Overview of ridge regression estimators in survey sampling

Camelia Goga and Muhammad Ahmed Shehzad
IMB, Université de Bourgogne, DIJON - France
email : *camelia.goga*@u-bourgogne.fr,
$Muhammad - Ahmed.Shehzad$@u-bourgogne.fr

December 2, 2011

Abstract

The ridge regression is a biased estimation method used to circumvent the instability in the regression estimators obtained by ordinary least squares method in the presence of multicollinearity. This method has been used in survey sampling in order to cope with negative or extremely large weights resulted when a very large number of calibration or balancing constraints was imposed. In this paper, we give a review and some new interpretations of the ridge-type estimators in a survey sampling framework.

Key Words: calibration, model-based estimators, model-assisted estimators, multicollinearity, penalized regression.

1 Introduction

Regression techniques are widely used in practice due to their large and ease applicability. They are based on ordinary least squares method. Nevertheless, in presence of multicollinearity of data, this estimator can have extremely large variance even if it has the desirable property of being the minimum variance estimator in the class of linear unbiased estimators (the Gauss-Markov theorem). Biased estimators have been suggested to cope with problem and the ridge regression is one of them. *Hoerl and Kennard* (1970) suggest in a seminal paper the ridge regression and show that for suitable values of the penalty parameter, the ridge estimator has smaller mean squared error that the ordinary least squares estimator. The method has been applied in many fields. The book of *Vinod and Ullah* (1981) gives a comprehensive description on this topic as well as many examples.

In a survey sampling setting, weighted estimators using auxiliary information are built in order to give precise estimations about parameters of interest such as totals, means, ratio and so on. Usually, these weighted estimators are equivalent to regression estimators but it happens that, in

the presence of a large amount of information, the weights are very unstable, negative or very large. Moreover, data may contain many zeros or the sample sizes may be smaller than the number of auxiliary variables as for the small domains causing problems of matrix invertibility.

The paper is structured as follows. Section 2 recalls the construction of the ridge estimator for the regression coefficient as introduced by *Hoerl and Kennard,* (1970) in a classical regression setting. At this occasion, we give the equivalent interpretations of this estimator such as the constrained minimization problem and the Bayesian point of view. The ridge estimator depends on a penalty parameter that controls the trade-off between the bias and variance of the estimator. We recall briefly the ridge trace as a method to find the penalty parameter. Section 3 gives a detailed presentation of the application of ridge principle in survey sampling. This presentation includes the derivation of the penalized estimators under the model-based approach given in section 3.1 as well as under the calibration approach, section 3.2. Section 3.3 exhibits the partial calibration or balancing. When we attribute a prior on previous estimations, we may use the Bayesian interpretation to construct ridge regression type estimator. Deville (1999) considered it as a calibration on an uncertain source. We describe it in section 3.4. Finally, section 3.5 gives the statistic properties of the class of penalized estimators and we finish by concluding remarks and some further work.

2 Ridge regression

Let $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_p)$ be a $n \times p$ matrix of standardized known regressors $\mathbf{X}_i = (X_{ki})_{k=1}^n$ for all $i = 1, \ldots, p$. Consider the following linear model,

$$\mathbf{y} = \mathbf{1}_n' \boldsymbol{\beta}_0 + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{y} = (y_k)_{k=1}^n$ is the $n \times 1$ vector of observations and $\boldsymbol{\varepsilon} = (\varepsilon_k)_{k=1}^n$ is the $n \times 1$ vector of errors. We assume that $\mathbf{X}$ is a non-stochastic matrix of regressors with $\mathbf{X}'\mathbf{X}$ of full rank matrix (i.e the rank of $\mathbf{X}$ is p). We suppose also that the errors ε_k are independent with zero mean and variance $\mathrm{Var}(\varepsilon_k) = \sigma^2$ for all $k = 1, \ldots, n$.

The ordinary least squares (OLS) estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ minimizes the error sum of squares (ESS),

$$ESS = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

yielding the following estimator,

$$\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

2.1 Multicollinearity, ill-conditioning and consequences on the OLS estimator

Zero or no dependence among the explanatory variables is one of the assumptions of classical linear regression model. The subject of multicollinearity is widely referred to the situation where there is either exact or approximately exact linear relationship among the explanatory variables (*Gujarati*, 2002).

Gunst and Mason (1977) discriminate between the existence and the degree of the multicollinearity found in the auxiliary variables. They state that *the closer the linear combinations between the columns of* $\mathbf{X}$ *are to zero, the stronger are the multicolinearities and the more damaging are their effects on the least squares estimator.* It should be kept in mind while detecting the multicollinearity that the question should be of the degree/intensity of multicollinearity and not of kind of the multicollinearity. Small eigenvalues and their corresponding eigenvectors help to identify the multicollinearities. Let $\lambda_1, \ldots, \lambda_p$ be the eigenvalues of $\mathbf{X}'\mathbf{X}$ in decreasing order,

$$\lambda_{max} = \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p = \lambda_{min} > 0$$

and their corresponding eigenvectors $\mathbf{V}_1, \ldots, \mathbf{V}_p$. If we write (*Gunst and Mason*, 1977),

$$\boldsymbol{\lambda}_j = \mathbf{V}_j'\mathbf{X}'\mathbf{X}\mathbf{V}_j = (\mathbf{X}\mathbf{V}_j)'(\mathbf{X}\mathbf{V}_j),\ j = 1, ..., p$$

we obtain that for small eigenvalues λ_j of $\mathbf{X}'\mathbf{X}$,

$$(\mathbf{X}\mathbf{V}_j)'(\mathbf{X}\mathbf{V}_j) \approx 0 \ \Rightarrow \mathbf{X}\mathbf{V}_j \approx 0$$

which means that there is an approximately linear relationship between the columns of $\mathbf{X}$. The elements of the corresponding eigenvector $\mathbf{V}_j$ allow to identify the coefficients used in the linear dependency.

The multicollinearity is one form of ill-conditioning. More general, a measure of ill-conditioning is the *conditioning number* K given by $K = \sqrt{\lambda_{max}/\lambda_{min}}$. For $\lambda_{min} \to 0$, we have $K \to \infty$, and so, a large K implies an ill-conditioned matrix $\mathbf{X}$.

The multicollinearity or the ill-conditioning of $\mathbf{X}$ have serious consequences on the OLS estimator. The mean square error (MSE) of any estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is given by

$$\text{MSE}(\hat{\boldsymbol{\beta}}) = E((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}))$$

and for the OLS estimator $\hat{\boldsymbol{\beta}}_{OLS}$, it becomes

$$\text{MSE}(\hat{\boldsymbol{\beta}}_{OLS}) = \sigma^2 \text{Trace}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2 \sum_i^p \frac{1}{\lambda_i} \tag{2}$$

The above expression implies that the smaller the eigenvalues are, the greater are the variance of $\hat{\boldsymbol{\beta}}_{OLS}$ and the average value of the squared distance from $\hat{\boldsymbol{\beta}}_{OLS}$ to $\boldsymbol{\beta}$. This results in wider confidence intervals and therefore leads to accept more often the *Null Hypothesis* (i.e. the true population coefficient is zero). Moreover, in case of ill-conditionning, the OLS solution is unstable meaning that the regression coefficients are sensitive to small changes in the y or $\mathbf{X}$ data (see *Marquardt and Snee*, 1975 and *Vinod and Ullah*, 1981). *Hoerl and Kennard* (1970) discuss the case when the least square coefficients can be both too large in absolute value and incorrect with respect to sign. Roundoff errors tend to occur into least square calculations while $\mathbf{X}'\mathbf{X}$ is computed. Obviously, any error in $\mathbf{X}'\mathbf{X}$ may be maximized during the calculations of $\boldsymbol{\beta}$ or any other related computations. The danger of roundoff errors in $\mathbf{X}'\mathbf{X}$ is particularly magnified when (a) the determinant of $\mathbf{X}'\mathbf{X}$ is close to zero and/or (b) $\mathbf{X}'\mathbf{X}$ has the elements substantially different in order of magnitude.

Methods dealing with such data consist in (1) using priori information (Bayesian approach), (2) omitting highly collinear variables, (3) obtaining additional or new data and (4) using ridge regression. These methods can be used individually or together depending upon the countered situation. Our discussion however remains limited towards the ridge regression which is an important tool to deal with multicollinearity.

2.2 Definition of the ridge estimator

Ridge regression was first used by *Hoerl and Kennard* (1962) and then by *Hoerl and Kennard* (1970) as a solution to the biased estimation for nonorthogonal data problems. As a purpose to control instability linked to the least squares estimates, *Hoerl and Kennard* (1962) and *Hoerl and Kennard* (1968) suggested an alternative estimate of the regression coefficients as obtained by adding a positive constant k to the diagonal elements of the least square estimator $\hat{\boldsymbol{\beta}}_{OLS}$,

$$\hat{\boldsymbol{\beta}}_k = (\mathbf{X}'\mathbf{X} + k\mathbf{I}_p)^{-1}\mathbf{X}'\mathbf{y} \tag{3}$$

where $\mathbf{I}_p$ is the p-dimensional identity matrix. Since the constant k is arbitrary, we obtain a class of estimators $\hat{\boldsymbol{\beta}}_k$ for the regression estimator $\boldsymbol{\beta}$ rather than a unique estimator. For $k = 0$, we obtain the OLS estimator and as $k \to \infty$, $\hat{\boldsymbol{\beta}}_k \to 0$, the null vector.

The relationship between the ridge estimator and the OLS estimator is given by (*Hoerl and Kennard*, 1970),

$$\hat{\boldsymbol{\beta}}_k = (\mathbf{I}_p + k(\mathbf{X}'\mathbf{X})^{-1})^{-1}\hat{\boldsymbol{\beta}}_{OLS}.$$

Let consider again the latent eigenvalues of $(\mathbf{X}'\mathbf{X})^{-1}$, $\lambda_1, \ldots, \lambda_p$ with the corresponding eigenvectors $\mathbf{V}_1, \ldots, \mathbf{V}_p$. Hence, the OLS estimator may be written as

$$\hat{\boldsymbol{\beta}}_{OLS} = \sum_{j=1}^{p} \frac{\mathbf{V}_j'\mathbf{X}'\mathbf{y}}{\lambda_j}\mathbf{V}_j.$$

The fact of adding a small constant to the diagonal of $\mathbf{X}'\mathbf{X}$ will have as consequence the increase of its eigenvalues with the same quantity and dramatically decrease in this way the conditioning number K. So, the matrix $\mathbf{X}'\mathbf{X} + k\mathbf{I}$ has eigenvalues $\lambda_1 + k, \ldots, \lambda_p + k$ with the same eigenvectors $\mathbf{V}_1, \ldots, \mathbf{V}_p$ and the ridge estimator may be written as follows

$$\hat{\boldsymbol{\beta}}_k = \sum_{j=1}^{p} \frac{\mathbf{V}_j'\mathbf{X}'\mathbf{y}}{\lambda_j + k}\mathbf{V}_j.$$

The effect of the smallest eigenvalues may not be entirely eliminated by this estimator $\hat{\boldsymbol{\beta}}_k$ but their effect on the parameter estimates are significantly lessened. *Hoerl and Kennard* (1970) show also that for $k \neq 0$, the length of the ridge estimator $\hat{\boldsymbol{\beta}}_k$ is shorter than that of $\hat{\boldsymbol{\beta}}_{OLS}$, namely $\hat{\boldsymbol{\beta}}_k'\hat{\boldsymbol{\beta}}_k < \hat{\boldsymbol{\beta}}'_{OLS}\hat{\boldsymbol{\beta}}_{OLS}$.

