Shrinkage averaging estimation

Michael Schomaker

Statistical Papers

ISSN 0932-5026
Volume 53
Number 4

Stat Papers (2012) 53:1015-1034
DOI 10.1007/s00362-011-0405-2

Springer

Stat Papers (2012) 53:1015–1034
DOI 10.1007/s00362-011-0405-2

REGULAR ARTICLE

Shrinkage averaging estimation

Michael Schomaker

Received: 3 December 2010 / Revised: 13 September 2011 / Published online: 2 October 2011

Abstract We discuss the impact of tuning parameter selection uncertainty in the context of shrinkage estimation and propose a methodology to account for problems arising from this issue: Transferring established concepts from model averaging to shrinkage estimation yields the concept of shrinkage averaging estimation (SAE) which reflects the idea of using weighted combinations of shrinkage estimators with different tuning parameters to improve overall stability, predictive performance and standard errors of shrinkage estimators. Two distinct approaches for an appropriate weight choice, both of which are inspired by concepts from the recent literature of model averaging, are presented: The first approach relates to an optimal weight choice with regard to the predictive performance of the final weighted estimator and its implementation can be realized via quadratic programming. The second approach has a fairly different motivation and considers the construction of weights via a resampling experiment. Focusing on Ridge, Lasso and Random Lasso estimators, the properties of the proposed shrinkage averaging estimators resulting from these strategies are explored by means of Monte-Carlo studies and are compared to traditional approaches where the tuning parameter is simply selected via cross validation criteria. The results show that the proposed SAE methodology can improve an estimators' overall performance and reveal and incorporate tuning parameter uncertainty. As an illustration, selected methods are applied to some recent data from a study on leadership behavior in life science companies.

Keywords Tuning parameter selection uncertainty · Model averaging · Model selection · Optimal weight choice · Cross validation

M. Schomaker (✉)
Centre for Infectious Diseases and Epidemiology Research, University of Cape Town, Anzio Road, Observatory, 7925 Cape Town, South Africa
e-mail: michael.schomaker@uct.ac.za

Mathematics Subject Classification (2000) 62

1 Introduction

Shrinkage estimation has become an important technique for the statistical modeling of data. Since the breakthrough of the Ridge estimator in the early 1970s (Hoerl et al. (1970)) various other approaches such as the Lasso (Tibsharani (1996)), SCAD (Fan and Li (2001)), the Elastic Net (Zou and Hastie (2005)), VISA (Radchenko and James (2008)), and numerous variations thereof were proposed to improve prediction accuracy, interpretability and stability of shrinkage estimation for particular applications. It is common to many of these estimators that the amount of shrinkage is controlled via constraints that impose bounds on the model parameters. Typically these bounds are specified through tuning parameters which are selected most widely via cross validation (Stone (1974), Golub etal. (1979)) or other criteria (see, e.g., Wang et.al. (2009) and the references therein). The choice of these tuning parameters relates to the choice for a particular statistical model and in this respect the task of tuning parameter selection may be viewed as a problem of model selection.

However, although tuning parameter selection via popular model selection criteria has proved to be useful for various situations (see Valdés-Sosa et.al. (2005), Oh et al. (2004), Tibsharani (1996) among others), it is without any doubt also a data-driven procedure which introduces additional uncertainty into the process of statistical modeling: typically, tuning parameters are not given a priori and hence, in reality, the properties of shrinkage estimators depend on the way these parameters have been selected in addition to the stochastic nature of the underlying model. It is a question of high concern whether any tuning parameter selection process affects the quality of shrinkage estimators in terms of stability, efficiency and variability or not.

In fact, it is well-known from the literature of model averaging that estimators post model selection are not necessarily stable, and the uncertainty associated with the selection process often yields estimators and confidence intervals of poor overall quality, see, e.g., Wang et.al. (2009b) for an overview of arguments and literature from a frequentist point of view. A review of the work conducted over the past few years suggests that neglecting model selection uncertainty typically induces a sort of double whammy: First, the variability associated with estimators will be underestimated and the corresponding confidence intervals tend to be too overoptimistic (Burnham and Anderson (2002), Hjort and Claeskens (2003)); second, estimators post model selection often have a poor mean square (prediction) error performance (Hansen (2007), Hansen and Racine (2011)).

Model averaging is an alternative to model selection and aims in combining estimates from a set of competing models to incorporate model uncertainty into the conclusions about the unknown parameters. This strategy has proved to be successful in many situations and typically overcomes the problems related to model selection (Hoerl et.al. (1999), Magnus et al. (2010), Wang et.al. (2009b), Hjort and Claeskens (2003)). Most often the risk in estimation can be improved by model averaging because it provides a kind of insurance against selecting a very poor model. In this article, we view tuning parameter selection as a problem of model selection and transfer some

interesting ideas from a model averaging background to the framework of shrinkage estimation. Treating the tuning parameter selection process as a source of uncertainty, and combining shrinkage estimators based on different tuning parameters in the spirit of model averaging, has the intention to

- improve stability of shrinkage estimators and make them more secure against an inadequate tuning parameter choice,
- improve the predictive performance of shrinkage estimators,
- provide estimates that reflect selection uncertainty, e.g., via suitable standard errors.

In detail, we clarify questions related to tuning parameter uncertainty by means of Monte Carlo experiments and the application of two distinct approaches: both of them rely on the idea that a *weighted* combination of shrinkage estimators with different tuning parameters may not only reveal and account for selection uncertainty but also help to improve and stabilize their overall performance in comparison with the standard approach where a "winning" tuning parameter yields the final estimate. These two approaches differ in the way the weights are chosen and reflect different paradigms currently discussed in the modern model averaging literature (see also Hjort and Claeskens (2006), Hansen (2007), Buchholz et al. (2008), Wan and Zhang (2010), Schomaker and Wan (2010), Wan et.al. (2010), Liang etal. (2011) and the references therein). The first approach relates to an optimal weight choice with regard to the predictive performance of the final weighted estimator: the weights are chosen in such a way that a weighted cross validation criterion is minimized. This idea is inspired by the work of Hansen and Racine (2011) and may be viewed as a generalization of these approaches to the context where different tuning parameters compete against each other instead of different covariates. The second approach is somewhat related to a bootstrap procedure described in Buckland et al. (1997) and considers weighting in terms of plausibility rather than optimality: the data are bootstrapped and the relative frequency of cases where a certain tuning parameter is chosen corresponds to a measure of importance which may be used for the construction of weighted shrinkage estimators.

The article proceeds with an introduction into the proposed concept of shrinkage averaging estimation (SAE) in Sect. 2. Several important issues related to this new approach, such as an adequate weight choice, are discussed there. Section 3 investigates the finite sample properties of selected Ridge estimators through a variety of Monte Carlo experiments. A real-data example from a study on leadership behavior in life science companies is presented in Sect. 4. Sect. 5 is devoted to discussions, extensions and simulations related to Lasso-type estimators. Numerical investigations, also in high-dimensioning data settings, give further insight into the perspectives of the proposed methodology, followed by concluding remarks in Sect. 6.

2 Shrinkage averaging estimation

Consider some data $\mathcal{D}$ consisting of independent observations $(y_i, \mathbf{x}_i)$, $i = 1, \ldots, n$. Denote by $\mathbf{y} = (y_1, \ldots, y_n)'$ the $n \times 1$ vector of response values, let $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ be the $1 \times p$ vector containing the i^{th} observation of each of the p covariates, define $\mathbf{X}_j = (x_{1j}, \ldots, x_{nj})'$ as the $n \times 1$ vector of values of the j^{th} covariate and term

$\mathbf{X} = (\mathbf{x}_1', \ldots, \mathbf{x}_n')'$ as the matrix of all covariates. Now, consider the conditional mean $\mu_i = \mathbb{E}(y_i|\mathbf{x}_i)$ so that $y_i = \mu_i + \epsilon_i$, $\mathbb{E}(\epsilon_i = 0)$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)'$ and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)'$. Assume without loss of generality that $\boldsymbol{\mu} = \mathbf{X}\beta, \boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$. Any estimator $\hat{\boldsymbol{\beta}}$ that relies on shrinkage estimation may be termed as $\hat{\boldsymbol{\beta}}_{\text{SE}}(\hat{\lambda})$ and is of primary interest in this section. To relate the response with a fixed set of regressors, one may consider a certain estimator such as the Ridge estimator,

$$\hat{\boldsymbol{\beta}}_{\text{RE}}(\lambda) = \arg\min \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}, \tag{1}$$

or the Lasso estimator,

$$\hat{\boldsymbol{\beta}}_{\text{LE}}(\lambda) = \arg\min \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}. \tag{2}$$

The complexity parameter $\lambda \geq 0$ tunes the amount of shrinkage and is typically estimated via the generalized cross validation criterion (GCV) or any other cross validation criterion (CV). The larger the value of λ, the greater the amount of shrinkage since the estimated coefficients are shrunk towards zero. The general notation $\hat{\boldsymbol{\beta}}_{\text{SE}}(\hat{\lambda})$ indicates that inference for shrinkage estimation is a two-step procedure whereby the first step relates to the choice of some tuning or complexity parameter λ and the second step refers to the estimation of the model parameters $\boldsymbol{\beta}$ conditional on the chosen $\hat{\lambda}$.

