf{x}_k^T)^{-1} \sum_{k\in U} q_k \mathbf{x}_k y_k$, in the sense that

$$\frac{1}{N}(\hat{t}_{yw}^{\text{cal}} - t_y) = \frac{1}{N}(\tilde{t}_{y,\mathbf{x}}^{\text{diff}} - t_y) + o_{\text{p}}(n^{-1/2}).$$

The above result is true regardless of the choice of the distance. In particular, the calibration estimator is consistent and ADU with asymptotic variance equal to

$$A\mathbb{V}_p(\hat{t}_{yw}^{\text{cal}}) = \sum_{k\in U}\sum_{k\in U}(\pi_{kl} - \pi_k\pi_l)d_k d_l(y_k - \mathbf{x}_k^T\tilde{\boldsymbol{\beta}}_{\mathbf{x}}(\mathbf{q}^{-1}))(y_l - \mathbf{x}_l^T\tilde{\boldsymbol{\beta}}_{\mathbf{x}}(\mathbf{q}^{-1})). \tag{1.18}$$

Hence, the calibration estimator will improve the HT estimator, namely $A\mathbb{V}_p(\hat{t}_{yw}^{\text{cal}}) \leq \mathbb{V}_p(\hat{t}_{yd})$, if the residuals $y_k - \mathbf{x}_k^T\tilde{\boldsymbol{\beta}}_{\mathbf{x}}(\mathbf{q}^{-1})$ are small. This happens if the auxiliary variable is strongly correlated with the study variable. Nevertheless, if the study variable is non linearly related to the auxiliary variables, or the study parameter is more complicated (for example, the Gini index), then the Deville and Särndal's suggested method should be modified. Several suggestions exist in the literature. Goga and Ruiz-Gazen [2014c] propose in a recent work a new calibration method that can handle efficiently and in a simple manner the estimation of nonlinear parameters or nonlinear superpopulation models. More details are given in Section 2.2.3.
Another issue with the calibration estimator is when the number p of auxiliary variables is too large. In this situation, several difficulties may arise. Section 3.2 gives a detailed presentation of these difficulties as well as the suggested methods to overcome them.

Model-assisted approach

In the model-assisted approach as described in Särndal et al. [1992], we assume that the finite population of the y_k values are realizations from an infinite superpopulation model ξ relating the auxiliary information $\mathbf{x}_k$ to y_k as follows:

$$\xi: \quad y_k = \mathbf{x}_k^T\boldsymbol{\beta} + \varepsilon_k, \quad k \in U. \tag{1.19}$$

The error terms ε_k are centered, namely $\mathbb{E}_\xi(\varepsilon_k) = 0$ with variance $\mathbb{V}_\xi(\varepsilon_k) = v_k$. Let $\mathbf{V}$ be the $N \times N$ diagonal variance matrix with diagonal elements v_k.
The model-assisted estimator has its origins in the generalized difference estimator suggested by Cassel et al. [1976]

$$\begin{aligned} t_{y,\mathbf{x}}^{\text{diff}} &= \hat{t}_{yd} - (\hat{t}_{\mathbf{x}d} - t_{\mathbf{x}})^T\boldsymbol{\beta}, \\ &= \sum_{k\in s}\frac{y_k - \mathbf{x}_k^T\boldsymbol{\beta}}{\pi_k} + \sum_{k\in U}\mathbf{x}_k^T\boldsymbol{\beta}, \end{aligned} \tag{1.20}$$

where $\boldsymbol{\beta}$ is the true regression coefficient. It is in fact the difference between the HT estimator $\hat{t}_{yd}$ and the bias of $\hat{t}_{yd} - t_y$ under the model ξ. It can be also seen as the prediction of t_y under the model ξ plus a design-bias adjustment. In practice, we never know the true $\boldsymbol{\beta}$, thus we have to build an estimator of it. Generally, this estimator is obtained using a two-step procedure:

we estimate first $\boldsymbol{\beta}$ under the model ξ by:

$$\tilde{\boldsymbol{\beta}}_{\mathbf{x}}(\mathbf{v}) = \left(\sum_{k\in U} v_k^{-1}\mathbf{x}_k\mathbf{x}_k^T\right)^{-1} \sum_{k\in U} v_k^{-1}\mathbf{x}_k y_k, \tag{1.21}$$

and next, we estimate $\tilde{\boldsymbol{\beta}}_{\mathbf{x}}(\mathbf{v})$ by using the sampling design:

$$\hat{\boldsymbol{\beta}}_{\mathbf{x}}(\mathbf{d},\mathbf{v}) = \left(\sum_{k\in s} d_k v_k^{-1}\mathbf{x}_k\mathbf{x}_k^T\right)^{-1} \sum_{k\in s} d_k v_k^{-1}\mathbf{x}_k y_k. \tag{1.22}$$

Plugging $\hat{\boldsymbol{\beta}}_{\mathbf{x}}(\mathbf{d},\mathbf{v})$ in (1.20) yields the generalized regression estimator (GREG) or the model-assisted estimator:

$$\hat{t}_{yw}^{\text{greg}} = \hat{t}_{yd} - (\hat{t}_{\mathbf{x}d} - t_{\mathbf{x}})^T \hat{\boldsymbol{\beta}}_{\mathbf{x}}(\mathbf{d},\mathbf{v}). \tag{1.23}$$

The calibration estimator obtained with the chi-squared distance is equal to the GREG estimator if $q_k = v_k^{-1}$ for all $k \in s$, and for this reason, this choice of q_k will be made in the following.

The GREG estimator is model-unbiased, namely $\mathbb{E}_\xi(\hat{t}_{yw}^{\text{greg}} - t_y) = 0$, and ADU, regardless if the model is true or not. This property may be seen as a robustness property. Fuller [2002] gives a comprehensive review of the GREG estimator and of its asymptotic properties.

The GREG estimator is asymptotically equivalent to the difference estimator from (1.17) with asymptotic variance given by (1.18). From a model point of view, the variance of the GREG estimator is small if the predicted values $\mathbf{x}_k^T\tilde{\boldsymbol{\beta}}_{\mathbf{x}}(\mathbf{v})$ are close enough to the y_k's, namely if the model ξ stated in (1.19) explains well the study variable. If the linear regression model does not fit the data well, there is no improvement over the HT estimator and non-parametric models may be used instead (see section 2.2).

Proposition 1.6. *(Särndal [1980]) Suppose the superpopulation model ξ holds, with variance matrix $\mathbf{V} = diag(v_k)_{k\in s}$ satisfying $v_k = \mathbf{c}^T\mathbf{x}_k$ for some $p-$dimensional vector $\mathbf{c}$. Then, the GREG estimator is the population total of the estimated predictions:*

$$\hat{t}_{yw}^{\text{greg}} = \left(\sum_{k\in U} \mathbf{x}_k^T\right) \hat{\boldsymbol{\beta}}_{\mathbf{x}}(\mathbf{d},\mathbf{v}).$$

Model-based approach

Using a model-based strategy, we look for a linear predictor of t_y being the sum of the sample y_k's values plus a predictor of the sum of non sampled units:

$$\hat{t}_{yw} = \sum_{k\in s} y_k + \sum_{k\in U-s} \mathbf{x}_k^T\boldsymbol{\beta} = \sum_{k\in s} w_{ks} y_k, \tag{1.24}$$

with weights w_{ks} derived by applying the Best Linear Unbiased (BLU) prediction based on the underlying super-population model ξ (Royall [1976]). When the errors ε_k are uncorrelated (as in our model 1.19), then the BLU predictor is the predictor minimizing the variance of $\hat{t}_{yw}$ under the model ξ, namely $\mathbb{E}_\xi(\hat{t}_{yw} - t_y)^2$, among the linear model-unbiased estimators $\hat{t}_{yw}$ of t_y, namely $\mathbb{E}_\xi(\hat{t}_{yw} - t_y) = 0$. Royall [1976] determined the BLU predictor in the case of correlated units. The solution is given by

$$w_{ks}^{\text{blu}} = 1 - v_k^{-1}\mathbf{x}_k^T \left(\sum_{k\in s} v_k^{-1}\mathbf{x}_k\mathbf{x}_k^T\right)^{-1} (\hat{t}_{\mathbf{x}d} - t_{\mathbf{x}}), \quad k \in s \tag{1.25}$$

and the BLU predictor $\hat{t}_{yw}^{\text{blu}}$ is given by

$$\hat{t}_{yw}^{\text{blu}} = \sum_{k\in s}(y_k - \mathbf{x}_k^T\hat{\boldsymbol{\beta}}_{\mathbf{x}}(\mathbf{1}_s, \mathbf{v})) + \sum_{k\in U}\mathbf{x}_k^T\hat{\boldsymbol{\beta}}_{\mathbf{x}}(\mathbf{1}_s, \mathbf{v}), \tag{1.26}$$

where $\hat{\boldsymbol{\beta}}_{\mathbf{x}}(\mathbf{1}_s, \mathbf{v}) = (\sum_{k\in s} v_k^{-1}\mathbf{x}_k\mathbf{x}_k^T)^{-1}\sum_{k\in s} v_k^{-1}\mathbf{x}_k y_k$ and $\mathbf{1}_s$ is the n-dimensional vector of ones. The variance (under the model ξ) of the BLU predictor is

$$\mathbb{V}_\xi(\hat{t}_{yw}^{\text{blu}}) = \sum_{k\in U-s} v_k + \left(\sum_{k\in U-s}\mathbf{x}_k\right)\left(\sum_{k\in s} v_k^{-1}\mathbf{x}_k\mathbf{x}_k^T\right)^{-1}\left(\sum_{k\in U-s}\mathbf{x}_k^T\right).$$

Remark that if the model-variance $\mathbf{V}$ satisfies the property given in Proposition (1.6), then the BLU predictor is also the total of predictions: $\hat{t}_{yw}^{\text{blu}} = \left(\sum_{k\in U}\mathbf{x}_k\right)^T\hat{\boldsymbol{\beta}}_{\mathbf{x}}(\mathbf{1}_s, \mathbf{v})$. Valliant [2009] consider that "models that satisfy the variance condition $v_k = \mathbf{c}^T\mathbf{x}_k$ for some $p-$dimensional vector $\mathbf{c}$, play a key role in robustness and optimality".

The BLU-weights may be obtained also as the minimizers of a distance function to the unit weights while satisfying the same constraints (1.13) (see Valliant et al. [2000], Guggemos and Tillé [2010]) :

$$\mathbf{w}_s^{\text{blu}} = \text{argmin}_{\mathbf{w}}\sum_{k\in s} v_k(w_k - 1)^2 \quad \text{subject to}\sum_{k\in s} w_{ks}^{\text{blu}}\mathbf{x}_k = t_{\mathbf{x}}, \tag{1.27}$$

where v_k is the variance of the error-term from the superpopulation model given by (1.19). The constraints result from the model-unbiasedness condition of the weighted estimator $\hat{t}_{yw}$. So, we can see the BLU weights w_{ks}^{blu} as the closest weights to the unit weights according to the following chi-square distance:

$$\Upsilon_s(\mathbf{w}) = \sum_{k\in s} v_k(w_k - 1)^2, \tag{1.28}$$

and satisfying the calibration constraint.

Unlike the model-assisted approach, estimators built in a model-based approach are model-dependent and may suffer from large bias if the model is misspecified. To protect against model-misspecification, Royall and Herson [1973] suggested to select a sample s such that is

balanced on power of the auxiliary information, namely

$$\frac{1}{n}\sum_{k\in s} x_k^{j-\frac{1}{2}} = \frac{\overline{x}^{(j)}}{\overline{x}^{(1/2)}}, \quad j = 1, \ldots, J \tag{1.29}$$

and $\overline{x}^{(j)} = \sum_{k\in U} x_k^j/N$. Exact methods for obtaining balanced samples are described in Chauvet and Tillé [2006] and Tillé [2006]. The estimator $\hat{t}_{yw}^{\text{blu}}$ is ADU only in particular situations (Särndal and Wright [1984], Brewer et al. [1988]).

As presented above, the calibration estimator $\hat{t}_{yw}^{\text{cal}}$ as well as the BLU predictor $\hat{t}_{yw}^{\text{blu}}$, may be obtained as minimizers of some distance function. In order to give a unified presentation of the penalized estimators in both calibration and model-based frameworks as presented in section 3.2, let $\mathbf{a} = (a_k)_{k\in s}$ and $\mathbf{b} = (b_k)_{k\in s}$ denote two pre-specified sets of positive numbers. Let Υ_s be the chi-squared distance (Chambers [1996])

$$\Upsilon_s((\mathbf{a}, \mathbf{b}); \mathbf{w}) = \sum_{k\in s} b_k \frac{(w_k - a_k)^2}{a_k}. \tag{1.30}$$

The optimum weights $\mathbf{w}_s^{\text{opt}}(\mathbf{a}, \mathbf{b}) = (w_{ks}^{\text{opt}}(\mathbf{a}, \mathbf{b}))_{k\in s}$ are the ones that satisfy

$$\mathbf{w}_s^{\text{opt}}(\mathbf{a}, \mathbf{b}) = \text{argmin}_{\mathbf{w}} \Phi_s((\mathbf{a}, \mathbf{b}); \mathbf{w})$$

subject to

$$\sum_{k\in s} w_{ks}^{\text{opt}}(\mathbf{a}, \mathbf{b})\mathbf{x}_k = t_{\mathbf{x}}.$$

We can remark that for $(\mathbf{a}, \mathbf{b}) = (\mathbf{d}, \mathbf{v})$, we obtain the calibration weights $\mathbf{w}_s^{\text{cal}} = \mathbf{w}_s^{\text{opt}}(\mathbf{d}, \mathbf{v})$ and for $(\mathbf{a}, \mathbf{b}) = (\mathbf{1}_s, \mathbf{v})$, we obtain the BLU weights, $\mathbf{w}_s^{\text{blu}} = \mathbf{w}_s^{\text{opt}}(\mathbf{1}_s, \mathbf{v})$.

Chapter 2

Estimation of nonlinear parameters and of their variance by a functional approach

The estimation of nonlinear parameters in finite populations has become a crucial problem in many recent surveys. For example, in the European Statistics on Income and Living Conditions (EU-SILC) survey, several indicators for studying social inequalities and poverty are considered; these include the Gini index, the at-risk-of-poverty rate, the quintile share ratio and the low-income proportion. Thus, deriving estimators and confidence intervals for such indicators is particularly useful.

Usually, such parameters are estimated using the sampling weights $d_k, k \in s$. Then, the variance of the so obtained estimators can not be computed by using the simple HT variance formula given in (1.2). Variance estimation of nonlinear parameters is a issue that has already been addressed in several papers. There exist mainly two approaches: resampling methods and linearization methods. Resampling methods include jackknife, balanced repeated replication and the bootstrap and they can all be very computationally intensive. Moreover, and unlike linearization methods, resampling methods can only be applied to specific sampling design. For unequal probability sampling designs, they may run into difficulties (Wolter [2007]). We refer to the recent review of resampling methods done by Gershunskaya et al. [2009]. As for the linearization methods, the well-known Taylor method can be used for nonlinear but continuously differentiable functions of totals such as ratios. Books of Wolter [2007] and Särndal et al. [1992] gave comprehensive descriptions of the method with application on the case of the ratio, the regression coefficient and the empirical distribution function. Binder [1983] suggested the estimating equations approach in order to obtain the asymptotic variance of estimators defined as solution of estimating equations. This method has been used by Binder and Kovacevic [1995] and Kovacevic and Binder [1997] for several measures of income inequalities. We mention also Shao [1994] for L-estimators. Recently, Deville [1999a] suggested the functional linearization

by means of the influence function and Demnati and Rao [2004] a modified Taylor method. We mention also the very recent work of Wang and Opsomer [2011] treating non differentiable parameters.

In presence of auxiliary information, it may be of interest to take it into account in order to improve nonlinear parameter estimation. However, this issue has been mainly addressed for estimating totals and means as presented briefly in Chapter 1. In multipurpose surveys, the goal is to use weights not depending on the study variable so that they can be used for estimating different totals or nonlinear parameters. But how should be computed these weights to improve most of parameter estimators? When estimating a nonlinear parameter, the linear models may not be the best choice and nonparametric models are preferred. In such conditions, the variance estimation is a difficult issue. During my PhD, I suggested a class of nonparametric model-assisted estimators based on B-spline regression (Goga [2005]) for the estimation of finite population totals. Lately, and in collaboration with A. Ruiz-Gazen, we extended the method to penalized B-spline regression for finite population totals as well as for nonlinear parameters (Goga and Ruiz-Gazen [2014a], Goga and Ruiz-Gazen [2014b]). Our main contribution consists firstly in developing a new system of survey weights, based on nonparametric regression, and that can be used to estimate any non-linear parameter that is associated with any study variable of the survey. And secondly, we show by using the influence function linearization that under mild assumptions, the nonparametric substitution estimator is asymptotically equivalent to a nonparametric generalized estimator for the total of an artificial variable depending on the study parameter called linearized variable.

When survey data are collected from several samples selected at different moments of time for example, it is of interest to study how parameters change over time. Estimating the change in the Gini index between two periods of time is one particular example. Work concerning temporal change mainly deals with the estimation of totals or means; we mention Särndal et al. [1992], Hidiroglou [2001] and Berger [2004] among others. However, as far as I know, the estimation of nonlinear parameters with multiple-samples has not been addressed. In this multiple-sample setting, several difficulties/questions arise such as: how can be used the information from the different samples and in particular, the one from the overlapping samples, to estimate efficiently the study parameter? What is the optimal overlapping? How is computed and estimated the variance for estimators of nonlinear parameters ? During my PhD, I was concerned with all these issues and I developed a class of composite estimators for the estimation of finite population totals. In collaboration with A. Ruiz-Gazen and J.-C. Deville, we extended this work to the nonlinear case and the two-sample case (Goga et al. [2006], Goga et al. [2009]). With two overlapping samples, there exist three non-overlapping samples which naturally lead us to consider three-variate functionals and their associated partial influence functions. Again, our main contribution was to suggest a composite system of weights which takes into account the disjoint samples. We showed by using the partial influence function linearization that under mild assumptions, the composite substitution estimator is asymptotically equivalent to a composite estimator for the total of the linearized variables. Chauvet et al. [2008] and Chauvet and Goga [2013b] give an extension of the Gross [1980]'s bootstrap for the two-sample framework.

This chapter is structured as follows. In Section 2.1, I will give the main result concerning the functional linearization as obtained in Goga and Ruiz-Gazen [2014a]. However, the presentation exhibited here is slightly different from the one adopted in Goga et al. [2009] and Goga and Ruiz-Gazen [2014a] and gives, in particular, a justification of some results introduced by Deville [1999a]. Section 2.2 extends the result for taking into account auxiliary information as developed in Goga and Ruiz-Gazen [2014a], Goga and Ruiz-Gazen [2014b] and gives a calibration point of view (Goga and Ruiz-Gazen [2014c]) as well as a small application using data extracted from the French Labor Force Survey from 1999-2000. Finally, Section 2.3 concludes and gives several directions for further work.

2.1 Basics of the method

Consider now a parameter Φ which is more complicated than a total or a mean. Broadly speaking, linearization techniques consist in obtaining an expansion of an estimator $\widehat{\Phi}$ of Φ as follows

$$\widehat{\Phi} - \Phi \simeq \sum_{k \in s} d_k u_k - \sum_{k \in U} u_k = \hat{t}_{ud} - t_u, \tag{2.1}$$

where u_k is a kind of artificial variable called *the linearized variable* of Φ by Deville [1999b]. The way it is derived depends on the type of linearization method used which could include Taylor series (Särndal et al. [1992]), estimating equations Binder [1983] or influence function (Deville [1999b]) approaches.

The right term of (2.1) is the difference between the HT estimator and the parameter it estimates, namely the total of the variable u_k over the population U. By consequence, the variance of the right-term is easily obtained and given by

$$\sum_{k \in U} \sum_{k \in U} (\pi_{kl} - \pi_k \pi_l) d_k d_l u_k u_l. \tag{2.2}$$

The main difficulties stand in finding the conditions under which the expansion (2.1) is true, allowing in this way to obtain the asymptotic variance of $\widehat{\Phi}$, and secondly, in computing the linearized variables u_k, for $k \in U$.

Goga and Ruiz-Gazen [2014a] aimed to provide a general method for the estimation of Φ by considering the functional approach introduced in survey sampling theory by Campbell [1980] and developed later by Deville [1999b]. As we will see, this approach is useful for constructing new estimators of Φ as well as for computing their asymptotic variance.

Each unit $k \in U$ is associated with y_k which may be scalar or vector. For simplicity, we consider in this section that $y_k \in \mathbb{R}$. In Chapter 3, the variable $\mathcal{Y}$ lies in a more general space, namely

the space $L^2[0,\mathcal{T}]$ of squared integrable function on $[0,\mathcal{T}]$.
The methodology consists in considering a measure defined on $(\mathbb{R},\mathcal{B}_{\mathbb{R}})$, where $\mathcal{B}_{\mathbb{R}}$ are the Borel sets on $\mathbb{R}$, by

$$M=\sum_{k\in U}\delta_{y_k} \tag{2.3}$$

where δ_{y_k} is the Dirac measure at the point y_k. This means that M assigns unity mass on each point y_k with $k\in U$ and zero mass elsewhere. The total mass of M is equal to N, the population size and its definition takes into account the units $k\in U$ as well as the study variable $\mathcal{Y}$.

Let $\mathcal{M}$ be the linear space generated by the set of measures given by (2.3). Consider functionals $T:\mathcal{M}\to\mathcal{G}$ where $\mathcal{G}$ is some arbitrary space. In this chapter, $\mathcal{G}=\mathbb{R}$ or $\mathcal{G}=\mathbb{R}^p$ and in Chapter 3, $\mathcal{G}=L^2[0,\mathcal{T}]$. Furthermore, we suppose that the study parameter Φ may be written as a functional of M,

$$\Phi \;=\; T(M). \tag{2.4}$$

Definition 2.1. A linear parameter Φ is a parameter which may be written as

$$\Phi=T(M)=\int a(y)dM(y), \tag{2.5}$$

where a is a function of y.

In survey sampling theory, we deal very often with totals which are in fact the simplest example of linear functionals as defined in (2.5). For example, the total of $\mathcal{Y}$ may be written as $t_y=\int ydM(y)$. Any linear combination of totals will be also a linear functional. For this reason, functionals are considered with respect to the measure M as given in (2.3) instead of being with respect to $M/N=\sum_{k\in U}\delta_{y_k}/N$, the empirical distribution, as is the custom in classical statistics. The mean of $\mathcal{Y}$ defined by $\overline{y}=\sum_{k\in U}y_k/N$ is also a linear parameter if the population size N is not estimated. The HT variance estimator $\mathbb{V}_p(\hat{t}_{yd})$ given in (1.2) may be also written in a functional form by considering a slightly different measure M as considered in Langel and Tillé [2013] and Goga and Ruiz-Gazen [2014b].

Definition 2.2. A nonlinear parameter Φ is a parameter $\Phi=T(M)$ which can not be written as in (2.5).

There are many examples of nonlinear parameters, the simplest one is the ratio between two finite population totals $R=\sum_{k\in U}y_k/\sum_{k\in U}z_k$. However, I will consider below two other examples of nonlinear parameters: the Gini index studied in Goga et al. [2009], Goga and Ruiz-Gazen [2014a] and the odds-ratio studied in Goga and Ruiz-Gazen [2014b]. Remark that many examples of nonlinear parameters are considered in Deville [1999a], such as the correlation coefficient and the eigenvalues and eigenvectors of a matrix.

Example 1. The Gini coefficient (Gini [1914]) is one of the most known concentration measure often desired in economical studies. If $\mathcal{Y}$ denotes a quantitative positive variable (for example, the income) and $F(\cdot)$ denotes its distribution function defined on $]-\infty,\infty[$, the Gini coefficient is

$$G = \frac{1}{2}\frac{\int\int |v-u| dF(u)dF(v)}{\int u dF(u)},$$

provided $\int u dF(u) \neq 0$. The Gini coefficient measures the dispersion of a quantitative positive variable within a population. Statistical institutes generally make use of the Gini coefficient to evaluate the income inequalities of a country at different periods of time, or of different countries at the same time. In the last decades, the Gini coefficient has also been considered in various fields such as biology (Graczyk [2007]), environment (Groves-Kirkby et al. [2009]) or astrophysics (Lisker [2008]).

In finite populations, the Gini index (Nygard and Sandström [1985]) is given by (after neglecting the term $1/N$):

$$\mathrm{G} = \frac{\sum_U y_k \left(2F(y_k) - 1\right)}{t_y} = \frac{\int (2F(y) - 1) y dM(y)}{\int y dM(y)} \tag{2.6}$$

where $F(y) = \int \mathbf{1}_{\{\eta \leq y\}} dM(\eta) / \int dM(y) = \sum_U \mathbf{1}_{\{y_k \leq y\}}/N$ is the empirical distribution function.

There is an extensive literature on variance estimation for the Gini estimator with observations obtained from survey data and the very recent paper of Langel and Tillé [2013] gives a comprehensive review and comparison of these works. Sandstrom et al. [1988] listed possible variance estimators for a general sampling design, including a jackknife variance estimator. Linearization variance estimation was studied by Kovacevic and Binder [1997], and Berger [2008] demonstrated the equivalence between linearization and a generalized jackknife technique first suggested by Campbell [1980]. Qin et al. [2010] proposed bootstrap and empirical likelihood based confidence intervals for the Gini coefficient. They studied these methods both theoretically and empirically in the particular case of stratified with replacement simple random sampling.

Example 2. The odds ratio measure is used in health and social surveys where the odds of a certain event is to be compared between two populations. The odds-ratio, which we denote by OR, can be used to quantify the association between the levels of a response variable $\mathcal{Y}$ and a risk variable $\mathcal{Z}$. Let $p_i = P(y_i = 1 | Z = z_i)$ and consider the logistic regression

$$\mathrm{logit}(p_i) = \log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 z_i.$$

Then, the odds-ratio is (Agresti [2002]):

$$\begin{aligned} \mathrm{OR} &= \frac{\mathrm{odds}(Y=1|Z=z_i+1)}{\mathrm{odds}(Y=1|Z=z_i)} \\ &= \exp \beta_1. \end{aligned}$$

The maximum likelihood estimator of $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ is the solution of the following estimating equation:

$$T(M;\boldsymbol{\beta}) = \int \mathbf{t}(\boldsymbol{\beta}) dM = \sum_{k \in U} \mathbf{t}_k(\boldsymbol{\beta}) = 0, \tag{2.7}$$

where $\mathbf{t}_k(\boldsymbol{\beta}) = \mathbf{z}_k(y_k - \mu(\mathbf{z}_k^T\boldsymbol{\beta}))$ with $\mu(\mathbf{z}_k^T\boldsymbol{\beta}) = \exp(\mathbf{z}_k^T\boldsymbol{\beta})(1 + \exp(\mathbf{z}_k^T\boldsymbol{\beta}))^{-1}$ and $\mathbf{z}_k = (1, z_k)^T$. Iterative methods may be used to compute $\boldsymbol{\beta}$. Korn and Graubard [1999] (p. 169-170) advocate the use of weights to estimate the odds ratios. With a categorical variable $\mathcal{Z}$, the odds-ratio has a simpler form and may be derived from a contingency table: $\text{OR} = \dfrac{N_{00}N_{11}}{N_{01}N_{10}}$, where N_{00}, N_{01}, N_{10} and N_{11} are the population counts associated to the contingency table.

More generally, Binder [1983] studied design-based estimators for parameters defined as solution of estimating equation as given in (2.7) and Rao et al. [2002] suggested using poststratification to improve the estimation of such parameters. Remark also, that the Gini index may be obtained as a solution of an estimating equation (Kovacevic and Binder [1997]).

Let consider now the estimation of the parameter Φ by using the sample s. In order to do that, we estimate first the measure M by the HT estimator defined below.

Definition 2.3. The HT estimator of the measure M is the estimator which assigns the sampling weight d_k to each point y_k with $k \in U$ and zero elsewhere, namely

$$\widehat{M}_d \quad = \quad \sum_{k \in s} d_k \delta_{y_k} = \sum_{k \in U} I_k d_k \delta_{y_k}. \tag{2.8}$$

Note that $\widehat{M}_d$ is a random measure of total mass equal to $\hat{N}_d = \sum_{k \in s} d_k$. The estimator of Φ based on the sampling weights d_k is obtained by the plug-in method.

Definition 2.4. The HT substitution estimator for Φ is obtained by plugging $\widehat{M}_d$ into (2.4):

$$\widehat{\Phi}_d = T(\widehat{M}_d).$$

In Section 2.2, an estimator of M which takes into account auxiliary information by nonparametric models is suggested and Goga et al. [2009] suggested an extension to a temporal framework.

The study parameters are considered with respect to measures M and $\widehat{M}_d$, but in order to obtain the asymptotic properties, we will consider their normalized measures $\dfrac{M}{N}$ and $\dfrac{\widehat{M}_d}{N}$. Suppose the following assumptions upon the functional T.

Assumption F1. The functional T is α-homogeneous, in that there is a real number $\alpha \geq 0$, dependent on T such that $T(rM) = r^{\alpha}T(M)$ for any real r.
We assume also that $\lim_{N\to\infty} N^{-\alpha}T(M) < \infty$.

Many study parameters may be written as homogeneous functionals. For example, a total t_y is 1-homogeneous, a ratio R is 0-homogeneous. The same is true for the Gini index while the HT variance $\mathbb{V}_p(\hat{t}_{yd})$ is 2-homogeneous (see Goga and Ruiz-Gazen [2014b]).
Functional expansion of T was first suggested by von Mises [1947] around F a distribution function. Heuristically speaking, for any G belonging to a neighborhood of F, the functional expansion of T in G is a Taylor-type expansion of T :

$$T(G) - T(F) = T(F; G - F) + \text{Rem}_T(G, F), \tag{2.9}$$

where $\text{Rem}_T(G, F)$ is the reminder term associated to T and depending on G, F and $T(F; G-F)$ is the Gâteaux derivative of T at F in the direction of G and defined by

$$T(F; G - F) = \lim_{\varepsilon\to 0} \frac{T(F + \varepsilon(G - F)) - T(F)}{\varepsilon}. \tag{2.10}$$

The weakest type of differentiability we assume is the Gateaux differentiability. Then, this derivative $T(F; \cdot)$ is linear in the second argument. In robust statistics, the interest is about the asymptotic behavior of functionals T with respect to F_N the empirical distribution function. In this case, replacing G with F_N and F with $\mathcal{F}$, the true distribution function, we obtain

$$\begin{aligned} T(F_N) - T(\mathcal{F}) &= T(\mathcal{F}; F_N - \mathcal{F}) + \text{Rem}_T(F_N, \mathcal{F}) \\ &= \frac{1}{N}\sum_{k=1}^{N} T(\mathcal{F}; \delta_{y_k} - \mathcal{F}) + \text{Rem}_T(F_N, \mathcal{F}) \end{aligned} \tag{2.11}$$

where δ_y is the Dirac mass in y. Using relation (2.10), we obtain that

$$T(\mathcal{F}; \delta_{y_k} - \mathcal{F}) = \lim_{\varepsilon\to 0} \frac{T((1 - \varepsilon)\mathcal{F} + \varepsilon\delta_{y_k}) - T(\mathcal{F})}{\varepsilon}$$

which is the definition of the influence function given by Hampel [1974] in robust statistics. This function played an important role in the theory of robust estimation due to the work by Hampel [1974] who remarked that, for large N, the influence function measures the effect on the parameter Φ of contaminating $\mathcal{F}$ with an infinitesimal mass on the observation y. Moreover, the influence function provides the asymptotic variance of the estimator of Φ when the y_k's are independent and identically distributed variables.

In a survey sampling setting, we take $G = \dfrac{\widehat{M}_d}{N}$ and $F = \dfrac{M}{N}$ in (2.10). Hence, we get by using the linearity of the derivative:

$$\begin{aligned} T\left(\frac{M}{N}; \frac{\widehat{M}_d - M}{N}\right) &= T\left(\frac{M}{N}; \frac{1}{N}\sum_{k\in U}(d_k I_k - 1)\delta_{y_k}\right) \\ &= \sum_{k\in U}(d_k I_k - 1) T\left(\frac{M}{N}; \frac{\delta_{y_k}}{N}\right). \end{aligned}$$

By using again the definition of the derivative (2.10) and the α-homogeneity of T (assumption F1), we get

$$\begin{aligned} T\left(\frac{M}{N}; \frac{\delta_{y_k}}{N}\right) &= \lim_{\varepsilon\to 0} \frac{T(\frac{M}{N} + \frac{\varepsilon\delta_{y_k}}{N}) - T(\frac{M}{N})}{\varepsilon} \\ &= N^{-\alpha} \lim_{\varepsilon\to 0} \frac{1}{\varepsilon}\left(T(M + \varepsilon\delta_{y_k}) - T(M)\right). \end{aligned}$$

Hence,

$$T\left(\frac{M}{N}; \frac{\widehat{M}_d}{N} - \frac{M}{N}\right) = N^{-\alpha} \sum_{k\in U}(d_k I_k - 1) \lim_{\varepsilon\to 0} \frac{1}{\varepsilon}\left(T(M + \varepsilon\delta_{y_k}) - T(M)\right). \tag{2.12}$$

The quantity of interest is now $\lim_{\varepsilon\to 0} \dfrac{1}{\varepsilon}\left(T(M + \varepsilon\delta_{y_k}) - T(M)\right)$ which corresponds also to a directional derivative. This derivative is referred to as the influence function by Deville [1999b]:

Definition 2.5. The influence function of T is given by:

$$IT(M, y) = \lim_{\varepsilon\to 0} \frac{T(M + \varepsilon\delta_y) - T(M)}{\varepsilon}$$

where δ_y is the Dirac measure at point y.

Following similar steps as in robust statistics leads us naturally to the slightly different definition of the influence function as suggested by Deville [1999a] in a finite population setting.

Definition 2.6. The linearized variable u_k, for all $k \in U$ of Φ are the values of the influence function IT computed at $y = y_k$, namely

$$u_k = IT(M, y_k), \quad k \in U.$$

Hence, in a finite population setting, relations (2.9) and (2.12) give

$$N^{-\alpha}(T(\hat{M}_d) - T(M)) = N^{-\alpha}(\sum_{k\in s} d_k u_k - \sum_{k\in U} u_k) + \mathrm{Rem}_T\left(\frac{\hat{M}_d}{N}, \frac{M}{N}\right). \tag{2.13}$$

Remark 1. *It is worth mentioning also that (2.12) remains valid as long as the estimator of the measure M may be written as a weighted sum of y_k's population values. The Hansen and Hurvitz estimator of M (Hansen and Hurvitz [1943]), usually used with sampling designs with replacement, may be also written in this form (Chauvet and Goga [2013a]) as well as the nonparametric estimator of M (Goga and Ruiz-Gazen [2014a]) presented in more detail in Section (2.2).*

Remark 2. *For the total $t_y = \int y dM(y)$, the linearized variable u_k is simply y_k and the expansion (2.12) becomes*

$$N^{-1}(\hat{t}_{yd} - t_y) = N^{-1}(\sum_{k\in s} d_k u_k - \sum_{k\in U} u_k).$$

As remarked by Beaumont et al. [2013], the influence function is not related to the sampling design or to the estimator of Φ. By consequence, it cannot be used to measure how robust a sampling design or an estimator is to outlying values. The influence function, and more exactly, the linearized variables $u_k, k \in U$ are quantities which serve as a tool for computing the asymptotic variance of $\hat{\Phi}_d$. The expression of u_k depends on the parameter of interest Φ and they are unknown even for the sampled individuals. It is worth mentioning that the linearized variables play an important role in the estimation of the parameter Φ. To improve the quality of $\hat{\Phi}_d$ at the sampling stage with for example, proportional to size sampling designs or at the estimation stage with GREG-type estimators, the statistician should take into account the expression of u_k. Section (2.2) and Section (3.1.7) present these issues.

Deville [1999b] provides many practical rules for computing u_k for rather complicated parameters Φ. In particular, an important property of α-homogenous functionals is that

$$\sum_{k\in U} IT(M, y_k) = \alpha T(M),$$

so that for functionals 0-homogeneous, we obtain that $\sum_{k\in U} IT(M, y_k) = 0$, condition usually assumed in robust statistics. Moreover, it is straightforward to obtain that for α-homogenous functionals, the influence function is $(\alpha - 1)$-homogeneous, namely

$$IT\,(rM, y) = r^{\alpha-1} IT(M, y).$$

To avoid theoretical difficulties encountered in the computation of u_k, numerical approximations of u_k by jackknife are possible (Davison and Hinkley [1997]).

Let consider again the Gini index from Example 1. The expression of the linearized variable u_k, $k \in U$ for the Gini index (Binder and Kovacevic [1995], Deville [1999a]) is:

$$u_{k,G} = 2F(y_k)\frac{y_k - \overline{y}_{k,<}}{t_y} - y_k\frac{1+G}{t_y} + \frac{1-G}{N} \tag{2.14}$$

where G is the Gini index, F the empirical distribution function, $\overline{y}_{k,<}$ the mean of y_j lower than y_k and t_y the total of the y_k on U.

Concerning the regression coefficient of the logistic regression from Example 2, the linearized variable $\mathbf{u}_{k,\boldsymbol{\beta}}$ is a two-dimensional vector given by (Goga and Ruiz-Gazen [2014b]):

$$\mathbf{u}_{k,\boldsymbol{\beta}} = -\mathbf{J}^{-1}(\boldsymbol{\beta}) \cdot \mathbf{z}_k(y_k - \mu(\mathbf{z}_k^T\boldsymbol{\beta})), \tag{2.15}$$

where $\mathbf{J}(\boldsymbol{\beta})$ is the Jacobian with respect to $\boldsymbol{\beta}$ of the functional T defined in (2.7).

Consider now the remainder term from the expansion given by (2.9). The goal is to obtain that this remainder is of order $o_p(n^{-1/2})$ and suppose for that the additional assumption of Fréchet differentiability of the functional T. In absence of this strong assumption, von Mises [1947] and Serfling [1980] recommend computing higher-order derivatives of Φ and prove that they are going to zero in probability.