Let study now the statistical properties of the ridge estimator. It is important to note that the ridge estimator $\hat{\boldsymbol{\beta}}_k$ is a biased estimator of $\boldsymbol{\beta}$ unless $k = 0$. The bias is given by

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}_k) - \boldsymbol{\beta} &= -k(\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\boldsymbol{\beta} \\ &= -k\sum_{j=1}^{p} \frac{(\mathbf{V}_j'\boldsymbol{\beta})\mathbf{V}_j}{\lambda_j + k} \end{aligned}$$

which depends on the unknown $\boldsymbol{\beta}$ and on k. It appears that $\hat{\boldsymbol{\beta}}_k$ can be used to improve the mean square error of the OLS estimator, and the magnitude of this improvement increases with an increase in spread of the eigenvalue spectrum. The ridge regression comes up with the objective of developing *stable* set of coefficients which will do a reasonable job for predicting future observations. *Conniffe and Stone* (1973) however criticized the $\hat{\boldsymbol{\beta}}_k$ since its properties depend on the non-stochastic choice of k. *Hoerl and Kennard* (1970) and *Hoerl, Kennard and Baldwin* (1975) show that an improvement

of the MSE can be obtained using $\hat{\boldsymbol{\beta}}_k$. Consider for that the MSE of $\hat{\boldsymbol{\beta}}_k$,

$$\begin{aligned} \text{MSE}(\hat{\boldsymbol{\beta}}_k) &= \sigma^2 \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j + k)^2} + k^2 \sum_{j=1}^{p} \frac{(\mathbf{V}_j' \boldsymbol{\beta})^2}{(\lambda_j + k)^2} \\ &= \text{Trace}(\text{Var}(\hat{\boldsymbol{\beta}}_k)) + (\text{Bias}(\hat{\boldsymbol{\beta}}_k))'(\text{Bias}(\hat{\boldsymbol{\beta}}_k)) \\ &= A + B \end{aligned} \tag{4}$$

Theorem 1 *(existence theorem, Hoerl and Kennard, 1970) There always exists a $k > 0$ such that*

$$MSE(\hat{\boldsymbol{\beta}}_k) < MSE(\hat{\boldsymbol{\beta}}) = \sigma^2 \sum_{j}^{p} \frac{1}{\lambda_j}.$$

Moreover, the above inequality is valid for all $0 < k < k_{max} = \frac{\sigma^2}{\alpha_{max}^2}$ where α_{max} is the largest value of $(\mathbf{V}_1, \ldots, \mathbf{V}_p)\boldsymbol{\beta}$.

The proof is based on the fact that the variance term A from relation (4) is a continuous, monotonically decreasing function of k and the squared bias term B is a continuous, monotonically increasing function of k. Their first derivatives are always non-positive and non-negative, respectively. Thus, a necessary condition to prove the theorem is to show that it always exists a $k > 0$ such that the first derivative of $\text{MSE}(\hat{\boldsymbol{\beta}}_k)$ is non-positive. This is possible for all $0 < k < \sigma^2/\alpha_{max}^2$.
However, *Theobald* (1974) criticized the MSE criteria used by *Hoerl and Kennard* (1970) and suggested a more general criteria. *Theobald* (1974) suggested minimizing the weighted mean square error (WMSE) defined by

$$WMSE(\hat{\boldsymbol{\beta}}) = E\left((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{W}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right)$$

for any non-negative definite matrix $\mathbf{W}$. For $\mathbf{W} = \mathbf{I}$ the identity matrix, we obtain the MSE criteria. He showed that minimizing the WMSE, for all non-negative definite matrix $\mathbf{W}$ is equivalent to minimizing the mean square error matrix (MMSE),

$$MMSE(\hat{\boldsymbol{\beta}}) = E\left((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\right).$$

Theorem 2 *(Theobald, 1974) The ridge estimator $\hat{\boldsymbol{\beta}}_k$ is better than $\hat{\boldsymbol{\beta}}_{OSL}$ in the sense that $MMSE(\hat{\boldsymbol{\beta}}_k) - MMSE(\hat{\boldsymbol{\beta}}_{OSL})$ is a positive-definite matrix whatever*

$$0 < k < \tilde{k}_{max} = \frac{2\sigma^2}{\boldsymbol{\beta}'\boldsymbol{\beta}}.$$

Vinod and Ullah (1981) give a different proof for the Theobald's result. A necessary and sufficient condition for $MMSE(\hat{\boldsymbol{\beta}}_k) - MMSE(\hat{\boldsymbol{\beta}}_{OSL})$ to be a positive-definite matrix given by *Swindel and Chapman*, (1973) is

$$0 < k < 2/[-\min(0, \eta)]$$

where η is the minimum eigenvalue of $(\mathbf{X}'\mathbf{X})^{-1} - (\boldsymbol{\beta}\boldsymbol{\beta}'/\sigma^2)$.

The ridge trace

We can remark that the $\hat{\boldsymbol{\beta}}_k$ depends upon the unknown parameter k which makes it impossible to calculate. *Hoerl and Kennard* (1970) suggested *the ridge trace* method to acquire the suitable value for the ridge parameter k. The *ridge trace* is a graphical tool that plots the components of the ridge regression coefficient $\hat{\boldsymbol{\beta}}_k$ versus k. It aims at finding an appropriate value of k which provides a set of coefficients $\hat{\boldsymbol{\beta}}_k$ with smaller MSE than that of the least squares solution $\hat{\boldsymbol{\beta}}_{OLS}$. The instability of the ridge trace indicates intercorrelations among regressors arising from multicollinearity and hence, the ridge solutions in the unstable region of ridge trace are more seriously affected by multicollinearity. *Marquardt and Snee* (1975) consider the ridge trace as one of the major advantages of the ridge regression. It is clear that this method do not yield a single automatic solution to the estimation problem, but rather, a family of solutions. Various rules for choosing k have been suggested in the literature (see *Vinod and Ullah*, 1981).

Conniffe and Stone (1973) criticized the ridge trace method because it may need iterations over a relatively long range of k and doubt the lack of improvement of the least squares estimator via any particular choice of k. Instead direct examination of eigenvalues may be preferred.

2.2.1 Other interpretations of the ridge regression estimator

The ridge regression estimator as a solution of a constrained minimization problem

The ridge estimator can also be seen as a solution of constrained optimization problem. *Hoerl and Kennard* (1970) consider the error sum of squares due to any estimate $\tilde{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$,

$$\begin{aligned} ESS(\tilde{\boldsymbol{\beta}}) &= (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) \\ &= ESS(\hat{\boldsymbol{\beta}}_{OLS}) + (\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{OLS})'\mathbf{X}'\mathbf{X}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{OLS}) \end{aligned}$$

which achieves its minimum only when $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{OLS}$. Relation (2) proves that on the average the

distance between $\boldsymbol{\beta}$ and $\hat{\boldsymbol{\beta}}_{OLS}$ increases with the presence of ill-conditioning in $\mathbf{X}'\mathbf{X}$ but without an appreciable increase in the error sum of squares. *Hoerl and Kennard* (1970) therefore, require finding the estimator $\tilde{\boldsymbol{\beta}}$ of minimum length that belongs to the hyperellipsoid centered at the OLS estimator and defined by the equation $\left(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{OLS}\right)' \mathbf{X}'\mathbf{X} \left(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{OLS}\right) = \Phi = \text{constant}$. Figure 1 illustrate the geometry of the ridge regression when $\boldsymbol{\beta} = (\beta_1, \beta_2)'$ is a two-dimensional parameter (*Marquart and Snee*, 1975). We can remark that $\hat{\boldsymbol{\beta}}_k$ is the shortest vector that gives a residual sum of squares as small as the Φ value anywhere on the small ellipse.

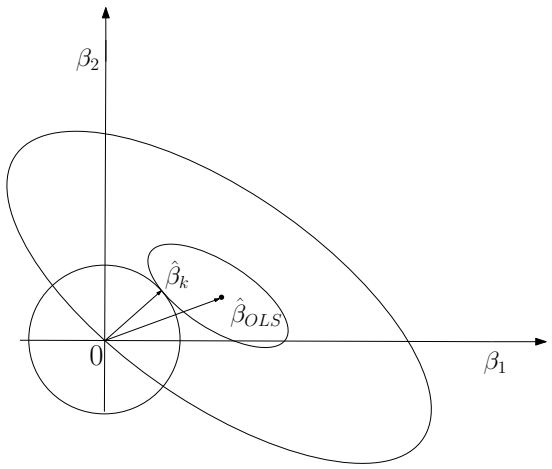

Figure 1: Geometry of ridge regression

In an equivalent way, we may minimize $ESS(\tilde{\boldsymbol{\beta}})$ for a fixed length of $\tilde{\boldsymbol{\beta}}$ say r. This is equivalent to finding the ellipse contour that is as close as possible to the circle centered in zero of ray equal to r. Using the Lagrangian principle (*Izenman*, 2008), the optimization problem may be presented as

$$\min_{\tilde{\boldsymbol{\beta}}} (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) + k(\tilde{\boldsymbol{\beta}}'\tilde{\boldsymbol{\beta}} - r^2)$$

or equivalently,

$$\min_{\tilde{\boldsymbol{\beta}}: ||\tilde{\boldsymbol{\beta}}||^2 \leq r^2} (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) \tag{5}$$

where $||\cdot||$ is the Euclidian norm. In order to attribute the same influence of the constraint from (5), it is advisable to standardize the regressors. With no-standardized variable, one may use some other norm (*Kapat and Goel*, 2010) or the generalized ridge regression when each diagonal element of $\mathbf{X}'\mathbf{X}$ is modified differently (*Hoerl and Kennard*, 1970).