It is clear that in many statistical applications a set of tuning parameters, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_L)$, produces L different shrinkage estimators $\hat{\boldsymbol{\beta}}_{\text{SE}}(\lambda_i) = (\hat{\beta}_{1,\text{SE}}(\lambda_i), \ldots, \hat{\beta}_{p,\text{SE}}(\lambda_i))'$, $i = 1, \ldots, L$, and that the $\lambda_i \in \boldsymbol{\lambda}$ which minimizes a certain criterion is the "winning" λ_i upon which the final estimator will be conditioned. In this respect, an unconditional shrinkage estimator is an estimator that combines the L different estimators over the set of $\boldsymbol{\lambda}$. If each estimator $\hat{\boldsymbol{\beta}}_{\text{SE}}(\lambda_i)$ maintains a specific weight w_{λ_i}, then a *shrinkage averaging estimator* takes the form

$$\hat{\hat{\boldsymbol{\beta}}}_{\text{SAE}} = \sum_{i=1}^{L} w_{\lambda_i} \hat{\boldsymbol{\beta}}_{\text{SE}}(\lambda_i) = \mathbf{w}_{\lambda} \hat{\boldsymbol{B}}_{\text{SE}}, \tag{3}$$

where $\lambda_i \in [0, c]$, $c > 0$ is a suitable constant, $\hat{\boldsymbol{B}}_{\text{SE}} = (\hat{\boldsymbol{\beta}}_{\text{SE}}(\lambda_1), \ldots, \hat{\boldsymbol{\beta}}_{\text{SE}}(\lambda_L))'$ is the $L \times p$ matrix of the estimators, e.g., Lasso or Ridge, $\mathbf{w}_{\lambda} = (w_{\lambda_1}, \ldots, w_{\lambda_L})$ is an $1 \times L$ weight vector, $\mathbf{w}_{\lambda} \in \mathcal{W}$ and $\mathcal{W} = \{\mathbf{w}_{\lambda} \in [0, 1]^L : \mathbf{1}'\mathbf{w}_{\lambda} = 1\}$.

Any standard procedure that singles out a "winning" tuning parameter, such as the generalized cross validation criterion GCV, is a special case of (3) by assigning a value of 1 to a particular w_{λ_i} and 0 to all other w_{λ_i}'s. For instance, the weight choice

$$\hat{w}_{\hat{\lambda}_i}^{\text{GCV}} = \begin{cases} 1, & \text{if } \hat{\lambda}_i = \arg\min_{\boldsymbol{\lambda}} \text{GCV}(\boldsymbol{\lambda}) \\ 0, & \text{else} \end{cases} \tag{4}$$

relates to the current practice of choosing the tuning parameter which minimizes the GCV.

In fact, shrinkage tuning parameters are continuous parameters and hence a 'real' unconditional estimator with respect to λ would be an estimator that integrates λ out, i.e., $\int w_\lambda \boldsymbol{\beta}_{\mathrm{SE}}(\lambda) d\lambda$. Using the weighted sum over the discrete set $\boldsymbol{\lambda}$, as indicated in (3), assumes an adequate choice and size of this set. Our presentation of simulation and data results give further insight on how the discrete set may look in a specific analysis, what are the implications in practical situations and how stable numerical results really are.

Note, that although the concept of a weighted combination of shrinkage estimators is similar to that of model averaging there are important differences: Especially, in Shrinkage Averaging Estimation the estimator $\hat{\bar{\boldsymbol{\beta}}}_{\mathrm{SAE}}$ averages over a particular sequence of λ_i of size L for a *fixed* set of regressors while model averaging estimators of the form $\hat{\boldsymbol{\beta}}_{\mathrm{MAE}} = \sum_{\kappa=1}^{k} w_\kappa \hat{\boldsymbol{\beta}}_\kappa$ average over a sequence of models, whereby typically each estimator of a particular model refers to a certain subset of regressors $\mathbf{X}_\kappa \subset \mathbf{X}$.

In the following two sections we discuss the weight choice for (3) based on concepts and arguments borrowed from the model averaging literature.

2.1 Cross-validation optimal weighting

It is well-known that many shrinkage estimators, also the Ridge and Lasso estimator, aim in improving the prediction accuracy compared to standard procedures such as ordinary least squares estimation. Therefore it may be desirable to construct a weight choice that enables the use of stable SA-estimators with good predictive performance. In many fields of statistical modeling, cross validation serves as an estimate of the expected prediction error and any weighted estimator that minimizes an estimated out-of-sample prediction error seems to be suitable in this context.

Now, consider $1 < k \leq n$ disjunct partitions of the data $\mathcal{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_k\}$ where $\mathcal{D}_\kappa$, $\kappa = 1, \ldots, k$, denotes the κ^{th} subset of the observations. Thus, each subset $\mathcal{D}_\kappa$ consists of the $\frac{n}{k}$ observations $(y_i, \mathbf{x}_i)$, $i = 1 + \frac{(\kappa-1)n}{k}, \ldots, \frac{n}{k} + \frac{(\kappa-1)n}{k}$. A k-fold cross validation procedure allows the estimation of

$$\hat{\boldsymbol{\mu}}_{\lambda_i,\kappa} = \mathbf{X}_\kappa \{\hat{\boldsymbol{\beta}}_{\mathrm{SE}}(\lambda_i) | \mathbf{X}_{-\kappa}\} \tag{5}$$

which is the estimated conditional mean of the κ^{th} subset of observations due to the corresponding covariates $\mathbf{X}_\kappa$, any given λ_i , and the shrinkage estimator $\hat{\boldsymbol{\beta}}_{\mathrm{SE}}(\lambda_i)|\mathbf{X}_{-\kappa}$ based on the data $\mathcal{D}_{-\kappa}$ that do not contain the observations $\mathcal{D}_\kappa$. Obviously, the k-fold cross validation $n \times 1$ vector of the conditional mean is $\hat{\boldsymbol{\mu}}_{\lambda_i} = (\hat{\boldsymbol{\mu}}'_{\lambda_i,1}, \ldots, \hat{\boldsymbol{\mu}}'_{\lambda_i,k})'$. Now, for a given tuning parameter λ_i and a certain k, define the cross-validation residuals

$$\begin{aligned} \tilde{\boldsymbol{\epsilon}}_k(\lambda_i) &= (\tilde{\epsilon}_1, \ldots, \tilde{\epsilon}_n | k, \lambda_i)' \\ &= (\tilde{\boldsymbol{\epsilon}}'_1, \ldots, \tilde{\boldsymbol{\epsilon}}'_k | k, \lambda_i)' \end{aligned} \tag{6}$$

where $\tilde{\boldsymbol{\epsilon}}_\kappa = y_\kappa - \hat{\boldsymbol{\mu}}_{\lambda_i,\kappa}$, $\kappa = 1, \ldots, k$. Consequently, the weighted shrinkage based cross-validation residuals are

$$\begin{aligned}\tilde{\boldsymbol{\epsilon}}_\kappa(w) &= \sum_{i=1}^{L} w_{\lambda_i}\tilde{\boldsymbol{\epsilon}}_k(\lambda_i)\\ &= \mathbf{E}_k\mathbf{w}_\lambda{}'\end{aligned} \tag{7}$$

where $\mathbf{E}_k = (\tilde{\boldsymbol{\epsilon}}_k(\lambda_1), \ldots, \tilde{\boldsymbol{\epsilon}}_k(\lambda_L))$ is the $n \times L$ matrix of the cross-validation residuals for the L competing tuning parameters. Define $\tilde{\mathbf{E}}_k = \mathbf{E}_k'\mathbf{E}_k$, then a weighted cross validation criterion is

$$\begin{aligned}\text{OCV}_k &= \frac{1}{n}\tilde{\boldsymbol{\epsilon}}_\kappa(w)'\tilde{\boldsymbol{\epsilon}}_\kappa(w)\\ &\propto \mathbf{w}_\lambda\mathbf{E}_k'\mathbf{E}_k\mathbf{w}_\lambda{}' = \mathbf{w}_\lambda\tilde{\mathbf{E}}_k\mathbf{w}_\lambda{}'.\end{aligned} \tag{8}$$