Assumption F2. The functional T is Fréchet differentiable at $M \in \mathcal{M}$ if there exists a functional $T(M;\Delta)$ linear in Δ such that

$$\lim_{\tilde{M}\to M} \frac{T(\tilde{M}) - T(M) - T(M;\tilde{M}-M)}{d(\tilde{M},M)} = 0,$$

where d is a distance between $\tilde{M}$ and M.

This definition is due to Huber [1981] (see also Huber and Ronchetti [2009]). Serfling [1980] and van der Vaart [1998]) considered the above definition with a norm instead of the distance d. The advantage of Huber's version of the Fréchet derivative is that it can be applied to a functional without extension to a vector space. Remark also, that if T is Fréchet differentiable, then it is Gateaux differentiable and the two derivatives coincide. This means that the derivative $T(M;\tilde{M}-M)$ from definition F2 is computed according to (2.10).
Then, the remainder from (2.9) depends on some distance function between $\frac{M}{N}$ and an estimator of this measure, in our case the HT estimator $\widehat{M}_d$:

$$N^{-\alpha}(T(\widehat{M}_d) - T(M)) = N^{-\alpha}(\sum_{k\in s} d_k u_k - \sum_{k\in U} u_k) + o\left(d\left(\frac{\widehat{M}_d}{N}, \frac{M}{N}\right)\right). \tag{2.16}$$

Several metrics d such as the Prohorov, the bounded Lipschitz or the total variation distance have been suggested in robust statistic (Huber and Ronchetti [2009], van der Vaart [1998]). As remarked by Serfling [1980], the choice of the distance d should be done with respect to two conflicting goals: this distance should be small enough so that the Fréchet differentiability would be easier to obtain but not too small since $n^{1/2}d$ should be bounded in probability. Goga

and Ruiz-Gazen [2014a] considered the total variation (pseudo) distance defined by:

$$d_{\text{tv}}(M_1, M_2) = \sup_{h \in \mathcal{H}} \left| \int h \, dM_1 - \int h \, dM_2 \right|$$

with $\mathcal{H} = \{h : \mathbb{R} \to \mathbb{R} | \sup_x |h(x)| \leq 1\}$. In general, this distance is too large in the sense that the requirement that $n^{1/2} d = O_p(1)$ is not fulfilled (see Dudley [2002]). Fortunately, used in this survey sampling context, $d_{\text{tv}}\left(\frac{\widehat{M}_d}{N}, \frac{M}{N}\right) = O_p(n^{-1/2})$ (lemma 2.7 from below). In Section 2.2, the result is extended to the nonparametric estimator $\widehat{M}_{np}/N$ (lemma 2.13).

Lemma 2.7. *(Goga and Ruiz-Gazen [2014a]) Assume assumptions (S1) and (S2). Then,*

$$d_{\text{tv}}\left(\frac{\widehat{M}_d}{N}, \frac{M}{N}\right) = O_p(n^{-1/2}).$$

Proof. Let $h \in \mathcal{H}$. Thus, for all $k \in U$, $|h(y_k)| \leq 1$ uniformly in $h \in \mathcal{H}$ and

$$\int h \, d\widehat{M}_d - \int h \, dM = \sum_{k \in s} d_k h(y_k) - \sum_{k \in U} h(y_k) = \sum_{k \in U} (d_k I_k - 1) \, h(y_k).$$

Following the same lines as in Breidt and Opsomer [2000], we have:

$$\begin{aligned} \mathbb{E}_p \left| \int h \, d\widehat{M}_d/N - \int h \, dM/N \right|^2 &= N^{-2} \mathbb{V}_p \left(\sum_{k \in s} d_k h(y_k) \right) \\ &\leq \left(\frac{1 - \tilde{\lambda}}{\tilde{\lambda} N} + \frac{n \max_{k \neq l \in U} |\pi_{kl} - \pi_k \pi_l|}{\tilde{\lambda}^2 n} \right) \frac{1}{N} \sum_{k \in U} h^2(y_k) = O(n^{-1}) \end{aligned}$$

uniformly in h by assumptions (S1) and (S2) and the fact that $h \in \mathcal{H}$. □

Note also that the same proof of lemma 2.7 works to prove that the bounded-Lipschitz d_{BL} distance (Huber [1981], Huber and Ronchetti [2009]) between $\frac{\widehat{M}_d}{N}$ and $\frac{M}{N}$ is also of order $O_p(n^{-1/2})$. The bounded-Lipschitz distance is defined as the total variation distance but the sup is considered over all bounded Lipschitz functions implying that $d_{\text{BL}}(M_1, M_2) \leq d_{\text{tv}}(M_1, M_2)$. This distance is attractive since it metrizes the weak topology in $\mathcal{M}$, whereas the total variation doesn't (Huber [1981], Huber and Ronchetti [2009]). Nevertheless, it is possible for a functional to be Fréchet differentiable with respect to the total variation distance whereas it may not be with respect to the bounded-Lipschitz distance.

Consider the following additional assumptions on the linearized variables u_k, for all $k \in U$.

Assumption V3.

(a) $\overline{\lim}_{N \to \infty} \frac{1}{N} \sum_{k \in U} (N^{1-\alpha} u_k)^2 < \infty$

(*b*) $\sup_{k\in U}|N^{1-\alpha}u_k| < C$ with C a positive constant not depending on N.

The stronger assumption V3 (b) is needed in Section 2.2. If the linearized variable is a vector or a matrix, then these assumptions may be reformulated by considering the Euclidian norm or some matrix norm. Remark that these assumptions are not very restrictive. Cardot et al. [2010a] show that they are satisfied for elements of the functional principal components analysis (see also Section 3.1), Chaouch and Goga [2010] for the geometric quantiles while Goga and Ruiz-Gazen [2014b] show that they are satisfied for the odds-ratio.

Putting together lemma 2.7 and relation (2.16), the theorem providing the functional linearization is obtained.

Theorem 2.8. *(Functional linearization, Goga and Ruiz-Gazen [2014a]). Make assumptions (F1), (F2) and (S1), (S2). Then, the HT substitution estimator* $\widehat{\Phi}_d$ *fulfills*

$$\begin{aligned} N^{-\alpha}\left(\widehat{\Phi}_d - \Phi\right) &= N^{-\alpha}\left(\sum_{k\in s} d_k u_k - \sum_{k\in U} u_k\right) + o_p(n^{-1/2}) \\ &= N^{-\alpha}(\hat{t}_{ud} - t_u) + o_p(n^{-1/2}) \end{aligned} \tag{2.17}$$

Moreover, if the linearized variable $u_k, k \in U$ *satisfy assumption V3 (a), then* $N^{-\alpha}\left(\widehat{\Phi}_d - \Phi\right) = O_p(n^{-1/2})$.

By consequence, $\widehat{\Phi}_d$ is ADU and consistent with respect to the sampling design. Remark also, that given the particular form of the measure M and of its estimator, the first term from the right-left side of (2.17) may be written into an integral form. More exactly, we have

$$\sum_{k\in s} d_k u_k - \sum_{k\in U} u_k = \int IT(M,y)d\widehat{M}_d(y) - \int IT(M,y)dM(y)$$

and for functionals 0-homogeneous, $\int IT(M,y)dM(y) = 0$.

I finish this section by giving a short comment about the relationship of the results presented above with the theory of empirical process which has received lately a large success in classical statistics. As already remarked at the beginning of this section, $\frac{M}{N}$ is the empirical distribution and the empirical distribution function defined by :

$$F_N(t) = \frac{1}{N}\sum_{k\in U} \mathbf{1}_{\{y_k \le t\}}$$

is a function of the empirical distribution. As a function of $t \in \mathbb{R}$, F_N is a process belonging to $\mathcal{D}[0,1]$, the space of cadlag functions: continue *à droite, limite à gauche*. The Donkster's theorem gives the convergence in distribution of the empirical process $\sqrt{N}(F_N - \mathcal{F})$ in $\mathcal{D}[0,1]$.

In a survey sampling setting, this approach is rather new and realized for the moment only for few sampling designs, the ones for which the independence assumption of observations is guaranteed. This is the case for example for the with replacement samplings and the Bernoulli or Poisson sampling designs (Breslow and Wellner [2007], Saegusa and Wellner [2013]). Also, it is worth mentioning that, the results are obtained at the price of numerous assumptions and lengthy proofs.

In classical statistics, nonlinear functionals Φ are usually considered with respect to the empirical distribution function F_N. The functional delta method is a powerful method to linearize such functionals. The remainder term is considered with respect to the sup norm between F_N and the true distribution function $\mathcal{F}$ which is going to zero thanks to the Glivenko-Cantelli's theorem. Fernholz [1983] and van der Vaart [1998] considered the case of Hadamard differentiable functionals Φ. Nevertheless, in a survey sampling setting, this approach has been applied only in the particular case of simple random sampling without replacement by Motoyama and Takahashi [2008] and many additional assumptions are supposed in order to check the tightness condition of the processus F_N (see 3.3 for the definition of tightness) needed to prove the consistency of the design-based estimator of the empirical distribution function.

Asymptotic normality

In robust statistics, the first-order term from the expansion (2.11) is the mean of the Hampel's influence functions computed at y_k which are independent and identically distributed (since y_k are identically distributed variables). By the central limit theorem, the asymptotic normality of $T(F_N)$ is obtained as soon as the remainder term $\text{Rem}_T(F_N, \mathcal{F})$ is going to zero in probability. Unfortunately, in a survey sampling framework, we cannot use the same arguments as in robust statistics in order to obtain the asymptotic normality of $T(\widehat{M}/N)$. But we can use the fact that the first-order term is the error between the HT estimator $\sum_{k\in s} d_k u_k$ and the total $\sum_{k\in U} u_k$. Then, if the linearized variable $N^{1-\alpha}u_k$ satisfies regularity assumptions from Section 1.4, then $N^{-\alpha}\left(\widehat{\Phi}_d - \Phi\right)$ is also asymptotically normal distributed with asymptotic variance equal to the HT variance of $\sum_{k\in s} d_k u_k$, namely

$$A\mathbb{V}_p(\widehat{\Phi}_d) = \sum_{k\in U}\sum_{k\in U}(\pi_{kl} - \pi_k\pi_l)d_k d_l u_k u_l. \tag{2.18}$$

Work is actually in progress in order to check the validity of assumptions from Section 1.4 for u_k computed for different parameters of interest.

Variance estimation

From 2.18, we can see that the asymptotic variance of $\widehat{\Phi}_d$ can not be computed since the double sums are on the population and the linearized variables u_k are unknown even for the sampled individuals. The quantities u_k are estimated by $\hat{u}_k$ and they are plugged-in the HT variance estimator from (1.3) leading to the following variance estimator of $\widehat{\Phi}_d$:

$$\widehat{V}(\widehat{\Phi}_d) = \sum_{k\in s}\sum_{k\in s}\frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}} d_k d_l \hat{u}_k \hat{u}_l. \tag{2.19}$$

Remark 3. *As remarked in Goga et al. [2009] and Langel and Tillé [2013], the estimator of u_k are in fact the influence function of T computed with respect to the measure estimator $\widehat{M}_d$, $u_k = IT(\widehat{M}_d; y_k)$ and coincides with the estimator obtained by using the approach of Demnati and Rao [2004]. In a recent work, Escobar and Berger [2013] suggest a new jackknife replicate variance estimator based on numerical approximation of $\hat{u}_k$.*

The variance estimator $\widehat{V}(\widehat{\Phi}_d)$ is no-longer unbiased for $A\mathbb{V}_p(\widehat{\Phi}_d)$, but it can be proven to be consistent under assumptions on the sampling design and the linearized variables. Let the following decomposition:

$$\widehat{V}(\widehat{\Phi}_d) - A\mathbb{V}_p(\widehat{\Phi}_d) = \left(\widehat{V}(\widehat{\Phi}_d) - \widehat{A\mathbb{V}_p}(\widehat{\Phi}_d)\right) + \left(\widehat{A\mathbb{V}_p}(\widehat{\Phi}_d) - A\mathbb{V}_p(\widehat{\Phi}_d)\right),$$

where $\widehat{A\mathbb{V}_p}(\widehat{\Phi}_d)$ is the HT variance estimator with the true u_k, $k \in U$:

$$\widehat{A\mathbb{V}_p}(\widehat{\Phi}_d) = \sum_{k\in s}\sum_{k\in s}\frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}} d_k d_l u_k u_l.$$

Using theorem 1.5, the estimator $\widehat{A\mathbb{V}_p}(\widehat{\Phi}_d)$ is consistent for $A\mathbb{V}_p(\widehat{\Phi}_d)$ if the linearized variable $N^{(1-\alpha)}u_k$ satisfies assumption (V2) and the sampling design, assumptions (S1), (S2) and (S4). Taking into account these considerations, this result is presented most of the time as an assumption. I do adopt the same attitude in the next of this work. As for the first difference, we need additional assumptions on the estimators $\hat{u}_k$ of u_k.

Theorem 2.9. *(Variance estimation) Make assumptions (S1), (S2) and suppose that the linearized variable u_k satisfies assumption V3(a) and $nN^{-2\alpha}\sum_{k\in U}(\hat{u}_k - u_k)^2 = o_p(1)$. If the HT variance estimator $\widehat{A\mathbb{V}_p}(\widehat{\Phi}_d)$ is consistent for $A\mathbb{V}_p(\widehat{\Phi}_d)$, then*

$$\frac{n}{N^{2\alpha}}(\widehat{V}(\widehat{\Phi}_d) - A\mathbb{V}_p(\widehat{\Phi}_d)) = o_p(1).$$

Again, these assumptions on $\hat{u}_k$ are satisfied for elements of the functional principal components analysis (Cardot et al. [2010a]), for the geometric quantiles (Chaouch and Goga [2010]) and for the odds-ratio (Goga and Ruiz-Gazen [2014b]).

2.2 Taking into account the auxiliary information through nonparametrics

As stated in Section 1.5, auxiliary information is useful for improving on the estimation of a total in terms of efficiency and many recent works have been dedicated to this item by considering parametric (Särndal et al. [1992]) or nonparametric models: Breidt and Opsomer [2000] proposed local linear estimators and Breidt et al. [2005] and Goga [2005] used nonparametric spline regression. Improving the estimation of nonlinear parameters by taking into account auxiliary information was less addressed. Deville [1999a] suggested estimating Φ by considering calibration weights from (1.15) and Berger and Skinner [2003] applied this method for the estimation of the low-income proportion. Särndal et al. [1992] treat briefly the estimation of the ratio by using a linear model. A special attention was addressed to the estimation of the empirical distribution function and quantiles (Dorfman [2009]).

Roughly speaking, when examining (2.1) :

$$\widehat{\Phi} - \Phi \simeq \sum_{k \in s} d_k u_k - \sum_{k \in U} u_k,$$

we can see that, if we estimate in an efficient way $t_u = \sum_{k \in U} u_k$, namely the variance given in (2.2) is small, we will achieve a small approximate variance and good precision for $\widehat{\Phi}$.

When estimating a total, note that the asymptotic variance of the GREG estimator depends on the residuals of the study variable on the auxiliary variable (see 1.18). Because linearized variables may have complicated mathematical expressions, fitting a linear model onto a linearized variable may not be the most appropriate choice. This may occur even if the study and the auxiliary variables have a clear linear relationship, as illustrated in the example given in Goga and Ruiz-Gazen [2014a]. In that example, a dataset of size 1000, extracted from the French Labor Force Survey was considered and y_k (the wages of person k in 2000) was the study variable and x_k (the wages of person k in 1999) the auxiliary variable. We considered the problem of estimating the Gini index. The linearized variable $u_{k,G}$, $k \in U$ for the Gini index is given in (2.14):

$$u_k = 2F(y_k)\frac{y_k - \overline{y}_{k,<}}{t_y} - y_k\frac{1+G}{t_y} + \frac{1-G}{N}$$

where G is the Gini index, F the empirical distribution function, $\overline{y}_{k,<}$ the mean of y_j lower than y_k and t_y the total of the y_k on U. It is a complex function of the study variable y_k, $k \in U$. In the left (resp. right) graphic of Figure 2.1, the study variable y_k is plotted (resp. the linearized variable u_k) on the y-axis and the auxiliary variable x_k is plotted on the x-axis. The relationship between the study variable and the auxiliary variable is almost linear; however the relationship between the linearized variable of the Gini index and the auxiliary information is no longer linear. The consequence of this is that we cannot increase the efficiency of estimating

a Gini index if we take the auxiliary information into account through a linear GREG estimator. Therefore, nonparametric models should be preferred to estimate nonlinear parameters Φ.

The class of estimators proposed in Goga and Ruiz-Gazen [2014a] is based on a nonparametric model-assisted approach. A calibration approach may be used also and it is presented in section (2.2.3). Interestingly, the estimators can be written as a weighted sum of the sampled observations, allowing a unique weight variable that can be used to estimate any nonlinear parameter associated with any study variable of the survey. Having a unique system of weights is very important in multipurpose surveys such as the EU-SILC survey.

In the next, nonparametric weights $\mathbf{w}_s^{\text{np}} = (w_{ks}^{\text{np}})_{k\in s}$ not depending on the study variable will be constructed in order to estimate efficiently the total t_u.

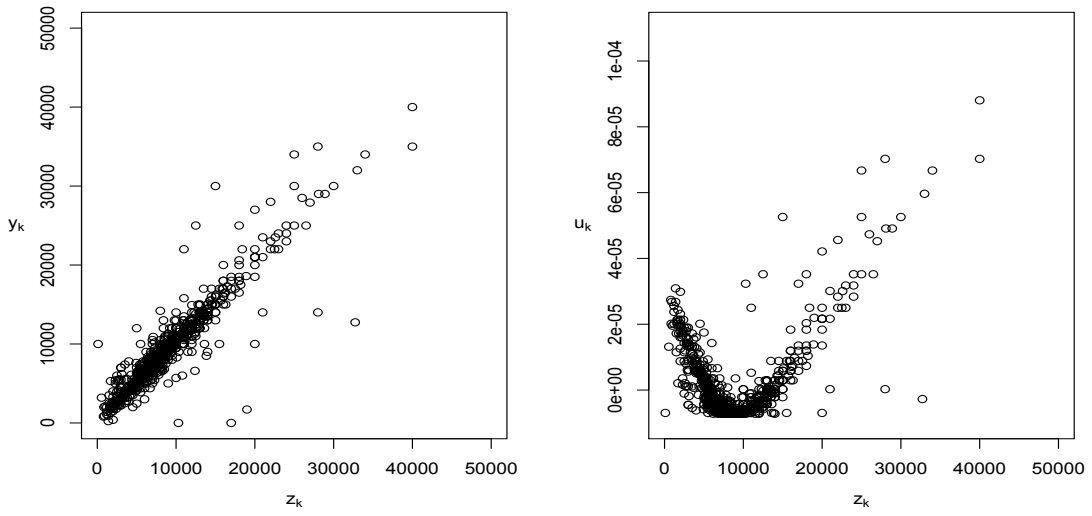


FIGURE 2.1: Left plot: y_k: the wages of person k in 2000 against z_k: the wages of person k in 1999. Right plot: u_k: linearized variable of the Gini index for the wages in 2000 for person k against z_k: the wages of person k in 1999.

2.2.1 Penalized B-spline estimators for nonlinear parameters

We suppose that the auxiliary information is given by the univariate variable $\mathcal{X}$ of values x_k, known for all the population units $k \in U$. We have complete auxiliary information. We suppose without loss of generality that all x_k have been normalized and lie in $[0, 1]$.

Let the superpopulation model relating the auxiliary information x_k to the linearized variable u_k given by

$$\xi' : \quad u_k = g(x_k) + \eta_k. \tag{2.20}$$

Remark that it is not really a model since we do not observe the linearized variables u_k. It can be viewed as a tool used to construct weights. If the linearized variable u_k is a linear combination of study variables (as for the ratio R) and if linear models fit these variables well, then a linear model will also fit u_k well. In this case, we can consider the linear regression function $g(x_k) = \mathbf{x}_k'\boldsymbol{\beta}$ which yields the generalized regression estimator (GREG) of t_u extensively studied by Särndal et al. [1992]. The GREG estimator is efficient if the model fits the data well, but if the model is misspecified, the GREG estimator exhibits no improvement over the HT estimator (from the variance point of view) and may even lead to a loss of efficiency. This is the case of the Gini index for instance and one way of guarding against model failure is to use nonparametric regression which does not require a predefined parametric mathematical expression for g.

Concerned by the estimation of finite population totals, Goga [2005] used B-spline functions to approximate the unknown regression function and this work was extended later by Goga and Ruiz-Gazen [2014a] to penalized B-spline regression. Goga and Ruiz-Gazen [2014a] proved the asymptotic results of estimators for finite population totals as well as for nonlinear parameters. In the following, I will present only results concerning the estimation of nonlinear parameters.

Ruppert et al. [2003] state that *"a spline is a thin strip of flexible timber. A mathematical spline is so named because of the analogy of a flexible function able to adapt to the data"*. The spline functions are used in statistics because of their flexibility to model nonlinear trends in the data that are difficult to be handled parametrically. Using higher-degree polynomials to explain nonlinear trends in the data has a number of disadvantages such as the high oscillatory behavior of the approximating polynomial (Agarwal and Studden [1980]). Moreover, it may be a difficult task to find the degree of the polynomial to be used and sometimes, data cannot be approximated by a single polynomial. Instead, it is more advisable to use spline functions because, besides their numerical stability and local behavior that is less dependent on their behavior elsewhere, they present ease of implementation and interpretability.

For a fixed $m > 1$, the set $S_{K,m}$ of spline functions of order m, with K equidistant interiors knots $0 = \xi_0 < \xi_1 < \ldots < \xi_K < \xi_{K+1} = 1$ is the set of piecewise polynomials of degree $m-1$ that are smoothly connected at the knots (Zhou et al. [1998]):

$$S_{K,m} = \{t \in C^{m-2}[0,1] : t(z) \quad \text{is a polynomial of degree } (m-1) \text{ on each interval} \quad [\xi_i, \xi_{i+1}]\}.$$

For $m = 1$, $S_{K,m}$ is the set of step functions with jumps at knots. For each fixed set of knots, $S_{K,m}$ is a linear space of functions of dimension $q = K + m$. A basis for this linear space is provided by the B-spline functions (Schumaker [1981], Dierckx [1993]) $B_1, \ldots, B_q$ defined by

$$B_j(x) = (\xi_j - \xi_{j-m}) \sum_{l=0}^{m} \frac{(\xi_{j-l} - x)_+^{m-1}}{\Pi_{r=0, r\neq l}^{m}(\xi_{j-l} - \xi_{j-r})}$$

where $(\xi_{j-l} - x)_+^{m-1} = (\xi_{j-l} - x)^{m-1}$ if $\xi_{j-l} \geq x$ and zero, otherwise. There is no general rule giving the exact number of knots but it should be large enough to have enough points between knots. Ruppert et al. [2003] recommend that no more than 30-40 knots should be used and

they give a simple rule for choosing K. As for the degree m, Ruppert et al. [2003] recommend $m = 3$ or $m = 4$. Figure 2.2 exhibits the six B-spline basis functions for $K = 3$ interior knots and $m = 3$.

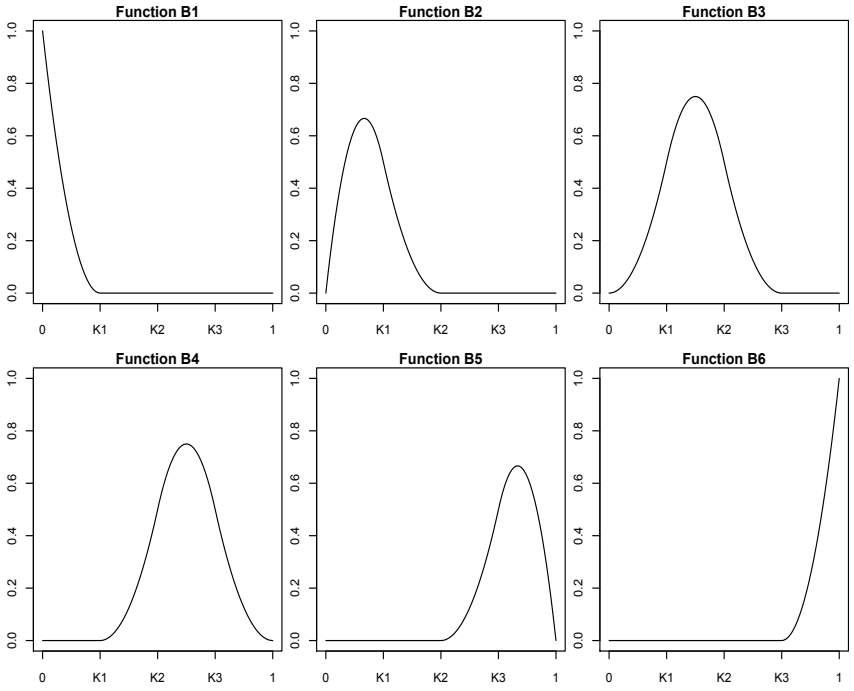


FIGURE 2.2: B-spline basis functions for $K = 3$ interior knots and $m = 3$

For all $j = 1, \ldots, q$, each function B_j has the knots $\xi_{j-m}, \ldots, \xi_j$ with $\xi_r = \xi_{\min(\max(r,0),K+1)}$ for $r = j - m, \ldots, j$ (Zhou et al. [1998]) which means that its support consists of a small, fixed, finite number of intervals between knots. Moreover, B-splines are positive functions with a total sum equal to unity:

$$\sum_{j=1}^{q} B_j(x) = 1 \ , \qquad x \in [0, 1]. \tag{2.21}$$

For the same order m and the same knot location, one can use the truncated power basis (Ruppert and Carroll [2000]) given by $1, x, x^2, \ldots, x^{m-1}, (x - \xi_1)_+^{m-1}, \ldots, (x - \xi_K)_+^{m-1}$. The B-spline and the truncated power bases are equivalent in the sense that they span the same set of spline functions $S_{K,m}$ (Dierckx [1993]). Nevertheless, as indicated by Ruppert et al. [2003], "*the truncated power bases have the practical disadvantage that they are far from orthogonal*", which leads to numerical instability especially if a large number of knots are used.

One way to overcome the issue of knot number, is to consider many knots and to constrain their influence by introducing a penalty. To estimate the unknown regression function g at the population level, Goga and Ruiz-Gazen [2014a] use spline approximation and a penalized

least squares criterion. We define the spline basis vector of dimension $q \times 1$ as $\mathbf{b}^T(x_k) = (B_1(x_k), \ldots, B_q(x_k))$, $k \in U$. The penalized spline estimator of $g(x_k)$ is given by

$$\tilde{g}_{u,k}(\lambda) = \mathbf{b}^T(x_k)\tilde{\boldsymbol{\theta}}_u(\lambda) \tag{2.22}$$

with $\tilde{\boldsymbol{\theta}}_u(\lambda)$ as the least squares minimizer of

$$\sum_{k=1}^{N}(u_k - \mathbf{b}^T(x_k)\boldsymbol{\theta})^2 + \lambda \int_0^1 [(\mathbf{b}^T(t)\boldsymbol{\theta})^{(\ell)}]^2 dt, \tag{2.23}$$

where $^{(\ell)}$ represents the ℓ-th derivate with $\ell \leq m-1$. The solution of (2.23) is a ridge-type estimator,

$$\tilde{\boldsymbol{\theta}}_u(\lambda) \quad = \quad \left(\sum_{k\in U} \mathbf{b}(x_k)\mathbf{b}^T(x_k) + \lambda \mathbf{D}_\ell\right)^{-1} \sum_{k\in U} \mathbf{b}(x_k)u_k \tag{2.24}$$

where $\mathbf{D}_\ell$ is the squared L^2 norm applied to the ℓ-th derivative of $\mathbf{b}^T\boldsymbol{\theta}$. Because the derivative of a B-spline function of order m may be written as a linear combination of B-spline functions of order $m-1$, for equidistant knots we obtain that $\mathbf{D}_\ell = K^{2\ell}\nabla_\ell'\mathbf{R}\nabla_\ell$ where the matrix $\mathbf{R}$ has elements $R_{ij} = \int_0^1 B_i^{(m-\ell)}(t)B_j^{(m-\ell)}(t)dt$ with $B_i^{(m-\ell)}$ as the B-spline function of order $m-\ell$ and ∇_ℓ as the matrix corresponding to the ℓ-th order difference operator (Claeskens et al. [2009]).

The amount of smoothing is controlled by $\lambda > 0$. The case $\lambda = 0$ results in an unpenalized B-spline estimator the asymptotic properties of which have been extensively studied in the literature (Agarwal and Studden [1980], Burman [1991], and Zhou et al. [1998], among others). The case $\lambda \to \infty$ is equivalent to fitting a $(\ell-1)$-th degree polynomial. The theoretical properties of penalized splines with $\lambda > 0$, have been studied only recently by Cardot [2002a], Cardot [2002b], Hall and Opsomer [2005], Kauermann et al. [2009] and Claeskens et al. [2009].

The pseudo [1] design-based estimators of $\tilde{g}_{u,k}$ are

$$\hat{g}_{u,k}(\lambda) = \mathbf{b}^T(x_k)\hat{\boldsymbol{\theta}}_u(\lambda) \tag{2.25}$$

where

$$\hat{\boldsymbol{\theta}}_u(\lambda) \quad = \quad \left(\sum_{k\in s} d_k\mathbf{b}(x_k)\mathbf{b}^T(x_k) + \lambda \mathbf{D}_\ell\right)^{-1} \sum_{k\in s} d_k\mathbf{b}(x_k)u_k \tag{2.26}$$

[1] This appellation is used since $\hat{g}_{u,k}(\lambda)$ can not be computed because u_k is unknown.

is the pseudo design-based estimator of $\tilde{\boldsymbol{\theta}}_u(\lambda)$. Finally, the penalized B-spline model-assisted estimator of $t_u = \sum_{k\in U} u_k$ is as follows:

$$\begin{aligned}\hat{t}_{uw}^{\mathrm{np}}(\lambda) &= \sum_{k\in s} d_k(u_k - \hat{g}_{u,k}(\lambda)) + \sum_{k\in U} \hat{g}_{u,k}(\lambda) \\ &= \sum_{k\in s} d_k u_k - \left(\sum_{k\in s} d_k \mathbf{b}(x_k) - \sum_{k\in U} \mathbf{b}(x_k)\right)^T \hat{\boldsymbol{\theta}}_u(\lambda). \end{aligned} \tag{2.27}$$

This indicates that $\hat{t}_{uw}^{\mathrm{np}}(\lambda)$ may be written as a GREG estimator that uses the vectors $\mathbf{b}(x_k)$ as regressors of dimension $q\times 1$ with q going to infinity and a ridge-type regression coefficient $\hat{\boldsymbol{\theta}}_u(\lambda)$. The following proposition states that the HT estimator of the residuals $u_k - \hat{g}_{u,k}(\lambda)$ is zero. This property is a consequence of the fact that the B-spline functions satisfy $\sum_{j=1}^{q} B_j(x) = 1$.

Proposition 2.10. *1. The nonparametric estimator $\hat{t}_{uw}^{\mathrm{np}}(\lambda)$ is the total of predictions $\hat{g}_{u,k}(\lambda)$:*

$$\hat{t}_{uw}^{\mathrm{np}}(\lambda) = \sum_{k\in U} \hat{g}_{u,k}(\lambda) = \sum_{k\in s} w_{ks}^{\mathrm{np}}(\lambda) u_k,$$

with weights given by

$$w_{ks}^{\mathrm{np}}(\lambda) = d_k \mathbf{b}^T(x_k) \left(\sum_{k\in s} d_k \mathbf{b}(x_k)\mathbf{b}^T(x_k) + \lambda \mathbf{D}_\ell\right)^{-1} \sum_{k\in U} \mathbf{b}(x_k), \quad k \in s. \tag{2.28}$$

2. The weights $w_{ks}^{\mathrm{np}}(\lambda)$ satisfy the calibration constraints:

$$\sum_{k\in s} w_{ks}^{\mathrm{np}}(\lambda) x_k^j = \sum_{k\in U} x_k^j, \quad j = 0, \ldots, \ell - 1. \tag{2.29}$$

Note the similarity with the GREG weights obtained in the case of a linear model when the variance of errors is linearly related to the auxiliary information (see proposition 1.6). The above calibration constraints mean that polynomials of degree lower than ℓ are left unchanged by the penalization. In particular, $\sum_{k\in s} w_{ks}^{\mathrm{np}}(\lambda) = N$ and $\sum_{k\in s} w_{ks}^{\mathrm{np}}(\lambda) x_k = \sum_{k\in U} x_k$.

Regression splines: For $\lambda = 0$, we obtained the unpenalized B-spline estimator studied by Goga [2005]. Based on assumptions regarding the sampling design and the variable $\mathcal{Y}$, (assumptions (S1), (S2) and (V3b)) and assumptions regarding the distribution of $\mathcal{X}$ and the knots number (assumptions (B1), (B2) from below), Goga [2005] proved that the B-spline estimator for the total t_y is ADU and consistent regardless the smoothness of the regression function of the underlying nonparametric model. We note that for $m = 1$, the estimator $\hat{t}_{uw}^{\mathrm{np}}(0)$ becomes the well-known poststratified estimator.

Penalized splines using truncated polynomial basis functions : Breidt et al. [2005] used the truncated polynomial basis functions and a penalized least square criterion for estimating finite populations totals. The estimator obtained by Breidt et al. [2005] is equivalent to the estimator $\hat{t}_{uw}^{\mathrm{np}}(\lambda)$ after a suitable changing of the penalization matrix. In fact, Claeskens et al.

[2009] exhibits the relationship between the penalization matrix obtained with a truncated polynomial basis functions and the matrix $\mathbf{D}_\ell$ from our case.

Let us resume the above construction: we improve the estimation of t_u by considering the nonparametric estimator $\hat{t}_{uw}^{\text{np}}(\lambda)$ with the weights $w_{ks}^{\text{np}}(\lambda)$ given by (2.28). The same weights may be used to estimate the nonlinear parameter $\Phi = T(M)$ by using the functional approach described in section 2.1.

Definition 2.11. The nonparametric estimator of M is given by

$$\widehat{M}^{\text{np}}(\lambda) = \sum_{k \in s} w_{ks}^{\text{np}}(\lambda)\delta_{y_k}$$

Definition 2.12. The nonparametric substitution estimator for Φ is obtained by plugging $\widehat{M}^{\text{np}}$ into (2.4):

$$\widehat{\Phi}^{\text{np}}(\lambda) = T(\widehat{M}^{\text{np}}(\lambda)).$$

Let illustrate the computation of $\widehat{\Phi}^{\text{np}}$ using again the Gini index (Example 1) and parameters defined by implicit equations, such as the odds-ratio (Example 2). The nonparametric estimator for G given in (2.6) is obtained by simply replacing M with $\widehat{M}_{np}$. Hence,

$$\widehat{\text{G}}^{\text{np}} = \frac{\sum_s w_{ks}^{\text{np}}(2\hat{F}^{\text{np}}(y_k) - 1)y_k}{\sum_s w_{ks}^{\text{np}} y_k}, \tag{2.30}$$

where $\hat{F}^{\text{np}}(y) = \dfrac{\int \mathbf{1}_{\{\eta \leq y\}} d\hat{M}^{\text{np}}(\eta)}{\int d\hat{M}^{\text{np}}(y)} = \dfrac{\sum_{k \in s} w_{ks}^{\text{np}} \mathbf{1}_{\{y_k \leq y\}}}{\sum_{k \in s} w_{ks}^{\text{np}}}$.

Let Φ be defined as the unique solution of an implicit estimating equation $\sum_{k \in U} t_k(\Phi) = 0$ that may be written in a functional form as $\int t(\Phi) dM = 0$. We replace M with $\widehat{M}^{\text{np}}$ and the nonparametric sample-based estimator of Φ is the unique solution of the sample-based estimating equation $\int t(\Phi) d\widehat{M}^{\text{np}} = \sum_{k \in s} w_{ks}^{\text{np}} t_k(\widehat{\Phi}^{\text{np}}) = 0$. The regression coefficient given in (2.7) and used for the estimation of the odds-ratio may be estimated by using this method.

2.2.2 Asymptotic properties

The nonparametric estimator $\widehat{\Phi}^{\text{np}}$ is doubly nonlinear, with nonlinearity due to the parameter Φ and nonlinearity due to the nonparametric estimation. Our main goal is to approximate $\widehat{\Phi}^{\text{np}}$ by using a linear estimator (HT type) which will allow us to compute the asymptotic variance of $\widehat{\Phi}^{\text{np}}$. Roughly speaking, this approximation will be accomplished in two steps: first, we will linearize Φ as:

$$\widehat{\Phi}^{\text{np}} - \Phi \simeq \hat{t}_{uw}^{\text{np}}(\lambda) - t_u$$

and next, we will linearize the nonparametric estimator $\hat{t}_{uw}^{\text{np}}(\lambda)$ obtained in step one by a generalized difference-type estimator :

$$\hat{t}_{uw}^{\text{np}}(\lambda) - t_u \simeq \tilde{t}_{u,x}^{\text{diff}}(\lambda) - t_u,$$

where $\tilde{t}_{u,x}^{\text{diff}}(\lambda)$ will be specified later (formula 2.31 from below).

The following assumptions on B-spline functions are needed in order to obtain the main result of this section.

Assumptions on B-splines

Assumption B1. Assume that there exists a distribution function $Q(x)$ with strictly positive density on $[0,1]$ such that $\sup_{x\in[0,1]} |Q_N(x) - Q(x)| = o(K^{-1})$, with $Q_N(x)$ the empirical distribution of $(x_i)_{i=1}^N$.

Assumption B2. Assume that the number of interior knots satisfies:

(a) $K = o(N)$;

(b) $K = O(n^a)$ with $0 < a < 1/3$.

Assumption B3. The penalty parameter λ is such that $K_\ell = (K+m-\ell)(\lambda\tilde{c})^{1/(2\ell)}N^{-1/(2\ell)} < 1$ where $\tilde{c} = c(1+o(1))$ with c a constant that depends only on ℓ and the design density.