Bayesian or Mixed Regression Interpretation of Ridge Coefficients

The Bayesian approach treats the parameter $\boldsymbol{\beta}$ as a random variable with a prior probability density which may be based on some subjective prior information about $\boldsymbol{\beta}$. The goal is to determine the posterior probability density of $\boldsymbol{\beta}$ which is done by combining the prior probability density with the sample information given by the likelihood function. A ridge estimator can be seen also as a Bayes estimator when $\boldsymbol{\beta}$ takes a suitable normal prior distribution with mean $\boldsymbol{\beta}_0$ and variance covariance matrix $\sigma_\beta^2\boldsymbol{\Omega}$ (*Vinod and Ullah*, 1981, *Izenman*, 2008). *Vinod and Ullah* (1981) advocate that the Bayesian interpretation of the ridge regression coefficient $\hat{\boldsymbol{\beta}}_k$ implies deriving the prior distribution of $\boldsymbol{\beta}$ for which $\hat{\boldsymbol{\beta}}_k$ is the posterior mean. They also state that the Bayesian methods imply that the posterior mean is the optimal estimator when using the MSE as expected loss. We consider the model given in (1) with the following supplementary assumptions: the errors $\boldsymbol{\epsilon}$ are normally distributed with mean zero and variance covariance matrix $\sigma^2\mathbf{I}_p$ with σ^2 a known constant and $\mathbf{I}_p$ is the p dimensional identity matrix. In other words, $\mathbf{y}$ is normally distributed $N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_p)$. We suppose that the prior normal distribution on $\boldsymbol{\beta}$ is also normal with known mean $\boldsymbol{\beta}_0$ and known variance $\sigma_\beta^2\boldsymbol{\Omega}$. The posterior density of $\boldsymbol{\beta}$ is therefore normal with mean $\boldsymbol{\beta}^*$ as follows

$$\begin{aligned} \boldsymbol{\beta}^* &= (\mathbf{X}'\mathbf{X} + \alpha\boldsymbol{\Omega}^{-1})^{-1}(\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}_{OLS} + \alpha\boldsymbol{\Omega}^{-1}\boldsymbol{\beta}_0) \quad (6) \\ &= \boldsymbol{\beta}_0 + (\mathbf{X}'\mathbf{X} + \alpha\boldsymbol{\Omega}^{-1})^{-1}\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}_0) \quad (7) \end{aligned}$$

of variance covariance matrix given by $\sigma^2\boldsymbol{\Omega}^* = \sigma^2(\mathbf{X}'\mathbf{X} + \alpha^2\boldsymbol{\Omega}^{-1})^{-1}$ and $\alpha = \sigma^2/\sigma_\beta^2$ is the ratio of the variances. Relations (6) or (7) show that if the prior information is useless, i.e. $\sigma_\beta^2 \to \infty$, then $\alpha \to 0$ and $\boldsymbol{\beta}^* = \hat{\boldsymbol{\beta}}_{OLS}$. On the other hand, for $\sigma_\beta^2 \to 0$, we have $\boldsymbol{\beta}^* = \boldsymbol{\beta}_0$. *Vinod and Ullah* (1981) remark that the estimator $\boldsymbol{\beta}^*$ given by formula (6) may be written as a weighted matrix combination of the OLS or the maximum likelihood estimator $\hat{\boldsymbol{\beta}}_{OLS}$ and the prior mean $\boldsymbol{\beta}_0$,

$$\boldsymbol{\beta}^* = \mathbf{A}\hat{\boldsymbol{\beta}}_{OLS} + (\mathbf{I}_p - \mathbf{A})\boldsymbol{\beta}_0 \quad (8)$$

where $\mathbf{A}$ is given by

$$\begin{aligned} \mathbf{A} &= \left(\mathrm{Var}(\hat{\boldsymbol{\beta}}_{OLS})^{-1} + \alpha\mathrm{Var}(\boldsymbol{\beta}_0)^{-1}\right)^{-1}\mathrm{Var}(\hat{\boldsymbol{\beta}}_{OLS})^{-1} \quad (9) \\ &= \mathbf{I}_p - \mathrm{Var}(\hat{\boldsymbol{\beta}}_{OLS})\left(\mathrm{Var}(\hat{\boldsymbol{\beta}}_{OLS}) + \alpha^{-1}\mathrm{Var}(\boldsymbol{\beta}_0)\right)^{-1} \quad (10) \end{aligned}$$

So, the normalized weights for $\hat{\boldsymbol{\beta}}_{OLS}$ and $\boldsymbol{\beta}_0$ are given by the precision matrix. The same result is obtained if one desires to compute the best estimator from the minimum variance point of view of

$\boldsymbol{\beta}$ being a matrix combination of $\hat{\boldsymbol{\beta}}_{OLS}$ and $\boldsymbol{\beta}_0$ namely,

$$\mathbf{A} = \text{argmin}_{\tilde{\mathbf{A}}} \text{Var}\left(\tilde{\mathbf{A}}\hat{\boldsymbol{\beta}}_{OLS} + (\mathbf{I}_p - \tilde{\mathbf{A}})\boldsymbol{\beta}_0\right)$$

One can remark from (6), that for $\boldsymbol{\Omega} = k^{-1}\mathbf{I}_p$ and $\boldsymbol{\beta}_0 = 0$, we get the ordinary ridge estimator $\hat{\boldsymbol{\beta}}_k$ given by (3). As *Vinod and Ullah* (1981) remarked, some Bayesians feel that this prior is unrealistic and a non null prior mean should be used but in absence of prior knowledge on $\boldsymbol{\beta}_0$, it is often conservative to shrink towards the zero vector. When a prior knowledge about $\boldsymbol{\beta}_0$ exists, then one shrinks the ridge estimator toward this known prior. Nevertheless, the drawback is that different choices of the prior lead to different ridge estimators.

It is worth mentioning that the Bayes estimator of $\boldsymbol{\beta}$ given by (7) corresponds to the estimator of the regression coefficient for the mixed regression model (*Vinod and Ullah, 1981*),

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ \boldsymbol{\beta} &= \boldsymbol{\beta}_0 + \boldsymbol{\eta} \end{aligned}$$

with $E(\boldsymbol{\eta}) = 0$ and $\text{Var}(\boldsymbol{\eta}) = \sigma_\beta\boldsymbol{\Omega}$. Conditionally on $\boldsymbol{\beta}_0$, the value of $\boldsymbol{\beta}^*$ given by (6) is then obtained by minimization with respect of $\boldsymbol{\beta}$ of

$$\frac{1}{\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \frac{1}{\sigma_\beta^2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)'\boldsymbol{\Omega}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)$$

Even if the two approaches lead to the same solution, the interpretations are different. In the Bayesian model, $\boldsymbol{\beta}$ is a random variable, whereas in the mixed regression model it is not.

We have supposed in this section that the regressors $\mathbf{X}_j$, for $j = 1, \ldots, p$ are standardized and the vector of errors $\boldsymbol{\varepsilon}$ is homoscedastic. With heteroscedasticity in the model (1), namely for $\text{Var}(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{V}$, where $\boldsymbol{V} = \text{diag}(v_1^2, v_2^2, \ldots, v_n^2)$ is the known positive definite variance-covariance matrix, one can easily transform the heteroscedastic model into a homoscedastic one by multiplying the model (1) by a matrix $\mathbf{G}$ satisfying the condition $\mathbf{G}'\mathbf{G} = \boldsymbol{V}^{-1}$. *Trenkler* (1984) discusses the performance of biased estimators in the linear regression model under the heteroscedasticity assumption.

In what concerns the standardization of the regressors, the problem is more delicate and it is not always very obvious when one should standardize the $\mathbf{X}$-variables. The standardization is not necessary for most theoretical results (*Vinod and Ullah*, 1981). However, it is advisable to

standardize data before computing the ridge estimator specially when there are large variations between regressors and they are measured in different scales. An additional advantage of the standardization is that it makes the numerical magnitude of the components of β comparable with each other. As *Kapat and Goel* (2010) remarked, different solutions for the ridge estimator $\hat{\boldsymbol{\beta}}_k$ may be obtained depending on the nature of the regressors, standardized or not, and on the constrained norm. Thus, it is important to distinguish between the solutions of these problems in order to avoid confusion.

3 Use of ridge principle in surveys

In this section, we undertake a detailed presentation of the use of ridge principle in survey sampling setting. Even if results are somewhat similar, the way they are derived is different from the classical statistics and this is mostly due the fact that in survey sampling framework, the main goal is to make inference about a function of $\mathbf{y}$ and not on the vector $\mathbf{y}$ or equivalently on the regression coefficient, $\boldsymbol{\beta}$. The simplest case, which will be considered here, is the estimation of the finite population total

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

of the variable of interest $\mathcal{Y}$ of values y_k. Here, U denotes a finite population containing N elements,

$$U = \{a_1, \ldots, a_k, \ldots, a_N\} = \{1, \ldots, k, \ldots, N\}$$

with the supposition that a population unit is identifiable uniquely by its label k. Furthermore, a sample s of size n is selected from U and the vector $\mathbf{y}$ is known only on the sample individuals. Usually, the finite population total t_y is estimated by a weighted estimator $\hat{t}_w$,

$$\hat{t}_w = \sum_{s} w_k y_k \tag{11}$$

where the weights w_k are derived usually using auxiliary information by means of a superpopulation model (model-based or model-assisted approach) or by calibration. Usually, with multipurpose surveys, weights should not depend on the study variable in order to estimate means or totals of a very large number of variables. They should also be positive and depend only on the auxiliary information. The weights necessarily should produce internally consistent estimators and if they are suitably chosen, these weights will produce estimators with smaller variance than the estimators without using the weights.

The idea of ridge estimation was used for the first time in a survey sampling framework in order to eliminate negative or extremely large weights obtained when a too restrictive condition of unbiasedness was imposed. The latter situations may cause inefficient results rather than improving the estimators. So, weights are crucial in survey sampling theory. From (11), the weights vector $\mathbf{w}_s = (w_k)_{k\in s}$ is the unknown parameter to be found. The role of $\boldsymbol{\beta}$ is taken now by $\mathbf{w}_s$. In sections 3.1 and 3.2 we give in detail the derivation of ridge weights in survey sampling as solutions of penalized optimization problems (section 2.2.1). The same estimators may be obtained by using a superpopulation linear model depending on a parameter and the class of model-based or model-assisted estimators for the finite population totals. This way of computing ridge estimators in survey sampling is the direct application of ridge principle from the classical regression described in section 2.2 and we present it below. When we attribute a prior on previous estimations, we may use the Bayesian interpretation to construct ridge regression type estimator. Deville (1999) considered it as a calibration on an uncertain source. We describe it in section 3.4.

Suppose that the relationship between the variable of interest $\mathcal{Y}$ and the auxiliary variables $\mathcal{X}_1, \ldots, \mathcal{X}_p$ is given by a superpopulation model denoted by ξ in the survey literature:

$$\xi : \quad \mathbf{y} = \mathbf{X}\beta + \boldsymbol{\varepsilon}. \tag{12}$$

The explicative variables are not standardized now. In order to distinguish the population from the sample, let $\mathbf{y} = (y_1, \ldots, y_N)'$ be a $N \times 1$ vector of and let $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_p)$ be the $N \times p$ matrix with $\mathbf{x}_k' = (X_{k1}, \ldots, X_{kp})$ as rows. The errors ε_k, for all $k \in U$ are independent one of each other, of mean zero and variance $\mathrm{Var}(\varepsilon_k) = \sigma^2 v_k^2$. Let $\mathrm{Var}_\xi(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{V}$ with $\mathbf{V} = \mathrm{diag}(v_k^2)_{k\in U}$ and v_k are positive known constants.