The SAE $\hat{\tilde{\boldsymbol{\beta}}}_{\text{SAE}} = \mathbf{w}_\lambda\hat{\boldsymbol{B}}_{\text{SE}}$ that minimizes (8) can be considered optimal in the sense of prediction accuracy and therefore the weight choice

$$\hat{\mathbf{w}}_\lambda^{\text{OCV}} = \arg\min_{\mathbf{w}_\lambda \in \mathcal{W}} \text{OCV}_k \tag{9}$$

yields the *Optimal Cross Validation Shrinkage Averaging Estimator*

$$\hat{\tilde{\boldsymbol{\beta}}}_{\text{OCV}} = \hat{\mathbf{w}}_\lambda^{\text{OCV}}\hat{\boldsymbol{B}}_{\text{SE}} \tag{10}$$

and as a special case of (10) the *Optimal Cross Validation Ridge Averaging Estimator* and the *Optimal Cross Validation Lasso Averaging Estimator*, respectively,

$$\hat{\tilde{\boldsymbol{\beta}}}_{\text{OCV}}^{\text{Ridge}} = \hat{\mathbf{w}}_\lambda^{\text{OCV}}\hat{\boldsymbol{B}}_{\text{RE}}\,, \qquad \hat{\tilde{\boldsymbol{\beta}}}_{\text{OCV}}^{\text{Lasso}} = \hat{\mathbf{w}}_\lambda^{\text{OCV}}\hat{\boldsymbol{B}}_{\text{LE}}. \tag{11}$$

Due to the inequality restriction $\mathbf{w}_\lambda \in \mathcal{W}$, $\mathcal{W} = \{\mathbf{w}_\lambda \in [0,1]^L : \mathbf{1}'\mathbf{w}_\lambda = 1\}$ which requires the weight vector to lie on the $\mathbb{R}^L$ unit simplex and sum up to one, there is no closed form solution for the determination of (9). Thus, since the criterion OCV_k is a quadratic function of the weight vector $\mathbf{w}_\lambda$, the optimal weights $\hat{\mathbf{w}}_\lambda^{\text{OCV}}$ may obtained via quadratic programming for which numerical algorithms are well-developed and available within almost every statistical and mathematical software package. For instance, the analyses in this paper are conducted with the `quadprog` package of the statistical software *R* (R Development Core Team (2010)).

It is worthwhile to note that optimal shrinkage averaging procedures as proposed above bear resemblance to other methodologies discussed in recent statistical literature: As mentioned above, the work of Hansen and Racine (2011) inspired our concept of optimal weighting in the sense of combining estimators with respect to an improvement of the mean squared prediction error. However, although the "Jackknife Model Averaging" (JMA) approach introduced by Hansen and Racine (2011) shares several basic concepts with the SAE methodology, such as the use of weighted residuals to obtain a suitable weight vector, it also differs in many aspects: First of all, Hansen and Racine (2011) investigate model averaging schemes mainly in conjunction with the classical linear regression model and not within a shrinkage

framework. Therefore, as a second issue, their weight vector relates to candidate models which refer to a certain subset of regressors $\mathbf{X}_k \subset \mathbf{X}$ instead of different tuning parameters $\lambda_i \in \boldsymbol{\lambda}$. Third, the authors draw their attention only on leave-one-out cross validation for the reason of some selected theoretical aspects, while this work also considers k-fold cross validation. Even though leave-one-out cross validation can be viewed to be optimal in some sense (Shao (1997)) it is well-known that for several practical purposes 5- or 10-fold cross-validation may be preferred (see, e.g., Breiman and Spector (1992), Kohavi (1995) and Hastie etal. (2001) for discussions on that issue).

2.2 Bootstrap-based weighting and visualization

The perspective of the former section relates to the view that an inadequate tuning parameter choice may corrupt the predictive performance (mean squared prediction error) of any shrinkage estimator and hence a correction for this issue may be the crucial point. However, another point of view directly addresses the first issue of the double whammy of model selection uncertainty: the underestimation of variability. Typically, regarding this problem, the model averaging literature suggests to construct a measure of importance corresponding to each candidate estimator, for instance any model selection criterion such as the AIC, and to use this measure for the construction of weights. These weights (e.g., exponential AIC- or BIC-weights) then can be used for corrected estimators, including corrected standard errors that indicate the higher variability due to model selection uncertainty. The justification for these procedures are based on bayesian or likelihood related arguments (e.g., exponential BIC-weights may be viewed as a coarse approximation of the posteriori probability of any specific model to be correct) and it is clear that in the current context of shrinkage estimation a direct adaption of this concept may not be suitable. Hence, the procedure described below characterizes the plausibility of an estimator not directly through criteria-based weights – but uses selection criteria indirectly via a resampling experiment to reveal the impact of tuning parameter selection uncertainty and construct shrinkage averaging weights.

The standard error approximation formula (13) is a direct application of the formula given in Buckland et al. (1997) and accounts for both the variability of the model parameters given a certain tuning parameter (the first term under the root) and the variability associated *between* the estimates over the $\boldsymbol{\lambda}$ sequence (the second term under the root). The justification for (13) relies on the assumptions that the weights are non-random and that the estimators are perfectly correlated and we refer the reader to Burnham and Anderson (2002) for a detailed derivation. Although these assumptions seem to be strong at first, there is numerous work (e.g., Candolo et al. (2003), Burnham and Anderson (2002)) which indicates that (13) can be an adequate approximation within some typical model averaging situations. We will explore the finite sample performance of (13) in the context of shrinkage estimation uncertainty briefly in Sect. 3. Note that the approximation-formula for the standard error can also be applied for the optimal cross validation SAE of Sect. 2.1.

Shrinkage Averaging Estimation via Bootstrap Weighting

Step 1: Generate B bootstrap samples $\mathcal{D}^1, \ldots, \mathcal{D}^B$ with replacement.

Step 2: Calculate the shrinkage estimators $\hat{\boldsymbol{\beta}}_{\text{SE}}(\lambda_i)$, $i = 1, \ldots, L$, and their corresponding estimated tuning parameter $\hat{\lambda}_i \in \boldsymbol{\lambda}$ for each sample. The tuning parameter may be selected via some criterion Γ in each bootstrap sample. Let $b(\lambda_i, \Gamma)$ denote the amount of bootstrap samples for which λ_i is chosen through Γ.

Step 3: Consider the sequence of tuning parameters $\boldsymbol{\lambda}$ and the relative frequency $w_{\lambda_i}^{\text{BW}} = b(\lambda_i, \Gamma)/B$ related to each possible $\lambda_i \in \boldsymbol{\lambda}$. A shrinkage averaging estimator based on bootstrap weights is given by

$$\hat{\tilde{\boldsymbol{\beta}}}_{\text{BW}} = \sum_{i=1}^{L} w_{\lambda_i}^{\text{BW}} \hat{\boldsymbol{\beta}}_{\text{SE}}(\lambda_i). \tag{12}$$

Step 4: Estimate the standard error for each $\hat{\tilde{\beta}}_{j,\text{BW}} \in \hat{\tilde{\boldsymbol{\beta}}}_{\text{BW}}$ via

$$\widehat{\text{s.e.}}(\hat{\tilde{\beta}}_{j,\text{BW}}) = \sum_{i=1}^{L} w_{\lambda_i}^{\text{BW}} \sqrt{\widehat{\mathbb{V}\text{ar}}\{\hat{\beta}_{j,\text{SE}}(\lambda_i)\} + \{\hat{\beta}_{j,\text{SE}}(\lambda_i) - \hat{\tilde{\beta}}_{j,\text{BW}}\}^2} \tag{13}$$

3 Simulation studies

As with all procedures developed, it is desirable to investigate their properties based on a finite amount of data. We do so here by means of a Monte-Carlo study with attention confined to a broader class of ridge estimators at first. In Sect. 5 we also consider the Lasso estimator, the Random Lasso estimator, and test their performance in both low- and high-dimensional data-settings.