These assumptions are classical in nonparametric regression (Agarwal and Studden [1980], Burman [1991], Zhou et al. [1998]); (B1) means that asymptotically, there is no sub-interval in $[0,1]$ without points x_k and (B2) ensures that the dimension of the B-spline basis goes to infinity but not too fast when the population and the sample sizes go to infinity. Assumption (B3) concerns the penalty λ as used by Claeskens et al. [2009] which guarantees the invertibility of $\sum_{k\in U}\mathbf{b}(x_k)\mathbf{b}^T(x_k) + \lambda\mathbf{D}_\ell$ and of the matrix $\sum_{k\in s} d_k\mathbf{b}(x_k)\mathbf{b}^T(x_k) + \lambda\mathbf{D}_\ell$, whatever the sample s is (Goga and Ruiz-Gazen [2014a]).

The first linearization step is a first-order expansion of $\widehat{\Phi}^{\text{np}}$ with the remainder going to zero and it is the extension of theorem 2.8 to the nonparametric case. In order to obtain it, we need to evaluate again the distance d_{tv} between the nonparametric estimator $\frac{\widehat{M}^{\text{np}}(\lambda)}{N}$ and the true $\frac{M}{N}$. Let $h \in \mathcal{H}$ and denote $h_k = h(y_k)$. We have that for all $k \in U$, $|h_k| \leq 1$ uniformly in $h \in \mathcal{H}$. Then

$$\begin{aligned}\int h d\widehat{M}^{\text{np}}(\lambda) - \int h dM &= \sum_{k\in s} w_{ks}^{\text{np}}(\lambda)h_k - \sum_{k\in U} h_k \\ &= \sum_{k\in s} d_k(h_k - \hat{g}_{h,k}(\lambda)) + \sum_{k\in U}\hat{g}_{h,k}(\lambda) - \sum_{k\in U} h_k,\end{aligned}$$

where $\hat{g}_{h,k}(\lambda)$ is obtained from (2.25) for u_k replaced by h_k. Following the same steps as in the proof of lemma 2.7, we need to evaluate

$$\begin{aligned} & \mathbb{E}_p \left| \int h \, d\left(\frac{\widehat{M}^{\text{np}}(\lambda)}{N}\right) - \int h \, d\left(\frac{M}{N}\right) \right| \\ \leq \; & \mathbb{E}_p \left| \frac{1}{N} \sum_{k \in U} (d_k I_k - 1)\,(h_k - \tilde{g}_{h,k}(\lambda)) \right| + \mathbb{E}_p \left| \frac{1}{N} \sum_{k \in U} (d_k I_k - 1)\,(\tilde{g}_{h,k}(\lambda) - \hat{g}_{h,k}(\lambda)) \right|, \end{aligned}$$

where $\tilde{g}_{h,k}(\lambda) = \mathbf{b}^T(x_k)\tilde{\boldsymbol{\theta}}_h(\lambda)$ and $\tilde{\boldsymbol{\theta}}_h(\lambda)$ is obtained from (2.24) for u_k replaced by h_k. From the proof of lemma 2.7, we see that the first term from the right-side is of order $O(n^{-1/2})$ uniformly in h if

$$(A^*) \quad \frac{1}{N} \sum_{k \in U} \tilde{g}^2_{h,k}(\lambda) = O(1) \quad \text{uniformly in h} \in \mathcal{H}.$$

If the second term from the right-side satisfies

$$(A^{**}) \quad \mathbb{E}_p \left| \frac{1}{N} \sum_{k \in U} (d_k I_k - 1)\,(\tilde{g}_{h,k}(\lambda) - \hat{g}_{h,k}(\lambda)) \right| = O(n^{-1/2}) \quad \text{uniformly in h} \in \mathcal{H},$$

then the total variation distance between $\widehat{M}^{\text{np}}(\lambda)/N$ and M/N will be of order $O_p(n^{-1/2})$. We can remark that the above reasonings do not make use of the fact that the penalized B-spline regression has been used. As a consequence, the convergence of a nonparametric estimator of M will be ensured if (A^*) and (A^{**}) are fulfilled (Goga and Ruiz-Gazen [2014a]).
Using classical assumptions from a B-spline framework and mild assumptions regarding the sampling design, Goga and Ruiz-Gazen [2014a] prove that (A^*) and (A^{**}) are fulfilled in the case of penalized B-spline regression (lemma 2.13 from below).

Lemma 2.13. *Make assumptions (S1) and (S2) on the sampling design and (B1)-(B3) on the B-spline functions. Then:*

1. $||\tilde{\boldsymbol{\theta}}_h(\lambda)|| = O(K^{1/2})$ *uniformly in* $h \in \mathcal{H}$, *where* $||\cdot||$ *is the usual Euclidian norm.*
2. $\mathbb{E}_p||\tilde{\boldsymbol{\theta}}_h(\lambda) - \hat{\boldsymbol{\theta}}_h(\lambda)||^2 = O\left(\frac{K^3}{n}\right)$ *uniformly in* $h \in \mathcal{H}$.

In particular, (A^{**}) $\mathbb{E}_p \left| \frac{1}{N} \sum_{k \in U} (d_k I_k - 1)\,(\tilde{g}_{h,k}(\lambda) - \hat{g}_{h,k}(\lambda)) \right| = o\left(n^{-1/2}\right)$ *uniformly in* h.

Lemma 2.14. *Under the assumptions of lemma 2.13, we get*

$$d_{\text{tv}}\left(\frac{\widehat{M}^{\text{np}}(\lambda)}{N}, \frac{M}{N}\right) = O_p(n^{-1/2}).$$

The following matrix norms will be used in the proofs of these two lemmas: $||\cdot||_\infty$ defined for a matrix $\mathbf{A} = (a_{i,j})_{i,j=1}^q$ by $||\mathbf{A}||_\infty = \max_{i=1}^q \sum_{j=1}^q |a_{ij}|$ and $||\cdot||_2$, the spectral norm defined by $||\mathbf{A}||_2 = \sup_{\mathbf{a},||\mathbf{a}||=1} \mathbf{a}^T\mathbf{A}^T\mathbf{A}\mathbf{a}$.

The proofs of these two lemmas are based on results from nonparametric statistic and two lemmas from Goga [2005]. More exactly, Zhou et al. [1998] showed that

$$\left|\left|N^{-1}\sum_{k\in U}\mathbf{b}(x_k)\mathbf{b}^T(x_k)\right|\right|_2 = O(K^{-1}) \quad \text{and}$$

$$\left|\left|N\left(\sum_{k\in U}\mathbf{b}(x_k)\mathbf{b}^T(x_k)\right)^{-1}\right|\right|_\infty = O(K).$$

Claeskens et al. [2009] extended this result for penalized splines showing that

$$\left|\left|N\left(\sum_{k\in U}\mathbf{b}(x_k)\mathbf{b}^T(x_k)+\lambda\mathbf{D}_\ell\right)^{-1}\right|\right|_\infty = O(K).$$

Proof. (of lemma 2.13) We have

$$\tilde{\boldsymbol{\theta}}_h(\lambda) = \left(N^{-1}(\sum_{k\in U}\mathbf{b}(x_k)\mathbf{b}^T(x_k)+\lambda\mathbf{D}_\ell)\right)^{-1}\left(\sum_{U}\mathbf{b}(x_k)h_k/N\right)$$

By using that $\sup_{k\in U}|h(y_k)| \leq 1$ and following the same lines as in Goga [2005], we get that $||\sum_U \mathbf{b}(x_k)h_k/N|| = O(K^{-1/2})$ uniformly in h. Hence, $||\tilde{\boldsymbol{\theta}}_h(\lambda)|| = O(K^{1/2})$ uniformly in h. For point 2, the convergence of $\hat{\boldsymbol{\theta}}_h(\lambda)$, the difficult part was to bound the inverse of the matrix $\sum_{k\in s} d_k\mathbf{b}(x_k)\mathbf{b}^T(x_k)+\lambda\mathbf{D}_\ell$. Goga and Ruiz-Gazen [2014a] showed that

$$\mathbb{E}_p\left|\left|N\left[\left(\sum_{k\in s}d_k\mathbf{b}(x_k)\mathbf{b}^T(x_k)+\lambda\mathbf{D}_\ell\right)^{-1}-\left(\sum_{k\in U}\mathbf{b}(x_k)\mathbf{b}^T(x_k)+\lambda\mathbf{D}_\ell\right)^{-1}\right]\right|\right|_2^2 = O\left(\frac{K^4}{n}\right).$$

□

Proof. (of lemma 2.14) In order to prove that the distance between $\widehat{M}^{\text{np}}/N$ and M/N is going to zero, it suffices to show that relations (A^*) and (A^{**}) are fulfilled. We have that $\tilde{g}_{h,k}(\lambda) = \mathbf{b}^T(x_k)\tilde{\boldsymbol{\theta}}_h(\lambda)$. Hence,

$$\frac{1}{N}\sum_{k\in U}\tilde{g}^2_{h,k}(\lambda) \leq ||\tilde{\boldsymbol{\theta}}_h(\lambda)||^2\left|\left|\frac{1}{N}\sum_{k\in U}\mathbf{b}(x_k)\mathbf{b}^T(x_k)\right|\right|_2 = O(1),$$

uniformly in h, by using $||N^{-1}\sum_{k\in U}\mathbf{b}(x_k)\mathbf{b}^T(x_k)||_2 = O(K^{-1})$ (lemma 6.2 from Zhou et al. [1998]). The property (A^{**}) results since we have

$$\begin{aligned}\mathbb{E}_p\left|\frac{1}{N}\sum_{k\in U}(d_kI_k-1)\left(\tilde{g}_{h,k}(\lambda)-\hat{g}_{h,k}(\lambda)\right)\right| &\leq \sqrt{\mathbb{E}_p\left|\left|\frac{1}{N}\sum_{k\in U}(d_kI_k-1)\,\mathbf{b}^T(x_k)\right|\right|^2\cdot\mathbb{E}_p||\tilde{\boldsymbol{\theta}}_h(\lambda)-\hat{\boldsymbol{\theta}}_h(\lambda)||^2}\\ &= o(n^{-1/2})\end{aligned}$$

for $K=O(n^a)$ with $a<1/3$. □

Lemma 2.14 and theorem 2.8 provide the first linearization step, by which the nonparametric estimator $\widehat{\Phi}^{\text{np}}(\lambda)$ is asymptotically equivalent to a nonparametric model-assisted estimator $\hat{t}^{\text{np}}_{uw}(\lambda)$ for $\sum_{k\in U}u_k$.

Theorem 2.15. *(First linearization step) (Goga and Ruiz-Gazen [2014a]) Make assumptions (F1), (F2) on the functional T and assumptions (S1), (S2) on the sampling design. Assume in addition that (B1)-(B3). Then, the nonparametric substitution estimator $\widehat{\Phi}^{\text{np}}$ fulfills*

$$\begin{aligned}N^{-\alpha}\left(\widehat{\Phi}^{\text{np}}(\lambda)-\Phi\right) &= N^{-\alpha}(\hat{t}^{\text{np}}_{uw}(\lambda)-t_u)+o_p(n^{-1/2})\\ &= N^{-\alpha}\left(\sum_{k\in s}w^{\text{np}}_{ks}(\lambda)u_k-\sum_{k\in U}u_k\right)+o_p(n^{-1/2}).\end{aligned}$$

In theorem 2.16 from below, we obtain that $\hat{t}^{\text{np}}_{uw}(\lambda)$ is asymptotically equivalent to the generalized difference estimator given by:

$$\tilde{t}^{\text{diff}}_{u,x}(\lambda)=\sum_{k\in s}d_k(u_k-\tilde{g}_{u,k}(\lambda))+\sum_{k\in U}\tilde{g}_{u,k}(\lambda), \tag{2.31}$$

where $\tilde{g}_{u,k}(\lambda)=\mathbf{b}^T(x_k)\tilde{\boldsymbol{\theta}}_u(\lambda)$ with $\tilde{\boldsymbol{\theta}}_u(\lambda)$ given by (2.23).

Theorem 2.16. *(Second linearization step) (Goga and Ruiz-Gazen [2014a]) Make assumptions from theorem 2.15. Assume in addition that the linearized variable u_k satisfies V3 (b). Then, the nonparametric estimator $\hat{t}^{\text{np}}_{uw}(\lambda)$ is asymptotically equivalent to the generalized difference estimator $\tilde{t}^{\text{diff}}_{u,x}(\lambda)$ given in (2.31) in the sense that*

$$N^{-\alpha}(\hat{t}^{\text{np}}_{uw}(\lambda)-t_u)=N^{-\alpha}(\tilde{t}^{\text{diff}}_{u,x}(\lambda)-t_u)+o_p(n^{-1/2}).$$

Combining both theorems, we obtain that the nonparametric estimator $\widehat{\Phi}^{\text{np}}(\lambda)$ is asymptotically equivalent to the nonparametric generalized difference estimator $\tilde{t}^{\text{diff}}_{u,x}(\lambda)$ and in particular:

$$N^{-\alpha}(\widehat{\Phi}^{\text{np}}(\lambda)-\Phi)=O_p(n^{-1/2}),$$

implying that it is ADU and consistent for Φ. Moreover, the asymptotic variance of $\widehat{\Phi}^{\text{np}}(\lambda)$ is given by:

$$\begin{aligned} A\mathbb{V}_p(\Phi^{\text{np}}) &= \sum_{k\in U}\sum_{k\in U}(\pi_{kl}-\pi_k\pi_l)d_kd_l(u_k-\tilde{g}_{u,k}(\lambda))(u_l-\tilde{g}_{u,l}(\lambda)) \\ &= \sum_{k\in U}\sum_{k\in U}(\pi_{kl}-\pi_k\pi_l)d_kd_l(u_k-\mathbf{b}^T(x_k)\tilde{\boldsymbol{\theta}}_u(\lambda))(u_l-\mathbf{b}^T(x_l)\tilde{\boldsymbol{\theta}}_u(\lambda)). \end{aligned} \tag{2.32}$$

The asymptotic variance is in fact the HT variance for the residuals $u_k-\mathbf{b}^T(x_k)\tilde{\boldsymbol{\theta}}_u(\lambda)$ of the linearized variable u_k under the model ξ' given in (2.20). This variance is similar to the asymptotic variance, given in (1.18), of the GREG estimator for the total t_y built under the linear model (1.19). The smallest the residuals $u_k-\mathbf{b}^T(x_k)\tilde{\boldsymbol{\theta}}_u(\lambda)$, $k\in U$ are, the best the estimator Φ^{np} for Φ is. Considering nonparametric models ξ' as in (2.20) and B-spline regression provide good prediction for rather complicated u_k and lead to low residuals $u_k-\mathbf{b}^T(x_k)\tilde{\boldsymbol{\theta}}_u(\lambda)$. Nevertheless, unlike the GREG estimators derived under a linear model, nonparametric model-assisted estimators need x_k to be known for all the individuals from the population. Goga and Ruiz-Gazen [2014a] suggested a variance estimator for Φ^{np} and gave assumptions under which the suggested variance estimator is consistent for $A\mathbb{V}_p(\Phi^{\text{np}})$.

Remark 4. *In the case of the estimation of finite population total $t_y=\sum_{k\in U}y_k$, than $\hat{t}_{yw}^{\text{np}}(\lambda)$, obtained for y_k instead of u_k, is the nonparametric estimator of t_y. Theorem 2.16 gives than the asymptotic behavior of this estimator.*

Remark 5. *In the case of the estimation of t_y for B-spline functions of order $m=1$ and $\lambda=0$, we obtain $\hat{t}_{yw}^{\text{np}}(0)$, the poststratified estimator of t_y with a number of post-strata going to infinity. In this context, theorem 2.16 provides a detailed theoretical justification of the consistency of the poststratified estimator, result claimed without proof in Deville [1999a].*

Remark 6. *A very important thing to be noticed is that the above asymptotic results are obtained without supplementary assumptions upon the smoothness of the regression function g.*

2.2.3 A calibration point of view for the unpenalized case

As Särndal [2007] stated, Deville and Särndal's calibration is a different point of view even if it leads (asymptotically) to an estimator equal to the model-assisted or GREG estimator derived under a linear model (see Chapter 1 for a brief review of the method). This method is very popular and used in statistical institutes. However, when the parameter of interest is not a total, an obvious question rises: which are the calibration constraints now? And for the moment, several authors tackled this issue and tried to answer this question (Särndal [2007]). For example, Harms and Duchesne [2006] suggested a calibration method for quantiles and Plikusas [2006] for the ratio or the covariance. Another calibration approach known as *model-calibration* has been introduced by Wu and Sitter [2001] for estimating means and totals. If

a known nonlinear model fits the data, Wu and Sitter [2001] suggested to calibrate on the predictions under the suggested model. If the regression function is unknown, Montanari and Ranalli [2005] suggested using local polynomial regression to estimate it and calibration on the estimated predictions. This approach has been called *nonparametric calibration*. However, all these approaches have important drawbacks: the sampling weights depend on the parameter to estimate (Harms and Duchesne [2006], Plikusas [2006]) or on the study variable (Wu and Sitter [2001], Montanari and Ranalli [2005]) entailing a loss of the multipurpose property.

The Deville and Särndal's method is based on an implicit underlying assumption that the relationship between the study and the auxiliary variable is linear. The main goal is to find calibration weights when this relationship is no longer linear and/or the study parameter is more complex than totals or means. In order to accomplish it, the calibration constraint must be changed while keeping the property that the obtained weights do not depend on the study variable or parameter.

A simple way to overcome all these difficulties is to consider calibration on the vector of the B-spline basis functions $\mathbf{b} = (B_1, \ldots, B_q)^T$ (Goga and Ruiz-Gazen [2014c]). More exactly, the B-spline calibration weights $\mathbf{w}_s^{\text{cal,np}} = (w_{ks}^{\text{cal,np}})_{k\in s}$ minimize the chi-squared distance $\Upsilon_s(\mathbf{w})$ from (1.14) to the HT weights:

$$\mathbf{w}_s^{\text{cal,np}} = \text{argmin}_{\mathbf{w}} \Upsilon_s(\mathbf{w})$$

subject to

$$\sum_{k\in s} w_{ks}^{\text{cal,np}} \boldsymbol{b}(x_k) = \sum_{k\in U} \boldsymbol{b}(x_k). \tag{2.33}$$

One can deduce (see also 1.15):

$$w_{ks}^{\text{cal,np}} = d_k - d_k q_k \boldsymbol{b}^T(x_k) \left(\sum_{k\in s} \frac{q_k \boldsymbol{b}(x_k) \boldsymbol{b}^T(x_k)}{\pi_k} \right)^{-1} \left(\sum_{k\in s} d_k \boldsymbol{b}(x_k) - \sum_{k\in U} \boldsymbol{b}(x_k) \right), \quad k \in s.$$

The vector of weights $\mathbf{w}_s^{\text{cal,np}}$ depends only on the auxiliary variable they offer a great adaptability with respect to the study parameter. In particular, the weights $\mathbf{w}_s^{\text{cal,np}}$ satisfy

$$\sum_{k\in s} \mathbf{w}_s^{\text{cal,np}} x_k^\ell = \sum_{k\in U} x_k^\ell, \quad \text{for} \quad \ell = 1, \ldots, q.$$

The estimator $\Phi^{\text{cal,np}}$ built by using these weights will be approximated by

$$\hat{t}_{uw}^{\text{cal,np}} = \sum_{k\in s} w_{ks}^{\text{cal,np}} u_k = \sum_{k\in s} d_k u_k - \left(\sum_{k\in s} d_k \mathbf{b}(x_k) - \sum_{k\in U} \mathbf{b}(x_k) \right)^T \hat{\boldsymbol{\beta}}_u(\mathbf{q}^{-1}),$$

where $\hat{\boldsymbol{\beta}}_u(\mathbf{q}^{-1}) = \left(\sum_{k\in s} d_k q_k \mathbf{b}(x_k)\mathbf{b}^T(x_k)\right)^{-1} \sum_{k\in s} d_k q_k \mathbf{b}(x_k) u_k$ with $\mathbf{q} = (q_k)_{k\in s}$. The above estimator is in fact the classical calibration estimator for the total $t_u = \sum_{k\in U} u_k$ as in (1.16).

Note also the similarity with the B-spline model-assisted estimator $\hat{t}_{uw}^{\text{np}}$ given in (2.27) and computed for $\lambda = 0$. In fact the two estimators coincide, $\hat{t}_{uw}^{\text{np}}(0) = \hat{t}_{uw}^{\text{cal,np}}$, if we take $q_k = 1$ for all $k \in s$ in the chi-square distance Υ_s as stated in the proposition 2.17 from below.

The method we suggest is different from the nonparametric model-calibration as suggested by Montanari and Ranalli [2005] for estimating means or totals. They considered calibration on the estimated predictions obtained by using local polynomial regression. One way to extend their method for an arbitrary parameter Φ and B-spline regression is given in the following.
The B-spline model-calibration weights $\mathbf{w}_s^{\text{mcal}} = (w_{ks}^{\text{mcal}})_{k\in s}$ minimize the chi-square distance $\Upsilon_s(\mathbf{w})$ to the HT weights and subject to:

$$\sum_{k\in s} w_{ks}^{\text{mcal}} \hat{g}_{u,k} = \sum_{k\in U} \hat{g}_{u,k}, \tag{2.34}$$

where $\hat{g}_{u,k} = \mathbf{b}^T(x_k)\hat{\boldsymbol{\theta}}_u$ is the estimated prediction of u_k under the model 2.20 with $\hat{\boldsymbol{\theta}}_u$ obtained from (2.26) for $\lambda = 0$. The Montanari and Ranalli's weights are given by:

$$w_{ks}^{\text{mcal}} = d_k - d_k q_k \hat{g}_{u,k} \left(\sum_{k\in s} d_k q_k \hat{g}_{u,k}^2\right)^{-1} \left(\sum_{k\in s} d_k \hat{g}_{u,k} - \sum_{k\in U} \hat{g}_{u,k}\right), \quad k \in s. \tag{2.35}$$

The estimator Φ^{mcal} built by using these weights will be approximated by

$$\hat{t}_{uw}^{\text{mcal}} = \sum_{k\in s} w_{ks}^{\text{mcal}} u_k = \sum_{k\in s} d_k u_k - \left(\sum_{k\in s} \hat{g}_{u,k} - \sum_{k\in U} \hat{g}_{u,k}\right) \frac{\sum_{k\in s} d_k q_k u_k \hat{g}_{u,k}}{\sum_{k\in s} d_k q_k \hat{g}_{u,k}^2}.$$

Wu and Sitter [2001] showed that the ratio $\sum_{k\in s} d_k q_k u_k \hat{g}_{u,k} / \sum_{k\in s} d_k q_k \hat{g}_{u,k}^2$ is not equal to 1 for nonlinear models and Montanari and Ranalli [2005] showed the same property for local polynomial regression. This means that their model-calibration estimator is different from the model-assisted estimator.
An important drawback with this approach is that the weights depend on u_k. Besides the loss of the multi-purpose property, the weights cannot be computed in this case since u_k is not known. We could have considered calibration on $\hat{g}_{\hat{u},k} = \mathbf{b}^T(x_k)\hat{\boldsymbol{\theta}}_{\hat{u}}$ obtained for an estimator of $u_k, k \in s$ but the weights so obtained would still have depended on the linearized variable and besides, the asymptotic results would have been even more complicated.
By using the fact that the estimated prediction $\hat{g}_{u,k}$ are linear combination of $\mathbf{b}(x_k)$, we deduce that the weights $\mathbf{w}_s^{\text{cal,np}}$ also satisfy the calibration constraints given in (2.34). This is a very important property of $\mathbf{w}_s^{\text{cal,np}}$ since they have much simpler expression and besides, do not depend on u_k. However, as stated in the following proposition, there is a particular case for which weights $\mathbf{w}_s^{\text{cal,np}}$ and $\mathbf{w}_s^{\text{mcal}}$ coincide, and are equal in this case to $\mathbf{w}_s^{\text{np}}(0)$.

Proposition 2.17. *If $q_k = 1$, for all $k \in s$, then $w_{ks}^{\text{mcal}} = w_{ks}^{\text{cal,np}} = w_{ks}^{\text{np}}(0)$, for $k \in s$.*

This property results from the particular relationship between the B-spline approach and the multivariate linear model. Unlike nonlinear and local polynomial model calibration estimators

as exhibited in Wu and Sitter [2001] and Montanari and Ranalli [2005]), the B-spline model-calibration estimator built for $q_k = 1$ is equal to the B-spline model-assisted estimator proving that this property is true for models that are even more general than the multivariate linear model.

Calibration on the B-spline functions as suggested in (2.33) consists in considering q calibration constraints, where q is the dimension of the B-spline basis. Recall that $q = K + m$ where K is the number of interior knots and m is the order of the B-spline functions. As already mentioned, low values of m are usually taken ($m = 2$ or $m = 3$). Considering a large number K of interior knots leads to a large q, so many calibration constraints. Many difficulties may arise in these situations as presented in Section 3.2. To overcome these difficulties, we can use the penalized calibration and obtain the penalized B-splines weights $\mathbf{w}_s^{\text{np}}(\lambda)$ from (2.2.1).

Application of the suggested method on data extracted from the French Labor Force surveys of 1999 and 2000 (results are not reported here) shows the good behavior of the suggested B-spline calibration estimator. This new calibration approach is a promising one and work on this topic is actually in progress. It concerns among others, the extension to multivariate auxiliary information and relationship with model-based estimators (Goga and Ruiz-Gazen [2014c]).

2.2.4 Application on the French Labour Force Survey from 1999-2000

Let us consider a data set from the French Labor Force surveys of 1999 and 2000 as considered by Goga et al. [2009] and Goga and Ruiz-Gazen [2014a]. The data consist of the monthly wages (in euros) of 19,378 wage-earners who were sampled in both years. The study variable y_k (resp. the auxiliary variable z_k) is the wage of person k in 2000 (resp. 1999).

The parameters to estimate here include the mean and the Gini index for the wages in 2000 using the wages in 1999 as auxiliary information. Goga and Ruiz-Gazen [2014a] has considered also the poverty rate and Goga and Ruiz-Gazen [2014b] the odds-ratio but results are not reported here. A simple random sampling without replacement of sizes $200, 500$ and respectively, 1000 is considered and the following estimators for each parameter:
- the Horvitz-Thompson estimator (HT), which does not incorporate any auxiliary information,
- poststratified estimators (POST) with a different number of strata bounded at the empirical quantiles for 1999 wages,
- the GREG estimator (GREG), which takes into account the 1999 wages as auxiliary information using a simple linear model,
- B-spline estimators (BS(m) where m denotes the spline order), which take into account the wages from 1999 as auxiliary information by using a nonparametric model with different numbers of knots (K) located at the quantiles of the empirical distribution for wages from 1999. The $m = 2$ and $m = 3$ orders are considered. The poststratified estimator is an example of a B-spline estimator with order $m = 1$. The number of strata correspond to the number of interior knots K plus one.

The performance of an estimator $\hat{\theta}$ for a parameter θ is evaluated by computing the ratio of root mean squared errors in percentage with respect to the HT estimator:

$$\text{RRMSE} = 100 \times \sqrt{\sum_{i=1}^{I}(\hat{\theta}_i - \theta)^2} / \sqrt{\sum_{i=1}^{I}(\hat{\theta}_{i,d} - \theta)^2}$$

for $I = 3000$ simulations. Results are reported in Table 2.1.

Not surprisingly, for complex parameters, the largest efficiency gain is observed when the B-spline estimators are compared to the HT estimator without auxiliary information. Because the wages from 2000 are almost linearly related to the wages from 1999, considering the B-spline estimator instead of the GREG estimator does not improve the performance of the mean estimation. However, regarding the Gini index, the incorporation of auxiliary information using GREG estimators does not improve efficiency compared to the HT estimator while using a B-spline approach improves the results especially for spline functions of order $m = 2$. When comparing the POST estimator with the BS(2) and BS(3) estimators, we notice that there is quite a large gain in efficiency when order $m = 2$ is used instead of $m = 1$, while there is an efficiency loss when $m = 3$ is used instead of $m = 2$, especially for sample sizes smaller than 1,000. Moreover, for $m = 2$ and $m = 3$, the results do not depend heavily on the number of knots and are similar for K between 2 and 4 while for the poststratified estimator, there are large variations in the results, regardless of whether we consider 3 or 5 strata.
Goga and Ruiz-Gazen [2014a] have also evaluated the coverage probabilities. They obtain that valid inference can be carried out using B-spline estimators as long as the spline order is not too high, especially when the sample size is not very large. No problems are detected for B-splines of order $m = 1$ and order $m = 2$ even when the sample size is $n = 200$; however for $m = 3$ and $n = 200$, the coverage probabilities for the Gini index estimation are approximately 75% which is quite far from the 95% nominal probability. This result indicates that for a moderate sample size, the variance may be underestimated when the order of the splines is larger than two. The results are not given for $m = 4$ but we have observed that the problem worsens when we increase the order of the splines. This is not really surprising due to double linearization and nonparametric estimation.
Based on this example, Goga and Ruiz-Gazen [2014a] do not recommend using high order values for B-spline regression, especially when the sample sizes are smaller than 500. However, choosing $m = 2$ instead of $m = 1$ (which corresponds to poststratification) leads to a clear improvement in terms of efficiency for complex parameters such as the Gini index or the low-income proportion, and they recommend this choice.

TABLE 2.1: RRMSE of GREG and POST, BS(2) and BS(3) for the mean and the Gini index

Parameter	n	GREG	POST	BS(2)	BS(3)
			$K = 2$ - $K = 4$	$K = 2$ - $K = 4$	$K = 2$ - $K = 4$
Mean	200	38	71 - 63	38 - 37	39 - 41
	500	40	73 - 65	40 - 39	38 - 39
	1,000	40	73 - 66	40 - 40	38 - 39
Gini index	200	96	92 - 80	53 - 53	70 - 70
	500	93	93 - 85	50 - 50	59 - 56
	1,000	92	93 - 86	49 - 48	55 - 51

2.3 Conclusion and perspectives

I have presented in this chapter the estimation with survey sampling designs of nonlinear parameters in a finite population framework. The main concern was about giving a unified presentation by means of the functional approach which is a powerful method to linearization of such parameters as well as developing rigorous justifications of the asymptotic results. Throughout the chapter, the theory has been illustrated on two nonlinear parameters: the Gini index and the odds-ratio. However, applications of results exhibited here are numerous and Chapter 3.1 is concerned, among others, with the estimation of elements of the functional principal component analysis or the functional median.

A general class of substitution estimators that allows to take into account complete auxiliary information is suggested. Through a nonparametric B-spline regression and a model-assisted approach, a unique system of weights is constructed and it can be used to estimate efficiently any nonlinear study parameter that is associated with any study variable of the survey. A very important feature of nonparametric weights based on B-spline regression is their great similarity with the parametric weights build under a linear model. Moreover, Goga and Ruiz-Gazen [2014a] conclude that in order to estimate efficiently a larger class of parameters, it is enough to consider slightly more complicated basis of functions than in the case of the linear model. This can be achieved for a low order of B-spline basis ($m = 2$ or $m = 3$) and few interior knots. The functions of the B-spline basis are easy to compute by using for instance the `transreg` procedure in the SAS software (SAS Institute, 2010) or the `splines` package from the R software (R Core Team, 2012). Finally, a new calibration approach is suggested and shown to have many attractive properties.

Work presented in this chapter may be extended and continued following several directions. The asymptotic behavior of the substitution estimators computed under complex sampling designs, such as the two-stage sampling, is actually in progress (Chauvet and Goga [2013b]). This work also considered a two-sample or temporal setting and the variance estimators obtained by functional linearization are compared with those obtained by using bootstrap methods for estimating the variance. The extension of bootstrap methods to the two-sample case presents many difficulties and drawbacks with respect to the linearization.

Concerning the nonparametrics framework, several directions for future works are possible. We mention the extension of the nonparametric approach to multivariate auxiliary information as well as the study of the computation of the penalty parameter λ. Another direction of research would be to borrow the idea used in B-spline calibration in the context of balanced sampling. For the moment, balanced sampling has been suggested to improve the estimation of the total. It may be of interest to extend it to the estimation of an arbitrary nonlinear parameter Φ. A research project on this topic is planned in collaboration with G. Chauvet and A. Ruiz-Gazen.

Chapter 3

Estimation with survey sampling techniques in presence of large datasets: functional and high dimensional data

Nowadays, with the spread of automatic process for data collection as well as increasing storage capacities, it is not unusual anymore to have to analyse large data sets. Audience curves or electricity load curves are two exemples of such data. More exactly, the ERDF (Electricité et Réseaux de France) plans to install more than 30 millions of smart meters in every household and company. These meters will be able to send individual electricity consumption measures at very fine time scales. The discretization scheme is very fine so that the statistical units can be considered as functions of time. We can use the tools of functional data analysis to describe the data and build statistical models. Even if some of these tools have been first proposed in the 1970s in Deville [1974] and Dauxois and Pousse [1976], these methods only really begun to spread twenty years ago with the increase of computer performances as well as storage capacities. The reader may refer to Ramsay and Silverman [2005] and Ferraty and Vieu [2006] for an overview of the different techniques developed in the statistical literature in functional data analysis as well as examples of application.

Nevertheless, in the presence of technical and budgetary constraints due to limited bandpass or storage costs of huge databases, the analysis of the whole dataset may be very difficult or even, not possible. In Chiky [2009], it is shown that if we are only interested in simple indicators, such as total or mean trajectories, even very simple survey sampling techniques, such as simple random sampling without replacement, are attractive alternatives to signal compression techniques since they permit to obtain precise estimates at a reasonable cost. Motivated by this new setting, several papers combined recently functional data analysis and sampling theory. Cardot and Josserand [2011] and Cardot et al. [2013b] considered the uniform convergence of the HT

estimator of the mean curve. An important issue with this new type of datasets, is how to build asymptotic confidence bands with desired coverage rates (Cardot and Josserand [2011]). Cardot et al. [2013c] made a comparaison, in terms of precision of the estimators for the mean of electricity consumption, of different approaches that can take auxiliary information into account. They also compare the width of the confidence bands. The conclusion of the empirical study was that incorporating the auxiliary information in the sampling design or at the estimation stage improves a lot the performance of estimators. In particular, the width of confidence bands is greatly reduced. Theoretical justification of these results are established in Cardot et al. [2013d] and Cardot et al. [2014b]. At the same time, Chaouch and Goga [2012] were interested in the estimation of robust parameters, such as the median curve and they studied the impact of different sampling designs and estimators on the estimation of this indicator.

Another aspect of large datasets is the possibility of having high-dimension sets of auxiliary information. In such conditions, the performance of estimators based on the whole auxiliary information may be damaged. During the 1990s, several authors suggested ridge-type estimators in a model-based approach as well as in a calibration approach in order to overcome the problems due to large datasets (Bardsley and Chambers [1984], Rao and Singh [1997]). However, these suggestions have received relatively little attention until now, when the amount of auxiliary information is becoming more and more large due to the increase of computer performances.

This chapter is structured as follows. Section 3.1 gives a presentation of results obtained for functional variables. The content of this chapter is mainly based on the review article submitted for a special issue for the "Journal de la SFdS" (Lardin-Puech et al. [2014]). After introducing the notations and parameters of interest in Section 3.1.1, the HT estimators as well as the substitution estimators are given in Sections 3.1.2-3.1.3. Consistency results are presented in Sections 3.1.4-3.1.6 and the HT estimators are improved in Section 3.1.7-3.1.8. Throughout this section, a test population of $N = 18902$ French companies for which the electricity consumption has been measured every half an hour over a period of two weeks, is used to illustrate the performance of the suggested estimators.

Section 3.2 treats the estimation of finite population totals by taking into account large datasets of auxiliary information. After presenting the difficulties risen in such conditions, Sections 3.2.1 and 3.2.2 give a detailed review of the penalized calibration as done in Goga and Shehzad [2014b]. A new class of penalized estimators based on principal components analysis is suggested in Sections 3.2.3-3.2.5. Asymptotic properties are established and a small application is conducted on electricity load curves.

3.1 Survey sampling designs for functional data

3.1.1 Notations and parameters of interest

We suppose that for each unit k from the population U, we can observe a deterministic function of time $Y_k = (Y_k(t))_{t\in[0,\mathcal{T}]}$[1] that belongs to some space of functions. Depending on the objective, this space will be either the space of continuous functions $C[0,\mathcal{T}]$ endowed with the sup norm or the Hilbert space $L^2[0,\mathcal{T}]$, *i.e.* the space of square integrable functions defined on the closed interval $[0,\mathcal{T}]$, equipped with the inner product $< f,g >= \int_0^{\mathcal{T}} f(t)g(t)dt$ and the induced norm $||f|| = (\int_0^{\mathcal{T}} f^2(t)d(t))^{1/2}$ for $f,g \in [0,\mathcal{T}]$.

For the illustration, consider a test population of $N = 18902$ French companies for which the electricity consumption has been measured every half an hour over a period of two weeks. A sample of 20 load curves extracted from the test dataset is drawn in Figure 3.1 as well as the mean and the median profiles.

In this functional setting, the statistician may be interested in estimating classical parameters of interest such as the total or the mean curve and their definition and interpretation are obtained easily by analogy with the non-functional case. The situation is more complicated for other parameters of interest, such as quantiles. The median may be defined in several manners for multivariate or functional data. Moreover, new functional parameters may be of interest now. When the aim is to build confidence bands the natural setting will be the space $C[0,\mathcal{T}]$ since we want to produce a confidence interval that is uniform in t. When the aim is to estimate the principal components or the geometric median, strict convexity of the norm of the underlying functional space as well as the existence of an inner product are required, so that the natural setting is to consider that the Y_k are elements of $L^2[0,\mathcal{T}]$.

We present below the functional parameters of interest that have been studied in a survey sampling setting. The simplest ones are total curve, defined as follows:

$$t_Y = \sum_{k\in U} Y_k$$

and the mean trajectory μ_N:

$$\mu_N = \frac{1}{N}\sum_{k\in U} Y_k. \tag{3.1}$$

The value of t_Y or μ_N in a measurement point $t \in [0,\mathcal{T}]$ is obtained directly as $t_Y(t) = \sum_{k\in U} Y_k(t)$ and $\mu_N(t) = \frac{1}{N}\sum_{k\in U} Y_k(t)$, respectively.

For such high dimensional data, other useful statistical indicators are given by the principal components that can exhibit the main modes of variation of the data around the mean curve (see

[1] In this section, I use capital letter for the study variables since we deal with functions now.