Some further notations are needed. Let $\mathbf{X}_s = (\mathbf{x}_k')_{k\in s}$, respectively $\mathbf{y}_s = (y_k)_{k\in s}$, be the restriction of $\mathbf{X}$, respectively of $\mathbf{y}$, on the sample s. Let also $\mathrm{Var}_\xi(\boldsymbol{\varepsilon}_s) = \sigma^2\mathbf{V}_s$ be the variance of $\boldsymbol{\varepsilon}_s$, the restriction of $\boldsymbol{\varepsilon}$ on the sample s, and $\mathrm{Var}_\xi(\boldsymbol{\varepsilon}_{\overline{s}}) = \sigma^2\mathbf{V}_{\overline{s}}$ be the variance of $\boldsymbol{\varepsilon}_{\overline{s}}$, the restriction of $\boldsymbol{\varepsilon}$ on $\overline{s} = U - s$. The population variance $\mathbf{V}$ may be written as

$$\mathbf{V} = \begin{pmatrix} \mathbf{V}_s & \mathbf{0}_{n\times(N-n)} \\ \mathbf{0}_{(N-n)\times n} & \mathbf{V}_{\overline{s}} \end{pmatrix}$$

Without auxiliary information, t_y is estimated by the *Horvitz and Thompson* (1952) estimator given by

$$\hat{t}_{y,d} = \sum_s d_k y_k = \sum_s \frac{y_k}{\pi_k} \tag{13}$$

where π_k is the first order inclusion probability of the individual $k \in U$. The model (12) is then used to improve the estimation of $\hat{t}_{y,d}$ by taking into account the auxiliary information given by $\mathbf{X}_1, \ldots, \mathbf{X}_p$.

Using the model ξ, one estimate the regression parameter $\boldsymbol{\beta}$ and after, plugs-in a *model based estimator*, abbreviated as MB below,

$$\hat{t}_{MB} = \sum_s y_k + \sum_{U-s} \mathbf{x}_k' \boldsymbol{\beta} \tag{14}$$

or in a *generalized difference estimator*, abbreviated as DIFF below,

$$\hat{t}_{DIFF} = \sum_s \frac{y_k}{\pi_k} - \left(\sum_s \frac{\mathbf{x}_k'}{\pi_k} - \sum_U \mathbf{x}_k' \right) \boldsymbol{\beta}. \tag{15}$$

This means that $\hat{t}_{MB}$ and $\hat{t}_{DIFF}$ rely on the estimation of the regression coefficient $\boldsymbol{\beta}$: best linear unbiased estimator of $\boldsymbol{\beta}$ for the MB (Royall, 1976) and the best design-based estimator of $\boldsymbol{\beta}$ for the MA (Särndal, 1980).

In a model-based setting and using the generalized least squares (GLS) estimation under the model ξ, the estimator of the regression coefficient $\boldsymbol{\beta}$ is obtained as solution of the optimization problem

$$\mathbf{(P1)}: \qquad \hat{\boldsymbol{\beta}}_{GLS,s} = \text{argmin}_{\boldsymbol{\beta}} (\mathbf{y}_s - \mathbf{X}_s \boldsymbol{\beta})' \mathbf{V}_s^{-1} (\mathbf{y}_s - \mathbf{X}_s \boldsymbol{\beta}) \tag{16}$$

yielding the estimator $\hat{\boldsymbol{\beta}}_{GLS,s} = (\mathbf{X}_s' \mathbf{V}_s^{-1} \mathbf{X}_s)^{-1} \mathbf{X}_s' \mathbf{V}_s^{-1} \mathbf{y}_s$ assuming that $(\mathbf{X}_s' \mathbf{V}_s^{-1} \mathbf{X}_s)^{-1}$ exists.The best linear unbiased estimator (BLUE) of t_y from the ξ-variance point of view is given by (Royall,1976)

$$\hat{t}_{BLUE} = \sum_s y_k + \sum_{U-s} \mathbf{x}_k' \hat{\boldsymbol{\beta}}_{GLS,s} \tag{17}$$

If the matrix $\mathbf{X}_s' \mathbf{V}_s^{-1} \mathbf{X}_s$ has eigenvalues close to zero, then it is advisable to perturb its diagonal before inverting it. We obtain the ridge estimator of $\boldsymbol{\beta}$ as follows

$$\hat{\boldsymbol{\beta}}_{MBR,s} = \left(\mathbf{X}_s' \mathbf{V}_s^{-1} \mathbf{X}_s + \mathbf{D} \right)^{-1} \mathbf{X}_s' \mathbf{V}_s^{-1} \mathbf{y}_s$$

where $\mathbf{D}$ is a $p \times p$ diagonal matrix with positive quantities on the diagonal. The ridge MB estimator given in (14) becomes

$$\hat{t}_{MBR} = \sum_s y_k + \left(\sum_{U-s} \mathbf{x}_k' \right) \hat{\boldsymbol{\beta}}_{MBR,s} \tag{18}$$

A similar reasoning may be used in a design-based approach. The design-based estimator $\hat{\boldsymbol{\beta}}_\pi$ of the regression coefficient $\boldsymbol{\beta}$ is the solution of the following optimization problem (Särndal, 1980),

$$\textbf{(P2)}: \qquad \hat{\boldsymbol{\beta}}_\pi = \text{argmin}_{\boldsymbol{\beta}} (\mathbf{y}_s - \mathbf{X}_s \boldsymbol{\beta})' \mathbf{V}_s^{-1} \boldsymbol{\Pi}_s^{-1} (\mathbf{y}_s - \mathbf{X}_s \boldsymbol{\beta})$$

where $\boldsymbol{\Pi}_s = \text{diag}(\pi_k)_{k \in s}$. This optimization problem yields the following estimator for $\boldsymbol{\beta}$,

$$\hat{\boldsymbol{\beta}}_\pi = \left(\mathbf{X}_s' \mathbf{V}_s^{-1} \boldsymbol{\Pi}_s^{-1} \mathbf{X}_s \right)^{-1} \mathbf{X}_s' \mathbf{V}_s^{-1} \boldsymbol{\Pi}_s^{-1} \mathbf{y}_s$$

and the total t_y is estimated by the well known GREG estimator (Särndal, 1980),

$$\hat{t}_{GREG} = \sum_s \frac{y_k}{\pi_k} - \left(\sum_s \frac{\mathbf{x}_k'}{\pi_k} - \sum_U \mathbf{x}_k' \right) \hat{\boldsymbol{\beta}}_\pi. \tag{19}$$

The ridge estimator of $\boldsymbol{\beta}$ becomes

$$\hat{\boldsymbol{\beta}}_{\pi,R} = \left(\mathbf{X}_s' \mathbf{V}_s^{-1} \boldsymbol{\Pi}_s^{-1} \mathbf{X}_s + \tilde{\mathbf{D}} \right)^{-1} \mathbf{X}_s' \mathbf{V}_s^{-1} \boldsymbol{\Pi}_s^{-1} \mathbf{y}_s \tag{20}$$

for some positive diagonal matrix $\tilde{\mathbf{D}}$ and plugging-in (15), we obtain

$$\hat{t}_{GREG,R} = \sum_s \frac{y_k}{\pi_k} - \left(\sum_s \frac{\mathbf{x}_k'}{\pi_k} - \sum_U \mathbf{x}_k' \right) \hat{\beta}_{\pi,R}. \tag{21}$$

The ridge estimator of $\boldsymbol{\beta}$ is ξ-biased but is more stable in presence of multicollinearity.

3.1 Ridge regression under the model-based approach

Bardsley and Chambers (1984) explored the relationship between the unbalanced samples and multicollinearity. We call a balanced sample a sample for which the following relation is satisfied

$$\sum_s w_k \mathbf{x}_k = \sum_U \mathbf{x}_k.$$

On the opposite situation, we have an unbalanced sample. As *Bardsley and Chambers* (1984) stated, in multipurpose sample surveys for which a large number of finite population totals or means are to be estimated, it is very difficult or even impossible to have a fully specified model underlying each variable. In such situations, balanced sampling may protect from model misspecification (Royall

and Herson, 1973).

In the model-based setting for unbalanced sample, exclusion of variables may increase the bias and inclusion of too many variables may result in a overspecified model and the estimates will be unstable and inefficient even if they are unbiased. Also these variables can linearly be related with each other, and hence can cause multicollinearity. The strategy suggested by *Bardsley and Chambers* (1984) is to consider as many variables as they exist but to relax the balancing condition which is in fact the unbiasedness condition of the estimator under the model. This is equivalent to deriving a biased estimator but with a smaller prediction error and this is why, it leads naturally to a ridge type estimator.

Bardsley and Chambers (1984) suggest finding the weights $\mathbf{w}_s = (w_k)_{k\in s}$ such that the weighted estimator $\hat{t}_w = \sum_s w_k y_k$ has minimum ξ-mean squared error among the class of bounded biased estimators,

$$\textbf{(P3)}: \quad \mathbf{w}_{MB,R} = \text{argmin}_{\mathbf{w}_s}(\mathbf{w}_s - \mathbf{1}_s)'\mathbf{V}_s(\mathbf{w}_s - \mathbf{1}_s) + \mathbf{B}'\mathbf{C}\mathbf{B} \tag{22}$$

where $\mathbf{B} = \sum_s w_k\mathbf{x}_k - \sum_U \mathbf{x}_k$ is the ξ-bias of $\hat{t}_w$, $\mathbf{C}$ is some diagonal cost matrix and $\mathbf{1}_s$ is the n-dimensional vectors of ones. The equality $\mathbf{B} = 0$ means that the estimator $\hat{t}_w$ is ξ-unbiased or that the design is *exactly balanced*. The optimization problem **(P3)** given by (22) means that we look for weights w_k that explain the best the vector $\mathbf{1}_s$ according to a specific metric and such that the weighted estimator is not very far away from the true total. The metric employed here uses the sample variance $\mathbf{V}_s$ as we are in the case of a model-based approach.