3.1 Outline

In the following four simulation examples, we consider the ordinary least squares estimate (OLS), the ridge estimator based on GCV tuning parameter selection (GCV), the ridge estimator based on 5-fold cross validation tuning parameter selection (CV_5), the optimal cross validation ridge averaging estimator (11) based on a 5-fold weighted cross validation criterion (OCV_5) and the SAE (12) that uses Bootstrap weights via $\Gamma = \text{GCV}$ and $B = 500$ (BW_{GCV}). We are especially interested in comparing

- $\underline{\text{OCV}_5 \leftrightarrow \text{CV}_5}$, i.e., can an optimal ridge averaging estimator based on a 5-fold weighted cross validation criterion (OCV_5) improve the prediction accuracy compared to the corresponding ridge estimator that results from a tuning parameter selection via 5-fold cross validation (CV_5)?
- $\underline{\text{BW}_{\text{GCV}} \leftrightarrow \text{GCV}}$, i.e., does the ridge averaging estimator based on bootstrap weights better incorporate the selection uncertainty than the ridge estimator whose tuning parameter is selected via GCV?

These questions relate to find out whether the proposed shrinkage averaging estimators fulfill the requirements they are created for. It is not our purpose to compare cross

validation and generalized cross-validation. As already mentioned before, this task has been discussed in the literature extensively and we refer the reader again to Breiman and Spector (1992), Kohavi (1995) and Hastie etal. (2001) and the references therein for an insight on that issue. Of course, also optimal ridge averaging estimators based on a weighted leave-one-out or 10-fold cross validation criterion (OCV_1, OCV_{10}) could have been considered as well as estimators on bootstrap weights that rely on other criteria than $\Gamma = \text{GCV}$ (BW_Γ); however, for a first insight on the questions imposed above, the current estimators will be satisfactory.

All four experiments rely on $\mathcal{R} = 1000$ simulation runs, whereby in each run the $n_{\text{train}} \times p + 1$ data are generated from the model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \sigma\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim N(0, 1).$$

The five estimators are compared by means of their median mean squared prediction error related to the test data of size $n_{\text{test}} \times p+1$ and the estimated mean squared error of the parameter vector. Moreover, both the quality of the standard errors and the impact of the tuning parameter selection uncertainty will be analyzed with respect to various issues. The $\boldsymbol{\lambda}$-sequence which is considered throughout the estimation process for all of the four ridge estimators corresponds to the values of $\lambda_0 = 0$ and furthermore $10^{\boldsymbol{\nu}}$, where $\boldsymbol{\nu}$ is an equally spaced vector of size 100 between -2 and 8. In all experiments only standardized covariate data are used for the estimation procedure.

- **Experiment 1** In this experiment we set $n_{\text{train}} = 50$, $n_{\text{test}} = 50$ and $p = 20$. The observations for the covariates are generated by the following distributions: $\mathbf{X}_1, \mathbf{X}_6, \mathbf{X}_{11}, \mathbf{X}_{16} \sim \text{N}(0.5, 1)$, $\mathbf{X}_2, \mathbf{X}_7, \mathbf{X}_{12}, \mathbf{X}_{17} \sim \log\text{N}(0.5, 0.5)$, $\mathbf{X}_3, \mathbf{X}_8, \mathbf{X}_{13}, \mathbf{X}_{18} \sim \text{Weibull}(1.75, 1.9)$, $\mathbf{X}_4, \mathbf{X}_9, \mathbf{X}_{14}, \mathbf{X}_{19} \sim \text{Bin}(1, 0.5)$ and $\mathbf{X}_5, \mathbf{X}_{10}, \mathbf{X}_{15}, \mathbf{X}_{20} \sim \text{Gamma}(0.25, 2)$. To model the dependency between the covariates we use a Clayton Copula with a copula parameter of 3 which indicates rather high correlation among the covariates and refers to a typical situation where the ridge estimator may be used. The *R*-package `copula` (Yan (2007)) provides an efficient tool to utilize this approach. The values of the response vector are realized via draws from $\mathbf{y} \sim \text{N}(\mu_1, \sigma)$ with $\mu_1 = \mathbf{X}\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_1 = (3, 3, 0, \ldots, 0)'$, $\sigma = 2.5$. This means, in this experiment we consider a small set of data with medium-to-high correlation within the covariates where only two out of the 20 variables influence the response vector.

- **Experiment 2** In this experiment we set $n_{\text{train}} = 40$, $n_{\text{test}} = 50$ and $p = 5$. The observations for the covariates are generated by the following distributions: $\mathbf{X}_1 \sim \text{N}(0.5, 1)$, $\mathbf{X}_2 \sim \log\text{N}(0.5, 0.5)$, $\mathbf{X}_3 \sim \text{Weibull}(1.75, 1.9)$, $\mathbf{X}_4 \sim \text{Bin}(1, 0.5)$ and $\mathbf{X}_5 \sim \text{Gamma}(0.25, 2)$. The copula parameter again is 3 and the values of the response vector are realized via draws from $\mathbf{y} \sim \text{N}(\mu_2, \sigma)$ with $\mu_2 = \mathbf{X}\boldsymbol{\beta}_2$, $\boldsymbol{\beta}_2 = (1.5, 1.25, 1, 0.75, 0.5)'$, $\sigma = 2.5$. Thus, in this experiment we consider a data-set where all of the 5 variables have some influence on the response vector.

- **Experiment 3** In this experiment we set $n_{\text{train}} = 150$, $n_{\text{test}} = 50$ and $p = 50$. The observations for the covariates are generated by the following distributions:

$\mathbf{X}_1, \ldots, \mathbf{X}_{50} \sim \mathrm{N}(0, 1)$. The copula parameter again is 3 and the values of the response vector are realized via draws from $\mathbf{y} \sim \mathrm{N}(\mu_3, \sigma)$ with $\mu_3 = \mathbf{X}\boldsymbol{\beta}_3$, $\boldsymbol{\beta}_3 = (3, 3, 3, 3, 1.5, 1.5, 1.5, 1.5, 0, \ldots, 0)'$, $\sigma = 2.5$. Consequently, in this experiment we consider the case of a multitude of covariates where 8 out of the 50 variables have influence on the response vector.

– **Experiment 4** This experiment is a variation of experiment 1 and we set $n_{\text{train}} = 50$, $n_{\text{test}} = 50$ and $p = 15$. The observations for the covariates are generated by the following distributions: $\mathbf{X}_1, \mathbf{X}_6, \mathbf{X}_{11} \sim \mathrm{N}(0.5, 1)$, $\mathbf{X}_2, \mathbf{X}_7, \mathbf{X}_{12} \sim \log \mathrm{N}(0.5, 0.5)$, $\mathbf{X}_3, \mathbf{X}_8, \mathbf{X}_{13} \sim \text{Weibull}(1.75, 1.9)$, $\mathbf{X}_4, \mathbf{X}_9, \mathbf{X}_{14} \sim \mathrm{Bin}(1, 0.5)$ and $\mathbf{X}_5, \mathbf{X}_{10}, \mathbf{X}_{15} \sim \text{Gamma}(0.25, 2)$. The copula parameter again is 3 and the values of the response vector are realized via draws from $\mathbf{y} \sim \mathrm{N}(\mu_4, \sigma)$ with $\mu_4 = \mathbf{X}\boldsymbol{\beta}_4$, $\boldsymbol{\beta}_4 = (6, 5.5, 5, 4.5, 4, \ldots, 1.5, 1, 0.5, 0, 0, 0)'$, $\sigma = 2.5$. This means in this experiment we consider a small set of data where 12 out of 15 variables influence the response vector to a different degree.