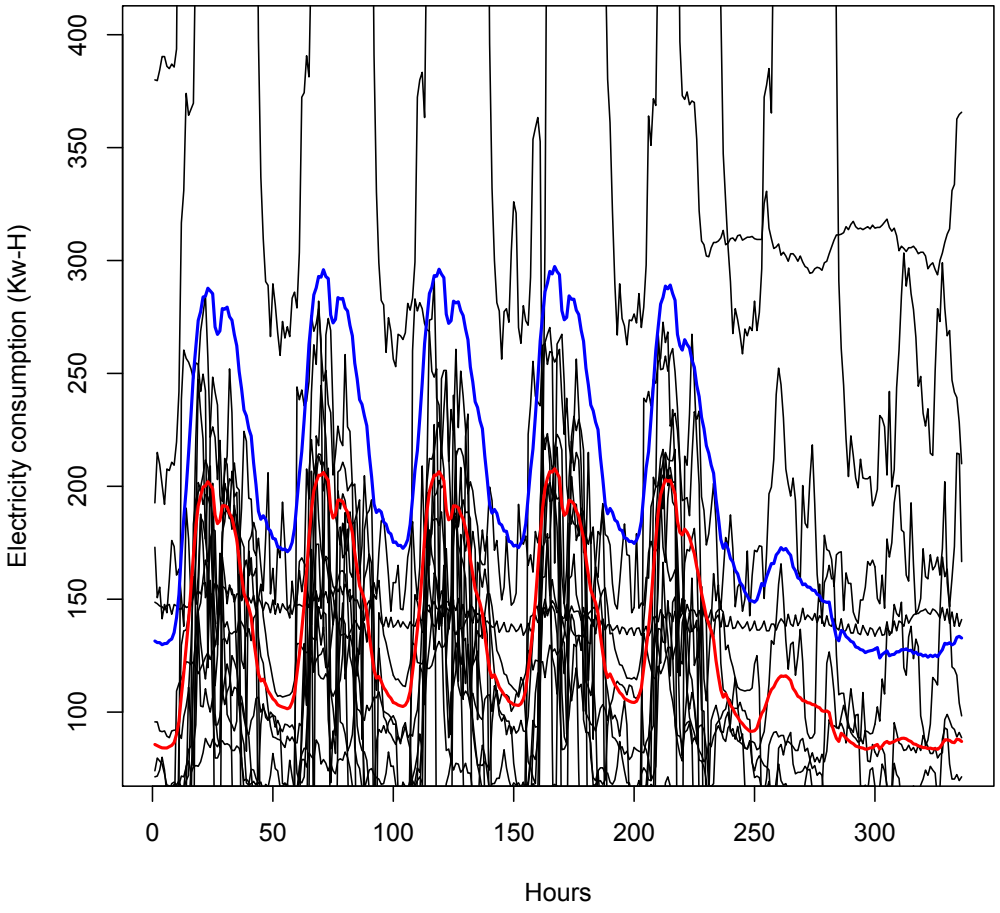


FIGURE 3.1: A sample of 20 electricity consumption curves measured every half an hour over a period of one week. The mean consumption curve in the population is drawn in bold blue line and the median curve in red one.

e.g Ramsay and Silverman [2005] and Cardot et al. [2010a]). To perform principal components analysis, it is first required to estimate the covariance function of the data at the population level. For r and t in $[0, \mathcal{T}]$, the covariance function $\gamma(r,t)$ between $(Y_k(r))_{k\in U}$ and $(Y_k(t))_{k\in U}$ is defined as follows:

$$\gamma(r,t) = \frac{1}{N}\sum_{k\in U}(Y_k(r) - \mu_N(r))(Y_k(t) - \mu_N(t)), \quad (r,t) \in [0,\mathcal{T}] \times [0,\mathcal{T}].$$

Then, the associated covariance operator $\mathbf{\Gamma}$ which maps $L^2[0,\mathcal{T}]$ to $L^2[0,\mathcal{T}]$ is defined by,

$$\mathbf{\Gamma}a(r) = \int_0^{\mathcal{T}} \gamma(r,t)a(t)dt, \tag{3.2}$$

for any function $a \in L^2[0,\mathcal{T}]$. The covariance operator has the following equivalent form:

$$\mathbf{\Gamma} = \frac{1}{N}\sum_{k\in U}(Y_k - \mu_N)\otimes(Y_k - \mu_N), \tag{3.3}$$

where the tensor product of two elements a and b of $L^2[0,\mathcal{T}]$ is the rank one operator such that $a\otimes b(y) =< a, y > b$ for all $y \in L^2[0,\mathcal{T}]$. The eigenvalues of Γ are non negative and supposed to be sorted in decreasing order $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_N \geq 0$. They satisfy:

$$\mathbf{\Gamma} v_j(t) = \lambda_j v_j(t), \quad j = 1,\ldots,N \tag{3.4}$$

where the eigenfunctions $v_j, j = 1,\ldots,N$ can be chosen to form an orthogonal system in $L^2[0,\mathcal{T}]$, namely $< v_j, v_{j'} >= 1$ if $j = j'$ and zero otherwise.

With high dimensional data, it is not uncommon to have outlying curves, such as consumers with very high levels of electricity consumption. In such a situation, it is advisable to consider indicators which are more robust to outlying data than the mean profile and the median is one of them. However, the notion of median can not be generalized easily to multivariate or functional data because of the lack of a natural ordering. There are several definitions of the median and we present here the one used by Kemperman [1987] and Gervini [2008] for functional data. Small [1990] gives a review of different definitions of the median with multidimensional data. With a finite population point of view, the median curve calculated from the elements $Y_1,\ldots,Y_N$ belonging to $L^2[0,\mathcal{T}]$ is defined by:

$$m_N = \text{argmin}_{y\in L^2[0,\mathcal{T}]}\sum_{k=1}^{N}\|Y_k - y\|. \tag{3.5}$$

For $Y_1,\ldots,Y_N \in \mathbb{R}^d$, m_N defined by the relation (3.5) arises as a natural generalization of the well-known characterization of the univariate median which can also be defined as $m_N = \arg\min_{y\in\mathbb{R}}\sum_{k=1}^{N}|Y_k - y|$ (see Koenker and Bassett [1978]). The median defined by (3.5) has been used for the first time at the beginning of the 20-th century. It was called the *spatial median* by Brown [1983] because, from a geometric point of view, the median is the point that minimizes the sum of distances to the points in the population. For example, Weber [1909] considered the following problem: a company wants to find the optimal location of its warehouse in order to serve the N customers with planar coordinates given by $Y_1,\ldots,Y_N$. It is also known as the Fermat-Weber point and Figure 3.2 gives the geometrical representation of the median with three bi-dimensional points. The median M is the point characterized by the fact that the three angles centered at M are equal.

The name of L_1-*median* was used by Small [1990] because the definition uses a L_1-criterion. Finally, Chaudhuri [1996] called it the *geometric median* because it may be seen as a particular case of the geometric quantiles whose definition uses the geometry of the data clouds by means of a direction and a magnitude.

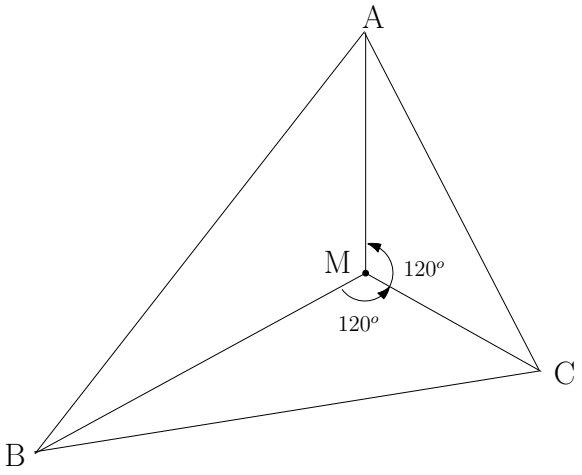


FIGURE 3.2: The median M of a uniform distribution with three atoms A, B and C in $\mathbb{R}^2$.

The median defined in (3.5) is global indicator of the centrality of the data, in the sense that it takes into account all the measurement instants. Chaudhuri [1996] showed that it is equivariant under orthogonal transformations and homogeneous scale transformations of the coordinates of the multivariate data. It possesses several robustness properties such as the breakdown point equal to 0.5 (Kemperman [1987], Lopuhaä and Rousseeuw [1987]), namely 50% of the data must be moved to infinity to force the median to do the same. As noted by Serfling [2002], the median defined by (3.5) and $Y_1, \ldots, Y_N \in \mathbb{R}^d$, m_N depends only on its direction towards Y_k. More exactly, m_N remains unchanged if the Y_k are moved outward along these rays; see also Ilmonen et al. [2012] for a recent review of the properties of the L_1-median.

If we assume that Y_k, for all $k = 1, \ldots, N$, are not concentrated on a line, the median exists and is unique (see Kemperman [1987]). If $m_N \neq Y_k$ for all $k = 1, \ldots, N$, then it is the unique solution of the following estimating equation:

$$\sum_{k=1}^{N} \frac{Y_k - y}{||Y_k - y||} = 0 \tag{3.6}$$

The median defined by (3.5) or (3.6) may be computed by using fast iterative algorithms such as Weiszfeld's algorithm (see Weiszfeld [1937] and Vardi and Zhang [2000]) for multivariate data or gradient algorithms (see Gervini [2008]) for sparse functional data. Note however that these algorithms may be time-consuming, especially if both the population size and the number of measurement instants are very large. To cope with this issue, Cardot et al. [2013a] suggest in a recent work to use recursive algorithms that are very fast and allow to compute the median when the data arrive sequentially. Alternatively, Chaouch and Goga [2012] suggest a weighted estimator of the L_1-median curve obtained by using only a sample drawn randomly from the population.

In a non-functional framework, Chaudhuri [1996] extends the definition given in (3.5) to geometric quantiles by using the geometry of data clouds. Chaouch and Goga [2010] investigated geometric quantiles with data collected from a complex survey and they suggested to use the quantile contours to detect outliers in multivariate data.

In a functional framework, the Chaudhuri [1996]'s definition indexes the quantiles by the elements $v \in L^2[0,\mathcal{T}]$ with $||v|| \leq 1$,

$$\mathcal{Q}(v) = \text{argmin}_{y \in L^2[0,\mathcal{T}]} \sum_{k=1}^{N} (||Y_k - y|| + < Y_k - y, v >).$$

In this way, functional quantiles are characterized by a direction and magnitude specified by $v \in L^2[0,\mathcal{T}]$ with $||v|| \leq 1$. Nevertheless, except the case $v = 0$, it is difficult to interpret the functional quantile defined in this way.

Considering again the electricity data presented in the Introduction we have plotted in Figure 3.3 the mean population curve as well as the L_1-median curve. As it can be seen, the median curve presents the same periodic patterns as the mean curve but with lower values. The difference comes from the fact that the distribution is very asymmetric, with a few curves with very large consumption levels.

3.1.2 The HT estimator for linear functional parameters

Without any auxiliary information, Cardot et al. [2008] and Cardot et al. [2010a] proposed to estimate the total curve t_Y by the (functional) HT estimator defined as follows[2]:

$$\hat{t}_Y = \sum_{k \in s} d_k Y_k = \sum_{k \in U} d_k Y_k \tag{3.7}$$

where π_k is not depending on t. The estimator $\hat{t}_{Y\pi}$ belongs to $L^2[0,\mathcal{T}]$ and its value at instant t, for $t \in [0,\mathcal{T}]$, is simply

$$\hat{t}_Y(t) = \sum_{k \in s} d_k Y_k(t).$$

Let us remark that the curves $Y_k(t)$ are considered as fixed with respect to the sampling design and it is the sample membership I_k that is random with respect to $p(\cdot)$. Using the fact that $\mathbb{E}_p(I_k) = \pi_k$, where $\mathbb{E}_p[\cdot]$ is the expectation with respect to the sampling design, we obtain easily that $\hat{t}_Y$ is design-unbiased for t_Y, namely $\mathbb{E}_p(\hat{t}_Y) = t_Y$.
An estimator of the mean curve μ_N is obtained by dividing by N the HT estimator $\hat{t}_{Y\pi}$, namely

$$\widehat{\mu} = \frac{1}{N} \hat{t}_Y. \tag{3.8}$$

[2] In this section, the subscript d from the definition of HT estimators is dropped off.

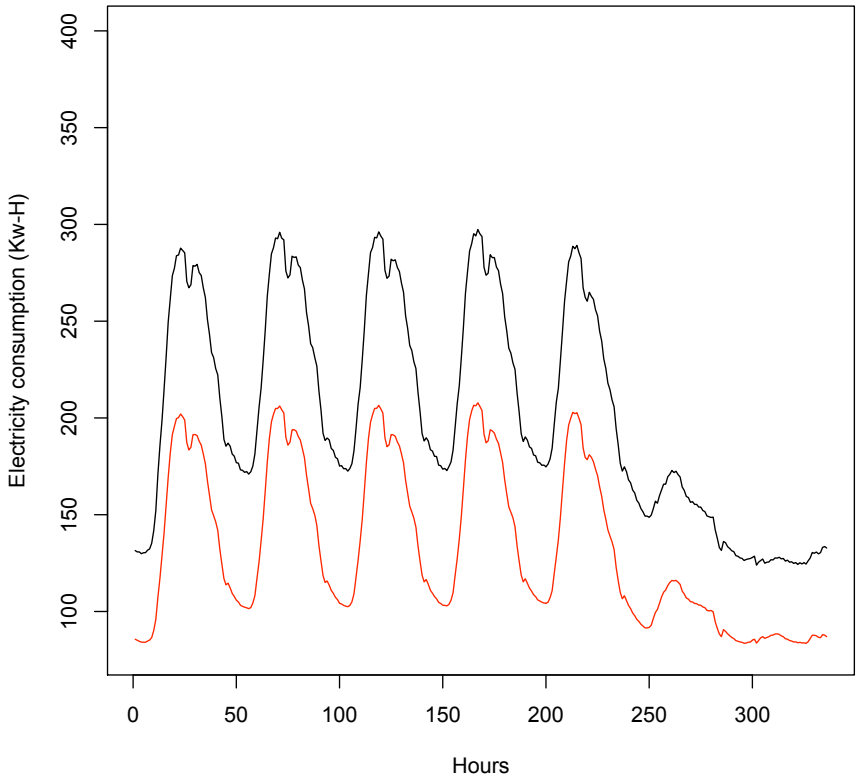


FIGURE 3.3: The L_1-median profile (in red) and the mean profile (in black) of the electricity consumption curves.

This estimator has been studied in Cardot et al. [2013c] and Cardot et al. [2013d].

The covariance between $\hat{t}_Y(r)$ and $\hat{t}_Y(t)$ computed with respect to the sampling design is derived easily by using the fact that $\text{Cov}_p(I_k, I_l) = \pi_{kl} - \pi_k\pi_l$ and it is given by a HT variance-type formula:

$$\gamma_p(r,t) = \sum_{k\in U}\sum_{l\in U}(\pi_{kl} - \pi_k\pi_l)d_k d_l Y_k(r)Y_l(t), \quad r,t \in [0,\mathcal{T}] \tag{3.9}$$

For $r = t$, we obtain the variance of $\hat{t}_Y(r)$. The covariance function $\gamma_p(r,t)$ is estimated unbiasedly with respect to the sampling design by:

$$\hat{\gamma}_p(r,t) = \sum_{k\in s}\sum_{l\in s}\frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}}d_k d_l Y_k(r)Y_l(t), \quad r,t \in [0,\mathcal{T}]. \tag{3.10}$$

Taking discretization effects into account

With real data, we generally do not observe $Y_k(t)$ at all instants t in $[0, \mathcal{T}]$ but only for a finite set of D measurement times, $0 = t_1 < ... < t_D = \mathcal{T}$. In functional data analysis, when the noise level is low and the grid of discretization points is fine, it is usual to perform a linear interpolation or to smooth the discretized trajectories in order to obtain approximations of the trajectories at every instant t (cf. Ramsay and Silverman [2005]). When there are no measurement errors and when the trajectories are regular enough, Cardot and Josserand [2011] showed, under weak regularity conditions, that linear interpolation can provide sufficiently accurate approximations of the trajectories to get efficient estimators of the mean trajectories. Thus, for each unit k in the sample s, we build the interpolated trajectory

$$Y_k^{(\mathrm{d})}(t) = Y_k(t_i) + \frac{Y_k(t_{i+1}) - Y_k(t_i)}{t_{i+1} - t_i}(t - t_i), \quad t \in [t_i, t_{i+1}], \tag{3.11}$$

and estimators can be constructed based on the interpolated values. For example, the estimator of t_Y based on the discretized observations is as follows:

$$\widehat{t}_Y^{(\mathrm{d})}(t) = \sum_{k \in s} d_k Y_k^{(\mathrm{d})}(t), \quad t \in [t_i, t_{i+1}]$$

and the mean trajectory by

$$\hat{\mu}^{(\mathrm{d})} = \frac{1}{N}\hat{t}_Y^{(\mathrm{d})}(t). \tag{3.12}$$

The covariance between $\widehat{t}_Y^{(\mathrm{d})}(t)$ and $\widehat{t}_Y^{(\mathrm{d})}(r)$ is then estimated by

$$\hat{\gamma}^{(\mathrm{d})}(r,t) = \sum_{k \in s}\sum_{l \in s} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl}} d_k d_l Y_k^{(\mathrm{d})}(r) Y_l^{(\mathrm{d})}(t), \quad r, t \in [0, \mathcal{T}]. \tag{3.13}$$

When the observations are corrupted by noise, Cardot et al. [2013b] proposed to replace the interpolation step by a smoothing step based on local polynomials. The smoothness of the mean estimator depends on a bandwidth whose value is selected by a cross-validation method that accounts for the sampling weights. They have shown on simulations that smoothing does really improve the accuracy of the HT estimator only when the noise level is high. On the other hand, smoothing can lead, for low and moderate levels of noise, to estimators that are outperformed by linear interpolation methods, specially when the value of the bandwidth is not selected effectively, which is the case if cross-validation is used curve by curve. Undersmoothing should be preferred.

3.1.3 Substitution estimators for non-linear parameters

The mean trajectory μ_N or the variance operator $\mathbf{\Gamma}$ are ratios of two finite population totals. The eigenvalues and eigenfunctions of $\mathbf{\Gamma}$ as well as the median trajectory m_N are also non-linear

functions of population totals as they are defined by the implicit equations (3.4) and (3.6), respectively.

To estimate these parameters, the strategy is simple and similar to the one used for real parameters and presented in Chapter 1. It consists in writing the parameter of interest as a functional T of the discrete measure M defined now on the space $L^2[0,\mathcal{T}]$ as suggested by Cardot et al. [2008] and Cardot et al. [2010a]:

$$M = \sum_{k\in U} \delta_{Y_k},$$

where δ_{Y_k} is the Dirac measure in Y_k with $k \in U$. All the non-linear parameters studied here can be written as functionals of M :

$$\mu_N = \frac{\int Y dM}{\int dM}, \tag{3.14}$$

$$\mathbf{\Gamma} = \frac{\int (Y-\mu_N)\otimes(Y-\mu_N) dM}{\int dM}. \tag{3.15}$$

The eigenvalues and eigenfunctions of Γ are also functionals of M as they are defined by the implicit equation (3.4). As for the median curve, consider the functional equal to the (Fréchet) derivative with respect to y of the objective function defined in (3.5):

$$T_{m_N}(M;y) = -\int \frac{Y-y}{||Y-y||} dM. \tag{3.16}$$

Then, the median is the unique solution of the implicit equation $T_{m_N}(M;m_N)=0$.

The measure M can be estimated by : [3]

$$\widehat{M} = \sum_{k\in s} d_k \delta_{Y_k},$$

and the estimators of $\mu, \mathbf{\Gamma}$ and of m_N respectively, are obtained by replacing M with $\hat{M}$ in their definition. These estimators are also called substitution estimators. More exactly, the mean trajectory μ_N is estimated by the Hájek-type estimator (see Cardot et al. [2010a]):

$$\widehat{\mu}_{Haj} = \frac{\hat{t}_Y}{\hat{N}}, \tag{3.17}$$

where $\hat{N} = \sum_s 1/\pi_k$ is the HT estimator of N and the variance operator $\mathbf{\Gamma}$ is estimated by

$$\widehat{\mathbf{\Gamma}} = \frac{1}{\hat{N}} \sum_{k\in s} \frac{(Y_k - \widehat{\mu}_{Haj})\otimes(Y_k-\widehat{\mu}_{Haj})}{\pi_k} = \frac{1}{\hat{N}}\sum_{k\in s}\frac{Y_k\otimes Y_k}{\pi_k} - \widehat{\mu}_{Haj}\otimes\widehat{\mu}_{Haj}.$$

The estimators $\hat{\lambda}_j$, $\hat{v}_j$ of λ_j, v_j are the eigenvalues and eigenfunctions of $\widehat{\mathbf{\Gamma}}$, namely

$$\widehat{\mathbf{\Gamma}}\hat{v}_j(t) = \hat{\lambda}_j \hat{v}_j(t), \quad t\in[0,\mathcal{T}]. \tag{3.18}$$

[3] The subscript d used in the notation of the HT estimators is dropped off in this section.

Considering now the median curve and assuming that all the Y_k, for $k \in s$ are not concentrated on a line, we obtain with (3.16), that m_N is estimated by $\hat{m}$, the unique solution of

$$\sum_{k\in s} d_k \frac{Y_k - \hat{m}}{\|Y_k - \hat{m}\|} = 0, \tag{3.19}$$

provided that $\hat{m} \neq Y_k$ for all $k \in s$ (see Chaouch and Goga [2012]).

3.1.4 Uniform consistency of the total or the mean curve estimators

For each fixed value of $t \in [0, \mathcal{T}]$, the estimator $\hat{t}_Y^{(\mathrm{d})}(t)$ is simply the estimator of a total of a real variable, so that under classical hypotheses on the sampling design and on the moments of $Y_k(t)$ for $k \in U$, it can be shown that it is consistent (see *e.g.* Chapter 1), namely

$$\text{for all} \quad \varepsilon > 0, \quad \lim_{N\to\infty} \mathbb{P}\left(\frac{1}{N}|\hat{t}_Y^{(\mathrm{d})}(t) - t_Y(t)| > \varepsilon\right) = 0,$$

as well as asymptotically Gaussian,

$$\frac{\sqrt{n}}{N}(\hat{t}_Y^{(\mathrm{d})}(t) - t_Y(t)) \to \mathcal{N}(0, \tilde{\gamma}_p(t))$$

where $\tilde{\gamma}_p(t) = \lim_{N\to\infty} \frac{n}{N^2}\gamma_p(t,t)$. Since $\hat{\mu}^{(\mathrm{d})} = \hat{t}_Y^{(\mathrm{d})}/N$, for ease of notation, we use $\hat{\mu}^{(\mathrm{d})}$ in the formulation of the following results.

In a functional setting, the interest is to get the uniform consistency of $\hat{\mu}^{(\mathrm{d})}$ as defined in the following definition.

Definition 3.1. An estimator $\hat{\theta}$ is uniformly consistent for θ if

$$\text{for all} \quad \varepsilon > 0, \quad \lim_{N\to\infty} \mathbb{P}\left(\sup_{t\in[0,\mathcal{T}]} |\hat{\theta}(t) - \theta(t)| > \varepsilon\right) = 0.$$

The suitable space for proving the uniform consistency is the space of continuous functions on $[0, \mathcal{T}]$ denoted by $C[0, \mathcal{T}]$ and equipped with its usual uniform distance:

$$\rho(f, g) = \sup_{t\in[0,\mathcal{T}]} |f(t) - g(t)| \quad \text{for } f, g \in C[0, \mathcal{T}].$$

We remark that the uniform consistency of $\hat{\mu}^{(\mathrm{d})}$ is in fact the convergence in probability (according to ρ) of $\hat{\mu}^{(\mathrm{d})}$ in $C[0, \mathcal{T}]$. One way to obtain the uniform consistency of an arbitrary estimator $\hat{\theta}$ is to show that $\mathbb{E}_p(\sup_{t\in[0,\mathcal{T}]} |\hat{\theta}(t) - \theta(t)|)$ is going to zero and next, to use the Markov inequality:

$$\mathbb{P}\left(\sup_{t\in[0,\mathcal{T}]} |\hat{\theta}(t) - \theta(t)| > \varepsilon\right) \leq \frac{\mathbb{E}_p(\sup_{t\in[0,\mathcal{T}]} |\hat{\theta}(t) - \theta(t)|)}{\varepsilon}.$$

In order to prove the uniform consistency, we need supplementary assumptions.

Assumptions on the regularity of trajectories

Assumption V4. There are two positive constants C_2 and C_3 and $1 \geq \beta > 1/2$ such that, for all N and for all $(r,t) \in [0,\mathcal{T}] \times [0,\mathcal{T}]$,

$$\frac{1}{N}\sum_{k\in U} Y_k(0)^2 < C_2 \quad \text{and} \quad \frac{1}{N}\sum_{k\in U}\{Y_k(t) - Y_k(r)\}^2 < C_3|t-r|^{2\beta}.$$

Assumption (V4) deal with the regularity of the trajectories and was already required in Cardot and Josserand [2011]. It can be remarked that the first part of assumption (V4) is similar to assumption (V1) from the non-functional case. In a functional setting, Hölder conditions are required to control the oscillations of the processus. Even if pointwise consistency, for each fixed value of t, can be proved without any condition on the Hölder coefficient β, this regularity condition is necessary to get a uniform convergence result. A counterexample is given in Hahn [1977] when $\beta \leq 1/2$. More precisely it is shown that the sample mean i.i.d copies of a uniformly bounded continuous random function defined on a compact interval may not satisfy the Central Limit Theorem in the space of continuous functions. The hypothesis $\beta > 1/2$ also implies that the trajectories of the residual processes ϵ_{kt}, see (3.35), are regular enough (but not necessarily differentiable).

If assumptions (S1), (S2) and (V4) hold and if the discretization scheme satisfies

$$\max_{i=\{1,..,D_N-1\}} |t_{i+1} - t_i|^{2\beta} = o(n^{-1}), \tag{3.20}$$

then, it is proven in Cardot and Josserand [2011] that the estimator of the mean curve $\hat{\mu}_d$ satisfies:

$$\mathbb{E}_p\left\{\sup_{t\in[0,\mathcal{T}]} |\hat{\mu}^{(\mathrm{d})}(t) - \mu_N(t)|\right\} = O(n^{-1/2}),$$

namely, it is asymptotically design-unbiased and uniformly consistent. Note that condition (3.20) ensures that the interpolation error is negligible compared to the sampling error. The proof uses maximal inequalities that can be found in van der Vaart and Wellner [2000] (see the proof of proposition 3.8).

In order to prove the uniform consistency of the variance function estimator, the assumption (S4) on the fourth-order inclusion probabilities and the following additional assumptions on the regularity of the trajectories are needed:

Assumption V5. There are two positive constants C_4 and C_5 and $1 \geq \beta > 1/2$ such that, for all N and for all $(r,t) \in [0,\mathcal{T}] \times [0,\mathcal{T}]$,

$$\frac{1}{N}\sum_{k\in U} Y_k(0)^4 < C_4 \quad \text{and} \quad \frac{1}{N}\sum_{k\in U}\{Y_k(t) - Y_k(r)\}^4 < C_5|t-r|^{4\beta}.$$

Cardot and Josserand [2011] have shown that the variance function estimator $\hat{\gamma}^{(\mathrm{d})}$ given by (3.13) is uniformly consistent:

$$\mathbb{E}_p\left(\sup_{t\in[0,\mathcal{T}]}\frac{1}{N^2}|\hat{\gamma}^{(\mathrm{d})}(t,t) - \gamma_p(t,t)|\right) = o(n^{-1}).$$

3.1.5 Asymptotic normality and confidence bands for the mean curve

For fixed $t \in [0,\mathcal{T}]$, establishing the asymptotic normality of the pointwise estimator $\sqrt{n}(\hat{\mu}^{(\mathrm{d})}(t) - \mu_N(t))$ falls down in the theory described in Section 1.4. Then, it is possible to construct asymptotic pointwise confidence intervals for $\mu_N(t)$:

$$\lim_{N\to\infty}\mathbb{P}\left(\mu_N(t) \in \left[\widehat{\mu}^{(\mathrm{d})}(t) \pm q_\alpha\,\frac{\widehat{\sigma}(t)}{\sqrt{n}}\right]\right) = 1-\alpha,$$

where $\alpha \in (0,1)$ and q_α is the quantile of order $1-\alpha/2$ of the standard normal distribution $\mathcal{N}(0,1)$.

The interest is to go further and to obtain the asymptotic distribution of $\sqrt{n}(\hat{\mu}^{(\mathrm{d})} - \mu_N)$ as an element of $C[0,\mathcal{T}]$ in order to be able to build asymptotic confidence bands. Recall first the definition of the convergence in distribution in $C[0,\mathcal{T}]$:

Definition 3.2. Let $(X_n)_n, X$ be random functions with values in $(C[0,\mathcal{T}],\rho)$. Then $(X_n)_n$ converges in distribution to X in $C[0,\mathcal{T}]$:

$$X_n \to_{\mathcal{D}} X \quad \text{in} \quad C[0,\mathcal{T}],$$

if for each $f : C[0,\mathcal{T}] \to \mathbb{R}$ continuous and bounded, we have:

$$\mathbb{E}(f(X_n)) \to \mathbb{E}(f(X)).$$

The terminology *weak convergence* is also employed (van der Vaart [1998], Billingsley [1968]). The following notion of tightness, by which it is disallowed any escape of mass, proves important both in the theory of convergence in distribution and in its applications.

Definition 3.3. A random function X from $C[0,\mathcal{T}]$ is called tight if for every $\varepsilon > 0$, there is a compact set $\mathbb{K} \subset C[0,\mathcal{T}]$ such that $P(X \in \mathbb{K}) > 1-\varepsilon$.

A sequence $(X_n)_n$ of real variables is tight if $(X_n)_n$ is bounded in probability, namely $X_n = O_p(1)$. For more general spaces, such as $C[0,\mathcal{T}]$, this property is more difficult to obtain since

it involves the characterization of compact sets in such spaces. Billingsley [1968] (theorem 8.2) uses the Arzelà-Ascoli's theorem to give a characterization of compactness in $C[0,\mathcal{T}]$. Other criteria for tightness (consequences of theorem 8.2) are given in Billingsley [1968]; in particular, theorem 12.3 gives a useful and simple criterion:

Theorem 3.4. *(Billingsley [1968]) The sequence $(X_n)_n$ is tight in $C[0,\mathcal{T}]$ if it satisfies these two conditions:*

1. *$(X_n(0))_n$ is tight and*
2. *There exist two constants $\gamma \geq 0$ and $\alpha > 1$ and a nondecreasing, continuous function F on $[0,\mathcal{T}]$ such that*

$$\mathbb{E}|X_n(t) - X_n(r)|^\gamma \leq |F(t) - F(r)|^\alpha, \quad \textit{for all} \quad t, r \in [0,\mathcal{T}].$$

As Billingsley [1968] states, the second condition from the above theorem stipulates "that the random functions X_n do not oscillate too violently".
Now, since the space $C[0,\mathcal{T}]$ is separable and complete, the convergence in distribution $C[0,\mathcal{T}]$ is equivalent to tightness and finite-dimensional convergence in distribution as obtained by Prohorov (see Billingsley [1968]):

Theorem 3.5. *A sequence of random functions $(X_n)_n$ from $C[0,\mathcal{T}]$ converges in distribution to a random function X, $X_n \to_{\mathcal{D}} X$, if and only if:*

1. *any finite-dimensional vector $(X_n(t_1),\ldots,X_n(t_r))_n$ converge in distribution to $(X(t_1),\ldots,X(t_r))$*
2. *$(X_n)_n$ is tight.*

In our setting, the sequence of random functions is given by $X_n = \sqrt{n}(\hat{\mu}^{(\mathrm{d})} - \mu_N)$. By assumption (V4), the trajectories Y_k are continuous as well as $Y_k^{(\mathrm{d})}$ by construction, implying that X_n belongs to $C[0,\mathcal{T}]$. By assuming the pointwise asymptotic normality of $\hat{\mu}$ and $N^{-1}\sum_{k\in U} Y_k^{2+\delta}(t) < \infty$ for all $t \in [0,\mathcal{T}]$ and some $\delta > 0$, Cardot and Josserand [2011] showed that $X_n = \sqrt{n}(\hat{\mu} - \mu_N)$ satisfies conditions 1 and 2 from theorem 3.5. It follows that

$$\sqrt{n}(\hat{\mu} - \mu_N) \to_{\mathcal{D}} Z \quad \text{in} \quad C[0,\mathcal{T}]$$

where Z is a Gaussian random function taking values in $C[0,\mathcal{T}]$ with mean 0 and covariance function $\tilde{\gamma}_p(r,t) = \lim_{N\to\infty} \frac{n}{N^2}\gamma_p(r,t)$. Next, if the discretization points are numerous enough (see condition (3.20)), it can be shown that $\sqrt{n}(\hat{\mu}^{(\mathrm{d})} - \mu_N) = \sqrt{n}(\hat{\mu} - \mu_N) + o(1)$ uniformly in t and as a consequence,

$$\sqrt{n}(\hat{\mu}^{(\mathrm{d})} - \mu_N) \to_{\mathcal{D}} Z \quad \text{in} \quad C[0,\mathcal{T}].$$

In a functional setting, we aim at building simultaneous confidence bands for μ_N of the form

$$\mathbb{P}\left(\mu_N(t) \in \left[\widehat{\mu}^{(\mathrm{d})}(t) \pm c_\alpha \frac{\widehat{\sigma}(t)}{\sqrt{n}}\right], \ \forall t \in [0,\mathcal{T}]\right) = 1 - \alpha, \tag{3.21}$$

where the coefficient c_α is unknown and depends on the desired level of confidence $1-\alpha$, and $\widehat{\sigma}(t) = \sqrt{\frac{n}{N^2}\hat{\gamma}^{(\mathrm{d})}(t,t)}$.

The calculation of c_α is based on the asymptotic distribution of $\hat{\mu}^{(\mathrm{d})}$ in $C[0,\mathcal{T}]$. Thus, for n large enough, we have that

$$\mathbb{P}\left(\mu_N(t) \in \left[\widehat{\mu}^{(\mathrm{d})}(t) \pm c_\alpha\, \widehat{\sigma}(t)\right],\ \forall t \in [0,\mathcal{T}]\right) \simeq \mathbb{P}\left(\sup_{t\in[0,\mathcal{T}]} \frac{|\widehat{Z}(t)|}{\widehat{\sigma}(t)} \leq c_\alpha\right)$$
$$\simeq \mathbb{P}\left(\sup_{t\in[0,\mathcal{T}]} \frac{|Z(t)|}{\sigma(t)} \leq c_\alpha\right)$$

where $\widehat{Z}$ is a zero mean Gaussian random function with covariance $\frac{n}{N^2}\hat{\gamma}^{(\mathrm{d})}$. The cut-off point c_α is the quantile of order $1-\alpha$ of $\sup_{t\in[0,\mathcal{T}]} |\widehat{Z}(t)|/\widehat{\sigma}(t)$ which can not be computed exactly since the distribution of the supremum of Gaussian processes is known only for few particular cases.

In a recent work, Cardot et al. [2013c] compared two methods for estimating the unknown cut-off point c_α. The first method relies on simulation of Gaussian processes and has been used in Degras [2011] in a non-sampling setting. This Monte Carlo method consists in simulating a Gaussian process $\widehat{Z}$ with zero mean and covariance function equal to $\hat{\gamma}^{(\mathrm{d})}$ in order to determine the distribution of its supremum and then estimate c_α. A rigorous mathematical justification of this technique has been given in Cardot et al. [2013b], Cardot et al. [2013d] and Cardot et al. [2014b].

The second method avoids the estimation of the variance of the mean estimator by using bootstrap techniques adapted to the functional case. The variance function $\gamma_p(r,t)$ and the value c_α are estimated from the bootstrap replications. Cardot et al. [2013c] used the bootstrap suggested by Gross [1980] for simple random sampling and its extensions to other sampling designs suggested by Chauvet [2007].

Using a slightly different population of load curves, Cardot et al. [2013c] compared these two methods for computing c_α. They conclude that the two methods give similar coverage rates which are very close to the desired nominal rates. Nevertheless, the bootstrap method is excessively time-consuming.

3.1.6 Some consistency results for the non-linear parameter estimators

The convergence has essentially been proven in the Hilbert space $L^2[0,\mathcal{T}]$ by extending the functional approach of Deville [1999b] to this space (Cardot et al. [2010a], Chaouch and Goga [2012]). The functionals defined above (equations 3.14, 3.15 and 3.16) are all 0-homogeneous (see assumption F1). The eigenvalues λ_j and eigenvectors v_j as well as the median curve may

be obtained as solutions of population estimating equations as described in Section 2.2.

Since all functionals considered in this section are 0-homogeneous and using theorem 2.8, a first-order von Mises expansion of the functional T may be given as follows:

$$T(\hat{M}) = T(M) + \sum_{k\in s} \frac{u_k}{\pi_k} - \sum_{k\in U} u_k + \text{Rem}_T\left(\frac{M}{N}, \frac{M}{N}\right), \tag{3.22}$$

where Rem_T is the reminder associated to the functional T and u_k is the linearized variable of T. Under the assumptions of theorem 2.8, the reminder is of order $\text{Rem}_T = o_p(n^{-1/2})$. As the eigenvalues $\hat{\lambda}_j$ and eigenvectors $\hat{\mathbf{v}}_j$ are defined as solutions of the implicit estimating equation 3.18, the implicit function theorem may be used to obtain the result. This method was used by Chaouch and Goga [2012] for the median curve. Cardot et al. [2010a] used the perturbation theory to obtain that, if assumptions (S1), (S2) are fulfilled and if $\sup_{k\in U} ||Y_k|| < \infty$, $\text{Rem}_\Gamma = o_p(n^{-1/2})$ and

$$\mathbb{E}_p||\hat{\mathbf{\Gamma}} - \mathbf{\Gamma}||^2_{HS} = O(n^{-1}),$$

where $||\cdot||_{HS}$ is the Hilbert-Schmidt [4] norm for operators. If all the nonnull eigenvalues $\lambda_j, j = 1, \ldots, N$ are distinct, then Rem_{λ_j} and Rem_{v_j} are also of order $o_p(n^{-1/2})$ and

$$\mathbb{E}_p(\sup_j |\hat{\lambda}_j - \lambda_j|)^2 = O(n^{-1}) \tag{3.23}$$

and for each fixed j,

$$\mathbb{E}_p||\hat{v}_j - v_j||^2 = O(n^{-1}). \tag{3.24}$$

Remark that the above results are still valid if the strong assumption that $\sup_{k\in U} ||Y_k|| < \infty$ is replaced by moment conditions on Y (see for example, the assumption (V4)).