The minimization problem from above can also be written as

$$\textbf{(P3')}: \quad \mathbf{w}_{MB,R} = \text{argmin}_{\mathbf{w}_s, ||\mathbf{B}||_C^2 \leq r^2}(\mathbf{w}_s - \mathbf{1}_s)'\mathbf{V}_s(\mathbf{w}_s - \mathbf{1}_s)$$

for the norm $||\mathbf{B}||_C^2 = \mathbf{B}'\mathbf{C}\mathbf{B}$ which means that we penalize large values of the bias $\mathbf{B}$. Solving this minimization problem, we obtain the weights $\mathbf{w}_{MB,R}$ given by

$$\mathbf{w}_{MB,R} = \mathbf{1}_s - \mathbf{V}_s^{-1}\mathbf{X}_s\left(\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{X}_s + \mathbf{C}^{-1}\right)^{-1}(\mathbf{1}_s'\mathbf{X}_s - \mathbf{1}_U'\mathbf{X})' \tag{23}$$

leading to the *model-based ridge* estimator $\hat{t}_{w,MB} = \mathbf{w}'_{MB,R}\mathbf{y}_s$,

$$\hat{t}_{w,MB} = \sum_s y_k + \left(\sum_{U-s}\mathbf{x}'_k\right)\hat{\beta}_{w,R} \tag{24}$$

with $\hat{\beta}_{w,R} = \left(\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{X}_s + \mathbf{C}^{-1}\right)^{-1}\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{y}_s$. Remark that we obtain an estimator similar to the one obtained in (14).

We have mentioned before that the weight vector $\mathbf{w}_s$ takes in a way the place of the regression coefficient $\boldsymbol{\beta}$. We have seen in section 2.2 that introducing a penalty parameter reduces the length of $\boldsymbol{\beta}$. The same result is true for the weight vector (*Bardsley and Chambers*, 1984). Consider the particular case $\mathbf{V}_s = \mathbf{I}_n$ and $\mathbf{C}^{-1} = k\mathbf{I}_p$,

$$\begin{aligned}\mathbf{w}_{MB,R}'\mathbf{w}_{MB,R} &\simeq \mathbf{1}_U'\mathbf{X}(\mathbf{X}_s'\mathbf{X}_s + k\mathbf{I}_p)^{-1}\mathbf{X}_s'\mathbf{X}_s(\mathbf{X}_s'\mathbf{X}_s + k\mathbf{I}_p)^{-1}\mathbf{X}'\mathbf{1}_U \\ &= \sum_{j=1}^{p}\eta_i^2\frac{\lambda_i}{(\lambda_i+k)^2}\end{aligned}$$

where λ_i, $i = 1,\ldots,p$ are the eigenvalues of $\mathbf{X}_s'\mathbf{X}_s$, $\boldsymbol{\eta} = (\eta_i)_{i=1}^p = \mathbf{P}\mathbf{X}'\mathbf{1}_U$ and $\mathbf{P}$ is the matrix of eigenvectors associated to the eigenvalues of $\mathbf{X}_s'\mathbf{X}_s$. Following the same arguments, we obtain that

$$\mathbf{w}_{MB}'\mathbf{w}_{MB} \simeq \mathbf{1}_U'\mathbf{X}(\mathbf{X}_s'\mathbf{X}_s)^{-1}\mathbf{X}'\mathbf{1}_U = \sum_{j=1}^{p}\eta_i^2\frac{1}{\lambda_i}$$

Since for any $k > 0$, we always have $\frac{1}{\lambda_i} > \frac{\lambda_i}{(\lambda_i+k)^2}$, we get that $\mathbf{w}_{MB,R}'\mathbf{w}_{MB,R} < \mathbf{w}_{MB}'\mathbf{w}_{MB}$. This proves that the scatter of ridge weights is smaller and more stable under perturbation of $\mathbf{X}_s$ than that of BLUE weights. This is in concordance with the ridge principle. We can see also that in the presence of multicollinearity, respectively of ill-conditionning, $\lambda_{min} = \min\lambda_i$ are close to zero, respectively the conditionning number $K = \sqrt{\lambda_{max}/\lambda_{min}}$ is very large, which entails negative or extremely large calibration weights.

It is worth mentioning two extreme values of $\hat{t}_{w,MB}$. As $\mathbf{C} \to \infty$ (i.e. infinite cost associated with the bias $\mathbf{B}$), we obtain $\mathbf{w}_{MB,R} = \mathbf{1}_s - \mathbf{V}_s^{-1}\mathbf{X}_s\left(\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{X}_s\right)^{-1}\mathbf{X}_s'\mathbf{V}_s^{-1}\mathbf{y}_s$ and $\hat{t}_{w,MB}$ is the minimum variance unbiased linear estimator (*Royall*, 1970). This means that the constraint $\mathbf{B} = 0$ is exactly satisfied. On the opposite case, as $\mathbf{C} \to 0$, we obtain $\mathbf{w}_{MB,R} = \mathbf{1}_s$ and $\hat{t}_{w,MB} = \sum_s y_k$ which is equivalent to removing the constraint from the optimization problem.

The derivation of model-based ridge estimator depends on the cost matrix $\mathbf{C}$. Considering that $\mathbf{C} = k^{-1}\mathbf{C}^*$, *Bardsley and Chambers* (1984) and *Chambers* (1996) use the ridge trace to determine the appropriate k. $\mathbf{C}^*$ is a fixed cost matrix providing a correct relative weighting of the components of the relative bias vector $(\mathrm{diag}(\mathbf{X}'\mathbf{1}_U))^{-1}\mathbf{B}$. This transformation was needed because of the large differences in scale between the predictors in $\mathbf{X}$ and it is a kind of standardization of variables.

3.2 Ridge under the calibration approach or penalized calibration

Without assuming a superpopulation model, one can use the *calibration method* (Deville and Särndal, 1992) which consists in deriving a weighted estimator

$$\hat{t}_w = \sum_s w_k y_k$$

with weights minimizing a pseudo-distance, subject to calibration constraints (i.e. all the auxiliary variable totals are exactly estimated). Usually a chi-square distance is used, $\sum_s \frac{(w_k - d_k)^2}{d_k q_k}$, yielding the calibration weights $\mathbf{w}_s^c = (w_k^c)_{k \in s}$

$$\textbf{(P4)}: \quad \mathbf{w}_s^c \;=\; \text{argmin}_{\mathbf{w}_s}(\mathbf{w}_s - \mathbf{d}_s)'\widetilde{\mathbf{\Pi}}_s(\mathbf{w}_s - \mathbf{d}_s) \quad \text{subject to} \quad (\mathbf{w}_s^c)'\mathbf{X}_s = \mathbf{1}_U'\mathbf{X}$$

where $\widetilde{\mathbf{\Pi}}_s = \text{diag}(q_k^{-1}d_k^{-1})_{k \in s}$ and q_k are positive constants. Most of the times, we consider $q_k = 1$ for all k. The calibration weights thus get the following shape,

$$\mathbf{w}_s^c = \mathbf{d}_s - \widetilde{\mathbf{\Pi}}_s^{-1}\mathbf{X}_s(\mathbf{X}_s'\widetilde{\mathbf{\Pi}}_s^{-1}\mathbf{X}_s)^{-1}(\mathbf{d}_s'\mathbf{X}_s - \mathbf{1}_U'\mathbf{X})'$$

For $q_k = 1/v_k^2$, the calibrated estimator $\hat{t}_{yw} = (\mathbf{w}_s^c)'\mathbf{y}_s$ is equal to the GREG estimator given by (19). Moreover, remark that in this case we have $\widetilde{\mathbf{\Pi}}_s = \mathbf{V}_s\mathbf{\Pi}_s$, which means that the optimization problem **(P2)** uses the inverse of the weight matrix employed in **(P4)**. For a more general distance function, *Deville and Särndal* (1992) show that under certain conditions the calibrated estimator is asymptotically equivalent to the model-assisted or GREG estimator $\hat{t}_{GREG}$. This equivalence is in the sense that $N^{-1}(\hat{t}_{yw} - \hat{t}_{GREG}) = O_p(n^{-1})$. This fact will consequently lead to the asymptotic equivalence of the variances of both estimators.

From a geometrical point of view, we search the weights w_k which explain the best the Horvitz-Thompson weights $d_k = 1/\pi_k$ and that lie in the orthogonal of the constraint space given by the kernel of the matrix $\mathbf{X}_s$. The constraint space is of dimension $n-p$, so increasing the number of auxiliary variables will decrease the number of degrees of freedom for w_k (*Guggemos and Tillé*, 2010). The similar reasoning by *Silva and Skinner* (1997) proved that by increasing the number of calibration variables after a certain number may increase the variance up to a harmful level. *Guggemos and Tillé* (2010) called it *over-calibration* and suggested not calibrating on those variables which are less correlated with the variables of interest.

Another issue with the calibration weights is the fact that they may not satisfy range restrictions (i.e. pre-specified lower and upper bounds) especially when the number of calibration or benchmark

constraints is large. Satisfying such condition is desirable especially for avoiding the inflation of the sampling error of estimates in small to moderate domains (*Beaumont and Bocci*, 2008). As *Deville and Särndal* (1992) stated, negative weights may occur when the chi-squared distance is employed. For the other distances used in their paper, the positiveness of weights is guaranteed but unrealistic or extreme weights may also occur. To cope with this issue, several modifications have been suggested in the literature. However, all these methods are iterative and may not yield a solution even if the range restriction are mild (*Rao and Singh*, 1997 and *Beaumont and Bocci*, 2008). This is more likely to happen when they are many constraints, and so multicollinearity of data, or when the sample size is small.

So, how to avoid negative or extremely large weights? *Chambers* (1996) and *Rao and Singh* (1997) answer this question by suggesting to relax the calibration constraints. Suppose we have non-negative constants C_j with $j = 1, ..., p$ representing the cost of the weighted estimator which does not satisfy the calibration equation. The cost C_j can also be the cost of the risk associated with the calibration equation not to be satisfied. The objective function can be given as,

$$\textbf{(P5)}: \quad \mathbf{w}_{R,s}^{c} = \text{argmin}_{\mathbf{w}_s}(\mathbf{w}_s - \mathbf{d}_s)'\widetilde{\mathbf{\Pi}}_s(\mathbf{w}_s - \mathbf{d}_s) + \frac{1}{\lambda}(\mathbf{w}_s'\mathbf{X}_s - \mathbf{1}_U'\mathbf{X})\mathbf{C}(\mathbf{w}_s'\mathbf{X}_s - \mathbf{1}_U'\mathbf{X})' \tag{25}$$

Rao and Singh (1997) consider the objective function without the constant λ. Writing the problem **(P5)** as a constrained optimization problem, puts into evidence that we lessen the calibration equation corresponding to those variables which are somehow unable to satisfy the calibration constraints but not too much since we penalize the large values of $\mathbf{w}_s'\mathbf{X}_s - \mathbf{1}_U'\mathbf{X}$. In this way we eliminate the possibility of having very large or negative weights. Simply, we can say that the ridge estimator performs as a variable selection tool.