3.2 Results

MSE-performance The estimated median mean squared prediction error and the estimated mean squared error for the model parameters for all four simulation experiments are presented in Table 1.

It can be seen that all Ridge estimators outperform the simple OLS estimator both with respect to the MMSPE and the MSE. The optimal cross validation ridge averaging estimator (OCV_5) is especially designed for the purpose to improve the predictive

Table 1 Estimated median mean squared prediction errors (MMSPE) and mean squared errors (MSE) for the simulation experiments

	Experiment 1		Experiment 2	
	MMSPE	MSE	MMSPE	MSE
OLS	12.711	0.741	8.029	0.372
OCV_5	10.482	0.403	7.687	0.168
CV_5	10.570	0.424	7.770	0.168
$\mathrm{BW}_{\mathrm{GCV}}$	11.255	0.503	7.660	0.189
GCV	10.462	0.399	7.049	0.167
	Experiment 3		Experiment 4	
	MMSPE	MSE	MMSPE	MSE
OLS	15.615	0.305	34.881	1.236
OCV_5	14.292	0.206	34.422	1.151
CV_5	14.357	0.206	34.568	1.163
$\mathrm{BW}_{\mathrm{GCV}}$	14.593	0.228	34.198	1.147
GCV	14.314	0.204	34.445	1.136

Table 2 Selected unconditional and empirical standard errors (in brackets) for simulation experiment 3

	$\mathbb{V}\text{ar}(\mathbf{X}_1)$		$\mathbb{V}\text{ar}(\mathbf{X}_2)$		$\mathbb{V}\text{ar}(\mathbf{X}_3)$		$\mathbb{V}\text{ar}(\mathbf{X}_4)$		$\mathbb{V}\text{ar}(\mathbf{X}_5)$		$\mathbb{V}\text{ar}(\mathbf{X}_6)$	
OLS	0.54	(0.58)	0.54	(0.58)	0.54	(0.57)	0.55	(0.57)	0.55	(0.55)	0.54	(0.55)
BW_{GCV}	0.49	(0.51)	0.49	(0.51)	0.49	(0.50)	0.49	(0.51)	0.48	(0.47)	0.47	(0.47)
GCV	0.36	(0.47)	0.36	(0.47)	0.36	(0.46)	0.36	(0.47)	0.35	(0.41)	0.35	(0.40)
OCV_5	0.55	(0.46)	0.54	(0.46)	0.55	(0.45)	0.55	(0.47)	0.43	(0.40)	0.42	(0.39)
CV_5	0.36	(0.47)	0.36	(0.48)	0.36	(0.47)	0.36	(0.49)	0.35	(0.41)	0.35	(0.41)
	$\mathbb{V}\text{ar}(\mathbf{X}_7)$		$\mathbb{V}\text{ar}(\mathbf{X}_8)$		$\mathbb{V}\text{ar}(\mathbf{X}_9)$		$\mathbb{V}\text{ar}(\mathbf{X}_{10})$		$\mathbb{V}\text{ar}(\mathbf{X}_{11})$		$\mathbb{V}\text{ar}(\mathbf{X}_{12})$	
OLS	0.54	(0.54)	0.54	(0.59)	0.54	(0.53)	0.55	(0.56)	0.54	(0.58)	...	...
BW_{GCV}	0.47	(0.46)	0.47	(0.50)	0.47	(0.44)	0.47	(0.48)	0.47	(0.49)	...	...
GCV	0.35	(0.40)	0.35	(0.43)	0.35	(0.37)	0.35	(0.40)	0.35	(0.41)	...	...
OCV_5	0.43	(0.39)	0.43	(0.42)	0.39	(0.36)	0.39	(0.39)	0.39	(0.40)	...	...
CV_5	0.35	(0.40)	0.35	(0.42)	0.35	(0.37)	0.35	(0.40)	0.35	(0.41)	...	...

accuracy and therefore to account for the second aspect of the double whammy arising from the neglect of model selection uncertainty. One can see that in all of the four experiments considered, the estimated median mean squared prediction error for the OCV_5 methodology is lower in comparison to the CV_5 approach. In two out of the four experiments also the estimated MSE is lower, in the other two it is about the same. By contrast, in most of the situations considered, the Ridge averaging estimator that uses bootstrap weights (BW_{GCV}) does not improve the performance of the ridge estimator based on a GCV tuning parameter selection with respect both to the MMSPE and the MSE.

Standard errors Now, we compare the averaged estimated unconditional standard error due to (13) with the empirical standard error resulting from the 1000 simulation runs. With regard to the traditional ridge estimators that rely on tuning parameter selection via GCV or CV_5, this formula simply corresponds to the sampling variance of the model parameters associated with the selected tuning parameter. For purpose of exposition, Table 2 presents the results for the first 12 variables of experiment 3.

The results for the variables $\mathbf{X_{13}} - \mathbf{X_{50}}$ are in line with those of the variables $\mathbf{X_9} - \mathbf{X_{12}}$ and the numbers produced for the other experiments lead to fairly similar statements as those from experiment 3. The table reveals some interesting insights into the methodologies considered:

- Examining a simple GCV and CV_5 methodology, it can be seen that for all variables the estimated standard error is always smaller than the empirical standard error. It is irrevocable that the true variability of the estimators is *underestimated* and therefore tuning parameter uncertainty has a considerable influence of the properties of the estimators.
- The BW_{GCV} approach produces estimates that are quite close to each other which indicates that the variance of this estimator is well-reflected by the use of (13) and tuning parameter selection uncertainty is covered adequately in this scenario.

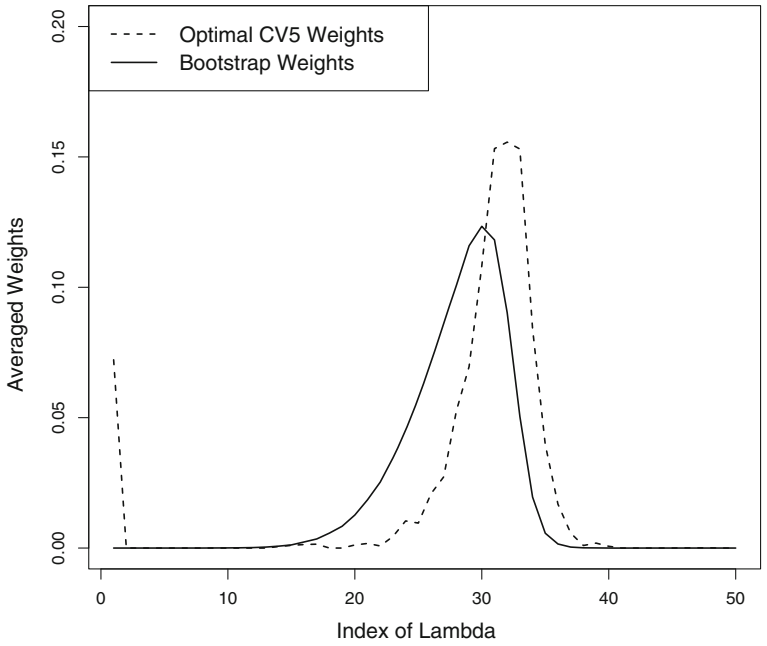

Fig. 1 Visualization of the impact of tuning parameter selection uncertainty via the averaged weights for the relevant part of the $\boldsymbol{\lambda}$-sequence in experiment 2

- The OCV_5 estimator does not seem to underestimate the variability, however, especially for the first four variables (which have the strongest influence on the outcome) one can observe an overestimation of the standard error.
- Comparing the empirical standard errors of all estimators, one can state that the OLS estimator possesses (as expected) the highest variability followed by the BW_{GCV} methodology and subsequent by the other three estimators. The high variability of the bootstrap-based estimator may explain the source of its comparatively poor MSE-performance.

Visualized uncertainty impact To get a deeper understanding of the behavior of the estimators and the tuning parameter selection uncertainty involved, we plot the averaged weights over the $\mathcal{R} = 1000$ simulation runs for each tuning parameter under consideration. We plot these averaged weights against the first 50 index numbers of the λ_i (not the real values) only for the reason to get a more clearly arranged picture. For purpose of exposition, results only of experiment 2 are presented in Fig. 1.