Consider now the linearized variables u_k for the different parameters of interest considered in this section. If $\sup_{k\in U} ||Y_k|| < \infty$, Cardot et al. [2010a] prove that the influence function of $\mathbf{\Gamma}$ exists and that the linearized variable is given by:

$$u_{k,\mathbf{\Gamma}} = \frac{1}{N}((Y_k - \mu) \otimes (Y_k - \mu) - \mathbf{\Gamma}), \quad k \in U.$$

If moreover, the nonnull eigenvalues of $\mathbf{\Gamma}$ are distinct, then the linearized variables of λ_j and v_j are:

$$u_{k,\lambda_j} = \frac{1}{N}(< Y_k - \mu, v_j >^2 - \lambda_j), \quad j = 1, \ldots N$$

[4] It is induced by the inner product between two operators Γ and Δ defined by $< \Gamma, \Delta >= \sum_{j=1}^{\infty} < \Gamma e_j, \Delta e_j >$ for any orthonormal basis $(e_j)_{j\leq 1}$ of $L^2[0, \mathcal{T}]$.

$$u_{k,v_j} = \frac{1}{N}\left(\sum_{l\neq j}\frac{< Y_k - \mu_N, v_j >< Y_l - \mu_N, v_l >}{\lambda_j - \lambda_l}v_l\right), \quad j = 1, \dots N$$

for all $k \in U$. Finally, if $N^{-1}\sum_{k\in U}\|Y_k - m_N\|^{-1} < \infty$ and $m_N \neq Y_k$ for all $k \in U$, then the linearized variable of the median curve is given by (see Chaouch and Goga [2012]):

$$u_{k,m_N} = \Delta^{-1}\left(\frac{Y_k - m_N}{\|Y_k - m_N\|}\right), \quad k \in U \tag{3.25}$$

where $\Delta = \sum_{k\in U}\frac{1}{||Y_k - m_N||}\left[\mathbf{I} - \frac{(Y_k - m_N)\otimes(Y_k - m_N)}{||Y_k - m_N||^2}\right]$ is the Jacobian operator of the functional T_{m_N} in equation (3.16) and where $\mathbf{I}$ is the identity operator defined by $\mathbf{I}y = y$.

One can remark that the linearized variables u_k are not even known for the sampled individuals, so we need to estimate them. Moreover, except for the case of the eigenvalues λ_j, u_k is a curve depending on $t \in [0, \mathcal{T}]$.

The expansion given in (3.22) is important since it allows the approximation of the substitution estimator $T(\hat{M})$ by the HT estimator $\sum_{k\in U} u_k$, provided that the reminder term is negligible, *i.e.* $\text{Rem}_T = o_p(n^{-1/2})$. Therefore, the asymptotic variance function of the substitution estimator $T(\hat{M})$ is the HT variance:

$$A\mathbb{V}_p(T(\hat{M}))(t) = \sum_{k\in U}\sum_{l\in U}(\pi_{kl} - \pi_k\pi_l)\frac{u_k(t)}{\pi_k}\frac{u_l(t)}{\pi_l}$$

and it can be estimated by

$$\hat{V}_p(T(\hat{M}))(t) = \sum_{k\in s}\sum_{l\in s}\frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}}\frac{\hat{u}_k(t)}{\pi_k}\frac{\hat{u}_l(t)}{\pi_l},$$

where $\hat{u}_k(t)$ is the estimator of $u_k(t)$. In order to prove that the variance function estimator is consistent in the sense that

$$n\{\hat{V}_p(T(\hat{M}))(t) - AV_p(T(\hat{M}))(t)\} = o_p(1),$$

an additional assumption on the fourth inclusion probabilities (assumption (S4)) is needed. We also suppose that the linearized variable estimator $\hat{u}_k$ is uniformly bounded and uniformly consistent for the true u_k. In the case of functional principal component analysis, we have that $||\hat{u}_{k,\mu_N}|| = O(n^{-1})$, $\hat{u}_{k,\lambda_j} = O(n^{-1})$, $||\hat{u}_{k,v_j}|| = O_p(n^{-1})$ and

$$\mathbb{E}_p||\hat{u}_{k,\mu_N} - u_{k,\mu_N}||^2 = O(N^{-3}),$$

$$\mathbb{E}_p(\hat{u}_{k,\lambda_j} - u_{k,\lambda_j})^2 = O(N^{-3}),$$

$$||\hat{u}_{k,v_j} - u_{k,v_j}||^2 = O(N^{-3}),$$

uniformly in k (see Cardot et al. [2010a]). To prove the consistency of $\hat{V}_p(\hat{v}_j)$, the covariance operator $\mathbf{\Gamma}$ is supposed to be of finite rank not depending on N. It is worth mentioning that the estimators of the linearized variables satisfy the general assumptions given in Goga et al. [2009] and given in theorem (2.9).
In the case of the median curve, the behavior of the variance estimator function was studied by Chaouch and Goga [2012] by means of simulations only.

3.1.7 Using auxiliary information at the sampling stage: stratified and πps sampling designs

If auxiliary information is used at the sampling stage, some changes are needed because the variables involved now are curves, otherwise the selection of the sample is realized from the sampling frame list as for classical multivariate surveys. For example, a simple random sampling without replacement (SRSWOR) consists of taking n elements from the list of N elements of the population and of recording the curve Y_k for each sampled individual k. The HT estimator of the mean curve μ_N is $\hat{\mu} = \frac{1}{n}\sum_{k\in s} Y_k$ with the covariance function given by:

$$\gamma_{SRSWOR}(r,t) = \left(\frac{1}{n} - \frac{1}{N}\right) S_{Y(r)Y(t),U}$$

where $S_{Y(r)Y(t),U} = \frac{1}{N-1}\sum_U (Y_k(r)-\mu_N(r))(Y_k(t)-\mu_N(t))$ is the population covariance function between $Y_k(r)$ and $Y_k(t)$. The estimator of the median is obtained from equation (3.19) for $\pi_k = n/N$ for all $k \in U$.

Stratified sampling with simple random sampling within strata (STRAT)

Suppose that the population is divided into H strata $U_1, ..., U_H$ of sizes $N_1, ..., N_H$ and a sample s_h of size n_h is drawn by a simple random sampling without replacement within the stratum U_h, $h = 1, ..., H$.

The mean curve estimator with stratified sampling is given by:

$$\widehat{\mu}_{\text{strat}}(t) = \sum_{h=1}^{H} \frac{N_h}{N}\left(\frac{1}{n_h}\sum_{k\in s_h} Y_k(t)\right), \ t \in [0,T], \tag{3.26}$$

with the covariance function given by:

$$\gamma_{\text{strat}}(r,t) = \frac{1}{N^2}\sum_{h=1}^{H} N_h^2\left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_{Y(r)Y(t),U_h} \ \ r,t \in [0,\mathcal{T}], \tag{3.27}$$

where $S_{Y(r)Y(t),U_h}$ is the population covariance function between $(Y_k(r))_{k\in U}$ and $(Y_k(t))_{k\in U}$ within each stratum U_h.

Stratified sampling can be also used to estimate any non-linear parameter of interest such as the eigenvalues λ_j and the eigenfunctions v_j for $j = 1, \ldots, N$ (see Cardot et al. [2010a]), or the median curve (see Chaouch and Goga [2012]). For example, to obtain the estimator $\hat{m}_{\text{strat}}$ of the median curve with a stratified sampling, one can use the sampling weights $\pi_k = n_h/N_h$ for all $k \in U_h, h = 1, \ldots H$ in equation (3.19) and solve the following estimation equation:

$$\sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{k \in s_h} \frac{Y_k - \hat{m}_{\text{strat}}}{\|Y_k - \hat{m}_{\text{strat}}\|} = 0. \tag{3.28}$$

The asymptotic variance function of $\hat{m}_{\text{strat}}$ is

$$AV_{\text{strat}}(\hat{m}_{\text{strat}})(t) = \sum_{h=1}^{H} N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right) S^2_{u_{m_N}(t), U_h}, \tag{3.29}$$

where $S^2_{u_{m_N}(t), U_h}$ is the population variance function of $u_{m_N}(t) = (u_{k,m_N}(t))_{k \in U_h}$ within stratum h and u_{k,m_N} is the linearized variable of the median curve given in equation (3.25). That is, the lower the variation of the linearized variable within stratum, the lower the asymptotic variance of $\hat{m}_{\text{strat}}$. In the latter situation, stratified sampling is efficient for estimating the median curve but may be poor for the estimation of other parameters. In such a situation, poststratification may be used (see Chaouch and Goga [2012]).

To choose the size n_h of the sample s_h, it is possible to use the proportional allocation $n_h = nN_h/N$, $h = 1, ..., H$ or the optimal allocation as suggested by Cardot and Josserand [2011]:

$$n_h = n \frac{N_h \sqrt{\int_0^{\mathcal{T}} S^2_{Y(r)Y(r), U_h} dr}}{\sum_{h=1}^{H} N_h \sqrt{\int_0^{\mathcal{T}} S^2_{Y(r)Y(r), U_h} dr}}, \quad h = 1, \ldots, H. \tag{3.30}$$

This allocation minimizes the mean variance of the stratified estimator:

$$\min_{(n_1, \ldots, n_H)} \int_0^{\mathcal{T}} \gamma_{\text{strat}}(t,t) dt \quad \text{subject to} \quad \sum_{h=1}^{H} n_h = n \quad \text{with} \quad n_h > 0, \quad \text{for } h = 1, \ldots, H.$$

This allocation is similar to that of the multivariate case when considering a total variance criterion (Cochran [1977]) and has the same interpretation, namely strata with higher variability should be sampled with a higher sampling rate than the other strata. In practice, $S^2_{Y(r)Y(r), U_h}$ are unknown for all $h = 1, ..., H$. An auxiliary variable $\mathcal{X}$ known for all individuals $k \in U$ and highly correlated with the interest variable can be used instead and the so-obtained allocation is called the x-optimal allocation.

Using the allocation given by (3.30) may not be optimal for estimating non-linear parameters of interest. In order to derive the optimal allocation for estimating the median, for example, one should minimize the asymptotic variance of $\hat{m}_{\text{strat}}(t)$. The so obtained allocation depends in this case on the linearized variable (see Chaouch and Goga [2012]).

Probability proportional-to-size sampling: πps sampling

Unequal probability designs are used in practice because they are usually more efficient than the equal probability designs. To estimate the mean curve, Cardot et al. [2013c] and Cardot et al. [2014b] consider the fixed-size without replacement designs and to estimate the median curve, Chaouch and Goga [2012] consider with replacement probability proportional-to-size designs. We give below a description of results obtained in the first case.

For a sampling design of fixed size n, it is possible to give the equivalent of the Yates and Grundy (Yates and Grundy [1953]) and Sen formula (Sen [1953]) in the functional case. The covariance $\gamma_p(r,t)$ of $\hat{t}_{Y\pi}$ between two instants r and t, verifies

$$\gamma_p(r,t) = -\frac{1}{2}\sum_{k\in U}\sum_{l\in U, l\neq k}(\pi_{kl}-\pi_k\pi_l)\left(d_kY_k(r)-d_lY_l(r)\right)\left(d_kY_k(t)-d_lY_l(t)\right). \tag{3.31}$$

Using equation (3.31), we clearly see that the covariance $\gamma_p(r,t)$ will be small if the first-order inclusion probabilities π_k are approximately proportional to $Y_k(t)$, for all instants $t \in [0,\mathcal{T}]$. Again, non-linear parameters may be estimated by using πps sampling designs. For example, the HT estimator for the median with πps design is obtained by using in (3.19) the π_k given by equation (1.6). Nevertheless, this design performs very poorly for the estimation of the median curve as can be seen in the application study presented in Section 3.1.9. This can be explained by the fact that the relationship between the linearized variable u_{k,m_N} and π_k is not linear as it can be remarked from (3.25) while Y_k is approximately proportional to π_k, which explains the the good performance of the πps sampling for estimating the mean curve. In order to improve the estimation of the median with a πps design, Goga [2014] suggests an estimator of m_N which consists in modifying the sampling weights $1/\pi_k$ by using a superpopulation model explaining the relationship between the u_k and π_k as follows:

$$u_{k,m_N}(t) = g(\pi_k,t)+\eta_{kt}, \quad k\in U$$

where g is unknown and the errors η_{kt} are centered. The function g can be estimated by using the B-spline regression as proposed by Goga and Ruiz-Gazen [2014a]. This leads to consider the following smoothed weights:

$$w_{ks}^{\mathrm{np}} = \frac{1}{\pi_k}\left(\sum_{l\in U}\mathbf{b}^T(\pi_l)\right)\left(\sum_{l\in s}\frac{\mathbf{b}(\pi_l)\mathbf{b}^T(\pi_l)}{\pi_l}\right)^{-1}\mathbf{b}(\pi_k), \quad k\in s$$

where $\mathbf{b} = (B_1,\ldots,B_q)^T$ is the vector of the B-spline basis of degree m and with K interior knots, $q = K+m$. The improved estimator of the median is obtained from (3.19) by replacing $1/\pi_k$ with the weights w_{ks}. In a model-based setting, Zheng and Little [2003] and Zheng and Little [2005] used a similar idea and penalized spline in order to estimate finite population totals with πps sampling designs. Work is actually in progress in order to obtain the asymptotic properties of this improved estimator of m_N.

Variance estimation and confidence bands with πps sampling

The covariance function γ_p given by (3.31) involves the second-order inclusion probabilities π_{kl} which are very difficult or even impossible to calculate for many πps designs. Recently, a functional Hájek approximation for the covariance function γ_p was suggested in Cardot et al. [2013c]. More exactly, suppose that the second-order inclusion probabilities satisfy the assumption (S3), namely:

$$\pi_{kl} = \pi_k \pi_l \left\{ 1 - \frac{(1-\pi_k)(1-\pi_l)}{D(\pi)} [1 + o(1)] \right\}$$

where $D(\pi) = \sum_{k \in U} \pi_k (1 - \pi_k)$ is supposed to tend to infinity when the population size N is growing to infinity. Then, we can approximate γ_p by the following covariance function γ_H which contains only the first-order inclusion probabilities:

$$\begin{aligned} \gamma_H(r,t) &= \sum_{k \in U} \pi_k (1-\pi_k) \left(\frac{Y_k(t)}{\pi_k} - R(t) \right) \left(\frac{Y_k(r)}{\pi_k} - R(r) \right) \\ &= \sum_{k \in U} \frac{1-\pi_k}{\pi_k} Y_k(t) Y_k(r) - \frac{1}{D(\pi)} \left(\sum_{k \in U} (1-\pi_k) Y_k(t) \right) \left(\sum_{l \in U} (1-\pi_l) Y_l(r) \right), \quad r,t \in [0,\mathcal{T}], \end{aligned} \tag{3.32}$$

where $R(t) = \dfrac{\sum_{k \in U} Y_k(t)(1-\pi_k)}{D(\pi)}$. This approximation appears to be very efficient when the sample size is large enough and the entropy of the sampling design is closed to the maximum entropy, in particular for the rejective sampling and the Sampford-Durbin sampling (see Cardot et al. [2014b]).

Using a slightly different population of load curves, the following estimator of the covariance function has been successfully used by Cardot et al. [2013c] to build confidence bands for the mean curve estimator:

$$\begin{aligned} \hat{\gamma}_H^{*(\mathrm{d})}(r,t) &= \sum_{k \in s} (1-\pi_k) \left(\frac{Y_k^{(\mathrm{d})}(t)}{\pi_k} - \hat{R}(t) \right) \left(\frac{Y_k^{(\mathrm{d})}(r)}{\pi_k} - \hat{R}(r) \right) \\ &= \sum_{k \in s} \frac{1-\pi_k}{\pi_k^2} Y_k^{(\mathrm{d})}(t) Y_k^{(\mathrm{d})}(r) - \frac{1}{\hat{D}(\pi)} \left(\sum_{k \in s} \frac{1-\pi_k}{\pi_k} Y_k^{(\mathrm{d})}(t) \right) \left(\sum_{l \in s} \frac{1-\pi_l}{\pi_l} Y_l^{(\mathrm{d})}(r) \right), \quad r,t \in [0,\mathcal{T}], \end{aligned} \tag{3.33}$$

where $\hat{R}(t) = \displaystyle\sum_{k \in s} \frac{Y_k^{(\mathrm{d})}(t)(1-\pi_k)}{\pi_k} / \hat{D}(\pi)$ and $\hat{D}(\pi) = \sum_{k \in s} (1-\pi_k)$. The simulation study has shown that the confidence bands had the desired coverage rates and their widths were greatly reduced compared to the ones obtained with simple random sampling without replacement. The estimator (3.33) is the functional version of the variance estimator suggested by Deville and Tillé [2005].

Note that it is also possible to consider the following covariance estimator:

$$\hat{\gamma}_H^{(\mathrm{d})} = \frac{\hat{D}(\pi)}{D(\pi)}\hat{\gamma}_H^{*(\mathrm{d})}, \tag{3.34}$$

which is a slightly modified functional analogue of the variance estimator proposed by Berger [1998a] in the real case. Assuming assumptions (S1) and (S2), it can be easily proven that $\lim_{N\to\infty} \frac{\hat{D}(\pi)}{D(\pi)} = 1$. This results implies that the covariance estimators $\hat{\gamma}_H^{*(\mathrm{d})}$ and $\hat{\gamma}_H^{(\mathrm{d})}$ have the same asymptotic behavior. However, it is a bit easier to prove the consistency of $\hat{\gamma}_{H,d}$.

Proposition 3.6. *(Cardot et al. [2014b]) Under assumptions (S1), (S3), (S4) and (V4), (V5) and if the discretization scheme satisfies*

$$\lim_{N\to\infty} \max_{i\in\{1,\ldots,D_N-1\}} |t_{i+1} - t_i| = 0,$$

then for all $r, t \in [0, \mathcal{T}]$,

$$\lim_{N\to\infty} n\mathbb{E}_p(|\hat{\gamma}_H^{(\mathrm{d})}(r,t) - \gamma_p(r,t)|) = 0,$$

and

$$\lim_{N\to\infty} n\mathbb{E}_p\left(\sup_{t\in[0,\mathcal{T}]} \frac{1}{N^2}|\hat{\gamma}_H^{(\mathrm{d})}(t,t) - \gamma_p(t,t)|\right) = 0.$$

The proof follows the same steps as in Cardot et al. [2013d]. It is shown first by using the same arguments as in Breidt and Opsomer [2000], that for all $t, r \in [0, \mathcal{T}]$, the estimator of the covariance function $\widehat{\gamma}_H^{(\mathrm{d})}(r,t)$ is pointwise convergent for $\gamma_p(r,t)$. In order to obtain the uniform consistency of the variance function estimator $\widehat{\gamma}_H^{(\mathrm{d})}$, it is shown that

$$\frac{n}{N^2}(\widehat{\gamma}_H^{(\mathrm{d})} - \gamma_p) \to_{\mathcal{D}} 0 \quad \text{in} \quad C([0,\mathcal{T}]).$$

Then, from the definition 3.2 of the convergence in distribution in $C([0,\mathcal{T}])$ and the boundedness and continuity of the sup functional, we obtain directly

$$\mathbb{E}_p\left(\sup_{t\in[0,\mathcal{T}]} \frac{n}{N^2}|\hat{\gamma}_H^{(\mathrm{d})}(t,t) - \gamma_p(t,t)|\right) \to 0.$$

In particular, they note that the errors due to the Hájek approximation is negligible. Further, as in Cardot et al. [2013d], in order to obtain the convergence in distribution of $n(\widehat{\gamma}_H^{(\mathrm{d})}(t,t) - \gamma_p(t,t))/N^2$, theorem 3.5 is used: the convergence of all finite linear combinations is shown first, which is easily obtained from the pointwise convergence, and next, it is checked that the sequence $n(\widehat{\gamma}_H^{(\mathrm{d})}(t,t) - \gamma_p(t,t))/N^2$ is tight by checking the two conditions given in theorem 3.4.

Remark 7. *It is worth mentioning that the properties of boundedness and continuity of the sup functional in the space* $C[0,\mathcal{T}]$ *are crucial for obtaining the asymptotic distribution. These properties are not true anymore in* $L^2[0,\mathcal{T}]$.

By using the approximations of the multiple inclusion probabilities given by Boistard et al. [2012], a sharper result can be obtained for the rejective sampling:

Proposition 3.7. *(Cardot et al. [2014b]) Under assumptions (S1), (S3) and (V4), (V5) and if the discretization schema satisfies*

$$\max_{i\in\{1,\ldots,D_N-1\}} |t_{i+1}-t_i|^{2\beta} = O(n^{-1}),$$

then for all $r,t \in [0,\mathcal{T}]$,

$$\mathbb{E}_p\left(\frac{1}{N^2}(\hat{\gamma}_H^{(\mathrm{d})}(r,t)-\gamma_p(r,t))\right)^2 = O(n^{-3}).$$

The accuracy of the proposed variance estimators has been evaluated by Cardot et al. [2014b] on the population of load curves considered before. They notice that even if this estimator generally provides good estimations of the true covariance function, for a few "bad" samples, its performances could be very poor. These bad performances, which fortunately occur in very rare occasions, are in fact due to a few individuals in the population that have both a very small inclusion probability π_k and a high consumption level Y_k. Further work is needed in order to build modified variance estimators that are more robust to the presence of influential individuals. More work is also needed to evaluate the performances of this variance estimator in the case of non-linear parameters.

3.1.8 Functional model-assisted estimator

In a recent work, Cardot et al. [2013d] suggested to improve the accuracy of the HT estimator $\widehat{\mu}$ of the mean curve $\mu_N(t)$ by a model-assisted estimator based on a functional linear model (see Faraway [1997]). This estimator can be seen as a direct extension, to the functional context, of the generalized regression estimator or GREG estimator studied in Robinson and Särndal [1983] and Särndal et al. [1992]. Its main advantage is that it only requires the knowledge of the total of the auxiliary variable at the population level.

As before, let $\mathcal{X}_1,...,\mathcal{X}_p$ be p real auxiliary variables and let also $\mathbf{x}_k = (x_{k1},...,x_{kp})^T$ be the value of the vector of auxiliary variables for the k-th individual from the population. The following superpopulation model ξ_t, also called functional linear model (see Faraway [1997]) is introduced:

$$\xi_t: \quad Y_k(t) = \mathbf{x}_k^T\boldsymbol{\beta}(t)+\epsilon_{kt}, \quad t\in[0,\mathcal{T}] \tag{3.35}$$

where $\boldsymbol{\beta}(t) = (\beta_1(t),\ldots,\beta_p(t))^T$ is the vector of functional regression coefficients, ϵ_{kt} are independent (across units) and centered continuous time processes, $\mathbb{E}_\xi(\epsilon_{kt})=0$, with covariance function $\mathrm{Cov}_\xi(\epsilon_{kt},\epsilon_{kr}) = \tilde{\Gamma}(t,r)$, for $(t,r)\in[0,\mathcal{T}]\times[0,\mathcal{T}]$.

The functional model-assisted or GREG estimator is given by (see Cardot et al. [2013d]):

$$\widehat{\mu}_{MA}(t) = \frac{1}{N}\sum_{k\in s} d_k Y_k(t) - \frac{1}{N}\left(\sum_{k\in s} d_k \mathbf{x}_k - \sum_{k\in U}\mathbf{x}_k\right)^T \widehat{\boldsymbol{\beta}}_{\mathbf{x}}(t), \quad t\in[0,\mathcal{T}], \tag{3.36}$$

where $\widehat{\boldsymbol{\beta}}_{\mathbf{x}}(t) = \widehat{\mathbf{G}}^{-1}\frac{1}{N}\sum_{k\in s} d_k \mathbf{x}_k Y_k(t)$ with $\widehat{\mathbf{G}} = \frac{1}{N}\sum_{k\in s} d_k \mathbf{x}_k \mathbf{x}_k^T$.

Remark 8. *We can note that the estimator $\widehat{\mu}_{MA}$ may be written as a weighted sum of Y_k, for $k\in s$ with weights not depending on the index t.*

When the matrix $\widehat{\mathbf{G}}$ is not well conditioned, Cardot et al. [2013d] proposed to replace $\widehat{\mathbf{G}}$ with the following regularized estimator:

$$\widehat{\mathbf{G}}_a = \sum_{j=1}^{p} \max(\lambda_{j,n}, a)\ \mathbf{v}_{jn}\mathbf{v}_{jn}^T$$

where $a>0$, $\lambda_{j,n}$ is the jth eigenvalue, $\lambda_{1,n}\geq\ldots\geq\lambda_{p,n}\geq 0$, and $\mathbf{v}_{jn}$ is the corresponding orthonormal eigenvector of the matrix $\widehat{\mathbf{G}}$. A similar idea was used by Bosq [2000] and Guillas [2001].
It is clear that $\widehat{\mathbf{G}}_a$ is always invertible and

$$\|\widehat{\mathbf{G}}_a^{-1}\|_2 \leq a^{-1}, \tag{3.37}$$

where $\|.\|_2$ is the spectral norm for matrices. Furthermore, if $\lambda_{p,n}\geq a$ then $\widehat{\mathbf{G}} = \widehat{\mathbf{G}}_a$. If $a>0$ is small enough, Cardot et al. [2013d] show under standard conditions on the moments of the variables $\mathcal{X}_1,\ldots,\mathcal{X}_p$ (assumptions (A1), (A2) from below) and on the first and second order inclusion probabilities (assumptions (S1), (S2)) that

$$\mathbb{P}(\widehat{\mathbf{G}}\neq\widehat{\mathbf{G}}_a) = \mathbb{P}(\lambda_{p,n}<a) = O(n^{-1})$$

(see Lemma A.1. from Cardot et al. [2013d]). In the following of this section, I will drop off the subscript a from $\widehat{\mathbf{G}}_a$ for ease of notation.

Finally, with interpolated values $Y_k^{(\mathrm{d})}$ as given in (3.11), the mean curve μ_N is estimated by

$$\widehat{\mu}_{MA}^{(\mathrm{d})}(t) = \frac{1}{N}\sum_{k\in s} d_k Y_k^{(\mathrm{d})}(t) - \frac{1}{N}\left(\sum_{k\in s} d_k \mathbf{x}_k - \sum_{k\in U}\mathbf{x}_k\right)^T \widehat{\boldsymbol{\beta}}_{\mathbf{x}}^{(\mathrm{d})}(t), \tag{3.38}$$

where $\widehat{\boldsymbol{\beta}}_{\mathbf{x}}^{(\mathrm{d})}(t) = \widehat{\mathbf{G}}^{-1}\frac{1}{N}\sum_{k\in s} d_k \mathbf{x}_k Y_k^{(\mathrm{d})}(t)$ and $t\in[t_i,t_{i+1}]$, $i=1,...,D_N$. Remark that the estimator $\widehat{\mu}_{MA}^{(\mathrm{d})}$ belongs to $C[0,\mathcal{T}]$ by construction.

Consistency of the functional model-assisted estimator

For each fixed $t \in [0, \mathcal{T}]$, the functional model-assisted estimator is asymptotically equivalent to the generalized difference estimator:

$$\tilde{\mu}(t) = \frac{1}{N}\sum_{k\in s} d_k Y_k(t) - \frac{1}{N}\left(\sum_{k\in s} d_k \mathbf{x}_k - \sum_{k\in U}\mathbf{x}_k\right)^T \tilde{\boldsymbol{\beta}}_{\mathbf{x}}(t), \tag{3.39}$$

where $\tilde{\boldsymbol{\beta}}_{\mathbf{x}}(t) = \mathbf{G}^{-1}\frac{1}{N}\sum_{k\in U}\mathbf{x}_k Y_k(t)$ and $\mathbf{G} = \sum_{k\in U}\mathbf{x}_k\mathbf{x}_k^T/N$. In particular, for each $t \in [0, \mathcal{T}]$,

$$\widehat{\mu}_{MA}^{(\mathrm{d})}(t) - \mu(t) = \widetilde{\mu}(t) - \mu(t) + o_p(n^{-1/2}). \tag{3.40}$$

Cardot et al. [2013d] prove the uniform consistency of $\widehat{\mu}_{MA}^{(\mathrm{d})}$ which requires showing first the uniform consistency of $\widehat{\boldsymbol{\beta}}_{\mathbf{x}}^{(\mathrm{d})}(t)$. We need the following assumptions on the auxiliary information.

Assumption A1. We assume that there is a positive constant C_4 such that for all $k \in U$, $||\mathbf{x}_k||^2 < C_4$.

Assumption A2. We assume that the matrix $\mathbf{G}$ is invertible and that the number a chosen before satisfies $||\mathbf{G}^{-1}||_2 < a^{-1}$.

Proposition 3.8. *(Cardot et al. [2013d]) Let assumptions (S1), (S2), (V4) and (A1), (A2) hold. If the discretization scheme satisfies*

$$\max_{i\in\{1,..,D_N-1\}} |t_{i+1} - t_i|^{2\beta} = o(n^{-1}),$$

then:

$$\mathbb{E}_p\left\{\sup_{t\in[0,\mathcal{T}]}\left\|\widehat{\boldsymbol{\beta}}_{\mathbf{x}}^{(\mathrm{d})}(t) - \tilde{\boldsymbol{\beta}}_{\mathbf{x}}(t)\right\|\right\} = O(n^{-1/2}) \tag{3.41}$$

and

$$\mathbb{E}_p\left\{\sup_{t\in[0,\mathcal{T}]} \mid \widehat{\mu}_{MA}^{(\mathrm{d})}(t) - \mu_N(t) \mid\right\} = O(n^{-1/2}). \tag{3.42}$$

Proof. The proof of (3.41) is similar to the one developed below. It consists in analyzing the interpolation error as well as the estimation error. We have

$$\sup_{t\in[0,\mathcal{T}]} \mid \widehat{\mu}_{MA}^{(\mathrm{d})}(t) - \mu_N(t) \mid \leq \sup_{t\in[0,\mathcal{T}]} \mid \widehat{\mu}_{MA}^{(\mathrm{d})}(t) - \widehat{\mu}_{MA}(t) \mid + \sup_{t\in[0,\mathcal{T}]} \mid \widehat{\mu}_{MA}(t) - \mu_N(t) \mid, \tag{3.43}$$

where

$$\widehat{\mu}_{MA}(t) = \frac{1}{N}\sum_{k\in s} d_k Y_k(t) - \frac{1}{N}\left(\sum_{k\in s} d_k \mathbf{x}_k - \sum_{k\in U}\mathbf{x}_k\right)^T \widehat{\boldsymbol{\beta}}_{\mathbf{x}}(t),$$

with $\widehat{\boldsymbol{\beta}}_{\mathbf{x}}(t) = \widehat{\mathbf{G}}^{-1}\frac{1}{N}\sum_{k\in s} d_k \mathbf{x}_k Y_k(t)$. Under the assumption on the repartition of the discretized points, we have

$$\sup_{t\in[0,\mathcal{T}]} \sqrt{n} \mid \widehat{\mu}_{MA}^{(\mathrm{d})}(t) - \widehat{\mu}_{MA}(t) \mid = o(1).$$

To bound the estimation error, we need to bound the following quantity:

$$\sup_{t,r\in[0,\mathcal{T}]} \sqrt{n}|\widehat{\mu}_{MA}(t) - \mu(t) - \widehat{\mu}_{MA}(r) + \mu(r)|.$$

This term is dealt with maximal inequalities, more exactly, the corollary 2.2.5 from van der Vaart and Wellner [2000] is used. According to this corollary, there is a constant $\mathbb{B} > 0$ such that

$$\left\{ \mathbb{E}_p \left(\sup_{t,r\in[0,\mathcal{T}]} \sqrt{n}\, |\widehat{\mu}_{MA}(t) - \mu(t) - \widehat{\mu}_{MA}(r) + \mu(r)| \right)^2 \right\}^{1/2} \leq \mathbb{B} \int_0^{\mathcal{T}} \sqrt{\mathbb{D}(x,d)} dx \qquad (3.44)$$

where $\mathbb{D}(x,d)$ is the packing number defined as the maximum number of points in $[0,\mathcal{T}]$ whose distance d between each pair is strictly larger than x (see van der Vaart and Wellner [2000]) and d is a semimetric defined by

$$d^2(r,t) = n\mathbb{E}_p|\widehat{\mu}_{MA}(t) - \mu(t) - \widehat{\mu}_{MA}(r) + \mu(r)|^2.$$

Cardot et al. [2013d] showed that there is a constant C such that

$$d^2(r,t) \leq C|t-r|^{2\beta}.$$

Hence, the packing number is bounded as follows: $\mathbb{D}(x,d) = O(x^{-1/\beta})$, implying that the integral from the right-hand side of (3.44) is finite when $\beta > 1/2$. □

Note that the interpolation error is negligible, compared to the sampling variability, under the additional assumption on the repartition of the discretization points.

Remark 9. *A shorter proof of the uniform consistency of* $\widehat{\mu}_{MA}^{(\mathrm{d})}$ *could have been obtained by using the definition 3.1 and the uniform consistency of* $\widehat{\boldsymbol{\beta}}_{\mathbf{x}}^{(\mathrm{d})}$ *given in (3.41). In fact, we have*

$$\widehat{\mu}_{MA}^{(\mathrm{d})} - \mu_N = \widetilde{\mu} - \mu_N - \frac{1}{N}\left(\sum_{k\in s} d_k \mathbf{x}_k - \sum_{k\in U} \mathbf{x}_k\right)^T (\widehat{\boldsymbol{\beta}}_{\mathbf{x}}^{(\mathrm{d})} - \tilde{\boldsymbol{\beta}}_{\mathbf{x}}) + \sum_{k\in s} d_k (Y_k^{(\mathrm{d})} - Y_k)$$

and

$$P\left(\sup_{t\in[0,\mathcal{T}]}\sqrt{n}|\widehat{\mu}_{MA}^{(\mathrm{d})}(t)-\mu_N(t)|>\varepsilon\right)\leq$$
$$P\left(\sup_{t\in[0,\mathcal{T}]}\sqrt{n}|\widetilde{\mu}(t)-\mu_N(t)|>\frac{\varepsilon}{3}\right)+P\left(\frac{\sqrt{n}}{N}\left|\sum_{k\in s}d_k\mathbf{x}_k-\sum_{k\in U}\mathbf{x}_k\right|^T\sup_{t\in[0,\mathcal{T}]}|\tilde{\boldsymbol{\beta}}_{\mathbf{x}}(t)-\hat{\boldsymbol{\beta}}_{\mathbf{x}}^{(\mathrm{d})}(t)|>\frac{\varepsilon}{3}\right)$$
$$+P\left(\sup_{t\in[0,\mathcal{T}]}\sqrt{n}|\sum_{k\in s}d_k(Y_k^{(\mathrm{d})}(t)-Y_k(t)|>\frac{\varepsilon}{3}\right)$$

which is going to zero under the assumptions on the discretization points and the fact that $\mathbb{E}_p\frac{\sqrt{n}}{N}\left|\sum_{k\in s}d_k\mathbf{x}_k-\sum_{k\in U}\mathbf{x}_k\right|=O(1)$ *under the assumptions on the design and the auxiliary information.*

Variance estimation and confidence bands

Result 3.42 allows us to approximate the covariance function of $\widehat{\mu}_{MA}^{(\mathrm{d})}$ between two instants r and t by the covariance of $\tilde{\mu}$:

$$\gamma_{\mathrm{MA}}(r,t) \quad = \quad \frac{1}{N^2}\sum_{k\in U}\sum_{l\in U}(\pi_{kl}-\pi_k\pi_l)\frac{Y_k(r)-\mathbf{x}_k^T\tilde{\boldsymbol{\beta}}(r)}{\pi_k}\frac{Y_l(t)-\mathbf{x}_l^T\tilde{\boldsymbol{\beta}}(t)}{\pi_l}. \tag{3.45}$$

The covariance estimator is given by

$$\widehat{\gamma}_{\mathrm{MA}}^{(\mathrm{d})}(r,t)=\frac{1}{N^2}\sum_{k,l\in s}\frac{\pi_{kl}-\pi_k\pi_l}{\pi_{kl}}\cdot\frac{Y_k^{(\mathrm{d})}(r)-\mathbf{x}_k^T\widehat{\boldsymbol{\beta}}_{\mathbf{x}}^{(\mathrm{d})}(r)}{\pi_k}\cdot\frac{Y_l^{(\mathrm{d})}(t)-\mathbf{x}_l^T\widehat{\boldsymbol{\beta}}_{\mathbf{x}}^{(\mathrm{d})}(t)}{\pi_l},\quad r,t\in[0,\mathcal{T}]. \tag{3.46}$$

It is proven in Cardot et al. [2013d] that the covariance estimator $\widehat{\gamma}_{\mathrm{MA}}^{(\mathrm{d})}$ is consistent and the variance function estimator is uniformly convergent. Thus, under additional asymptotic normality assumptions, it is also possible to build confidence bands with the Monte Carlo procedure described in Section 3.1.5.