The weights are given by

$$\mathbf{w}_{R,s}^{c} = \mathbf{d}_s - \widetilde{\mathbf{\Pi}}_s^{-1}\mathbf{X}_s(\mathbf{X}_s'\widetilde{\mathbf{\Pi}}_s^{-1}\mathbf{X}_s + \lambda\mathbf{C}^{-1})^{-1}(\mathbf{X}_s'\mathbf{d}_s - \mathbf{X}'\mathbf{1}_U) \tag{26}$$

which yield the ridge calibration estimator or the penalized calibration of the population total t_y,

$$\begin{aligned} \hat{\mathbf{t}}_{y,Rw} = (\mathbf{w}_{R,s}^{c})'\mathbf{y}_s &= \mathbf{d}_s'\mathbf{y}_s - (\mathbf{X}_s'\mathbf{d}_s - \mathbf{X}'\mathbf{1}_U)'\hat{\boldsymbol{\beta}}_\lambda \\ &= \hat{t}_{y,d} - \left(\hat{\mathbf{t}}_{x,d} - \mathbf{t}_x\right)'\hat{\boldsymbol{\beta}}_\lambda \end{aligned} \tag{27}$$

where $\hat{\boldsymbol{\beta}}_\lambda = (\mathbf{X}_s'\widetilde{\mathbf{\Pi}}_s^{-1}\mathbf{X}_s + \lambda\mathbf{C}^{-1})^{-1}\mathbf{X}_s'\widetilde{\mathbf{\Pi}}_s^{-1}\mathbf{y}_s$ and $\hat{\mathbf{t}}_{x,d}$ is the Horvitz-Thompson estimator for the total $\mathbf{t}_x$. This approach is equivalent to construct a GREG estimator of population total with the

regression coefficient estimated by a ridge estimator (*Hoerl and Kennard*, 1970). More precisely, $\hat{\boldsymbol{\beta}}_{\lambda}$ is in fact $\hat{\boldsymbol{\beta}}_{\pi,R}$ from (20) for $\lambda\mathbf{C}^{-1} = \tilde{\mathbf{D}}$ and $\widetilde{\boldsymbol{\Pi}}_s^{-1} = \mathbf{V}_s^{-1}\boldsymbol{\Pi}_s^{-1}$.

The ridge estimator given by (27) can be written as a linear combination of the Horvitz-Thompson estimator and the GREG estimator (*Rao and Singh*, 1997) as follows,

$$\hat{t}_{y,Rw} \quad = \quad (1-\alpha)\hat{t}_{y,d} + \alpha\hat{t}_{GREG}$$

where $\hat{t}_{GREG,R} = \hat{t}_{y,d} - \left(\hat{\mathbf{t}}_{x,d} - \mathbf{t}_x\right)' \hat{\boldsymbol{\beta}}_{\pi}$ is the GREG estimator given by (19) and α is given by,

$$\alpha \quad = \quad \mathbf{y}_s'\widetilde{\boldsymbol{\Pi}}_s^{-1}\mathbf{X}_s \left(\mathbf{X'}_s\widetilde{\boldsymbol{\Pi}}_s^{-1}\mathbf{X}_s + \lambda\mathbf{C}^{-1}\right)^{-1} (\mathbf{t}_x - \hat{\mathbf{t}}_{x,d}) \left[\mathbf{y}_s'\widetilde{\boldsymbol{\Pi}}_s^{-1}\mathbf{X}_s \left(\mathbf{X'}_s\widetilde{\boldsymbol{\Pi}}_s^{-1}\mathbf{X}_s\right)^{-1} (\mathbf{t}_x - \hat{\mathbf{t}}_{x,d})\right]^{-1}$$

As for the model-based approach, the Horvitz-Thompson as well as the GREG estimator are two limit values of $\hat{t}_{y,Rw}$. More exactly, consider relation (27) for a fixed cost matrix $\mathbf{C}$ and let λ vary from 0 to ∞. The ridge calibration estimator is a continuos function of λ. For $\lambda = 0$, then $\alpha = 1$ and an infinite cost is attributed to all constraints meaning that they are all exactly satisfied. It implies that $\hat{t}_{y,Rw}$ is the GREG estimator which is ξ-unbiased for the population total t_y. Ridge weights with strictly positive biasing parameter λ means that the weights do not satisfy exactly the calibration equations. In this case, the estimator $\hat{t}_{y,Rw}$ is ξ-biased but the weights $\mathbf{w}_{R,s}^c$ are more stable (Chambers, 1996) and implied a reduction in *MSE* (*Bardsley and Chambers, 1984*). Values of λ producing weights larger or equal to 1 are accepted by *Chambers* (1996).
As $\lambda \to \infty$, $\alpha \to 0$ and the ridge calibrated estimator $\hat{t}_{y,Rw}$ goes to the Horvitz-Thompson estimator. In this case, we do not use any of the auxiliary variables for the estimation of the finite population total of the variable of interest.

It is of interest to see how $\hat{t}_{y,Rw}$ changes when a specific cost C_j varies from 0 to ∞. The zero cost $C_j = 0$ means that the constraint corresponding to the total t_{X_j} is discarded and the large or infinite cost $C_j = \infty$, that the corresponding calibration constraint is exactly satisfied. In the latter situation, the weights are computed using (25) with the cost matrix $\mathbf{C}^{-1}$ having 0 on the j-th diagonal element.

Using the same justifications given by *Hoerl and Kennard* (1970) for obtaining the ridge regression coefficient as a solution of constrained minimization (see section 2.2.1), the weights $\mathbf{w}_{R,s}^c$

satisfying the optimization problem **(P5)** satisfy also the following optimization problem,

$$\begin{aligned}\textbf{(P6)}: \quad \mathbf{w}_{R,s}^{c} &= \text{argmin}_{\mathbf{w}_s}(\mathbf{w}_s'\mathbf{X}_s - \mathbf{1}_U'\mathbf{X})\mathbf{C}(\mathbf{w}_s'\mathbf{X}_s - \mathbf{1}_U'\mathbf{X})' + \lambda(\mathbf{w}_s - \mathbf{d}_s)'\widetilde{\mathbf{\Pi}}_s(\mathbf{w}_s - \mathbf{d}_s) \\ &= \text{argmin}_{\mathbf{w}_s, ||\mathbf{w}_s - \mathbf{d}_s||^2_{\widetilde{\mathbf{\Pi}}_s} \leq r^2}(\mathbf{w}_s'\mathbf{X}_s - \mathbf{1}_U'\mathbf{X})\mathbf{C}(\mathbf{w}_s'\mathbf{X}_s - \mathbf{1}_U'\mathbf{X})'\end{aligned}$$

which means that we find the smallest distance between the weighted estimator $\hat{t}_{y,Rw}$ and the total t_y with weights satisfying the range restrictions. The geometric interpretations are somehow similar to those given in section 2.2.1 and using Figure 1. *Beaumont and Bocci* (2008) give another justification of this result and suggest a bisection algorithm to find w_k knowing that the maximum value of $(\mathbf{w}_s'\mathbf{X}_s - \mathbf{1}_U'\mathbf{X})\mathbf{C}(\mathbf{w}_s'\mathbf{X}_s - \mathbf{1}_U'\mathbf{X})'$ is reached for $\mathbf{w}_s = \mathbf{d}_s$ while satisfying the range restrictions. The minimum value is reached for the GREG estimator but without respecting necessarily the restriction on $\mathbf{w}_s$. Nevertheless, as *Beaumont and Bocci* remarked, this algorithm may be time-consuming.

3.3 Partially penalized ridge or partially penalized calibration

In a model based approach, *Bardsley and Chambers* (1984) suggested to divide the p variables in the data matrix $\mathbf{X}$ into two sets of variables $\widetilde{\mathbf{X}}_1$ and $\widetilde{\mathbf{X}}_2$ based on the fact that variables in $\widetilde{\mathbf{X}}_1$ contain much more importance than the variables in $\widetilde{\mathbf{X}}_2$ in the sense that they can contribute more influentially in the estimation process. We may consider that the matrix $\mathbf{X}$ has the following expression after re-ordering the variables $\mathbf{X_1}, \ldots, \mathbf{X}_p$,

$$\mathbf{X} = \left(\widetilde{\mathbf{X}}_1, \widetilde{\mathbf{X}}_2\right)$$

where $\widetilde{\mathbf{X}}_1 = [\mathbf{X}_1, \ldots, \mathbf{X}_q]$ and $\widetilde{\mathbf{X}}_2 = [\mathbf{X}_{q+1}, \ldots, \mathbf{X}_p]$. The variables contained in $\widetilde{\mathbf{X}}_1$ may be related for example to socio-demographic criteria. *Bardsley and Chambers* (1984) attach the importance to the variables in terms of cost which are in fact penalties associated to the variables. Let $\mathbf{C}$ be the diagonal matrix of nonnegative costs which can measure the acceptable level of error while estimating the totals of variable from the $\mathbf{X}$ matrix,

$$\mathbf{C} = \begin{pmatrix} \mathbf{C}_1 & \mathbf{0}_{(q,p-q)} \\ \mathbf{0}_{(p-q,p)} & \mathbf{C}_2 \end{pmatrix}$$

where $\mathbf{C}_1$, respectively $\mathbf{C}_2$, is the relative diagonal cost matrix of size $q \times q$ associated to $\widetilde{\mathbf{X}}_1$, respectively of size $(p-q) \times (p-q)$ associated to $\widetilde{\mathbf{X}}_2$.

As discussed in the above section, allowing an infinite cost C_j means that the associated constraint is

exactly satisfied. *Bardsley and Chambers* (1984) consider the case when constraints corresponding to $\mathbf{X}_1, \ldots, \mathbf{X}_q$ are all exactly satisfied. This means $\mathbf{C}_1 = \infty$ and hence, weights may be derived using relation (23) with $\mathbf{C}_1^{-1} = \mathbf{0}_{(q\times q)}$. The weights using this *partially penalized ridge* regression and abbreviated as $\mathbf{w}_{ppr}$ below can be written as,

$$\mathbf{w}_{ppr} = \\ \mathbf{1}_s - \left(\mathbf{V}_s^{-1}\widetilde{\mathbf{X}}_{1s}, \quad \mathbf{V}_s^{-1}\widetilde{\mathbf{X}}_{2s}\right)\begin{pmatrix} \widetilde{\mathbf{X}}_{1s}'\mathbf{V}_s^{-1}\widetilde{\mathbf{X}}_{1s} & \widetilde{\mathbf{X}}_{1s}'\mathbf{V}_s^{-1}\widetilde{\mathbf{X}}_{2s} \\ \widetilde{\mathbf{X}}_{2s}'\mathbf{V}_s^{-1}\widetilde{\mathbf{X}}_{1s} & \widetilde{\mathbf{X}}_{2s}'\mathbf{V}_s^{-1}\widetilde{\mathbf{X}}_{2s} + \mathbf{C}_2^{-1} \end{pmatrix}^{-1}\begin{pmatrix} \widetilde{\mathbf{X}}_{1s}'\mathbf{1}_s - \widetilde{\mathbf{X}}_1'\mathbf{1}_U \\ \widetilde{\mathbf{X}}_{2s}'\mathbf{1}_s - \widetilde{\mathbf{X}}_2'\mathbf{1}_U \end{pmatrix} \tag{28}$$

where $\widetilde{\mathbf{X}}_{1s}$, respectively $\widetilde{\mathbf{X}}_{2s}$, is the sample restriction of $\widetilde{\mathbf{X}}_1$, respectively of $\widetilde{\mathbf{X}}_2$. Using a calibration approach, the weight are derived using the above formula with $\mathbf{V}_s$ replaced by $\widetilde{\mathbf{\Pi}}_s$ and $\mathbf{1}_s$ by $\mathbf{d}_s$. In particular, we have $\mathbf{w}_{ppr}'\widetilde{\mathbf{X}}_{1s} = \mathbf{1}_U'\widetilde{\mathbf{X}}_1$.