The examination of the averaged bootstrap weights (solid line) offers a very interesting insight on the distribution of the selected λ_i. The picture shows a considerable variation of the selected tuning parameter between λ_{20} and λ_{35} which corresponds to real value numbers between around 0.8 and 27.2. A look at the averaged optimal CV_5 weights (dashed line) reveals a new sight of the OCV methodology: In comparison to the BW_{GCV}, the OCV approach enforces, to a certain amount, a stronger shrinkage of the coefficients since the weights for the larger λ_i are also considerably higher. However, it is apparent that in addition also the weight for the smallest λ_i (which

produces the OLS estimate) is quite high; this offers the conclusion that the Optimal Cross Validation Ridge Averaging Estimator can improve its predictive performance in some situations by combining a certain ridge estimator with the OLS estimate (note that these are averaged weights and only in some of the 1000 simulations runs the smallest λ_i receives a weight greater than zero).

Summary, conclusions and remarks on the experiments Using Ridge Regression and selecting the tuning parameter by GCV or CV_5 yields (as to expect) estimates that outperform the simple OLS estimator. Due to tuning parameter selection uncertainty the variance of these estimators typically will be underestimated. Both SAE considered (OCV_5, BW_{GCV}), yield more realistic standard errors, indicating that selection uncertainty is covered adequately.

While the OCV_5 averaging estimator has a good performance with regard to the mean square prediction error, the BW_{GCV} cannot improve existing methods in three out of the four experiments. Further simulation studies in Sect. 5 confirm that cross-validation optimal weighting typically improves predictive performance, also beyond a simple ridge regression framework.

We remark that some sensitivity analyses on the Monte Carlo experiments indicate that other choices of the $\boldsymbol{\lambda}$-sequence do not affect the results at all and there exist various other scenarios with results similar to the ones reported above.

4 Analysis of leadership data

In this section, we consider some recent data of Klaußner (2007) who analyzes the effect of leadership behavior on the satisfaction of employees with respect to a companies "phase of life". For this purpose, the author collected data from $n = 101$ biotechnology companies via some well-established questionnaire. Klaußner (2007) performs a linear regression analysis and accounts for nine covariates which relate to widely accepted scores for different types of leadership : these variables are contingent reward (CR), management by exception active (MEA) and management by exception passive (MEP) who represent a kind of "give-and-take-basis leadership behavior", then, idealized influence attributed (IIA), idealized influence behavioral (IIB), inspirational motivation (IM), intellectual stimulation (IS) and individual consideration (IC) which stand for a broader, more stimulating sense of leadership, and furthermore Laisser-faire (L) that indicates the amount of passive-leadership. Another covariate is the companies phase of life which is a dummy-coded, 4-categorial variable (formation ($P1$), growth ($P2$), maturity ($P3$) and decline ($P4$)), while the response, which relates to the degree of satisfaction of the employee, is a score variable that is considered to be normally distributed in the appropriate literature. The analyses of Klaußner (2007) reveal not only medium to high correlation between the covariates but also considerable model selection uncertainty concerning the effects of a few variables (IC, MEA, L, Phase and the interaction of Phase and Laisser-faire (I1, I2, I3, I4)). Hence, the application of ridge regression may be suitable to get some more insights on the questions imposed by the author.

We again consider the OLS, GCV, CV_5, OCV_5 and BW_{GCV} estimators and due to the findings of Klaußner (2007) we consider all nine leadership score-variables,

Table 3 Parameter estimates of the Leadership Data and the corresponding standard errors (in brackets)

	IIA	IIB	IM	IS	IC
OLS	0.305 (0.092)	-0.077 (0.080)	-0.001 (0.084)	0.084 (0.088)	0.255 (0.103)
GCV	0.204 (0.047)	-0.019 (0.045)	0.020 (0.042)	0.088 (0.052)	0.184 (0.039)
BW_{GCV}	0.254 (0.082)	-0.045 (0.066)	0.005 (0.063)	0.085 (0.084)	0.219 (0.073)
CV_5	0.192 (0.042)	-0.012 (0.041)	0.025 (0.038)	0.088 (0.046)	0.175 (0.034)
OCV_5	0.189 (0.041)	-0.010 (0.040)	0.026 (0.037)	0.088 (0.045)	0.172 (0.033)
	CR	MEA	MEP	L	P2
OLS	0.092 (0.093)	-0.073 (0.066)	0.012 (0.067)	0.054 (0.180)	0.193 (0.158)
GCV	0.100 (0.047)	-0.063 (0.041)	-0.040 (0.035)	-0.080 (0.045)	0.067 (0.037)
BW_{GCV}	0.098 (0.075)	-0.075 (0.058)	-0.016 (0.049)	-0.058 (0.089)	0.093 (0.070)
CV_5	0.100 (0.042)	-0.058 (0.038)	-0.044 (0.032)	-0.079 (0.041)	0.065 (0.033)
OCV_5	0.099 (0.040)	-0.057 (0.037)	-0.045 (0.032)	-0.078 (0.039)	0.065 (0.032)
	P3	P4	I2	I3	I4
OLS	-0.056 (0.182)	0.117 (0.154)	-0.164 (0.174)	0.011 (0.211)	-0.359 (0.177)
GCV	-0.024 (0.024)	-0.060 (0.047)	-0.000 (0.036)	0.013 (0.041)	-0.140 (0.036)
BW_{GCV}	-0.048 (0.061)	-0.010 (0.088)	-0.041 (0.082)	0.032 (0.095)	-0.203 (0.092)
CV_5	-0.021 (0.021)	-0.064 (0.042)	0.004 (0.033)	0.009 (0.037)	-0.132 (0.031)
OCV_5	-0.020 (0.021)	-0.065 (0.041)	0.005 (0.032)	0.008 (0.035)	-0.130 (0.030)

the companies phase of life and the interaction of phase and Laisser-faire as potential covariates. The $\boldsymbol{\lambda}$-sequence corresponds to the values of $\lambda_0 = 0$ and furthermore $10^{\boldsymbol{\nu}}$, where $\boldsymbol{\nu}$ is an equally spaced vector of size 100 between -2 and 8. The results of our investigation are presented in Table 3.

One can see that there is a remarkable amount of shrinkage induced by all of the four ridge estimators. Independent of the methodology related to the treatment of the tuning parameter, the covariates IIA and IC are found to be fairly important. While the OLS, GCV, CV_5 and OCV_5 estimator produce numbers that indicate a moderate influence of Phase and Laisser-faire (e.g., through the main effects of L and $P2$ and also through the interaction effect $I4$), the BW_{GCV} estimator yields more conservative conclusions with respect to the main effects. Concerning the standard errors of the model parameters, the OLS estimator possesses the highest variability followed by the BW_{GCV} approach and subsequent by the other three estimators.

Figure 2 offers a deeper understanding on how exactly the estimators worked in the current example.

Selecting the tuning parameter via the GCV leads to the tuning parameter $\hat{\lambda}_{36}$, using the CV_5 criterion yields $\hat{\lambda}_{37}$ and the OCV_5 approach compromises between $\hat{\lambda}_{37}$ and $\hat{\lambda}_{38}$ with the weights $\hat{w}_{37} = 0.6931$ and $\hat{w}_{38} = 0.3069$, respectively, and therefore the corresponding estimator employs a slightly stronger shrinkage on the parameters. The BW_{GCV} estimator indicates considerable tuning parameter selection uncertainty and, in accordance with the simulation results, proposes weight choices that are smaller than those of the OCV_5 estimator.