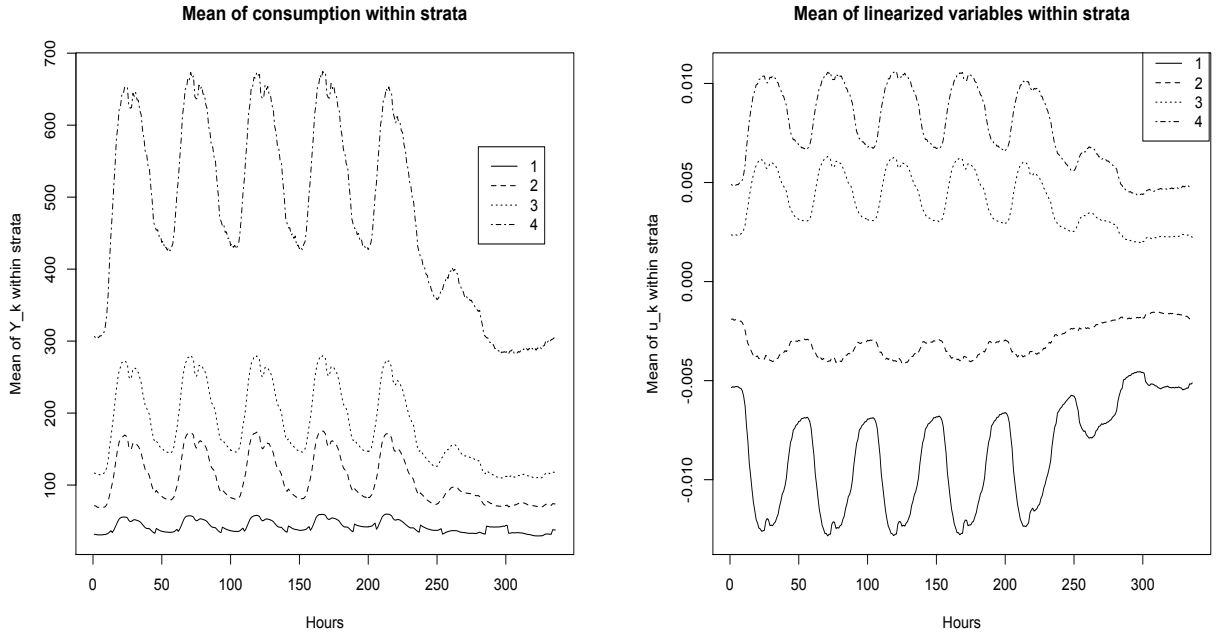
Note that previous model can be extended without difficulties for auxiliary variables that vary in time, so that we have for each unit of the sample $\mathbf{x}_k(t)=(x_{k1}(t),...,x_{kp}(t))^T$ for $t\in[0,\mathcal{T}]$. As in Cardot et al. [2010b] nonparametric models can also be considered by first reducing the dimension of the data with principal components, as described in Section 3.1.3, and then consider a single index or an additive model on the principal component scores.

3.1.9 An application to French electricity load curves

Consider the test population of $N = 18902$ French companies whose electricity consumption has been measured every half-hour over a period of two weeks as considered in Cardot and Josserand [2011] and Chaouch and Goga [2012]. The data recorded over the first week $\mathbf{X}_k$ are used as auxiliary information, while the data recorded over the second week $\mathbf{Y}_k$ are the study variable. More exactly, we have 336 instant measures per week and let $\mathbf{X}_k = (X_k(t_d))_{d=1}^{336}$ and $\mathbf{Y}_k = (Y_k(t_d))_{d=1}^{336}$. The goal is to estimate the mean curve μ_N and the median curve m_N by using a sample of size $n = 2000$ selected according to SRSWOR and STRAT designs.

The population is divided into $H = 4$ strata constructed according to the maximum level of $\mathbf{X}_k$ and based on the quartiles, so that all the strata have almost the same size. In Cardot et al. [2013c], two stratification of the population have been used (but not considered here): the first one was carried out using the k-means classification of the discretized trajectories $\mathbf{X}_k$ and the second one, by using the mean consumption during the first week (see formula 3.47).

The stratum 1, corresponds to consumers with low global consumption level, whereas stratum 4 corresponds to consumers with high global levels of consumption. We plot in Figure 3.4(a), the mean of $\mathbf{Y}_k$ within each stratum and in Figure 3.4(b), the mean of the linearized variable of the median $\mathbf{u}_{k,m_N} = (u_{k,m_N}(t_d))_{d=1}^{336}$ within each stratum. Note that the population of the linearized variable curves is also stratified.



(a) (b)

FIGURE 3.4: Stratification based on the consumption curve: (a) Mean of the consumption curve $\mathbf{Y}_k$ within each stratum. (b) Mean of the linearized variable $\mathbf{u}_{k,m_N}$ within each stratum.

To draw a STRAT sampling, we use the proportional allocation (PROP) and the x-optimal allocation (x-OPT) computed with respect to the consumption $\mathbf{X}_k$ recorded during the previous week. Table 3.1 gives the stratum sizes and the sample sizes for both types of allocation.

Stratum number	1	2	3	4
Stratum size N_h	4725	4726	4725	4726
PROP allocation	500	500	500	500
x-OPT allocation	126	212	333	1329

TABLE 3.1: Strata sizes, proportional and x-optimal allocations for a sample size of $n = 2000$.

In the following, several sampling designs of size $n = 2000$ and the HT estimator are compared. This includes the simple random sampling without replacement (SRSWOR), stratified sampling with SRSWOR within each stratum and the proportional allocation (STRAT+PROP) as well as the x-optimal allocation (STRAT+opt), and finally πps sampling. A πps sample of size $n = 2000$ is selected with first-order inclusion probabilities π_k proportional to the mean consumption recorded during the previous week:

$$x_k = \frac{1}{336} \sum_{d=1}^{336} X_k(t_d), \quad k \in U. \tag{3.47}$$

To draw such a sample, one may use the fast version of the cube algorithm (see Chauvet and Tillé [2006]) balanced on the vector of first-order inclusion probabilities $\boldsymbol{\pi} = (\pi_1, ..., \pi_N)$ with π_k given by (1.6) and x_k by (3.47).
In order to compare these designs, $I = 500$ samples are drawn. Tables 3.2 and 3.3 give statistics about the estimation errors computed according to the following loss criterion:

$$R(\hat{\theta}) = \int_0^{\mathcal{T}} |\hat{\theta}(t) - \theta(t)| dt \simeq \frac{1}{336} \sum_{d=1}^{336} |\hat{\theta}(t_d) - \theta(t_d)|, \tag{3.48}$$

with $\hat{\theta}$ an estimator of θ.

We can remark that clustering the space of functions by performing stratified sampling leads to an important gain compared to simple random sampling without replacement especially for the estimation of the mean curve. STRAT with proportional allocation gives slightly better results

	Mean	1^{st} quartile	median	3^{rd} quartile
SRSWOR	4.624	2.405	3.694	6.073
STRAT+PROP	3.731	2.116	3.041	4.803
STRAT+OPTIM	**2.507**	**1.605**	**2.198**	**3.128**

TABLE 3.2: Estimation errors (criterion 3.48) of the mean curve μ_N with SRSWOR and STRAT sampling.

	Mean	1^{st} quartile	median	3^{rd} quartile
SRSWOR	2.697	1.362	2.274	3.527
STRAT+PROP	**1.632**	**1.048**	**1.402**	**2.017**
STRAT+OPTIM	2.263	1.444	1.969	2.865

TABLE 3.3: Estimation errors (criterion 3.48) of the median curve m_N with SRSWOR and STRAT sampling.

	Mean	1^{st} quartile	median	3^{rd} quartile
πps for μ_N	1.816	1.447	1.709	2.081
πps for m_N	7.263	2.901	5.918	9.733
πps for m_N with B-spline	**1.947**	**1.364**	**1.711**	**2.209**

TABLE 3.4: Estimation errors (criterion 3.48) of the mean curve μ_N and the median curve m_N with the πps sampling.

for the estimation of the median than those obtained with the optimal allocation. This is due to the fact that the optimal allocation is computed by minimizing the variance of the estimator for the mean curve.

Table 3.4 gives the estimation errors of the mean and median curve with the πps sampling. We remark that this design performs very well for the estimation of the mean curve but very poorly for the estimation of the median. We give in Table 3.4 the estimation errors of the median estimator obtained by using the B-spline smoothed weights w_{ks}^{np} for $m = 3$, $K = 8$. We can remark that the performance of the πps design for estimating m_N was greatly improved.
Cardot et al. [2013c] also evaluated the performance of the functional model-assisted estimator. The test population did not contain any auxiliary information, so that the mean past consumption over the previous week x_k has been used as auxiliary information plus the intercept term. Their correlation with the current consumption is always very high (between 0.85 and 0.95), so that linear regression models are natural candidates for improving the HT estimator. The simulation results obtained showed that the functional model-assisted estimator improves a lot the efficiency of the HT estimator even if it does not performs as well as the πps or the stratified estimator. However, the main advantage of this estimator with respect to its competitors is that it only requires the knowledge of the total of the auxiliary information.

Simulation studies showed that the mean width of the confidence bands is greatly reduced (almost by half) by taking into account auxiliary information at the sampling stage (STRAT or πps) or at the estimation stage. Computing confidence bands of level $1-\alpha$ requires computing c_α, the quantile of order $1-\alpha$ of the supremum of Gaussian processes. Figure 3.5 shows an example of confidence band (continuous grey curves). We can remark that, in this situation, the true mean curve falls within the confidence band.

In order to avoid the estimation of the covariance function, $\hat{\gamma}^{(\text{d})}(r,t)$, Cardot et al. [2013c] also considered the bootstrap as suggested by Gross [1980] and adapted for each strategy described

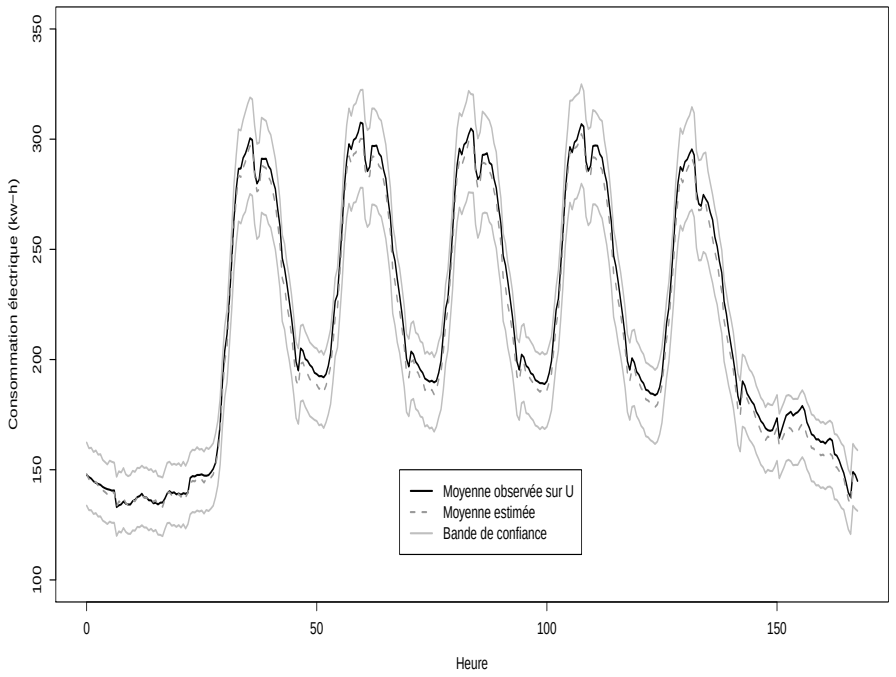


FIGURE 3.5: Confidence band (continuous grey curves) for the true mean load curve (continuous black curve)

above (Chauvet [2007]). The results on the coverage rates and the width of the confidence bands are very similar to those obtained with the Gaussian processes simulation method. However, the computation times with the bootstrap are much greater, by a factor of approximately 1 to 1000. The reason for this is that the entire bootstrap process (duplication of the population) which must be repeated at each simulation.

3.1.10 Conclusion and perspectives

Even if some work has already been done, there are still many fields to explore in the near future, at the frontier of survey sampling and functional data analysis.

With unequal probability sampling designs, Cardot et al. [2014b] noted that the HT estimator and his covariance estimator are not robust to the presence of atypical individuals. Such outlying data may not be uncommon in large samples and another interesting direction of research would be to consider correction techniques of the samplings weights of the most influential units of the sample (see *e.g.* Beaumont and Rivest [2009]) in order to get a more stable variance estimator. Some work is also needed to adapt what already exists to the functional context.

In their work, Cardot et al. [2013d] extended the results on uniform consistency obtained by Cardot and Josserand [2011] for the HT estimator of the mean, $\hat{\mu}^{(\mathrm{d})}$, to the functional model-assisted estimator of the mean $\widehat{\mu}_{MA}^{(\mathrm{d})}$. In particular, they proved the uniform consistency of the estimator $\widehat{\boldsymbol{\beta}}_{\mathbf{x}}^{(\mathrm{d})}$ of the regression coefficient. The parameter of interest $\boldsymbol{\beta}$ is an example of nonlinear parameter of totals of functional and multivariate variables. In would be interesting to give a general approach for obtaining linearization and asymptotic properties of such parameters by making use of the tightness property. As van der Vaart [1998] remarked, "weak convergence of random elements in metric spaces is intimately connected with compact sets, through Prohorov's theorem, Hadamard differentiability is the right type of differentiability in connection with delta method."

The space $C[0, \mathcal{T}]$ is unsuitable for the description of processes that may have jumps. An example of such data is given by the audience curves recorded at Médiamétrie (the French company for measuring the audience). Figure 3.6 displays the audience curves recorded every minute for a sample of 5 individuals. The suitable space for such random functions is $D[0, \mathcal{T}]$, the space of cadlag functions: continue *à droite, limite à gauche*. It would be interesting to see how results developed for random functions belonging to $C[0, \mathcal{T}]$ may be extended to $D[0, \mathcal{T}]$. Asymptotics in $D[0, \mathcal{T}]$ (extensively described in Billingsley [1968]) is complicated by the loss of some properties of $C[0, \mathcal{T}]$ such as the separability property with respect to the uniform distance ρ.

So far, the method of estimation combining functional data analysis and surveys techniques do not take into account the presence of non-response in individual curves. Trajectories with missing observations during some intervals of time may not be so rare because of transmission problems. In order to reconstruct the missing parts of the trajectories, classical methods of imputation (see Haziza [2009] for a review) can be applied, instant by instant. The disadvantage of these methods, which are essentially univariate, is that they do not take into account the history (the temporal correlation) of the individuals. Note also that a further difficulty arises from the fact that this history can also contain non-response. A second possibility would be to apply interpolation or smoothing techniques, by adapting to a survey sampling context previous works (see Staniswalis and Lee [1998]) in nonparametric estimation, on the missing part of the trajectories. This latter approach would allow to reconstruct individual trajectories by taking into account not only their history but also the shape of the other trajectories. First works on this topic have been done in De Moliner et al. [2014], Cardot et al. [2014a]. Further work is needed to build an imputation method that allows to impute the trajectories taking into account all the points of observation of the variable of interest for each individuals in our sample as well as auxiliary information. The nearest neighbor imputation technique (see Chen and Shao [2000], Shao and Wang [2008] and Beaumont and Bocci [2009]) by its nonparametric point of view and its simplicity seems to be a good candidate.

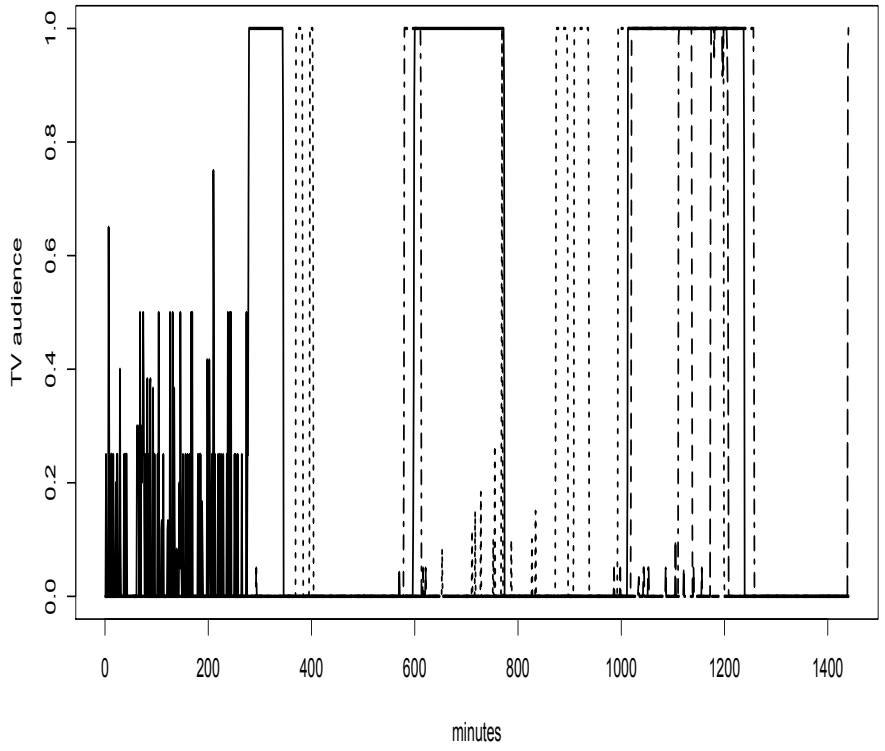


FIGURE 3.6: A sample of 5 audience curves. The audience is recorded every minute.

3.2 High-dimensional auxiliary information

Let consider now that a large number of auxiliary variables is available. Chapter 1 presents the different approches for taking into account the auxiliary information.

In a model-based approach, we have seen that the BLU estimators are model-dependent and may suffer from large bias if the model is misspecified. As a measure of protection from model erroneous specification, the balanced sampling (see relation 1.29) was suggested (Royall and Herson [1973]). Nevertheless, in multipurpose sample surveys, when a large number of finite population totals or means are to be estimated, it may not be always possible to have a balanced sample. So, many auxiliary variables should be included in order to have a fully specified model for each variable. But, the inclusion of too many variables may result in an over-specified model and the unbiased estimators (the BLU estimators) derived in these conditions are unstable and inefficient. This is why, penalized estimators have been suggested first in a model-based approach (Bardsley and Chambers [1984]) to overcome with the estimation issues for data obtained from unbalanced samples.

In a calibration approach, estimation in presence of a very large number of calibration variables was called *over-calibration* by Guggemos and Tillé [2010]. In a calibration setting, it is often desired that the values of the ratio between the calibration weights and the sampling weights lie between positive predefined upper and lower bounds, called also range restrictions and denoted by U and L. Not satisfying such conditions may be due to the chosen distance Υ_s and in order to cope with this issue, several modifications of the distance have been suggested in the literature (Deville and Särndal [1992]; Jayasuriya and Valliant [1996] and Singh and Mohl [1996]), but as Beaumont and Bocci [2008] remarked, these methods "are all iterative and may not yield a solution even if the range restrictions are mild". Now, if a large number of auxiliary variables is used, the calibration weights derived in this situation may be unstable and very large and the range restrictions are difficult to be satisfied. Moreover, auxiliary variables may be related linearly to each other, and hence can cause multicollinearity. In this case, the matrix $\sum_{k\in s} d_k \mathbf{x}_k \mathbf{x}_k^T$ is not a full rank matrix and the calibration weights cannot be computed directly with (1.15). A generalized inverse of the previous matrix should be used. Théberge [1999] suggested the minimum norm least squares method and the Moore-Penrose inverse matrix to derive weights in presence of multicollinearity among regressors.

Finally, let us analyse the impact of having many auxiliary variables on the variance of the calibration estimator. As Särndal [2007] states, the calibration should bring "the extra benefit of improved accuracy (lower variance and/or reduced nonresponse bias)". However, when a very large number p of auxiliary variables is used, this result is no longer true as it was remarked by Silva and Skinner [1997] in a simulation study. Recently, Chauvet and Goga [2013a] introduced the following two additional assumptions on the auxiliary information:

Assumption A3. $||\mathbf{x}_k||^2 = O(p)$ for all $k \in U$,

Assumption A4. there exist c and C positive constants such that $c < \lambda_{\min} < \lambda_{\max} < C$, where $\lambda_{\min}$ and $\lambda_{\max}$ are the smallest and the largest eigenvalue of $N^{-1}\sum_{k\in U}\mathbf{x}_k\mathbf{x}_k^T$,

When these hypotheses hold, they have showed the following result:

Proposition 3.9. *Make the assumptions (S1), (S2) on the sampling design, (V3) on the study variable and (A3), (A4) on the auxiliary variables.*

Then,

$$\frac{1}{N}(\hat{t}_{\mathbf{x}d} - t_{\mathbf{x}})^T(\hat{\boldsymbol{\beta}}_{\mathbf{x}}(\mathbf{d},\mathbf{v}) - \tilde{\boldsymbol{\beta}}_{\mathbf{x}}(\mathbf{v})) = O_{\mathrm{p}}\left(\frac{p^2}{n}\right),$$

where $\hat{\boldsymbol{\beta}}_{\mathbf{x}}(\mathbf{d},\mathbf{v})$ *is given by (1.22) and* $\tilde{\boldsymbol{\beta}}_{\mathbf{x}}(\mathbf{v})$ *by (1.21).*

This means that the error between the calibration estimator and the generalized difference estimator depends on the number p of the auxiliary variables. As a consequence, including more and more auxiliary variables may alter the performance of the calibration estimator. The asymptotic variance of the calibration estimator given in (1.18) remains valid only if $p = O(n^a)$ with $a < 1/2$. Otherwise, the extra-variability of $(\hat{t}_{\mathbf{x}d} - t_{\mathbf{x}})^T(\hat{\boldsymbol{\beta}}_{\mathbf{x}}(\mathbf{d},\mathbf{v}) - \tilde{\boldsymbol{\beta}}_{\mathbf{x}}(\mathbf{v}))$ should also be taken into account.

One way to avoid many of the above difficulties, is to choose only a subset of the auxiliary variables and to consider only the auxiliary variables that are expected to be the more relevant (Silva and Skinner [1997]; Skinner and Silva [1997]; Chambers et al. [1999]; Clark and Chambers [2008] and Chauvet and Goga [2013a]). Another alternative is to keep all the variables and to use different methods such as ridge regression, principal component regression, partial least squares or lasso methods that are frequently used in classical statistics to deal with high dimensional data.

3.2.1 Penalization in survey sampling by ridge regression

One way to circumvent the problems due to over-calibration is to relax the calibration constraints, meaning that the too restrictive requirement of being exactly satisfied as in (1.13) is dropped off and replaced by the requirement of being only approximatively satisfied. Nevertheless, the deviation between $\hat{t}_{\mathbf{x}w} = \sum_{k\in s} w_k\mathbf{x}_k$ and $t_{\mathbf{x}} = \sum_{k\in U}\mathbf{x}_k$ is controlled by means of a penalty.

A class of penalized estimators was suggested in a model-based setting and extended later by Chambers [1996] and by Rao and Singh [1997], Rao and Singh [2009] in a design-based (or model-assisted) setting. These approaches lead to a class of model-based or GREG-type estimators that use regression coefficients estimated by ridge-type estimators. Goga and Shehzad [2014b] give a recent review of the application of the ridge-regression in survey sampling and in

both the model-based and the design-based approaches.

Consider the chi-squared distance $\Upsilon_s((\mathbf{a},\mathbf{b});\mathbf{w})$ given in (1.30). In a model-based or design-based setting, the penalized weights $\mathbf{w}_s^{\text{pen}} = (w_{ks}^{\text{pen}})_{k\in s}$ are obtained as the solutions of the following penalized minimization problem:

$$\mathbf{w}_s^{\text{pen}}((\mathbf{a},\mathbf{b});\lambda) = \text{argmin}_{\mathbf{w}}\,\Upsilon_s((\mathbf{a},\mathbf{b});\mathbf{w}) + \lambda^{-1}(\hat{t}_{\mathbf{x}w} - t_{\mathbf{x}})^T\mathbf{C}(\hat{t}_{\mathbf{x}w} - t_{\mathbf{x}}), \tag{3.49}$$

where $\hat{t}_{\mathbf{x}w} = \sum_{k\in s} w_k\mathbf{x}_k = \mathbf{w}^T\mathbf{X}_s$, $\lambda > 0$ is a scale factor and $\mathbf{C} = \text{diag}(c_j)_{j=1}^p$ with $c_j \geq 0$ is a user-specified cost associated with the jth calibration constraint (Bardsley and Chambers [1984], Chambers [1996]). Beaumont and Bocci [2008] considered that c_j is "the cost associated with not satisfying the jth calibration constraint" and they suggested to take $c_j = 1/t_{X_j}$, where t_{X_j} is the population total of $\mathcal{X}_j$. In a calibration approach, Beaumont and Bocci [2008] have considered the penalized optimization problem (3.49) for more general distance functions Υ_s and they suggested a simply iteratively re-weighted chi-square algorithm to obtain the penalized weights in this case.

The optimization problem (3.49) means that we look for the weights $w_{ks}^{\text{pen}}, k \in s$ that best explain the vector $\mathbf{a}$ and such that the weighted estimator $\hat{t}_{\mathbf{x}w}$ is close enough to the true total $t_{\mathbf{x}}$. With this kind of penalization, the finite population total $t_{\mathbf{x}}$ is not estimated exactly anymore by $\hat{t}_{\mathbf{x}w}$ but large deviations are penalized by the cost matrix $\mathbf{C}$ and the penalty parameter λ.

Proposition 3.10. *The solution of (3.49) is*

$$w_{ks}^{\text{pen}}((\mathbf{a},\mathbf{b});\lambda) = a_k - a_k b_k^{-1}\mathbf{x}_k^T\left(\sum_{k\in s} a_k b_k^{-1}\mathbf{x}_k\mathbf{x}_k^T + \lambda\mathbf{C}^{-1}\right)^{-1}(\hat{t}_{\mathbf{x}a} - t_{\mathbf{x}}) \quad k \in s, \tag{3.50}$$

where $\hat{t}_{\mathbf{x}a} = \sum_{k\in s} a_k\mathbf{x}_k$. *The penalized estimator is given by*

$$\begin{aligned}\hat{t}_{yw}^{\text{pen}}((\mathbf{a},\mathbf{b});\lambda) &= \sum_{k\in s} w_{ks}^{\text{pen}}((\mathbf{a},\mathbf{b});\lambda)y_k \\ &= \sum_{k\in s} a_k y_k - \left(\sum_{k\in s} a_k\mathbf{x}_k - \sum_{k\in U}\mathbf{x}_k\right)^T\hat{\boldsymbol{\beta}}_{\mathbf{x}}((\mathbf{a},\mathbf{b});\lambda),\end{aligned} \tag{3.51}$$

where

$$\hat{\boldsymbol{\beta}}_{\mathbf{x}}((\mathbf{a},\mathbf{b});\lambda) = \left(\sum_{k\in s} a_k b_k^{-1}\mathbf{x}_k\mathbf{x}_k^T + \lambda\mathbf{C}^{-1}\right)^{-1}\sum_{k\in s} a_k b_k^{-1}\mathbf{x}_k y_k.$$

We remark that $\hat{\boldsymbol{\beta}}_{\mathbf{x}}((\mathbf{a},\mathbf{b});\lambda)$ is a ridge-type estimator of the regression coefficient $\boldsymbol{\beta}$ from the superpopulation model ξ as suggested by Hoerl and Kennard [1970] in classical statistics. In a calibration approach, namely for $\mathbf{a} = \mathbf{d}$ and $\mathbf{b} = \mathbf{v}$, we obtain a GREG-type estimator with a design-based ridge-type estimator $\hat{\boldsymbol{\beta}}_{\mathbf{x}}((\mathbf{d},\mathbf{v});\lambda)$:

$$\hat{t}_{yw}^{\text{pen}}((\mathbf{d},\mathbf{v});\lambda) = \hat{t}_{yd} - (\hat{t}_{\mathbf{x}d} - t_{\mathbf{x}})^T\hat{\boldsymbol{\beta}}_{\mathbf{x}}((\mathbf{d},\mathbf{v});\lambda).$$

In a model-based approach, namely for $\mathbf{a} = \mathbf{1}_s$ and $\mathbf{b} = \mathbf{v}$, we obtain a model-based type estimator:

$$\hat{t}_{yw}^{\text{pen}}((\mathbf{1}_s, \mathbf{v}); \lambda) = \sum_{k \in s} y_k + \left(\sum_{k \in U-s} \mathbf{x}_k \right)^T \hat{\boldsymbol{\beta}}_{\mathbf{x}}((\mathbf{1}_s, \mathbf{v}); \lambda)$$

with $\boldsymbol{\beta}$ estimated by a model-based ridge-type estimator $\hat{\boldsymbol{\beta}}_{\mathbf{x}}((\mathbf{1}_s, \mathbf{v}); \lambda)$. In fact, the same penalized GREG or model-based estimators would have been obtained, if we would have plugged-in the estimator $\hat{\boldsymbol{\beta}}_{\mathbf{x}}((\mathbf{d}, \mathbf{v}); \lambda)$ into the generalized difference estimator given in (1.20) or $\hat{\boldsymbol{\beta}}_{\mathbf{x}}((\mathbf{1}_s, \mathbf{v}); \lambda)$ into the model-based predictor (1.24). This is not surprising since the ridge-regression was suggested in classical statistics to robustify the ordinary least square estimator of $\boldsymbol{\beta}$ in presence of multicollinearity between the regressors from the superpopulation model ξ (Hoerl and Kennard [1970]).

In order to better understand the improvement brought by imposing a quadratic penalty in (3.49), consider in detail the following two quantities of interest: the distance Υ_s between the initial weights $\mathbf{a}$ and the final weights $\mathbf{w}_s^{\text{pen}}$, and secondly, the erreur between the weighted estimator $\hat{t}_{\mathbf{x}w}$ and the true total $t_{\mathbf{x}}$. As it will be noticed, the ridge-type estimators may be seen as a trade-off between full-calibrated and no-calibrated estimators.

The value of the (chi-)squared distance Υ_s between $\mathbf{w}_s^{\text{opt}}$ and $\mathbf{a}$ tends to be large if there is a small eigenvalue of $\sum_{k \in s} a_k b_k^{-1} \mathbf{x}_k \mathbf{x}_k^T$. Indeed, let $\hat{\alpha}_1 \geq \ldots \geq \hat{\alpha}_p > 0$ be the eigenvalues of $\sum_{k \in s} a_k b_k^{-1} \mathbf{x}_k \mathbf{x}_k^T$ considered in decreasing order with $\hat{\mathbf{v}}_j$ the corresponding orthonormal eigenvectors. We have

$$\begin{aligned} \Upsilon_s((\mathbf{a}, \mathbf{b}); \mathbf{w}_s^{\text{opt}}) &= \sum_{k \in s} b_k \frac{(w_{ks}^{\text{opt}} - a_k)^2}{a_k} \\ &= (\hat{t}_{\mathbf{x}d} - t_{\mathbf{x}})^T \left(\sum_{j=1}^{p} \frac{1}{\hat{\alpha}_j} \hat{\mathbf{v}}_j \hat{\mathbf{v}}_j^T \right) (\hat{t}_{\mathbf{x}d} - t_{\mathbf{x}}) \end{aligned}$$

which states clearly that the worse the conditioning of $\sum_{k \in s} a_k b_k^{-1} \mathbf{x}_k \mathbf{x}_k^T$, the more $\mathbf{w}_s^{\text{opt}}$ can be expected to be long and far away from $\mathbf{a}$. Consider now the (chi-)squared distance between $\mathbf{w}_s^{\text{pen}}$ and $\mathbf{a}$. The penalized matrix $\sum_{k \in s} a_k b_k^{-1} \mathbf{x}_k \mathbf{x}_k^T + \lambda \mathbf{C}^{-1}$ has the eigenvalues $\hat{\alpha}_1 + \lambda c_1^{-1}, \ldots, \hat{\alpha}_p + \lambda c_p^{-1}$ with the same eigenvectors $\hat{\mathbf{v}}_j$ (Hoerl and Kennard [1970]) and we can write:

$$\begin{aligned} \Upsilon_s((\mathbf{a}, \mathbf{b}); \mathbf{w}_s^{\text{pen}}) &= \sum_{k \in s} b_k \frac{(w_{ks}^{\text{pen}} - a_k)^2}{a_k} \\ &= (\hat{t}_{\mathbf{x}d} - t_{\mathbf{x}})^T \left(\sum_{j=1}^{p} \frac{\hat{\alpha}_j}{(\hat{\alpha}_j + \lambda c_j^{-1})^2} \hat{\mathbf{v}}_j \hat{\mathbf{v}}_j^T \right) (\hat{t}_{\mathbf{x}d} - t_{\mathbf{x}}). \end{aligned}$$

We always have

$$\frac{1}{\hat{\alpha}_j} > \frac{\hat{\alpha}_j}{(\hat{\alpha}_j + \lambda c_j^{-1})^2}$$

for any $\lambda > 0$ and $c_j > 0$ and we have that

$$\Upsilon_s((\mathbf{a}, \mathbf{b}); \mathbf{w}_s^{\text{pen}}) < \Upsilon_s((\mathbf{a}, \mathbf{b}); \mathbf{w}_s^{\text{opt}}),$$

namely, the distance between the penalized weights $\mathbf{w}_s^{\text{pen}}$ and the initial vector $\mathbf{a}$ is shorter than the distance between the optimal weights $\mathbf{w}_s^{\text{opt}}$ and $\mathbf{a}$. In particular, we obtain that the scatter of penalized weights $\mathbf{w}_s^{\text{pen}}$ is smaller than the one of $\mathbf{w}_s^{\text{opt}}$ (Bardsley and Chambers [1984]).

On the other side, the weighted error between a weighted estimator $\hat{t}_{\mathbf{x}w}$ and the true total $t_{\mathbf{x}}$ is given by

$$\begin{aligned} \Psi_s(\mathbf{w}) &= (\hat{t}_{\mathbf{x}w} - t_{\mathbf{x}})^T \mathbf{C} (\hat{t}_{\mathbf{x}w} - t_{\mathbf{x}}) = (\hat{t}_{\mathbf{x}w} - \hat{t}_{\mathbf{x}w}^{\text{opt}})^T \mathbf{C} (\hat{t}_{\mathbf{x}w} - \hat{t}_{\mathbf{x}w}^{\text{opt}}) \\ &= (\mathbf{w} - \mathbf{w}_s^{\text{opt}})^T \mathbf{X}_s \mathbf{C} \mathbf{X}_s^T (\mathbf{w} - \mathbf{w}_s^{\text{opt}}). \end{aligned}$$

Contours of constant Ψ_s are the surfaces of hyperellipsoids centered at $\mathbf{w}_s^{\text{opt}}$. The error Ψ_s attains the minimum value $\Psi_{\min} = 0$ for $\mathbf{w} = \mathbf{w}_s^{\text{opt}}$. We obtain in this case the full-calibrated estimator (in a design-based approach) or the full-balanced estimator (in a model-based approach). As mentioned at the beginning of this section, the weights $\mathbf{w}_s^{\text{opt}}$ may not lie between the range restrictions when a large number of auxiliary variables is used (in a design-based approach). On the other side, the weights $\mathbf{a}$ satisfy the range restrictions but yield the maximum of Ψ_s : $\Psi_{\max} = (\hat{t}_{\mathbf{x}a} - t_{\mathbf{x}})^T \mathbf{C} (\hat{t}_{\mathbf{x}a} - t_{\mathbf{x}})$ for $\mathbf{w} = \mathbf{a}$. So, the optimization problem given by (3.49) can be viewed as follows: we can move a little away from the minimum sum of squares point $\Psi_{\min}$ in a direction that will shorten the distance Υ_s. Then, the penalized weight vector is the value of $\mathbf{w}_s$ minimizing

$$\mathbf{w}_s^{\text{pen}}((\mathbf{a}, \mathbf{b}); \tilde{\lambda}) = \text{argmin}_{\mathbf{w}} \Upsilon_s((\mathbf{a}, \mathbf{b}); \mathbf{w}) \quad \text{subject to} \quad (\hat{t}_{\mathbf{x}w} - t_{\mathbf{x}})^T \mathbf{C} (\hat{t}_{\mathbf{x}w} - t_{\mathbf{x}}) = \Psi_0, \quad (3.52)$$

where Ψ_0 is determined from the fixed discrepancies δ_j (*e.g.* differences between $\hat{t}_{\mathbf{x}w}$ and the totals $t_{\mathbf{x}}$) and the costs c_j, $j = 1, \ldots, p$; $\tilde{\lambda}$ is the Lagrange multiplier used for solving the above optimization problem. The choice of discrepancies $\delta_j, j = 1, \ldots, p$ is discussed for example in Rao and Singh [1997].

A completely equivalent statement of the method is the following: we consider that the weights $\mathbf{w}_s$ are located at distance $\Upsilon_s((\mathbf{a}, \mathbf{b}); \mathbf{w}_s) = r_0^2$ from the initial weights $\mathbf{a}$, where r_0 is such that the range restrictions $L \leq \frac{w_k}{d_k} \leq U$ are satisfied for all $k \in U$. Then the penalized weight vector is the value of $\mathbf{w}_s$ that minimizes the distance Ψ between the estimates $\hat{t}_{\mathbf{x}w}$ and their totals $t_{\mathbf{x}}$:

$$\mathbf{w}_s^{\text{pen}}((\mathbf{a}, \mathbf{b}); \tilde{\lambda}_1) = \text{argmin}_{\mathbf{w}} (\hat{t}_{\mathbf{x}w} - t_{\mathbf{x}})^T \mathbf{C} (\hat{t}_{\mathbf{x}w} - t_{\mathbf{x}}) \quad \text{subject to} \quad \Upsilon_s((\mathbf{a}, \mathbf{b}); \mathbf{w}_s) = r_0^2, \quad (3.53)$$

where $\tilde{\lambda}_1$ is the Lagrange multiplier. For a fixed value of $\Psi = (\hat{t}_{\mathbf{x}w} - t_{\mathbf{x}})^T \mathbf{C} (\hat{t}_{\mathbf{x}w} - t_{\mathbf{x}})$, the Lagrange multiplier $\tilde{\lambda}_1$ may be found by means of Newton-Raphson algorithm. Beaumont and Bocci [2008] suggested the bisection algorithm to find the smallest value of Ψ such that the

range restrictions are satisfied, knowing that $\Psi_{\min} = 0$ and $\Psi_{\max} = (\hat{t}_{\mathbf{x}a} - t_{\mathbf{x}})^T \mathbf{C} (\hat{t}_{\mathbf{x}a} - t_{\mathbf{x}})$. Nevertheless, each iteration of the bisection algorithm needs the computation of the Lagrange multiplier and verification of the constraint. The combination of both algorithms may be time-consuming.