Now, if the cost matrix $\mathbf{C}_2$ also goes to infinity, then the constraints corresponding to variables in $\widetilde{\mathbf{X}}_2$ are also exactly satisfied. Hence, the estimator using the weights so derived is again nothing else than the best linear unbiased estimator $\hat{t}_{BLUE}$ given by (17) and derived under the model ξ that uses the whole matrix $\mathbf{X}$. Moreover, in the case $\mathbf{C}_2 \to \mathbf{0}_{(p-q,p-q)}$ the variables included in $\widetilde{\mathbf{X}}_2$ are discarded from the constraints and thus the model will include only the calibration variables from $\widetilde{\mathbf{X}}_1$,

$$\mathbf{w}_{ppr} \to \mathbf{w}_{ppr}^{(1)} = \mathbf{h} - \mathbf{V}_s^{-1}\widetilde{\mathbf{X}}_{1s}\left(\widetilde{\mathbf{X}}_{1s}'\mathbf{V}_s^{-1}\widetilde{\mathbf{X}}_{1s}\right)^{-1}(\mathbf{1}_s'\widetilde{\mathbf{X}}_{1s} - \mathbf{1}_U'\widetilde{\mathbf{X}}_1)'$$

The penalized estimator becomes the best unbiased estimator based under the restricted model that uses only the matrix $\widetilde{\mathbf{X}}_1$. Since $\hat{t}_{BLUE}$ based on the whole model ξ as well as on the restricted model with $\widetilde{\mathbf{X}}_1$ are two extreme estimators as $\mathbf{C}_2$ varies from ∞ to 0, *Bardsley and Chambers* (1984) called the estimator that uses weights $\mathbf{w}_{ppr}$ *an interpolated estimator* between the two extremes. So, the penalized ridge estimator may be considered as a trade-off between an over-specified model and an under-specified model.

One can show that the ridge weights $\mathbf{w}_{ppr}$ verifying the optimization problem **(P3)** with the inverse matrix cost

$$\mathbf{C}^{-1} = \begin{pmatrix} \mathbf{0}_{(q,q)} & \mathbf{0}_{(q,p-q)} \\ \mathbf{0}_{(p-q,p)} & \mathbf{C}_2^{-1} \end{pmatrix} \tag{29}$$

may be obtained as a solution of the following optimization problem

$$\begin{aligned} \textbf{(P7)}: \quad \mathbf{w}_{ppr} &= \text{argmin}_{\mathbf{w}}(\mathbf{w}_s - \mathbf{1}_s)'\mathbf{V}_s^{-1}(\mathbf{w}_s - \mathbf{1}_s) + (\mathbf{w}_s'\widetilde{\mathbf{X}}_{2s} - \mathbf{1}_U'\widetilde{\mathbf{X}}_2)\mathbf{C}_2^{-1}(\mathbf{w}_s'\widetilde{\mathbf{X}}_{2s} - \mathbf{1}_U'\widetilde{\mathbf{X}}_2)' \\ \mathbf{w}_s'\widetilde{\mathbf{X}}_{1s} &= \mathbf{1}_U'\widetilde{\mathbf{X}}_1 \end{aligned} \tag{30}$$

This kind of optimization problem was used by *Park and Yang* (2010) and *Guggemos and Tillé* (2010). Using the model ξ given by (12) with intercept, *Park and Yang* (2010) aim at estimating the mean $\overline{y}_U = \sum_U y_k/N$ of the variable of interest $\mathcal{Y}$ using a Hajek-type estimator. This means that they use a weighted estimator with weights that sum up to unity and being as close as possible to the Hajek weights,

$$\alpha_i = \frac{\pi_i^{-1}}{\sum_s \frac{1}{\pi_i}}$$

This means that the optimization problem **(P7)** is used with $\mathbf{1}_s$ replaced by $\boldsymbol{\alpha}_s = (\alpha_i)_{i\in s}$. They build two partially penalized estimators. In the first case, $\widetilde{\mathbf{X}}_1 = \mathbf{1}_U$ and in the second case, $\widetilde{\mathbf{X}}_1 = (\mathbf{1}_U, \mathbf{X}_2, \ldots, \mathbf{X}_q)$. Weights may be derived using relation (28). Slightly simplified formulas are obtained since $\mathbf{1}_s'\boldsymbol{\alpha}_s - \mathbf{1}_U'\mathbf{1}_U/N = 0$.

In a linear regression context, it is not very common to consider the penalty or the cost matrix $\mathbf{C}^{-1}$ given by (29). This is more likely to happen with a mixed model. *Guggemos and Tillé* (2010) consider the following mixed model

$$\xi': \quad \mathbf{y} = \widetilde{\mathbf{X}}_1\mathbf{B} + \widetilde{\mathbf{X}}_2\mathbf{u} + \boldsymbol{\eta}$$

and the calibration approach, namely we replace in the objective function from the optimization problem **(P7)** from (30) the matrix $\mathbf{V}_s$ by $\widetilde{\boldsymbol{\Pi}}_s$, respectively $\mathbf{1}_s$ by $\mathbf{d}_s$. *Guggemos and Tillé* consider also that the second term of the objectif function depends on a penalty parameter and they suggest the Fisher scoring algorithm to compute it. The value of the penalty parameter is obtained at the convergence of the Fisher scoring algorithm. They give also application of the penalized calibration for estimation of finite population totals in a small area context.

3.4 Calibration on uncertain auxiliary information

In presence of several extern estimations which may be considered as uncertain, *Deville* (1999) suggested another construction which uses in fact the Bayesian interpretation of the ridge estimator given in section 2.2.1. Consider that another estimation $\hat{t}_{x^*,d} = \mathbf{d}_s'\mathbf{X}^*$ based on the auxiliary information $\mathbf{X}^*$ is available from external sources such as previous surveys. We have also the

current estimation based on $\mathbf{X}$. We suppose that the variances of $\hat{t}_{x^*,d}$ and $\hat{t}_{x,d}$ are known and the covariance between the two sources is zero. We suppose also that the covariance between $\hat{t}_{x^*,d}$ and $\hat{t}_{y,d}$ is also zero. *Deville* looks for linear weighted estimators for t_y of the form

$$\hat{t}_w = \mathbf{d}_s'\mathbf{y}_s + (\mathbf{d}_s'\mathbf{X}_s^* - \mathbf{d}_s'\mathbf{X}_s)\boldsymbol{\beta} = \hat{t}_{y,d} + (\hat{t}_{x^*,d} - \hat{t}_{x,d})'\boldsymbol{\beta}. \tag{31}$$

The optimal value of the unknown parameter $\boldsymbol{\beta}$ is the one that minimizes the sampling variance of $\hat{t}_w$. We find

$$\boldsymbol{\beta}_{opt} = \left(\mathrm{Var}(\hat{t}_{x^*,d}) + \mathrm{Var}(\hat{t}_{x,d})\right)^{-1} \mathrm{Cov}(t_{y,d}, t_{x,d})$$

and the same value may be derived by using a variance minimization criteria as in *Montanari* (1987) plus a penalty term, namely

$$\mathbf{(P8)}: \quad \boldsymbol{\beta}_{opt} = \mathrm{argmin}_{\boldsymbol{\beta}}(\mathbf{y}_s - \mathbf{X}_s\boldsymbol{\beta})'\boldsymbol{\Delta}(\mathbf{y}_s - \mathbf{X}_s\boldsymbol{\beta}) + \boldsymbol{\beta}'\mathbf{X}_s'^*\boldsymbol{\Delta}\mathbf{X}_s^*\boldsymbol{\beta} \tag{32}$$

where $\boldsymbol{\Delta} = (\frac{\pi_{ij} - \pi_i\pi_j}{\pi_{ij}})_{i,j\in U}$. We remark that the penalty is now on the variance of $\hat{t}_{x^*,d}$.
The estimation of t_y given by (31) computed for $\boldsymbol{\beta} = \boldsymbol{\beta}_{opt}$ may be improved by replacing $\hat{t}_{x^*,d}$ with the best unbiased linear estimator of $\hat{t}_{x^*,d}$ and $\hat{t}_{x,d}$. This is equivalent to determine the posterior estimation knowing that the priori estimation given by the auxiliary information is $\hat{t}_{x^*,d}$ and the actual estimation is $\hat{t}_{x,d}$. One may use relation (8) to find the posterior estimation as

$$\hat{t}_{x,x^*}^{opt} = (\mathbf{I}_p - \mathbf{A})\,\hat{t}_{x^*,d} + \mathbf{A}\hat{t}_{x,d}$$

where $\mathbf{A}$ is a squared p-dimensional matrix given by

$$\mathbf{A} = \mathbf{I}_p - \mathrm{Var}(\hat{t}_{x,d})\left(\mathrm{Var}(\hat{t}_{x,d}) + \mathrm{Var}(\hat{t}_{x^*,d})\right)^{-1}.$$

Then, one can derive the estimator $\hat{t}_y^{opt}$ of t_y from relation (31) with $\hat{t}_{x^*,d}$ replaced with $\hat{t}_{x,x^*}^{opt}$,

$$\hat{t}_y^{opt} = \hat{t}_{y,d} + (\hat{t}_{x,x^*}^{opt} - \hat{t}_{x,d})'(\mathrm{Var}(\hat{t}_{x,d}))^{-1}\mathrm{Cov}(t_{y,d}, t_{x,d})$$

One can easily obtain that for $y_i = x_i$, we obtain $\hat{t}_{x,x^*}^{opt}$ or equivalently, the estimator is calibrated on $\hat{t}_{x,x^*}^{opt}$. If the variance covariance $\mathrm{Cov}(t_{y,d}, t_{x,d})$ is estimated by the usual Horvitz-Thompson estimator, $\hat{t}_y^{opt}$ is a linear estimator in y_i with weights w_i given by

$$w_i = d_i + (\hat{t}_{x^*,d} - \hat{t}_{x,d})'(\mathrm{Var}(\hat{t}_{x,d}))^{-1} z_i d_i,$$

where $z_i = \sum_{j \in s} \frac{\Delta_{ij}}{\pi_{ij}} \frac{x_j}{\pi_j}$. The main advantage of Deville's construction is that it does not need to determine a penalty parameter as it was the case before. All we need is the variance of the externe estimation.