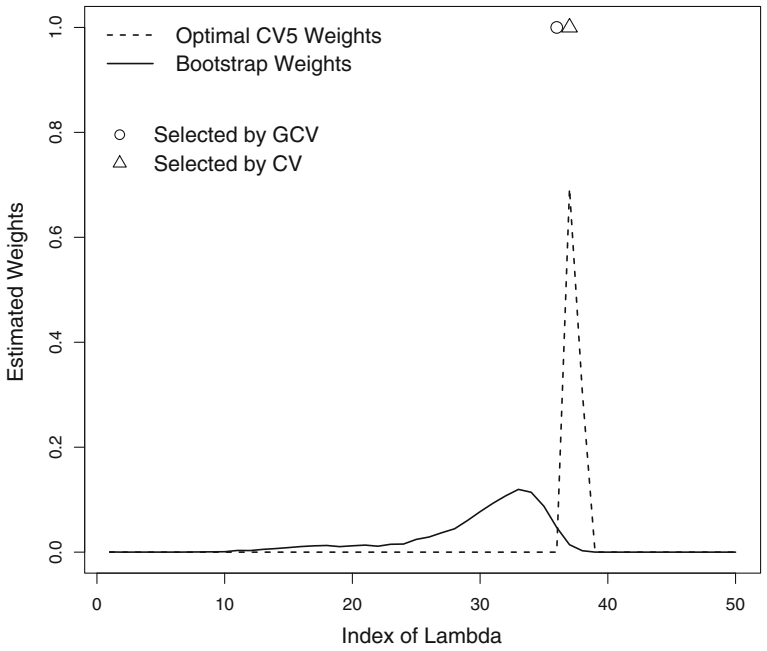

Fig. 2 Visualization of the impact of tuning parameter selection uncertainty via the estimated weights for a part of the $\boldsymbol{\lambda}$-sequence in the leadership data example

5 Lasso-type estimators and high-dimensional data

In this section we focus on the Lasso and the Random Lasso estimator (Wang et.al. (2011)), review experiments 1 and 2 from Sect. 3, and consider two new high-dimensional data settings, experiment 5 and 6.

The Random Lasso is a computationally intensive method to improve the lasso especially in high-dimensional data settings and situation where there are groups of correlated variables, maybe with different signs. It consists of two steps: In the first step B_1 Bootstrap samples are drawn and for each bootstrap sample q_1 variables are randomly selected and the lasso is applied. The average lasso estimator over the B_1 bootstrap sample yields an importance measure for each variable. In the second step B_2 bootstrap samples are drawn and q_2 variables are selected - with selection probability proportional to the corresponding importance measure. The lasso (or a variation thereof) is applied and the average over the B_2 bootstrap samples yields the final Random Lasso estimator. Simulations studies have shown that this estimator can be useful in both high- and low-dimensional data settings.

In general, the motivation behind the Lasso estimator is to have an interpretable estimator with good predictive performance, i.e., an estimator where several parameters can be set to zero to guarantee variable selection. Here, we compare the Lasso based on a tuning parameter selected on 5-fold cross validation (CV_5) with the averaged Lasso estimator based on the OCV weights (OCV_5). The OCV weights are especially attractive to Lasso-type estimators as they also aim in a good predictive performance

Table 4 Estimated median mean squared prediction errors (MMSPE) and mean squared errors (MSE) for the lasso estimators

	Experiment 1		Experiment 2	
	MMSPE	MSE	MMSPE	MSE
OCV_5	9.41	0.14	7.99	0.32
CV_5	9.53	0.14	8.08	0.33

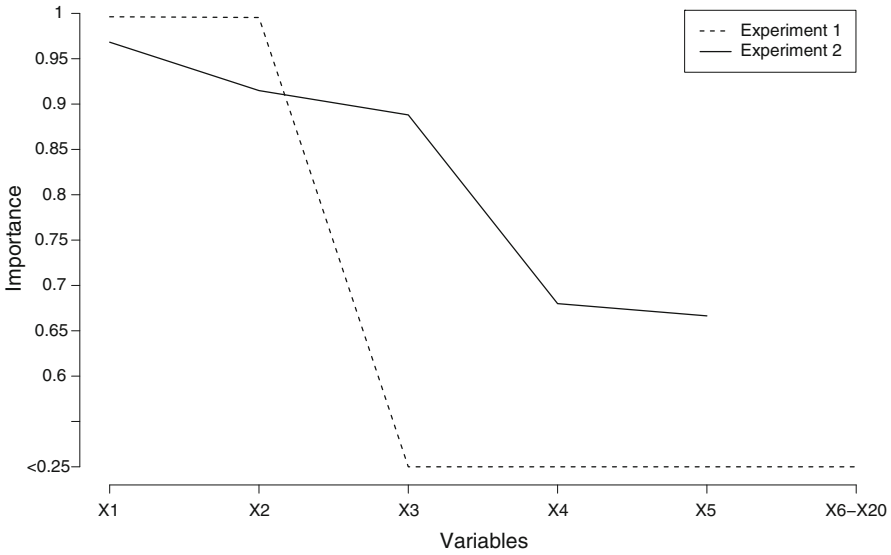

Fig. 3 Variable importance measure based on OCV-weights for the lasso-OCV estimator in experiments 1 and 2

and since often only very few estimators are averaged, variable selection can still be expected. The calculated OCV-weights may also serve as variable importance measure, giving additional insight into selection uncertainty: For each variable those weights $\hat{\mathbf{w}}_{\lambda_i}^{\text{OCV}}$ where the corresponding estimates $\hat{\boldsymbol{\beta}}_{\text{LE}}(\lambda_i)$ are non-zero are added up - to indicate how secure we are that there is an effect.

We now review experiments 1 and 2 from Sect. 3 and compare the Lasso estimator with the optimal cross validation lasso averaging estimator as indicated above. Table 4 shows the summary of the results. It can be seen that the OCV methodology improves the predictive performance, the MSE is about the same.

Figure 3 shows the importance measures for all variables averaged over the simulation runs. Recalling that in the first experiments the first 2 out of 20 variables were important, and in the second experiment all 5 variables to a different degree, one can see that the importance measure may serve as a useful sensitivity tool in data analysis.

Now we compare two Random Lasso estimators: The first estimator uses the Lasso estimate based on 5-fold cross validation for the estimation over the B_2 bootstrap samples (CV_5), the second estimator uses the Lasso-OCV estimator over the the B_2

Table 5 Random lasso experiments: Comparison of (1) Random Lasso using Lasso and 5-fold CV in the estimation step (CV_5) and (2) Random Lasso using Lasso and the OCV methodology (OCV_5) in the estimation step

	Experiment 1		Experiment 2		Experiment 5		Experiment 6	
	PP	MSE	PP	MSE	PP	MSE	PP	MSE
OCV_5	8.8	0.13	7.9	0.31	162.5	4.75	270.4	6.98
CV_5	29.9	0.71	18.5	0.91	166.7	4.94	288.5	7.56

The numbers represent the estimated predictive performance (PP) via the median mean squared prediction error and the estimated mean squared error (MSE)

bootstrap samples (OCV_5). In addition to the two experiments 1 and 2, two new high-dimensional data settings are considered:

- **Experiment 5** In this experiment we set $n_{\text{train}} = 50, n_{\text{test}} = 50$ and $p = 100$. The observations for the covariates are generated by the following distributions: $\mathbf{X}_1$ - $\mathbf{X}_{20} \sim \text{N}(0.5, 1)$, $\mathbf{X}_{21}$ - $\mathbf{X}_{40} \sim \log \text{N}(0.5, 0.5)$, $\mathbf{X}_{41}$ - $\mathbf{X}_{60} \sim$ Weibull(1.75, 1.9), $\mathbf{X}_{61}$ - $\mathbf{X}_{80} \sim \text{Bin}(1, 0.5)$ and $\mathbf{X}_{81}$ - $\mathbf{X}_{100} \sim$ Gamma(0.25, 2). To model the dependency between the covariates we use two independent Clayton Copulas (of size 40 and 60, respectively) with a copula parameter of 3 which indicates rather high correlation in two different groups of variables. The values of the response vector are realized via draws from $\mathbf{y} \sim \text{N}(\mu_5, \sigma)$ with $\mu_5 = \mathbf{X}\boldsymbol{\beta}_5$, $\sigma = 1.5$ and parameter vector $\boldsymbol{\beta}_5 = (5, 5, -5, -5, 0, \ldots, 0, \ldots, \ldots, 5, 5, -5, -5, 0, \ldots, 0)'$. This means, in this experiment we consider a high-dimensional set of data with two groups of variables where 20 out of 100 have influence on the response vector to a different degree.
- **Experiment 6** In this experiment we set $n_{\text{train}} = 50, n_{\text{test}} = 50$ and $p = 100$. The observations for the covariates are generated by the following distributions: $\mathbf{X}_1$ - $\mathbf{X}_{100} \sim \text{N}(0.5, 1)$. To model the dependency between the covariates we use two independent Clayton Copulas (of size 40 and 60, respectively) with a copula parameter of 1.5 which indicates medium correlation in two different groups of variables. The values of the response vector are realized via draws from $\mathbf{y} \sim \text{N}(\mu_6, \sigma)$ with $\mu_6 = \mathbf{X}\boldsymbol{\beta}_6$, $\boldsymbol{\beta}_1 = (-10, -9, \ldots, 9, 10, 0, \ldots, 0)'$, $\sigma = 1.5$. This means, in this experiment we consider a high-dimensional set of data with two groups of variables where 20 out of 100 have influence on the response vector to a different degree.