Goga and Shehzad [2014b] give a geometrical interpretation of the constrained optimization problems (3.52) and (3.53) and the way the penalized weights are found. This interpretation is the analogue of the Marquart and Snee [1975]'s geometrical interpretation given in the case of classical ridge-regression. Figure 3.7 exhibits this interpretation for the two-dimensional case $\mathbf{w}_s = (w_{1s}, w_{2s})'$ and for centered auxiliary variables, namely $t_{\mathbf{x}} = 0$.

Weight vectors satisfying $\hat{t}_{\mathbf{x}w}^T \mathbf{X}_s \mathbf{C} \mathbf{X}_s^T \hat{t}_{\mathbf{x}w} = \Psi$ lie on an ellipse centered at the origin and weight vectors satisfying $\Upsilon_s((\mathbf{a}, \mathbf{b}); \mathbf{w}_s) = r^2$ lie on a ellipse centered at $\mathbf{a}$. The largest ellipse centered at the origin in Figure 3.7 is obtained for $\Psi = \Psi_{\max}$ and the initial weight vector $\mathbf{a}$ belongs to this ellipse.

The constrained optimization problem (3.52) means that we fix $\Psi = \Psi_0$: weight vector $\mathbf{w}_s$ is located on the ellipse centered at the origin and of equation $(\hat{t}_{\mathbf{x}w} - t_{\mathbf{x}})^T \mathbf{C} (\hat{t}_{\mathbf{x}w} - t_{\mathbf{x}}) = \Psi_0$ (the small ellipse centered at the origin from Figure 3.7). To find the penalized weight vector $\mathbf{w}_s$, we grow up the ellipse centered at $\mathbf{a}$ and of equation $\Upsilon_s((\mathbf{a}, \mathbf{b}); \mathbf{w}_s) = r^2$ (the ellipse centered at $\mathbf{d}$ from Figure 3.7) until it touches the first ellipse.

The constrained optimization problem (3.53), means that we fix $r = r_0$, namely $\mathbf{w}_s$ is located on the ellipse centered at $\mathbf{a}$ and of equation $\Upsilon_s((\mathbf{a}, \mathbf{b}); \mathbf{w}_s) = r_0^2$ (the small ellipse centered at $\mathbf{a}$ from Figure 3.7). We find the ellipse contour centered at the origin and of equation $\hat{t}_{\mathbf{x}w}^T \mathbf{X}_s \mathbf{C} \mathbf{X}_s^T \hat{t}_{\mathbf{x}w} = \Psi$ as close as possible to the ellipse centered at $\mathbf{a}$ (the small ellipse centered at the origin from Figure 3.7). The penalized calibration weights $\mathbf{w}_s^{\text{pen}}$ is the vector at the first point where the ellipse contour centered at the origin touches the constraint ellipse.

In the context of empirical likelihood approach, Chen et al. [2002] considered discrepancies δ_j depending on a tuning parameter δ and they suggested also the bisection algorithm to find δ. The obtained solution minimises the error $\hat{t}_{\mathbf{x}w} - t_{\mathbf{x}}$ while satisfying the range restrictions and being as close as possible from the initial weights. Beaumont and Bocci [2008] established the link between the cost matrix $\mathbf{C}$ and the discrepancy matrix of Chen et al. [2002]. Beaumont and Bocci [2008] concluded also that "ridge calibration is better than the Chen et al. [2002] method if minimizing the distance Ψ, for a fixed matrix $\mathbf{C}$ of costs, is a desirable goal." If the discrepancies δ_j for $j = 1, \ldots, p$ are fixed first, then Rao and Singh [1997] and Rao and Singh [2009] established the link between c_j and δ_j and suggested a ridge-shrinkage method to find the penalized weights.

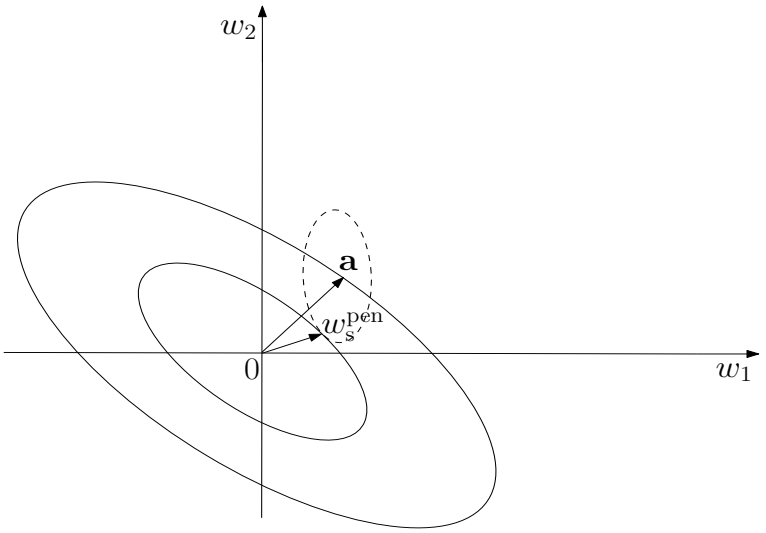


FIGURE 3.7: The small ellipse centered at the origin is the set of points in the (w_1, w_2)-plane where the sum of squares of residuals is equal to $\Psi > \Psi_{\min}$ and the largest ellipse is the set of points where the sum of squares of residuals is equal to $\Psi_{\max}$. The ellipse centered at $\mathtt{a}$ is the set of points situated at distance r.

3.2.2 Partial penalization

Assuming that $0 < c_j < \infty$ for all $j = 1, \ldots, p$, one can remark from (3.49), that for $\lambda \to 0$, we obtain $\mathbf{w}_s^{\text{pen}}((\mathbf{a}, \mathbf{b}); \lambda) \to \mathbf{w}_s^{\text{opt}}(\mathbf{a}, \mathbf{b})$ meaning that the constraints are exactly satisfied. The estimator $\hat{t}_{yw}^{\text{pen}}((\mathbf{a}, \mathbf{b}); \lambda)$ becomes the calibration estimator $\hat{t}_{yw}^{\text{cal}}$ for $(\mathbf{a}, \mathbf{b}) = (\mathbf{d}, \mathbf{v})$, and the BLU predictor $\hat{t}_{yw}^{\text{blu}}$ for $(\mathbf{a}, \mathbf{b}) = (\mathbf{1}_s, \mathbf{v})$. From a practical point of view, weights may be derived by taking $\lambda = 0$ in (3.50).
On the other side, for $\lambda \to \infty$, we obtain $\mathbf{w}_s^{\text{pen}}((\mathbf{a}, \mathbf{b}); \lambda) \to \mathbf{a}$ meaning that the auxiliary information is not taken into account. The estimator $\hat{t}_{yw}^{\text{pen}}((\mathbf{a}, \mathbf{b}); \lambda)$ becomes $\hat{t}_{yd} = \sum_{k \in s} d_k y_k$ for $\mathbf{a} = \mathbf{d}$ or $\hat{t}_{y1} = \sum_{k \in s} y_k$ for $\mathbf{a} = \mathbf{1}_s$. Again, from a practical point of view, weights may be derived by taking a very large value of λ in (3.50).

For fixed $\lambda > 0$, we can choose to penalize differently the p constraints. This is done by means of the costs c_j. If some auxiliary variables are not relevant, we can discard them by putting $\lambda^{-1} c_j = 0$ in (3.49). Rao and Singh [1997] gave a proof of this result. On the opposite situation, if some auxiliary variables are important for the survey study, then their population totals should be estimated exactly. This is accomplished by considering that $\lambda^{-1} c_j \to \infty$ in (3.49) for the jth constraint to be estimated exactly (see Beaumont and Bocci [2008] for a proof). From a practical point of view, one can put a very large value of c_j^{-1} for the nonbinding constraint in (3.50). The largest is the value of c_j^{-1}, the less is binding the jth constraint. With $c_j^{-1} = 0$ in (3.50), the corresponding j-th constraint is binding (*e.g.* its total t_{X_j} should be estimated exactly).

This kind of partial penalization was first suggested by Bardsley and Chambers [1984]. Let

write the data matrix $\mathbf{X} = (\widetilde{\mathbf{X}}_1, \widetilde{\mathbf{X}}_2)$ with $\widetilde{\mathbf{X}}_1$ containing the q binding variables, such as socio-demographic variables, and $\widetilde{\mathbf{X}}_2$ the other $p-q$ variables. The partial penalized weights are easily obtained from (3.50) for the inverse cost matrix $\mathbf{C}$ given by

$$\mathbf{C}^{-1} = \begin{pmatrix} \mathbf{0}_{(q,q)} & \mathbf{0}_{(q,p-q)} \\ \mathbf{0}_{(p-q,p)} & \mathbf{C}_2^{-1} \end{pmatrix}, \tag{3.54}$$

where $\mathbf{C}_2 = \text{diag}(c_j)_{j=q+1}^{p}$ is the diagonal cost matrix of size $(p-q)\times(p-q)$ associated with $\widetilde{\mathbf{X}}_2$. Let call $\mathbf{w}_s^{\text{ppen}}$ the partial penalized weights. Park and Yang [2008] and Guggemos and Tillé [2010] were concerned by the same issue but they considered a different optimization problem:

$$\begin{aligned} \mathbf{w}_s^{\text{ppen}} &= \text{argmin}_{\mathbf{w}} \Upsilon_s((\mathbf{a},\mathbf{b});\mathbf{w}) + \lambda(\hat{t}_{\mathbf{x}_2 w} - t_{\mathbf{x}_2})^T \mathbf{C}_2 (\hat{t}_{\mathbf{x}_2} - t_{\mathbf{x}_2 w}) \quad \text{subject to} \\ \hat{t}_{\mathbf{x}_1 w} &= t_{\mathbf{x}_1} \end{aligned} \tag{3.55}$$

where the subscript 1 refers to $\widetilde{\mathbf{X}}_1$ and subscript 2 to $\widetilde{\mathbf{X}}_2$. The above optimization problem states clearly that exact calibration on variables from $\widetilde{\mathbf{X}}_1$ is realized while this fact should be proven for the Bardsley and Chambers's optimization problem. But, the weights solution of (3.56) derived by Guggemos and Tillé [2010] are rather complicated and they are not reported here. Using matrix calculus, Goga and Shehzad [2014a] show in a recent paper that the optimization problem of Bardsley and Chambers's with $\mathbf{C}^{-1}$ given by (3.54) and the Guggemos and Tillé's optimization problem from (3.56) lead to the same vector of partially penalized weights. Goga and Shehzad [2014a] suggest a kind of combined strategy: the penalized weights minimize the Guggemos and Tillé's constrained optimization problem (3.56) and their expression are given by relation (3.50) with $\mathbf{C}^{-1}$ given by (3.54) obtained with Bardsley and Chambers's optimization problem. Guggemos and Tillé [2010] use the Fisher scoring algorithm to compute λ.

We mention also that Park and Yang [2008] were concerned by the estimation of the mean $\overline{y}_U = \sum_{k\in U} y_k/N$ by using a weighted estimator with weights summing up to unity and as close as possible to the Hájek [1971]'s weights. This is why, their estimator is obtained for $\mathbf{a} = (d_i/\sum_{k\in s} d_k)_{i\in s}$.

Asymptotic properties

Both model-based and design-based penalized estimators are biased under the model ξ. Bardsley and Chambers [1984] claimed that the model-based ridge estimator $\hat{t}_{yw}^{\text{pen}}((\mathbf{1}_s,\mathbf{v});\lambda)$ has smaller prediction variance than the BLU predictor $\hat{t}_{yw}^{\text{blu}}$ but they did not prove it. Using un intermediate result and arguments of Vinod and Ullah [1981], Bellhouse [1987] shows that under the model ξ with $v_k = \sigma^2$ for all k and the cost matrix $\mathbf{C}^{-1} = \kappa\mathbf{I}_p$ with κ satisfying $0 < \kappa < \frac{2\sigma^2}{\boldsymbol{\beta}^T\boldsymbol{\beta}}$:

$$\mathbb{E}_\xi\mathbb{E}_p(\hat{t}_{yw}^{\text{pen}}((\mathbf{1}_s,\mathbf{v});\lambda) - t_y)^2 < \mathbb{E}_\xi\mathbb{E}_p(\hat{t}_{yw}^{\text{blu}} - t_y)^2.$$

A similar result was proved by Theobald [1974] for classical ridge-regression. A necessary and sufficient condition for the ridge estimator $\hat{t}_{yw}^{\text{pen}}((\mathbf{1}_s,\mathbf{v});\lambda)$ to be more efficient than the BLU

predictor $\hat{t}_{yw}^{\text{blu}}$ is $0 < \kappa < 2/(-\min(0, \psi))$, where ψ is the minimum eigenvalue of $(\mathbf{X}_s^T\mathbf{X}_s)^{-1} - \frac{\boldsymbol{\beta}^T\boldsymbol{\beta}}{\sigma^2}$ (Swindel and Chapman [1973]). In a design-based setting, Park and Yang [2008] determined the optimal values of the penalty matrix C_2 from (3.54).

In a design-based framework, the concern is about the asymptotic properties of $\hat{t}_{yw}^{\text{pen}}((\mathbf{d}, \mathbf{v}); \lambda)$ with respect to the sampling design p. Rao and Singh [1997] stated that "an important requirement while relaxing benchmark constraints is that for given tolerance levels, the calibration method should ensure design consistency like the generalized regression method." Under broad assumptions, the design-based ridge estimator $\hat{\boldsymbol{\beta}}_{\mathbf{x}}((\mathbf{d}, \mathbf{v}); \lambda)$ tends in probability to

$$\tilde{\boldsymbol{\beta}}_{\mathbf{x}}(\mathbf{v}; \lambda) = \left(\sum_{k \in U} v_k^{-1} \mathbf{x}_k \mathbf{x}_k^T + \lambda \mathbf{C}^{-1} \right)^{-1} \sum_{k \in U} v_k^{-1} \mathbf{x}_k y_k,$$

and the ridge estimator $\hat{t}_{yw}^{\text{pen}}((\mathbf{d}, \mathbf{v}); \lambda)$ is asymptotically equivalent to

$$\tilde{t}_{y,\mathbf{x}}^{\text{diff}}(\lambda) \quad = \quad \hat{t}_{yd} - \left(\hat{t}_{\mathbf{x}d} - t_{\mathbf{x}}\right)^T \tilde{\boldsymbol{\beta}}_{\mathbf{x}}(\mathbf{v}; \lambda)$$

This implies that $\hat{t}_{yw}^{\text{pen}}((\mathbf{d}, \mathbf{v}); \lambda)$ is asymptotically design-unbiased and consistent under the broad assumptions used for the design-unbiasedness and consistency of the HT estimators $\hat{t}_{yd}$ and $\hat{t}_{\mathbf{x}d}$ (Rao and Singh [1997]; Théberge [2000]). Then, the asymptotic variance under the sampling design of $\hat{t}_{yw}^{\text{pen}}((\mathbf{d}, \mathbf{v}); \lambda)$ is the HT variance applied to the residuals $y_k - \mathbf{x}_k^T \tilde{\boldsymbol{\beta}}_{\mathbf{x}}(\mathbf{v}; \lambda)$.

3.2.3 Calibration on principal components

Cardot et al. [2014c] suggest another class of penalized calibration estimators which is based on principal component analysis (PCA). In multivariate statistics, PCA is one of the most popular techniques for reducing the dimension of a set of quantitative variables (see *e.g.* Jolliffe [2002]) by extracting most of the variability of the data by projection on a low dimension space. Principal component analysis consists in transforming the initial data set into a new set of a few uncorrelated synthetic variables, called principal components (PC), which are linear combinations of the initial variables with the largest variance. The principal components are "naturally" ordered, with respect to their contribution to the total variance of the data, and the reduction of the dimension is then realized by taking only the first few PCs. PCA is particularly useful when the correlation among the variables in the dataset is strong.
The method suggested by Goga et al. [2011], Shehzad [2012] and Cardot et al. [2014c] consists in reducing the number of auxiliary variables by considering a small number of PC's and by performing calibration on these new synthetic variables. The method is easy to put into practice with a software computer used for performing calibration, such as CALMAR used at the French National Statistical Institut (Insee), since with centered data, these new calibration variables are also centered.

Complete auxiliary information

We suppose without loss of generality that the auxiliary variables are centered, namely $t_{\mathbf{x}}/N = 0$.

We suppose also that the auxiliary information is complete, namely the p-dimensional vector $\mathbf{x}_k$ is known for all the units $k \in U$.
Let $\mathbf{X}$ be the $N \times p$ matrix having $\mathbf{x}_k^T, k \in U$ as rows. The variance-covariance matrix of the initial variables $\mathbf{X}_1, \ldots, \mathbf{X}_p$ is given by $N^{-1}\mathbf{X}^T\mathbf{X}$. Let $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$ be the eigenvalues of $N^{-1}\mathbf{X}^T\mathbf{X}$ considered in decreasing order with $\mathbf{v}_1, \ldots, \mathbf{v}_p$ the corresponding orthonormal eigenvectors,

$$\frac{1}{N}\mathbf{X}^T\mathbf{X}\mathbf{v}_j = \lambda_j \mathbf{v}_j, \quad j = 1, \ldots, p. \tag{3.56}$$

For $j = 1, \ldots, p$, the jth principal component, denoted by $\mathbf{Z}_j$, is defined as follows

$$\mathbf{Z}_j = \mathbf{X}\mathbf{v}_j = (z_{kj})_{k \in U}. \tag{3.57}$$

We will only consider the first r (with $r < p$) principal components, $\mathbf{Z}_1, \ldots, \mathbf{Z}_r$, which correspond to the r largest eigenvalues. In a survey sampling framework, the goal is not to give interpretations of these new variables $\mathbf{Z}_1, \ldots, \mathbf{Z}_r$ as it is the custom in PCA. These variables serve as a tool to obtain calibration weights which are more stable than the calibration weights that would have been obtained with the whole set of auxiliary variables.

More exactly, we want to find the principal component (PC) calibration estimator

$$\hat{t}_{yw}^{\mathrm{pc}}(r) = \sum_{k \in s} w_{ks}^{\mathrm{pc}}(r) y_k,$$

where the PC calibration weight vector $\mathbf{w}_s^{\mathrm{pc}}(r) = (w_{ks}^{\mathrm{pc}}(r))_{k \in s}$ is the solution of the optimization problem (1.12) and subject to

$$\sum_{k \in s} w_{ks}^{\mathrm{pc}}(r) \mathbf{z}_{kr} = \sum_{k \in U} \mathbf{z}_{kr},$$

where $\mathbf{z}_{kr}^T = (z_{k1}, \ldots, z_{kr})$ is the vector containing the values of the r first PCs computed for the k-th individual. Considering the chi-square distance function Υ_s, defined in (1.14), the PC calibration weights $w_{ks}^{\mathrm{pc}}(r)$'s are given by

$$w_{ks}^{\mathrm{pc}}(r) = d_k - d_k \mathbf{z}_{kr}^T \left(\sum_{k \in s} d_k \mathbf{z}_{kr} \mathbf{z}_{kr}^T \right)^{-1} (\hat{t}_{\mathbf{z}_r d} - t_{\mathbf{z}_r}),$$

where $\hat{t}_{\mathbf{z}_r d} = \sum_{k \in s} d_k \mathbf{z}_{kr}$ is the HT estimator of the total $t_{\mathbf{z}_r} = (0, \ldots, 0)$ since we have supposed that the original variables have mean zero so that the principal components are also centered variables. The total t_y is again estimated by a GREG-type estimator which uses $\mathbf{Z}_1, \ldots, \mathbf{Z}_r$ as auxiliary variables

$$\hat{t}_{yw}^{\mathrm{pc}}(r) = \hat{t}_{yd} - \left(\hat{t}_{\mathbf{z}_r d} - t_{\mathbf{z}_r}\right)^T \hat{\boldsymbol{\gamma}}_{\mathbf{z}}(r), \tag{3.58}$$

where $\hat{\boldsymbol{\gamma}}_{\mathbf{z}}(r)$ is given by [5] :

$$\hat{\boldsymbol{\gamma}}_{\mathbf{z}}(r) = \left(\sum_{k\in s} d_k \mathbf{z}_{kr}\mathbf{z}_{kr}^T\right)^{-1} \sum_{k\in s} d_k \mathbf{z}_{kr} y_k. \tag{3.59}$$

The PC calibration estimator $\hat{t}_{yw}^{\text{pc}}(r)$ depends on the number r of the PC variables and we can note that if $r = 0$, that is to say if we do not take auxiliary information into account, then $\hat{t}_{yw}^{\text{pc}}(0)$ is simply the HT estimator (or the Hájek estimator if the intercept term is included in the model) whereas if $r = p$, we get the calibration estimator which takes account all the auxiliary variables.

Calibration on second moment of PC's

Remark that, with complete auxiliary information, the totals of squares of the PCs are known since

$$\frac{1}{N}\mathbf{Z}_j^T\mathbf{Z}_j = \frac{1}{N}\sum_{k\in U} z_{kj}^2 = \lambda_j, \quad \text{for all} \quad j = 1, \ldots, p.$$

Cardot et al. [2014c] suggested considering additional calibration on the second moment of these PCs. The estimator derived in this way is expected to perform better than the estimator calibrated only on the first moment of the principal components (Särndal [2007], Ren [2000]). Nevertheless, calibration on the second moment of the PCs requires r additional calibration constraints.

A model-assisted point of view

As the ridge regression (Hoerl and Kennard [1970]), the principal component regression (PCR) is a biased estimation method of the coefficient of regression (Jolliffe [2002]) suggested to overcome the problem of multicollinearity between the regressors.

Consider again the superpopulation model given in (1.19) and let $\mathbf{G} = (\mathbf{v}_1, \ldots, \mathbf{v}_p)$. Then, the model ξ may be written in the equivalent form

$$\xi : \quad y_k = \mathbf{z}_k^T\boldsymbol{\gamma} + \varepsilon_k,$$

where $\boldsymbol{\gamma} = \mathbf{G}^T\boldsymbol{\beta}$ and $\mathbf{z}_k^T = (z_{k1}, \ldots, z_{kp})$ with z_{kj} the value of $\mathbf{Z}_j$ for the kth individual. Principal component regression consists in using a reduced model which uses as predictors only the r first PCs, $\mathbf{Z}_1, \ldots, \mathbf{Z}_r$, as follows

$$\xi_r : \quad y_k = \mathbf{z}_{kr}^T\boldsymbol{\gamma}(r) + \varepsilon_{kr}, \tag{3.60}$$

[5] Throughout this section, $q_k = 1$ for all $k \in U$ and the calibration approach is used; so, for ease of notation, the arguments $(\mathbf{q}, \mathbf{v})$ are dropped off from the expression of the estimators of the regression coefficient.

where $\boldsymbol{\gamma}(r)$ is a vector of r elements that are a subset of elements of $\boldsymbol{\gamma}$ and ε_{kr} is the appropriate error term of zero mean. Using least squares to estimate $\boldsymbol{\gamma}(r)$, we obtain

$$\tilde{\boldsymbol{\gamma}}_{\mathbf{z}}(r) = (\sum_{k\in U} \mathbf{z}_{kr}\mathbf{z}_{kr}^T)^{-1} \sum_{k\in U} \mathbf{z}_{kr} y_k \tag{3.61}$$

which in turn can be estimated, on a sample s, by the design-based estimator $\hat{\boldsymbol{\gamma}}_{\mathbf{z}}(r)$ given by (3.59). We can see now that the PC calibration estimator given in (3.58) is in fact a GREG-type estimator assisted by the reduced model ξ_r from (3.60). Note also that since the PC's are centered and uncorrelated, the matrix $\sum_{k\in U} \mathbf{z}_{kr}\mathbf{z}_{kr}^T$ is diagonal, with diagonal elements $\lambda_1 N, \ldots, \lambda_r N$.

When there is strong multicollinearity among the auxiliary variables, than the well-known ordinary least squares estimator of $\boldsymbol{\beta}$

$$\tilde{\boldsymbol{\beta}}_{\mathbf{x}} = \left(\frac{1}{N}\sum_{k\in U} \mathbf{x}_k\mathbf{x}_k^T\right)^{-1} \frac{1}{N}\sum_{k\in U} \mathbf{x}_k y_k,$$

is very sensitive to small changes in $\mathbf{x}_k$ and y_k and it has a very large variance (Hoerl and Kennard [1970]). To understand better how small eigenvalues affect the efficiency of $\tilde{\boldsymbol{\beta}}_{\mathbf{x}}$, Gunst and Mason [1977] write this estimator as follows:

$$\tilde{\boldsymbol{\beta}}_{\mathbf{x}} = \sum_{j=1}^{p} \frac{1}{\lambda_j}\mathbf{v}_j^T \left(\frac{1}{N}\sum_{k\in U} \mathbf{x}_k y_k\right) \mathbf{v}_j.$$

Approximating the covariance matrix $\frac{1}{N}\mathbf{X}^T\mathbf{X}$ by the rank r matrix $\left(\sum_{j=1}^{r} \lambda_j \mathbf{v}_j\mathbf{v}_j^T\right)$ leads to consider the following approximation to the regression estimator that is based on the r first principal components:

$$\tilde{\boldsymbol{\beta}}_{\mathbf{x}}^{\mathrm{pc}}(r) = \mathbf{G}_r\tilde{\boldsymbol{\gamma}}_{\mathbf{z}}(r) = \sum_{j=1}^{r} \frac{1}{\lambda_j}\mathbf{v}_j^T \left(\frac{1}{N}\sum_{k\in U} \mathbf{x}_k y_k\right) \mathbf{v}_j, \tag{3.62}$$

where $\mathbf{G}_r = (\mathbf{v}_1, \ldots, \mathbf{v}_r)$. This means that $\tilde{\boldsymbol{\beta}}_{\mathbf{x}}^{\mathrm{pc}}(r)$ is obtained by subtracting from $\tilde{\boldsymbol{\beta}}_{\mathbf{x}}$ the part of the data that belongs to the $p-r$ dimensional space with the smallest variance and by performing the regression in the r dimensional space that contain most of the variability of the data. Note that ridge-regression (Hoerl and Kennard [1970]) which is an alternative way of dealing with the multicollinearity issue, consists in adding a positive term λ to all eigenvalues $\lambda_j, j = 1, \ldots, p$. Both the ridge regression estimator and the principal components estimator $\tilde{\boldsymbol{\beta}}_{\mathbf{x}}^{\mathrm{pc}}(r)$ are biased for $\boldsymbol{\beta}$ under model ξ (Gunst and Mason [1977]).

The PC regression estimator $\tilde{\boldsymbol{\beta}}_{\mathbf{x}}^{\mathrm{pc}}(r)$ is estimated under the sampling design by

$$\hat{\boldsymbol{\beta}}_{\mathbf{x}}^{\mathrm{pc}}(r) = \mathbf{G}_r\hat{\boldsymbol{\gamma}}_{\mathbf{z}}(r), \tag{3.63}$$

where $\hat{\boldsymbol{\gamma}}_{\mathbf{z}}(r)$ is given by (3.59). Using relation (3.63) and the fact that $\mathbf{Z}_j = \mathbf{X}\mathbf{v}_j$, we obtain that

$$\left(\hat{t}_{\mathbf{z}_r d} - t_{\mathbf{z}_r}\right)^T \hat{\boldsymbol{\gamma}}_{\mathbf{z}}(r) = \left(\hat{t}_{\mathbf{x}d} - t_{\mathbf{x}}\right)^T \hat{\boldsymbol{\beta}}_{\mathbf{x}}^{\mathrm{pc}}(r). \tag{3.64}$$

So, $\hat{t}_{yw}^{\mathrm{pc}}(r)$ may be written in the following form

$$\hat{t}_{yw}^{\mathrm{pc}}(r) \quad = \quad \hat{t}_{yd} - \left(\hat{t}_{\mathbf{x}d} - t_{\mathbf{x}}\right)^T \hat{\boldsymbol{\beta}}_{\mathbf{x}}^{\mathrm{pc}}(r). \tag{3.65}$$

As a consequence, $\hat{t}_{yw}^{\mathrm{pc}}(r)$ may be seen as a GREG-type estimator assisted by the model ξ when $\boldsymbol{\beta}$ is estimated by $\hat{\boldsymbol{\beta}}_{\mathbf{x}}^{\mathrm{pc}}(r)$.

3.2.4 Partial calibration with principal components

The calibration estimator derived before does not allow to find the exact finite population totals of the initial variables $\mathbf{X}_j$, $j = 1, \ldots, p$. In practice, it is often desired to have this property satisfied for socio-demographical variables such as sex and socio-professional category. Let

$$\mathbf{X} = (\tilde{\mathbf{X}}_1, \tilde{\mathbf{X}}_2),$$

where $\tilde{\mathbf{X}}_1$ contains p_1 variables, with p_1 small, for which exact calibration is desired and $\tilde{\mathbf{X}}_2$ containing the remaining $p_2 = p - p_1$ variables. Cardot et al. [2014c] suggest to calibrate on the auxiliary variables from $\tilde{\mathbf{X}}_1$ and on the r_1 first PC's $\tilde{\mathbf{Z}}_j, j = 1, \ldots, r_1$ of $(\mathbf{I}_N - \mathbf{P}_{\tilde{\mathbf{X}}_1})\tilde{\mathbf{X}}_2$ with $\mathbf{P}_{\tilde{\mathbf{X}}_1} = \tilde{\mathbf{X}}_1(\tilde{\mathbf{X}}_1^T\tilde{\mathbf{X}}_1)^{-1}\tilde{\mathbf{X}}_1^T$. The calibration variables are $(\tilde{\mathbf{X}}_1, \tilde{\mathbf{Z}}_1, \ldots, \tilde{\mathbf{Z}}_{r_1})$ of zero totals and the partial principal component (PPC) calibration estimator of t_y is

$$\hat{t}_{yw}^{\mathrm{ppc}}(r) = \sum_{k \in s} w_{ks}^{\mathrm{ppc}}(r) y_k,$$

where the PPC calibration weights $w_{ks}^{\mathrm{ppc}}(r)$'s minimize (1.12), subject to

$$\sum_{k \in s} w_{ks}^{\mathrm{ppc}}(r) \begin{pmatrix} \tilde{\mathbf{x}}_k \\ \tilde{\mathbf{z}}_{kr_1} \end{pmatrix} = \sum_{k \in U} \begin{pmatrix} \tilde{\mathbf{x}}_k \\ \tilde{\mathbf{z}}_{kr_1} \end{pmatrix},$$

where $\tilde{\mathbf{x}}_k = (\tilde{x}_{k1}, \ldots, \tilde{x}_{kp_1})$ is the vector of the values of variables contained in $\tilde{\mathbf{X}}_1$ and $\mathbf{z}_{kr_1}^T = (\tilde{z}_{k1}, \ldots, \tilde{z}_{kr_1})$ is the vector of the values of $\tilde{\mathbf{Z}}_1, \ldots, \tilde{\mathbf{Z}}_r$ recorded for the k-th individual. Breidt and Chauvet [2012] use a similar technique at the sampling stage by considering penalized balanced sampling.

3.2.5 Calibration on estimated principal components

Cardot et al. [2014c] suggest calibration on estimated principal components to deal with the situation when the auxiliary information is not complete, namely we know the totals of $\mathcal{X}_j$ only.

As a consequence, the eigenvalues and eigenvectors of the variance-covariance matrix can not be computed.

Let $\hat{\mathbf{\Gamma}}$ be the design-based estimator of the variance-covariance matrix $\mathbf{\Gamma} = \frac{1}{N}\mathbf{X}^T\mathbf{X}$ given by

$$\hat{\mathbf{\Gamma}} = \frac{1}{\hat{N}}\sum_{k\in s}\frac{1}{\pi_k}(\mathbf{x}_k - \hat{\overline{\mathbf{X}}})(\mathbf{x}_k - \hat{\overline{\mathbf{X}}})^T = \frac{1}{\hat{N}}\sum_{k\in s}\frac{1}{\pi_k}\mathbf{x}_k\mathbf{x}_k^T - \hat{\overline{\mathbf{X}}}\hat{\overline{\mathbf{X}}}^T \tag{3.66}$$

where $\hat{N} = \sum_{k\in s}\frac{1}{\pi_k}$ and $\hat{\overline{\mathbf{X}}} = \frac{1}{\hat{N}}\sum_{k\in s}\frac{1}{\pi_k}\mathbf{x}_k$. Let $\hat{\lambda}_1 \geq \ldots \geq \hat{\lambda}_p$ be the eigenvalues of $\hat{\mathbf{\Gamma}}$ considered in decreasing order with $\hat{\mathbf{v}}_1, \ldots, \hat{\mathbf{v}}_p$ the corresponding eigenvectors. Cardot et al. [2014c] suggest estimating the unknown $\mathbf{Z}_j$ given in (3.57) as follows

$$\hat{\mathbf{Z}}_j = \mathbf{X}\hat{\mathbf{v}}_j.$$

Note that $\hat{\mathbf{Z}}_j = (\hat{z}_{kj})_{k\in U}$ is known only for the sampled units, but its population total $t_{\hat{\mathbf{Z}}_j} = \sum_{k\in U}\hat{z}_{kj}$, is equal to zero since $t_{\hat{\mathbf{Z}}_j} = t_{\mathbf{x}}^T\hat{\mathbf{v}}_j = 0, \quad j = 1, \ldots, p$. Consider now the first r estimated PC

$$\hat{\mathbf{Z}}_1, \ldots, \hat{\mathbf{Z}}_r$$

corresponding to the r largest eigenvalues $\hat{\lambda}_j$. Remark that the number of PC considered here may be different from the one considered in the section 3.2.3 but, for ease of notation, we will use the same r.

The estimated principal component (EPC) calibration estimator of t_y is

$$\hat{t}_{yw}^{\text{epc}}(r) = \sum_{k\in s} w_{ks}^{\text{epc}}(r)y_k,$$

where the EPC calibration weights w_{ks}^{EPC}'s are the solution of the following optimization problem (1.12) and subject to

$$\sum_{k\in s} w_{ks}^{\text{epc}}(r)\hat{\mathbf{z}}_{kr} = \sum_{k\in U}\hat{\mathbf{z}}_{kr},$$

where $\hat{\mathbf{z}}_{kr}^T = (\hat{z}_{k1}, \ldots, \hat{z}_{kr})$ is the vector of values of $\hat{\mathbf{Z}}_j$, $j = 1, \ldots, r$ recorded for the kth unit. With the chi-squared distance function Υ_s given by (1.14), the EPC calibration estimator for t_y is again a GREG-type estimator given by

$$\hat{t}_{yw}^{\text{epc}}(r) \quad = \quad \hat{t}_{yd} - \left(\hat{t}_{\mathbf{x}d} - t_{\mathbf{x}}\right)^T\hat{\boldsymbol{\beta}}_{\mathbf{x}}^{\text{epc}}(r) \tag{3.67}$$

where $\hat{\boldsymbol{\beta}}_{\mathbf{x}}^{\text{epc}}(r) = (\hat{\mathbf{v}}_1, \ldots, \hat{\mathbf{v}}_r)(\sum_{k\in s} d_k\hat{\mathbf{z}}_{kr}\hat{\mathbf{z}}_{kr}^T)^{-1}\sum_{k\in s} d_k\hat{\mathbf{z}}_{kr}y_k$.

Asymptotic properties

The same asymptotic framework from Isaki and Fuller [1982] is considered. Under mild assumptions on the sampling design and on the study and auxiliary variables, the estimators $\hat{t}_{yw}^{\text{pc}}(r)$ and $\hat{t}_{yw}^{\text{epc}}(r)$ are proven to be asymptotically equivalent to the generalized difference estimator

$$\tilde{t}_{y,\mathbf{x}}^{\text{diff}}(r) = \hat{t}_{yd} - \left(\hat{t}_{\mathbf{x}d} - t_{\mathbf{x}}\right)^T \tilde{\boldsymbol{\beta}}_{\mathbf{x}}^{\text{pc}}(r).$$

In particular, we obtain their consistency and ADU-ness with respect to the sampling design.

Proposition 3.11. *(Cardot et al. [2014c]) Make the assumptions (S1), (S2), (V3) and (A1). Suppose also that* $\lambda_r > \lambda_{r+1} \geq 0$. *Then,* $\hat{\boldsymbol{\gamma}}_{\mathbf{z}}(r) - \tilde{\boldsymbol{\gamma}}_{\mathbf{z}}(r) = O_p(n^{-1/2})$ *and*

$$N^{-1}(\hat{t}_{yw}^{\text{pc}}(r) - t_y) = N^{-1}\left(\tilde{t}_{y,\mathbf{x}}^{\text{diff}}(r) - t_y\right) + o_p(n^{-1/2}).$$

A similar result may be stated for the EPC calibration estimator.

Proposition 3.12. *(Cardot et al. [2014c]) Make the assumptions (S1), (S2), (V3) and (A1). Suppose also that* $\lambda_r > \lambda_{r+1} \geq 0$. *Then,* $\hat{\boldsymbol{\beta}}_{\mathbf{x}}^{\text{epc}}(r) - \tilde{\boldsymbol{\beta}}_{\mathbf{x}}^{\text{pc}}(r) = O_{\text{p}}(n^{-1/2})$ *and*

$$\frac{1}{N}(\hat{t}_{yw}^{\text{epc}}(r) - t_y) = \frac{1}{N}\left(\hat{t}_{y,\mathbf{x}}^{\text{diff}}(r) - t_y\right) + o_p(n^{-1/2}).$$

The proof follows the same lines as in Breidt and Opsomer [2000] and uses the consistency of $\hat{\lambda}_j$ and $\hat{\mathbf{v}}_j$ as proven in Cardot et al. [2010a] (relations 3.23 and 3.24). Remark that even if both estimators $\hat{t}_{yw}^{\text{pc}}(r)$ and $\hat{t}_{yw}^{\text{epc}}(r)$ have the same asymptotic variance given by

$$\mathbb{V}_p(\tilde{t}_{y,\mathbf{x}}^{\text{diff}}(r)) = \sum_{k\in U}\sum_{k\in U}(\pi_{kl} - \pi_k\pi_l)d_k d_l \left(y_k - \mathbf{x}_k^T\tilde{\boldsymbol{\beta}}_{\mathbf{x}}^{\text{pc}}(r)\right)\left(y_l - \mathbf{x}_l^T\tilde{\boldsymbol{\beta}}_{\mathbf{x}}^{\text{pc}}(r)\right) \tag{3.68}$$

the variance estimators are different since $\tilde{\boldsymbol{\beta}}_{\mathbf{x}}^{\text{pc}}(r)$ is estimated by $\hat{\boldsymbol{\beta}}_{\mathbf{x}}^{\text{pc}}(r)$ for complete information and by $\hat{\boldsymbol{\beta}}_{\mathbf{x}}^{\text{epc}}(r)$ otherwise. The asymptotic variance as well as the variance estimators may be written with respect to the principal components $\mathbf{Z}_j$ by using relation (3.64).