Deville (1999) give also a practical implementation and generalization to several external estimation.

3.5 Statistical properties of ridge estimators with survey data

Ridge-type estimators are biased estimators suggested in classical regression in order to diminish the model mean squares error. *Bardsley and Chambers* (1984) affirm that the model-based ridge estimator has smaller prediction variance than the best linear unbiased estimator $\hat{t}_{BLUE}$ but they do not give a rigorous proof. *Bellhouse* (1987) shows that a predictor $\hat{Y}^{(1)} = \sum_s y_k + (N-n)\hat{\mu}_s^{(1)}$ of the finite population total t_y is better than another predictor $\hat{Y}^{(2)} = \sum_s y_k + (N-n)\hat{\mu}_s^{(2)}$ with respect to the mean square error under the model ξ and the sampling design p if, for every sample s of fixed size n, $\hat{\mu}_s^{(1)}$ is better than $\hat{\mu}_s^{(2)}$ in the sense that

$$E_\xi(\hat{\mu}_s^{(1)} - \mu_{ns})^2 \leq E_\xi(\hat{\mu}_s^{(2)} - \mu_{ns})^2$$

where μ_{ns} is the unknown prediction of the non sampled mean of $\mathcal{Y}$. Using this result and the same arguments as in *Vinod and Ullah* (1981), one can get that for any penalty constant k satisfying $0 < k < 2\sigma^2/\boldsymbol{\beta}'\boldsymbol{\beta}$,

$$E_\xi E_p(\hat{t}_{w,MB} - t_y)^2 < E_\xi E_p(\hat{t}_{BLUE} - t_y)^2$$

where $\hat{t}_{w,MB}$ is the ridge model based estimator given by (24) for $\mathbf{C}^{-1} = k\mathbf{I}_p$ and $\hat{t}_{BLUE}$ is the best linear unbiased estimator given by (17). A necessary and sufficient condition for the ridge estimator to be more efficient than the least squares estimator is

$$0 < k < 2/[-\min(0, \psi)]$$

where ψ is the minimum eigenvalue of $(\mathbf{X}_s'\mathbf{X}_s)^{-1} - (\boldsymbol{\beta}\boldsymbol{\beta}'/\sigma^2)$ (*Swindel, and Chapman*, 1973). *Dunstan and Chambers* (1986) derived confidence intervals for finite population totals estimated using the ridge model-based procedure and robust model-based variance estimators.

In a design-based setting, the concern is about asymptotic properties of $\hat{t}_{y,Rw}$ given by (27) with respect to the sampling design p. As *Rao and Singh* (1997) stated, *"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."* The asymptotic design unbiasedness and

consistency of $\hat{t}_{y,Rw}$ are derived using the equivalence with GREG estimators even if $\hat{t}_{y,Rw}$ has been obtained as a solution of penalized calibration problems. Under broad assumptions (*Fuller*, 2002), the design-based ridge estimator $\hat{\boldsymbol{\beta}}_\lambda$ of $\boldsymbol{\beta}$ tends in probability to $\boldsymbol{\beta}_\lambda = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{C}^{-1})^{-1}\mathbf{X}'\mathbf{y}$ and the ridge estimator $\hat{t}_{y,Rw}$ is asymptotically equivalent to

$$\hat{t}_{y,Rw} \simeq \mathbf{d}_s'\mathbf{y}_s - (\mathbf{X}_s'\mathbf{d}_s - \mathbf{X}'\mathbf{1}_U)'\boldsymbol{\beta}_\lambda = \mathbf{d}_s'(\mathbf{y}_s - \mathbf{X}_s\boldsymbol{\beta}_\lambda) + \mathbf{1}_U'\mathbf{X}\boldsymbol{\beta}_\lambda$$

which implies that the $\hat{t}_{y,Rw}$ is asymptotically design unbiased and consistent under abroad assumptions that provide the design unbiasedness and consistency of the Horvitz-Thompson estimators $\mathbf{d}_s'\mathbf{y}_s$ and $\mathbf{d}_s'\mathbf{X}_s$ (*Rao and Singh*, 1997 and *Théberge*, 2000). The asymptotic variance under the sampling design may thus be deduced as being the Horvitz-Thompson variance applied to residuals $y_k - \mathbf{x}_k'\boldsymbol{\beta}_\lambda$.

4 Conclusion and extensions

In this paper, we have undertaken an overview of the applications of ridge-type estimators in survey sampling theory. Even if the paper of *Bardsley and Chambers* (1984) has not received much attention at the beginning, we assist now of an increasingly interest on this subject. This is mostly due to the fact that nowadays, we face more and more information and this kind of issue is more often encountered in practice than before.

To use this class of estimators, two practical issues should be treated carefully. The first one is the computation of the penalty parameter. Several algorithms have been suggested in the literature such as the ridge trace (*Bardsley and Chambers*, 1984), the Fischer scoring algorithm (*Guggemos and Tillé*, 2010) or the bisection algorithm (*Beaumont and Bocci*, 2008). *Beaumont and Bocci* (2008) compare the ridge calibration with the method of Chen *et al.* (2002) showing the superiority of the ridge calibration method. Nevertheless, it would be interesting having a comparison between all these algorithms.

Their is another important point that we would like to stress. All the papers dealing with ridge-type estimators in survey sampling give few details about the standardization of the auxiliary variables if any has been done. Or, as mentioned at the end of section 2.2.1, it is important to know what kind of standardization is used since different methods lead to different ridge estimators. The cost matrix used in the objective functions from the optimization problems **(P3)** and **(P5)** may be interpreted as a standardization matrix.

Finally, some other alternative methods for dealing with huge data sets must be investigated. We mention here the lasso methods which consist in considering a penalty with the absolute value instead of the euclidian norm. We are not aware of the existence of such application in survey sampling. The regression on principal component analysis is another interesting alternative. This method consists in considering the principal components of $\mathbf{X}'\mathbf{X}$ which reduce the number of auxiliary variables while keeping maximum of information. For huge survey data, Goga *et al.* (2011) suggest calibration on the set of these new variables which is in general of much smaller dimension than the initial one.

Bibliography

Bardsley, P. and Chambers, R.L. (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**, 5-20.

Bellhouse, D. R. (1987), Model-based estimation in finite population sampling, *Journal of the American Statistical Association*, **41**, 260-262.

Chambers, R.L. (1996), Robust Case-Weighting for Multipurpose Establishment Surveys, *Journal of Official statistics*, **12**, 3-32.

Chen, J., Sitter, R.R. and Wu, C. (2002), Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys, *Biometrika*, **89**, 230-237.

Conniffe, D. and Stone, J. (1973), A critical view of ridge regression, *The Statistician*, **22**, 181-187.

Deville, J.C. (1992), Constrained samples, conditional inference, weighting: three aspects of the utilization of auxiliary information, *In Proceedings of the Workshop on the Uses of Auxiliary Information in Surveys, rebro (Sweden).*

Deville, J.C. (1999), Calage simultané de plusieurs enquêtes, *Recueil du Symposium 99 de Statistique Canada*, mai, 1999.

Deville, J.-C., and Särndal, C.-E. (1992), Calibration estimation in survey sampling, *Journal of the American Statistical Association*, **418**, 376-382.

Dunstan, R. and Chambers, R.L. (1986), Model-based confidence intervals in multipurpose surveys, *Applied Statistics*, **35**, 276-280.

Fuller, W. A. (2002), Estimation par régression appliquée à l'échantillonnage, *Techniques d'enquêtes*, **28**, 5-25.

Goga, C., Muhammad-Shehzad, A. and Vanheuverzwyn, A. (2011), Principal Component Regression with Survey Data. Application on the French Media Audience, *Proceedings of the 58th World Statistics Congress of the International Statistical Institute, Dublin, Ireland.*

Guggemos, F. and Tillé, Y. (2010), Penalized calibration in survey sampling: Design-based estimation assisted by mixed-models, *Journal of Statistical Planning and Inference*, **140**, 3199-3212.

Gujarati, D. N., (2002), Basic Econometric, *4th Edition*, *New York: McGrath Hill.*

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.

Hoerl, A. E., and Kennard, D.J. (1962), Application of ridge analysis to regression problems, *Chemical Engineering Progress*, **58**, 54-59.

Hoerl, A. E., and Kennard, D.J. (1968), On regression analysis and biased estimation, *Technometrics*, **10**, 422-423.

Hoerl, A. E., and Kennard, D.J. (1970), Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics*, **12**, 55-67.

Hoerl, A. E., and Kennard, D.J. (1976), Ridge regression: Iterative estimation of the biasing parameter, *Communications in Statistics*, **5**, 77-88.

Hoerl, A. E., Kennard, D.J., and Baldwin, K. F. (1975), Ridge regression: Some simulations, *Communications in Statistics*, **4**, 105-123.

Horvitz, D.G. and Thompson, D.J. (1952), A generalization of sampling without replacement from a finite universe, *Journal of the American Statistical Association*, **47**, 663-685.

Izenman, A. J. (2008), Modern Multivariate Statistical Technique, *Springer.*

Kapat, P. and Goel, P.K. (2010), Anomalies in the Foundations of Ridge Regression: Some Clarifications, *International Statistical Review*, **78**, 209-215.

Kmenta, J. (1986), Elements of econometrics. *2th Edition*, *New York: McMillan.*

Marquardt, D. W. (1970), Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation, *Technometrics*, **12**, 591-612.

Marquardt, D. W., and Snee, R. D. (1975), Ridge regression in practice, *The American Statistician*, **29**, 3-20.

Montanari, G.E. (1987), Post-sampling efficient prediction in large-scale surveys, *International Statistical Review*, **55**, 191-202.

Park, M., and Yang, M. (2010), Ridge regression estimation for survey samples, *Communications in Statistics. Theory Methods*, **37**, 532-543.

Rao, J.N.K. and Singh, A. C. (1997), A ridge-shrinkage method for range-restricted weight calibration in survey sampling, *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 57-65.

Royall, R.M. (1976), The linear least squares 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. *Journal of the American Statistical Association*, **68**, 880-889.

Särndal, C.E. (1980), On π-inverse weighting versus best linear unbiased weighting in probability sampling, *Biometrika*, **67**, 639-650.

Silva, P.L.N. and Skinner, C. (1997), Variable selection for regression estimation in finite populations, *Survey Methodology*, **23**, 23-32.

Swindel, B.F. and Chapman, D.D. (1973), Good Ridge Estimators, Abstracts Booklet, Joint Statistical Meetings in New York City, 126.

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 Royal Statistical Society*, **36**, 102-106.

Trenkler, G. (1984), On the performance of biased estimators in the linear regression model with correlated or heteroscedastic errors, *Journal of Econometrics*, **25**, 179-190.

Vinod, H. D., and Ullah, A. (1981), Recent advances in regression methods, *Statistics: Textbooks and Monographs*, New York: Marcel Dekker Inc. **41**.