The results of experiments 1, 2, 5 and 6 are presented in Table 5. It can be seen that the OCV methodology yields a remarkable improvement within the Random Lasso methodology. Both the predictive performance and the MSE are better when using the averaging technique.

In their experiments, Wang et.al. (2011) also consider selection frequencies of important variables as a measure of quality. They define a important variable to be selected if the corresponding parameter estimate $\hat{\beta}_j$ is larger than $1/n$. In all our experiments the important variable selection frequencies are much higher for the

Random Lasso using the OCV_5 methodology in comparison to the standard Random Lasso.

6 Conclusions

The analyses and discussions of the present article provide a critical assessment of current practices related to an adequate tuning parameter choice for shrinkage estimators. Viewing tuning parameter selection as a problem of model selection enables the transfer and the implementation of ideas from recent model averaging literature. It has been shown that in many situations tuning parameter selection uncertainty is apparent, yielding in an underestimation of variability and doubts about efficiency.

The weighting methods presented in this paper reveal the intensity of selection uncertainty and show to be useful in various situations. Especially the proposed Optimal Cross Validation weighting scheme shows good performance with regard to predictive performance and stability. Results related to the Random Lasso simulations, also in high-dimensional data settings, are particularly promising in terms of future perspectives, being aware that some of the modern shrinkage estimators are computationally intensive and not necessarily stable. Moreover, using averaging weights to construct importance measures for variables may be useful in order to quantify selection uncertainty as indicated in the simulation studies of Sect. 5.

Of course, averaging over a set of estimators does not necessarily lead to improvements in every situation. One should acknowledge the findings from the model averaging literature, showing that model averaging performs mainly better than model selection in situations with at least modest selection uncertainty. In situations where the data supports one particular model strongly, averaging techniques may not improve overall performance of an estimator, see also Yuan and Yang (2005) and Zhang et al. (2011). The same applies for tuning parameter selection: The more insecure we are about the selection of a tuning parameter, the more promising an averaging technique. However, thinking about selection uncertainty at all, and applying sensitivity analysis, e.g., by means of the proposed methodologies, may turn out to be an important step forward.

Finally, it is worth mentioning that as with all Monte Carlo experiments, the results we have reported are tentative, and care must be exercised in attempting to generalize our conclusions to cases other than those investigated here. We have limited attention only to Ridge and Lasso estimators, and it would be interesting to broaden the analyses to other shrinkage estimators such as the the Elastic Net or other estimators that suffer from data-based decisions in connection with tuning parameters. Despite these limitations, this study has offered some interesting insights into some very practical questions on tuning parameter selection uncertainty that hopefully establish the foundations for fruitful discussions, but certainly warrant further studies on a subject that is well worth exploring.

Acknowledgements I thank Professor Alan Wan for various valuable comments that led to major improvement of this article. I also thank Alexander Klaußner for not only providing the data but also giving insightful thoughts on the interpretation of it.

References

Breiman L., Spector P (1992) Submodel selection and evaluation in regression: the X-random case. Int Stat Rev 60:291–319

Buchholz A, HollSnder N, Sauerbrei W (2008) On properties of predictors derived with a two-step bootstrap model averaging approach - A simulation study in the linear regression model. Comput Stat Data Anal 52:2778–2793

Buckland TS, Burnham PK, Augustin HN (1997) Model selection: an integral part of inference. Biometrics 53:603–618

Burnham K, Anderson D (2002) Model selection and multimodel inference. A practical information-theoretic approach. Springer, New York

Candolo C., Davison AC, DemTtrio CGB (2003) A note on model uncertainty in linear regression. Statistician 52:165–177

Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc 96:1348–1360

Golub HG, Heath M, Wahba G (1979) Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics 21:215–223

Hansen BE (2007) Least squares model averaging. Econometrica 75:1175–1189

Hansen BE, Racine J (2011) Jackknife model averaging. J Econ (forthcoming)

Hastie T, Tibsharani R, Friedman J (2001) Elements of statistical learning. Springer, New York

Hjort L, Claeskens G (2003) Frequentist model average estimators. J Am Stat Assoc 98:879–945

Hjort LN, Claeskens G (2006) Focussed information criteria and model averaging for Cox's hazard regression model. J Am Stat Assoc 101:1449–1464

Hoerl A, Kennard R (1970) Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12:55–67

Hoeting JA, Madigan D, Raftery AE, Volinsky CT (1999) Bayesian model averaging: a tutorial. Stat Sci 14:382–417

Klaußner A (2007) Phasenangepasste Führung von Wachstumsunternehmen (in German). International University Schloss Reichartshausen

Kohavi R (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. International Joint Conference on Artificial Intelligence (IJCAI) 1137–1143

Liang H, Zou GH, Wan ATK., Zhang X (2011) Optimal weight choice for frequentist model average estimators. J Am Stat Assoc. doi:10.1198/jasa.2011.tm09478

Magnus JR, Powell O, Prnfer P (2010) A comparison of two model averaging techniques with an application to growth empirics. J Econ 154:139–153

Oh H, Nychka D, Brown T, Charbonneau P (2004) Analysis of variable stars by robust smoothing. J R Stat Soc C 53:15–30

Radchenko P, James MG (2008) Variable inclusion and shrinkage algorithms. J Am Stat Assoc 103: 1304–1315

R Development Core Team (2010) R: A language and environment for statistical computing. R foundation for statistical computing. Vienna, Austria, ISBN:3-900051-07-0, http://www.R-project.org

Schomaker M, Wan ATK, Heumann C (2010) Frequentist model averaging with missing observations. Comput Stat Data Anal 54:3336–3347

Shao J (1997) An asymptotic theory for linear model selection. Stat Sinica 7:221–264

Stone M (1974) Cross-validatory choice and assessment of statistical predictions. J R Stat Soc B 36:111–147

Tibsharani R (1996) Regression shrinkage and selection via the Lasso. J R Stat Soc B 58:267–288

Valdés-Sosa P, Sßnchez-Bornot J, Lage-Castellanos A, Vega-Hernaindez M, Bosch-Bayard J, Melie-Garcia L, Canales-Rodriguez E (2005) Estimating brain functional connectivity with sparse multivariate autoregression. Philos Trans R Soc 360:969–981

Wan ATK, Zhang X (2009) On the use of model averaging in tourism research. Ann Tour Res 36:525–532

Wan ATK, Zhang X, Zou HG (2010) Least squares model averaging by Mallows criterion. J Econ 156: 277–283

Wang H, Li B, Leng C (2009a) Shrinkage tuning parameter selection with a diverging number of parameters. J R Stat Soc B 71:671–683

Wang H, Zhang X, Zou G (2009b) Frequentist model averaging: a review. J Syst Sci Complex 22:732–748

Wang S, Nan B, Rosset S, Zhu J (2011) Random Lasso. Ann Appl Stat 5:468–485

Yan J (2007) Enjoy the joy of copulas: with package copula. Journal of Statistical Software 21:1–21

Yuan Z, Yang Y (2005) Combining linear regression models: when and how? J Am Stat Assoc 100: 1215–1225

Zhang X., Wan A.T.K., Zhou SZ (2011) Focused information criteria, model selection and model averaging in a Tobit model with a non-zero threshold. J Busi Econ Stat. doi:10.1198/jbes.2011.10075

Zou H, Hastie T (2005) Regularization and variable selection via the Elastic Net. J R Stat Soc B 67:301–320