3.2.6 A small illustration on CER electricity data

Cardot et al. [2014c] illustrate the interest of using PC calibration on the Irish Commission for Energy Regulation (CER) Smart Metering Project that has been conducted in 2009-2010 (CER, 2011)[6]. In this project, which focuses on energy consumption and energy regulation, about 6000 smart meters have been installed in order to collect every half an hour, over a period of about two years, the electricity consumption of Irish residential and business customers.

[6] The data are available on request at the address: `http://www.ucd.ie/issda/data/commissionforenergyregulation/`

We evaluate the interest of employing reduction dimension techniques based on PCA by considering a period of 14 consecutive days and a population of $N = 6291$ smart meters (households and companies). Thus, we have for each unit k in the population $(2\times 7)\times 48 = 672$ measurement instants and we denote by $y_k(t_j), j = 1, \dots 672$ the data corresponding to unit k where $y_k(t_j)$ is the electricity consumption (in kW) associated to smart meter k at instant t_j. Our variable of interest is the total electricity consumption over the second week,

$$t_y = \sum_{k\in U} \sum_{j=336}^{672} y_k(t_j).$$

The auxiliary information is the load electricity curve of the first week. This means that we have $p = 336$ auxiliary variables, which are the consumption electricity levels at each of the $p = 336$ half hours of the first week. A sample of 5 auxiliary information curves is drawn in Figure 3.8. The condition number of $N^{-1}\mathbf{X}^T\mathbf{X}$ is 67055.78, which means that the matrix is really ill-conditioned and there are strong relationships between the calibration variables. Reducing the dimension should improve the performances of calibration.

To make comparisons we draw $I = 1000$ samples of size $n = 600$ (the sampling fraction is about 0.1) according to a simple random sampling design without replacement and we estimate the total consumption t_y over the second week with the following estimators : the HT estimator, denoted by $\hat{t}_{yd}$, the calibration estimator $\hat{t}_{yw}$ that takes account of all the $p = 336$ auxiliary variables plus the intercept term and finally, the estimated principal components calibration estimator $\hat{t}^{\text{epc}}_{yw}(r)$ that takes account of r estimated PC plus the intercept term. The dimension r plays the role of a tuning parameter.

In Figure 3.9, we represent the boxplot of the EPC weights for different values of r, the number of PCs. We clearly see that these weights have larger values and they also become more heterogeneous as the number r of principal components becomes large.

The accuracy of the estimators are evaluated by comparing their mean square errors to the mean square error of the calibration estimator $\hat{t}_{yw}$. The relative mean square error is defined as follows,

$$R(\widehat{\theta}) = \frac{\sum_{i=1}^{I}(\widehat{\theta}^{(i)} - t_y)^2}{\sum_{i=1}^{I}(\widehat{t}^{(i)}_{yw} - t_y)^2}, \tag{3.69}$$

where $\hat{\theta}$ is the HT estimator $\hat{t}_{yd}$ or the EPC calibration estimator $\hat{t}^{\text{epc}}_{yw}(r)$. The HT estimator conducts very bad, $R(\hat{t}_{yd}) = 23.3$. We computed this ratio for several values of r, (see Figure 3.10), starting from $r = 1$ to $r = 336$ which leads to the calibration estimator $\hat{t}_{yw}$ using all the auxiliary information. We remark that the ratio is roughly decreasing for $r = 9$ PC and then, it is increasing up to 1 when the number of PC is also increasing. So, with only 9 PC the variance of the EPC calibration estimator is almost half of the variance of the calibration estimator based on the whole auxiliary information.

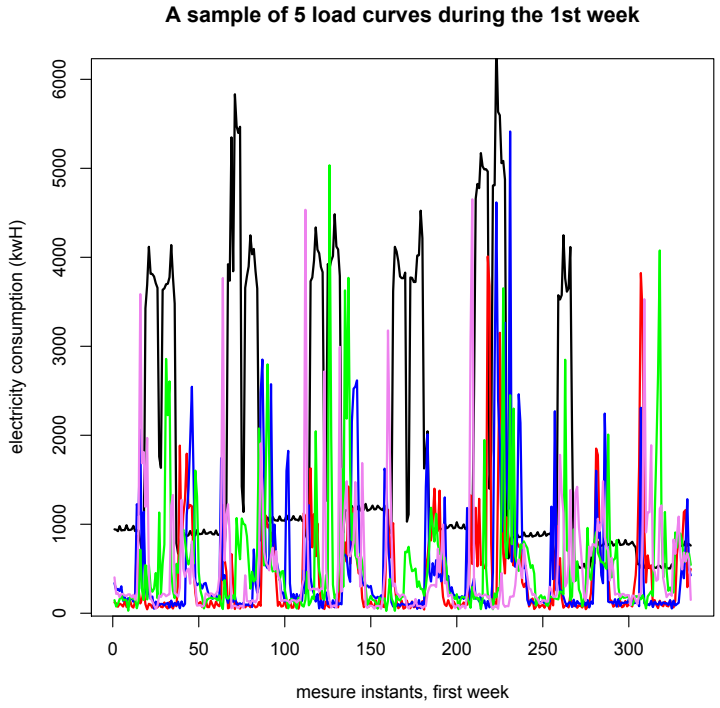


FIGURE 3.8: A sample of 5 electricity load curves observed during the first week.

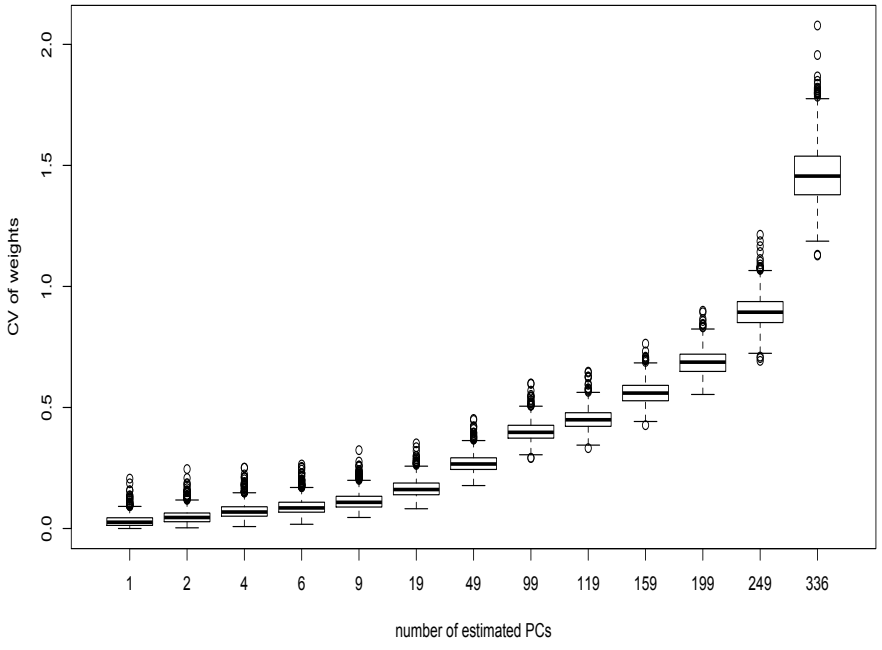


FIGURE 3.9: Distribution of the sampling weights for different values of the dimension r.

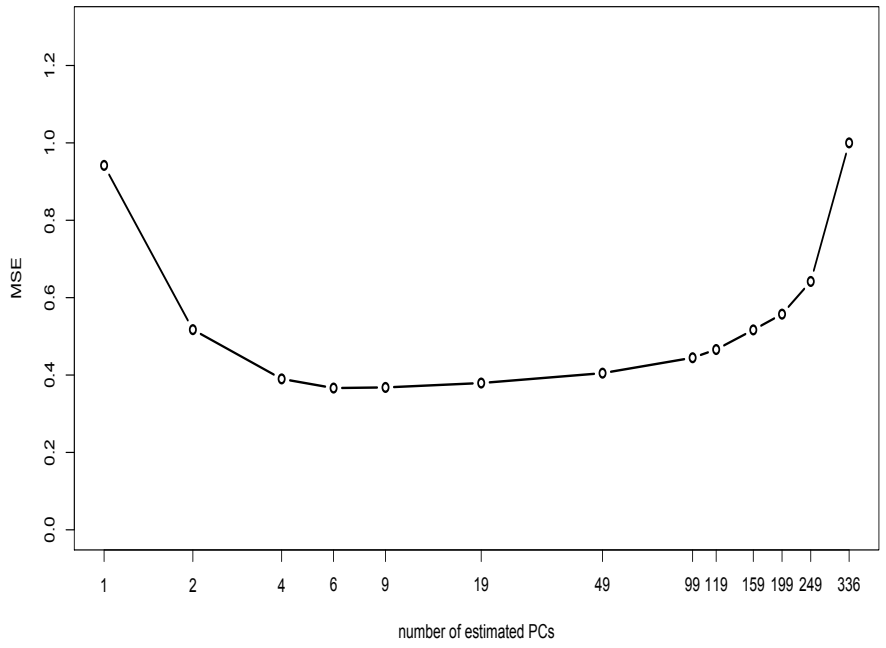


FIGURE 3.10: Evolution of the relative MSE, defined in (3.69), according to the dimension r.

3.2.7 Conclusion and perspectives

A simple dimension reduction technique based on principal components calibration has been studied in this article. It provides an effective technique for approximate calibration when the number of auxiliary variables is large and can improve significantly the estimation compared to calibration on the whole set of initial auxiliary variables. Furthermore this simple technique can also be modified so that calibration can be exact for a set of a few important auxiliary variables. We have also noted in the previous Section that a bad choice of the number of principal components which are used as calibration variables may not have dramatic consequences. Nevertheless, finding automatic data-driven procedures that could help in choosing a reasonable values for the dimension r is of real interest.

From a more theoretical point of view, it would be interesting to examine what happens when the number p of auxiliary variables is also allowed to tend to infinity when the sample size grows. Different situations about the asymptotic behavior of the smallest eigenvalue of matrix $\frac{1}{Np}\mathbf{X}^T\mathbf{X}$ may be distinguished. According to the fact that it tends to zero or not, different conclusions on how the number of principal components should be chosen may be drawn. This difficult problem is related to functional data analysis and inverse problems techniques.

In this work, it was supposed that the auxiliary information may be recorded for at least the sample individuals. Or, it may arrive that this information may be missing for some individuals.

Another direction for further work would be to use techniques conceived for conducting PCA with missing data and adapt them to the survey sampling setting.

Bibliography

Agarwal, G. G. and Studden, W. J. (1980). Asymptotic integrated mean square error using least squares and bias minimizing splines. *Ann. Statist.*, 8(6):1307–1325.

Agresti, A. (2002). *Categorical data analysis*. New York: Wiley, 2nd edition edition.

Bardsley, P. and Chambers, R. (1984). Multipurpose estimation from unbalanced samples. *Applied Statistics*, 33:290–299.

Beaumont, J.-F. and Bocci, C. (2008). Another look at ridge calibration. *Metron-International Journal of Statistics*, LXVI:260–262.

Beaumont, J.-F. and Bocci, C. (2009). Variance estimation when donor imputation is used to fill in missing values. *Canad. J. Statist.*, 37:400–416.

Beaumont, J.-F., Haziza, D., and Ruiz-Gazen, A. (2013). A unified approach to robust estimation in finite population sampling. *Biometrika*, 100:555–569.

Beaumont, J.-F. and Rivest, L.-P. (2009). Dealing with outliers in survey data. In Pfeffermann, D. and Rao, C., editors, *Handbook of Statistics*, volume 29A, pages 247–279. Elsevier.

Bellhouse, D. R. (1987). Model-based estimation in finite population sampling. *Journal of the American Statistical Association*, 41:260–262.

Berger, Y. (1998a). Rate of convergence to normal distribution for the Horvitz-Thompson estimator. *J. of Statistical Planning and Inference*, 67:209–226.

Berger, Y. G. (1998b). Rate of convergence for asymptotic variance of the Horvitz-Thompson estimator. *Journal of Statistical Planning and Inference*, 74:149–168.

Berger, Y. G. (2004). Variance estimation for measures of change in probability sampling. *Canad. J. Statist.*, 32:451–467.

Berger, Y. G. (2008). A note on the asymptotic equivalence of jackknife and linearization variance estimation for the Gini coefficient. *Journal of Official Statistics*, 24:541–555.

Berger, Y. G. and Skinner, C. J. (2003). Variance estimation for a low income proportion. *J. Roy. Statist. Soc. Ser. C*, 52(4):457–468.

Billingsley, P. (1968). *Convergence of Probability Measures*. John Wiley and Sons.

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51:279–292.

Binder, D. A. and Kovacevic, M. S. (1995). Estimating some measures of income inequality from survey data: an application of the estimating equations approach. *Survey Methodology*, 21:137–145.

Boistard, H., Lopuhaä, H. P., and Ruiz-Gazen, A. (2012). Approximation of rejective sampling inclusion probabilities and application to high order correlations. *Electron. J. Stat.*, 6:1967–1983.

Bosq, D. (2000). *Linear Processes in Function Spaces: Theory and Applications*, volume 149 of *Lecture notes in Statistics*. Springer Verlag, New-York.

Breidt, F., Claeskens, G., and Opsomer, J. (2005). Model-assisted estimation for complex surveys using penalized splines. *Biometrika*, 92:831–846.

Breidt, F.-J. and Chauvet, G. (2012). Penalized balanced sampling. *Biometrika*, 99:945–958.

Breidt, F.-J. and Opsomer, J.-D. (2000). Local polynomial regression estimators in survey sampling. *The Annals of Statistics*, 28(4):1023–1053.

Breidt, F.-J. and Opsomer, J. D. (2008). Endogeous post-stratification in surveys: classifying with a sample-fitted model. *The Annals of Statistics*, 36:403–427.

Breslow, N. and Wellner, J. A. (2007). Weighted likelihood for semiparametric models and two-phase stratified samples, with application to cox regression. *Scandinavian Journal of Statistics*, 34:86–102.

Brewer, K. and Hanif, M. (1983). *Sampling with unequal probabilities*. Springer-Verlag, New York.

Brewer, K., Hanif, M., and Tam, S. (1988). How nearly can model-based prediction and design-based prediction be reconciled? *Journal of the American Statistical Association*, 83:128–132.

Brown, B. (1983). Statistical use of the spatial median. *Journal of the Royal Statistical Society, B*, 45:25–30.

Burman, P. (1991). Regression function estimation from dependent observations. *Journal of Multivariate Analysis*, 36:263–279.

Campbell, C. (1980). A different view of finite population estimation. In Washington, D., editor, *Proc. Survey Res. Meth. Sect. Am. Statist. Assoc.*, pages 319–324.

Cardot, H. (2002a). Local roughness penalties for regression splines. *Comput. Statist.*, 17(1):89–102.

Cardot, H. (2002b). Spatially adaptive splines for statistical linear inverse problems. *J. Multivariate Anal.*, 81(1):100–119.

Cardot, H., Cénac, P., and Zitt, P.-A. (2013a). Efficient and fast estimation of the geometric median in hilbert spaces with an averaged stochastic gradient algorithm. *Bernoulli*, 19:18–43.

Cardot, H., Chaouch, M., Goga, C., and Labruère, C. (2008). Functional principal components analysis with survey data. In Dabo-Niang, S. and Férraty, F., editors, *Functional and Operatorial Statistics*. Springer-Verlag Heidelberg.

Cardot, H., Chaouch, M., Goga, C., and Labruère, C. (2010a). Properties of design-based functional principal components analysis. *J. of Statistical Planning and Inference*, 140:75–91.

Cardot, H., De Moliner, A., and Goga, C. (2014a). Estimating with kernel smoothers the mean of functional data in a finite population and in the presence of missing data. Preprint.

Cardot, H., Degras, D., and Josserand, E. (2013b). Confidence bands for Horvitz-Thompson estimators using sampled noisy functional data. *Bernoulli*, 19:2067–2097.

Cardot, H., Dessertaine, A., Goga, C., Josserand, E., and Lardin, P. (2013c). Comparison of different sample designs and construction of confidence bands to estimate the mean of functional data : an illustration on ellectricity consumption. *Survey Methodology*, 39(2):283–301.

Cardot, H., Dessertaine, A., and Josserand, E. (2010b). Semiparametric models with functional responses in a model assisted survey sampling setting. In Lechevallier, Y. and Saporta, G., editors, *Compstat 2010*, pages 411–420. Physica-Verlag, Springer.

Cardot, H., Goga, C., and Lardin, P. (2013d). Uniform convergence and asymptotic confidence bands for model-assisted estimators of the mean of sampled functional data. *Electronic Journal of Statistics*, 7:562–596.

Cardot, H., Goga, C., and Lardin, P. (2014b). Variance estimation and asymptotic confidence bands for the mean estimator of sampled functional data with high entropy unequal probability sampling designs. *Scandinavian J. of Statistics*, 41:516–534.

Cardot, H., Goga, C., and Shehzad, M.-A. (2014c). Calibration and partial calibration on principal components when the number of auxiliary variables is large. submitted.

Cardot, H. and Josserand, E. (2011). Horvitz-Thompson estimators for functional data: asymptotic confidence bands and optimal allocation for stratified sampling. *Biometrika*, 98:107–118.

Cassel, C., Särndal, C., and Wretman, J. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63:615–620.

Chambers, R. (1996). Robust case-weighting for multipurpose establishment surveys. *Journal of Official Statistics*, 12:3–32.

Chambers, R., Skinner, C., and Wang, S. (1999). Intelligent calibration. *Bulletin of the International Statistical Institute*, 58(2):321–324.

Chaouch, M. and Goga, C. (2010). Design-based estimation for geometric quantiles with application to outlier detection. *Computational Statistics and Data Analysis*, 54:2214–2229.

Chaouch, M. and Goga, C. (2012). Using complex surveys to estimate the L_1-median of a functional variable: application to electricity load curves. *International Statistical Review*, 80(1):40–59.

Chaudhuri, P. (1996). On a geometric notion of quantiles for multivariate data. *J. Amer. Statist. Assoc.*, 91:862–872.

Chauvet, G. (2007). *Méthods de bootstrap en population finie*. PhD thesis, Université de Rennes 2.

Chauvet, G. and Goga, C. (2013a). Bootstrap variance estimation and variable selection. in work.

Chauvet, G. and Goga, C. (2013b). Variance estimation of the Gini coefficient change: linearization versus bootstrap. In revision for *Survey Methodology*.

Chauvet, G., Goga, C., and Ruiz-Gazen, A. (2008). Estimation de la variance en présence de deux échantillons : linéarisation et bootstrap. In Guibert, P., Haziza, D., Ruiz-Gazen, A., and Tillé, Y., editors, *Méthodes de sondages*, pages 369–374. Dunod, Sciences Sup, Paris.

Chauvet, G. and Tillé, Y. (2006). A fast algorithm of balanced sampling. *Computational Statistics*, 21:53–61.

Chen, J. and Qin, J. (1993). Empirical likelihood estimation for finite populations and the effective use of auxiliary information. *Biometrika*, 80:107–116.

Chen, J. and Shao, J. (2000). Nearest neighbor imputation for survey data. *J. Official Statist.*, 16:113–132.

Chen, J., Sitter, R., and Wu, C. (2002). Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys. *Biometrika*, 89:230–237.

Chen, X.-H., Dempster, A. P., and Liu, J. S. (1994). Weighting finite population sampling to maximise entropy. *Biometrika*, 81:457–469.

Chiky, R. (2009). *Résumé de flux de données distribués*. PhD thesis, Sup Telecom, Paris.

Claeskens, G., Krivobokova, T., and Opsomer, J. D. (2009). Asymptotic properties of penalized spline estimators. *Biometrika*, 96(3):529–544.

Clark, R. and Chambers, R. (2008). Adaptive calibration for prediction of finite population totals. *Survey Methodology*, 34:163–172.

Cochran, W. (1977). *Sampling techniques.* John Wiley and sons, New York, 3rd edition.

Dauxois, J. and Pousse, A. (1976). *Les analyse factorielles en calcul des probabilités et en statistique : essai d'étude synthétique.* PhD thesis, Université Paul Sabatier, Toulouse.

Davison, A. and Hinkley, D. (1997). *Bootstrap methods and their application.* Cambridge Series in Statistical and Probabilistic Mathematics.

De Moliner, A., Cardot, H., and Goga, C. (2014). Estimation d'une courbe moyenne de consommation électrique par sondage en présence de valeurs manquantes. In *Actes de congrès de 46è Journées de Statistique, Rennes.*

Degras, D. (2011). Simultaneous confidence bands for non-parametric regression with functional data. *Statistica Sinica*, 21(4):1735–1765.

Demnati, A. and Rao, J. (2004). Linearization variance estimators for survey data. *Survey Methodology*, 30:17–26.

Deville, J. and Tillé, Y. (2004). Efficient balanced sampling: the cube algorithm. *Biometrika*, 91:893–912.

Deville, J. and Tillé, Y. (2005). Variance approximation under balanced sampling. *Journal of Statistical Planning and Inference*, 128:569–591.

Deville, J.-C. (1974). Méthodes statistiques et numériques de l'analyse harmonique. *Ann. Insee*, 15:3–104.

Deville, J.-C. (1999a). Simultaneous calibration of several surveys. In *Proceedings of Statistics Canada Symposium 99 of Statistics Canada*, pages 207–212.

Deville, J.-C. (1999b). Variance estimation for complex statistics and estimators: linearization and residual techniques. *Survey Methodology*, 25:193–203.

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

Dierckx, P. (1993). *Curves and Surface Fitting with Splines.* Oxford: Clarendon.

Dorfman, A. (2009). Inference on distribution functions and quantiles. In *Handbook of Statistics, vol. 29B.* Elsevier.

Dudley, R. (2002). *Real analysis and probability.* Cambridge studies in advanced mathematics.

Erdös, P. and Rényi, A. (1959). On the central limit theorem for samples from a finite population. *Publ. Math. Inst. Hungar. Acad. Sci.*, 4:49–61.

Escobar, E. and Berger, Y. G. (2013). A new replicate variance estimator for unequal probability sampling without replacement. *The Canadian Journal of Statistics*, 41:508–524.

Faraway, J. (1997). Regression analysis for a functional response. *Technometrics*, 39(3):254–261.

Fernholz, L. T. (1983). *Von Mises Calculus for Statistical Functionals*. Lecture notes in Statistics 19.

Ferraty, F. and Vieu, P. (2006). *Nonparametric functional data analysis. Theory and practice*. Springer Series in Statistics. Springer, New York.

Fuller, W.-A. (2002). Regression estimators for survey samples (with discussion). *Survey Methodology*, 28:5–23.

Fuller, W.-A. (2009). *Sampling Statistics*. John Wiley & Sons.

Gershunskaya, J., Jiang, J., and Lahiri, P. (2009). Resampling methods in surveys. In Pfeffermann, P. and Rao, C., editors, *Handbook of Statistics, vol. 29B*. Elsevier.

Gervini, D. (2008). Robust functional estimation using the spatial median and spherical principal components. *Biometrika*, 95:587–600.

Gini, C. (1914). Sulla misura della concentrazione e della variabilità dei caratteri. *Atti del R. Istituto Veneto di Scienze Lettere ed Arti.*

Goga, C. (2005). Réduction de la variance dans les sondages en présence d'information auxiliaire: une approche non paramétrique par splines de régression. *Canad. J. Statist.*, 33(2):163–180.

Goga, C. (2014). Improving the estimation of the functional median using survey data and B-spline modeling. in work.

Goga, C., Deville, J.-C., and Ruiz-Gazen, A. (2006). Linéarisation par la fonction d'influence pour les données issues de deux échantillons. In Lavallée, P. and Rivest, L.-P., editors, *Méthodes d'enquêtes et sondages. Pratiques européenne et nord-américaine*. Dunod, Sciences Sup, Paris.

Goga, C., Deville, J.-C., and Ruiz-Gazen, A. (2009). Use of functionals in linearization and composite estimation with application to two-sample data. *Biometrika*, 96(3):691–709.

Goga, C. and Ruiz-Gazen, A. (2014a). Efficient estimation of non-linear finite population parameters by using non-parametrics. *Journal of the Royal Statistical Society, B*, 76:113–140.

Goga, C. and Ruiz-Gazen, A. (2014b). Estimating the odds-ratio using auxiliary information. *Mathematical Population Studies (to appear).*

Goga, C. and Ruiz-Gazen, A. (2014c). Nonparametric B-spline calibration. in work.

Goga, C. and Shehzad, M.-A. (2014a). A note on partially penalized calibration. To appear in the Pakistan Journal of Statistics.

Goga, C. and Shehzad, M.-A. (2014b). Penalization in survey sampling: a unified point of view. in work.

Goga, C., Shehzad, M.-A., and Vanheuverzwyn, A. (2011). Principal component regression with survey data. Application on the French media audience. In Institute, I. S., editor, *Proceedings of the 58th World Statistical Congress, 2011, Dublin (Session CPS002)*, pages 3847–3852.

Graczyk, P. (2007). Gini coefficient: A new way to express selectivity of kinase inhibitors against a family of Gini coefficient: A new way to express selectivity of kinase inhibitors against a family of kinases. *Journal of Medicinal Chemistry*, 50:5773–5779.

Gross, S. (1980). Median estimation in sample surveys. *ASA Proceedings of Survey Research*, pages 181–184.

Groves-Kirkby, C., Denman, A., and Phillips, P. (2009). Lorenz curve and Gini coefficient: Novel tools for analysing seasonal variation of environmental radon gas. *Journal of Environmental Management*, 90:2480–2487.

Guggemos, F. and Tillé, Y. (2010). Penalized calibration in survey sampling: Design-based estimation assisted by mixed models. *J. of Statistical Planning and Inference*, 140:3199–3212.

Guillas, S. (2001). Rates of convergence of autocorrelation estimates for autoregressive Hilbertian processes. *Statistical and Probability Letters*, 55:281–291.

Gunst, R. F. and Mason, R. L. (1977). Biased estimation in regression: an evaluation using mean squared error. *Journal of the American Statistical Association*, 72:616–628.

Hahn, M. (1977). Conditions on sample-continuity and the central limit theorem. *Annals of Probability*, 5:351–360.

Hájek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Publ. Math. Inst. Hungar. Acad. Sci.*, 5:361–374.

Hájek, J. (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics*, 35:1491–1523.

Hájek, J. (1971). Comment on a paper by D. Basu. In *Foundations of Statistical Inference (eds. V.P. Godambe and D.A. Sprott)*, page 236. Toronto: Holt, Rinehart and Winston.

Hájek, J. (1981). *Sampling from a finite population.* Statistics: Textbooks and Monographs. Marcel Dekker, New York.

Hall, P. and Opsomer, J. D. (2005). Theory for penalised spline regression. *Biometrika*, 92(1):105–118.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.*, 69:383–393.

Hansen, M. and Hurvitz, W. (1943). On the theory of sampling from finite populations. *The Annals of Mathematical Statistics*, 14:333–362.

Harms, T. and Duchesne, P. (2006). On calibration estimation for quantiles. *Survey Methodology*, 32:37–52.

Haziza, D. (2009). Imputation and inference in the presence of missing data. In Pfeffermann, D. and Rao, C., editors, *Handbook of statistics*, volume 29A, pages 215–246. Elsevier.

Hidiroglou, M. A. (2001). Double sampling. *Survey Methodology*, 27:143–154.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67.

Horvitz, D. and Thompson, D. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47:663–685.

Huber, P. J. (1981). *Robust Statistics*. New-York, Wiley.

Huber, P. J. and Ronchetti, E. M. (2009). *Robust Statistics, Second Edition*. Wiley Series in Probability and Statistics.

Ilmonen, P., Oja, H., and Serfling, R. (2012). On invariant coordinate system (ics) functionals. *International Statistical Review*, 80:93–110.

Isaki, C.-T. and Fuller, W.-A. (1982). Survey design under the regression superpopulation model. *J. Amer. Statist. Assoc.*, 77:49–61.

Jayasuriya, B. and Valliant, R. (1996). An application of restricted regression estimation in a household survey. *Survey Methodology*, 22:127–137.

Jolliffe, I. T. (2002). *Principal component analysis*. Springer Series in Statistics. Springer-Verlag, New York, second edition.

Kauermann, G., Krivobokova, T., and Fahrmeir, L. (2009). Some asymptotic results on generalized penalized spline smoothing. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 71(2):487–503.

Kemperman, J. (1987). The median of a finite measure on a banach space. *In: Dodge, Y. (Ed.), Statistical Data Analysis Based on the L_1 Norm and Related Methods, North-Holland, Amesterdam*, pages 217–230.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica*, 46:33–50.

Korn, E. and Graubard, B. (1999). *Analysis of health surveys*. New York: John Wiley.

Kovacevic, M. and Binder, D. A. (1997). Variance estimation for measures of income inequality and polarization-the estimating equations approach. *Journal of Official Statistics*, 13:41–58.

Langel, M. and Tillé, Y. (2013). Variance estimation of the Gini index: revisiting a result several times published. *J. Roy. Statist. Soc. Ser. A*, 176:521–540.

Lardin-Puech, P., Cardot, H., and Goga, C. (2014). Analysing large datasets of functional data: a survey sampling point of view. To appear in the *Journal de la Soc. Franc. de Statis.*

Lisker, T. (2008). Is the Gini coefficient a stable measure on galaxy structure? *The Astrophysical Journal Supplement Series*, 179:319–325.

Lopuhaä, H. P. and Rousseeuw, P. (1987). Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. Technical Report 87-14, Faculty of Mathematics and Informatics, Delft University of Technology.

Marquart, D. W. and Snee, R. D. (1975). Ridge regression in practice. *The American Statistician*, 29:3–20.

Matei, A. and Tillé, Y. (2005). Evaluation of variance approximations and estimators in maximum entropy sampling with unequal probability and fixed sample size. *Journal of Official Statistics*, 21(4):543–570.

Montanari, G. E. and Ranalli, M. G. (2005). Nonparametric model calibration in survey sampling. *J. Amer. Statist. Assoc.*, 100:1429–1442.

Motoyama, H. and Takahashi, H. (2008). Smoothed versions of statistical functionals from a finite population. *Journal of Japan Statistical Society*, 38:399–412.

Nygard, F. and Sandström, A. (1985). The estimation of Gini and the entropy inequality parameters in finite population. *Journal of Official Statistics*, 4:399–412.

Park, M. and Yang, M. (2008). Ridge regression estimation for survey samples. *Communications in Statistics: Theory and Methods.*, 37:532–543.

Plikusas, A. (2006). Nonlinear calibration. In *Proceedings, Workshop on Survey Sampling, Venspils, Latvia. Riga: Central Statistical Bureau of Latvia.*

Qin, Y., Rao, J., and Wu, C. (2010). Empirical likelihood confidence intervals for the Gini measure of income inequality. *Economic Modelling*, 27:1429–1435.

Ramsay, J.-O. and Silverman, B.-W. (2005). *Functional Data Analysis.* Springer Series in Statistics, New York, second edition.

Rao, J. and Singh, A. C. (1997). A ridge-shrinkage method for range-restricted weight calibration in survey sampling. In *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

Rao, J. and Singh, A. C. (2009). Range restricted weight calibration for survey data using ridge regression. *Pakistan Journal of Statistics*, 25(4):371–384.

Rao, J., Yung, W., and Hidiroglou, M. A. (2002). Estimating equations for the analysis of survey data using poststratification information. *Sankhya: The Indian Journal of Statistics*, 64:364–378.

Ren, R. (2000). *Utilisation d'information auxiliaire par calage sur fonction de répartition.* PhD thesis, Université Paris IX Dauphine.

Robinson, P. M. and Särndal, C.-E. (1983). Asymptotic properties of the generalized regression estimator in probability sampling. *Sankhyā Ser. B*, 45(2):240–248.

Royall, R. M. (1976). The linear least square prediction approach to two-stage sampling. *Journal of the American Statistical Association*, 71:657–664.

Royall, R. M. and Herson, J. (1973). Robust estimation in finite populations. I. *J. Amer. Statist. Assoc.*, 68:880–889.

Ruppert, D. and Carroll, R. J. (2000). Spatially-adaptive penalties for spline fitting. *Aust. N. Z. J. Stat.*, 42:205–223.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric regression*, volume 12 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

Saegusa, T. and Wellner, J. A. (2013). Weighted likelihood estimation under two-phase sampling. *The Annals of Statistics*, 41:269–295.

Sandstrom, A., Wretman, J., and Walden, B. (1988). Variance estimators of the Gini coefficient - probability sampling. *Journal of Business and Economic Statistics*, 6:113–119.

Särndal, C.-E. (1980). On the π-inverse weighting best linear unbiased weighting in probability sampling. *Biometrika*, 67:639–650.

Särndal, C.-E. (2007). The calibration approach in survey theory and practice. *Survey Methodology*, 33:99–119.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model assisted survey sampling*. Springer Series in Statistics. Springer-Verlag, New York.

Särndal, C.-E. and Wright, R. (1984). Cosmetic form of estimators in survey sampling. *Scandinavian J. of Statistics*, 11:146–156.

Schumaker, L. L. (1981). *Spline Functions: Basic Theory*. New York: Wiley.

Sen, A.-R. (1953). On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 5:119–127.

Serfling, R. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley Series in Probability and Statistics.

Serfling, R. (2002). Quantile functions for multivariate analysis: approaches and applications. *Statistica Neerlandica*, 56:214–232.

Shao, J. (1994). L-statistics in complex survey problems. *Ann. Statist.*, 22:946–967.

Shao, J. and Wang, H. (2008). Confidence intervals based on survey data with nearest neighbor imputation. *Statist. Sinica*, 18:281–297.

Shehzad, M.-A. (2012). *Penalization and auxiliary information reduction methods in surveys.* PhD thesis, Université de Bourgogne.

Silva, P. and Skinner, C. (1997). Variable selection for regression estimation in finite populations. *Survey Methodology*, 23:23–32.

Singh, A. C. and Mohl, C. (1996). Understanding calibration estimators in survey sampling. *Journal of the American Statistical Association*, 22:107–115.

Skinner, C. and Silva, P. (1997). Variable selection for regression estimation in the presence of nonresponse. In *Proceedings of the Section on Survey Research Methods*. American Statistical Association.

Small, C. (1990). A survey of multidimensional medians. *International Statistical Review*, 58:263–277.

Staniswalis, J.-G. and Lee, J.-J. (1998). Nonparametric regression analysis of longitudinal data. *J. Amer. Statist. Assoc.*, 93:1403–1418.

Swindel, B. F. and Chapman, D. D. (1973). Good ridge estimators. Abstract booklet, Joint Statistical Meetings in New York City, 126.

Théberge, A. (1999). Extentions of calibration estimators in survey sampling. *Journal of the American Statistical Association*, 94:635–644.

Théberge, A. (2000). Calibration and restricted weights. *Survey Methodology*, 26:99–107.

Theobald, C. M. (1974). Generalizations of mean square error applied to ridge regression. *Journal of the Royal Statistical Society, B*, 36:103–106.

Thompson, M. (1997). *Theory of sample surveys.* Chapman and Hall, London.

Tillé, Y. (2006). *Sampling algorithms.* Springer Series in Statistics. Springer, New York.

Valliant, R. (2009). Model-based prediction of finite population totals. In Pfeffermann, D. and Rao, C., editors, *Handbook of Statistics, vol. 29B*. Elsevier.

Valliant, R., Dorfman, A., and Royall, R. M. (2000). *Finite Population Sampling and Inference.* Wiley Series in Probability and Statistics.

van der Vaart, A. and Wellner, J. A. (2000). *Weak Convergence and Empirical Processes. With Applications in Statistics.* Springer Verlag, New-York.

van der Vaart, A. W. (1998). *Asymptotic Statistics.* Cambridge University Press.

Vardi, Y. and Zhang, C.-H. (2000). The multivariate L_1-median and associated data depth. *Proc. Natl. Acad. Sci. USA*, 97(4):1423–1426.

Vinod, H. D. and Ullah, A. (1981). *Recent advances in regression methods.* Statistics: Textbooks and Monographs, New York: Marcel Dekker Inc. 41.

Vísek, J. A. (1979). Asymptotic distribution of simple estimate for rejective, Sampford and successive sampling. In *Contributions to Statistics*. Reidel, Dordrecht.

von Mises, R. (1947). On the asymptotic distribution of differentiable statistical functions. *Annals of Mathematical Statistics*, 18:309–348.

Wang, J. C. and Opsomer, J. D. (2011). On asymptotic normality and variance estimation for nondifferentiable survey estimators. *Biometrika*, 98(1):91–106.

Weber, A. (1909). *Uber Den Standard Der Industrien, Tubingen. English translation by C.J. Freidrich (1929). Alfred Weber's theory of location of industries.* Chicago University Press.

Weiszfeld, E. (1937). Sur le point pour lequel la somme des distances de n points donnés est minimum. *Tôhoku Mathematical Journal*, 43:355–386.

Wolter, K. (2007). *Introduction to Variance Estimation*. New York: Springer.

Wu, C. and Sitter, R. R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96:185–193.

Yates, F. and Grundy, P.-M. (1953). Selection without replacement from within strata with probability proportional to size. *J. Royal Statist. Soc., B*, 15:235–261.

Zheng, H. and Little, R. (2003). Penalized spline model-based estimation of the finite populations total from probability-proportional-to-size samples. *Journal of Official Statistics*, 19:99–117.

Zheng, H. and Little, R. (2005). Inference for the population total from propbability-proportional-to size samples based on predictions from a penalized spline nonparametric model. *Journal of Official Statistics*, 21:1–20.

Zhou, S., Shen, X., and Wolfe, D. (1998). Local asymptotics for regression splines and confidence regions. *The Annals of Statistics*, 26(5):1760–1782.