A focused mean squared error approach for selecting tuning parameters in penalized regression

Kristoffer H. Hellton*

Abstract

Penalized regression methods depending on one or more tuning parameters require fine-tuning to achieve optimal prediction performance. For ridge regression, which introduces an L_2 penalty on the regression coefficients, a range of tuning procedures have been developed, but in practice K-fold cross-validation has become the standard procedure. This paper explores a focused tuning approach for ridge regression where the tuning parameter is made dependent on the covariates of the *specific* observation to be predicted. The observation-specific tuning parameter is defined as the minimand of the empirical mean square prediction error, obtained by plugging in pilot estimates of the regression coefficients and error variance in the theoretical mean squared error expressions. Several pilot estimates are proposed, and we present risk expressions for the case of an OLS pilot. The focused ridge estimator is compared to standard ridge regression fine-tuned by cross-validation in simulations and a real data set.

Key Words: Tuning parameters, ridge regression, focused information criterion, focused model selection, personalized prediction.

1. Introduction

The technological development produces with increasing speed vast amounts of large data in fields from finance to genetics. The emerging Big Data era demands tailored statistical methods and at the same time broadens the possibilities for more detailed predictions. For high-dimensional data, where the data dimension p greatly exceeding the observations n, standard linear regression requires some form of penalization of the regression coefficients. Ridge regression penalizes the sum of squared regression coefficients,

$$\hat{\beta}(\lambda) = \arg\min_{\beta} \left\{ \|Y - X\beta\|_2^2 + \lambda \sum_{i=1}^{p} \beta_i^2 \right\},$$

equivalent to imposing an L_2 penalty on the vector of coefficients. Hoerl and Kennard (1970) originally introduces the penalty to ordinary least squares (OLS) regression to deal with data matrices close to not being of full rank, such that X^TX is not invertible. If X^TX becomes singular, the OLS solution is no longer unique. The L_2 penalty avoids this issue by ensuring the data matrix to always be of full rank (adding a constant $\lambda > 0$ to the zero eigenvalues of X^TX), a form of penalization also referred to as Tikhonov regularization.

Ridge regression has several appealing theoretical characteristics contributing to the good predictive performance. The L_2 penalty enforces a proportional shrinkage of all regression coefficients towards zero, introducing a bias, but lowering the variance of the predictor. Further, ridge regression shrinks positively correlated variables towards each other, a form of information pooling (Hastie et al., 2009), and penalizes the least informative directions,which together effectively improves predictions. Among regularized regression methods, ridge regression has been shown to asymptotically give better predictive performance (Frank and Friedman, 1993). In the field of genomics, which deals routinely with high-dimensional data, ridge regression has become a standard prediction tool. For

*Dept. of Mathematics, University of Oslo, Oslo, Norway

instance, Bøvelstad et al. (2007) showed that ridge regression has the overall best performance for predicting cancer survival based on microarray gene expression, giving lower prediction error compared to a range of other methods.

The tuning parameter, λ, controls the ridge penalty and tries to balance between overfitting to the training data and shrinking coefficients too much towards zero. To select a proper value for λ, an overwhelming range of procedures have been proposed, but cross-validation (CV) has emerged as the current canonical choice, in particular 5-fold or 10-fold cross-validation Hastie et al. (2009, p. 243). The procedure divides the data in K folds and predicts each fold by fitting a model on the remaining data. The resulting unbiased estimate of the prediction error can be calculated for a range of tuning parameters, with the value corresponding to the lowest error being chosen. Variations of the CV procedure include generalized and approximate cross-validation (Golub et al., 1979; Meijer and Goeman, 2013). Other examples of tuning methods include marginal maximum likelihood (Tran, 2009), bootstrapping (Delaney and Chatterjee, 1986), Bayesian methods (Zuliana and Perperoglou, 2016) and different versions of AIC (Boonstra et al., 2015).

Common for all the current selection procedures is that only *one* tuning parameter value is chosen for all further use and future predictions. However, the following thought experiment is possible: is if feasible to fine-tune the penalty parameter to target a specific set of covariates x_0? This introduces the idea of a focused tuning parameter λ_{x_0}, optimal for predicting a specific covariate x_0. Selecting tuning parameter(s) shares parallels with model selection, where the focused information criterion (FIC) has introduced the concept of addressing the quality of *the aim* of a statistical analysis (Claeskens and Hjort, 2008). Defining a final outcome of a fitted model, such as a specific prediction

$$\mu = x_0^T \beta,$$

stands in opposition to caring about general overall performance, like goodness of fit. This requires a clearly defined population quantity-of-interest μ, termed the focus parameter, for which it is possible to estimate the mean squared error (MSE) $\mathrm{MSE}_{\hat{\mu}}$ (or other risk measures). For ridge regression, the natural focus parameter is the new prediction, $x_0^T \beta$.

The outline of the paper is as follows: Section 2 introduces the general framework of the procedure; minimizing the mean squared error of the focus parameter as a function of λ. Section 3 explores the theoretical characteristics of the proposed estimators in the case of $p = 1$ and an orthogonal design matrix. Section 4 compares the prediction performance of the focused ridge estimator to standard ridge regression with cross-validation using simulations, and section 5 illustrates the method in a real data example.

2. Focused tuning in ridge regression

Suppose we have data $\{y_i, x_i\}$, comprising of n observations of a continuous outcome y_i and p-dimensional covariate vector x_i. We use the following notation; the outcome vector, Y, and the $n \times p$ data matrix X with rows x_i^T. Consider the standard linear regression model:

$$y_i = x_i^T \beta + \varepsilon_i, \quad i = 1, \ldots, n$$

where β is an unknown p-dimensional vector of regression coefficients and ε *iid* noise with zero mean, $\mathbb{E}\,\varepsilon_i = 0$, and variance, $\mathrm{Var}\,\varepsilon_i = \sigma^2$. The ordinary least squares (OLS) estimate of β, minimizing the least squares criterion, is given when $p < n$ and $X^T X$ is of full rank as

$$\tilde{\beta} = (X^T X)^{-1} X^T Y.$$

Ridge estimator minimizes instead a penalized least squares criterion (Hoerl and Kennard, 1970)

$$\hat{\beta}(\lambda) = \arg\min_{\beta} \left\{ \|Y - X\beta\|_2^2 + \lambda \sum_{i=1}^{p} \beta_i^2 \right\}, \tag{1}$$

with the explicit solution

$$\hat{\beta}(\lambda) = (X^T X + \lambda I_p)^{-1} X^T Y = (X^T X + \lambda I_p)^{-1} X^T X \tilde{\beta},$$

where the penalization parameter λ requires some form of fine-tuning.

Remark 1. The ridge estimate gives, in contrast to OLS, a unique solution even if X is not of full rank as in the high dimensional case ($p > n$).

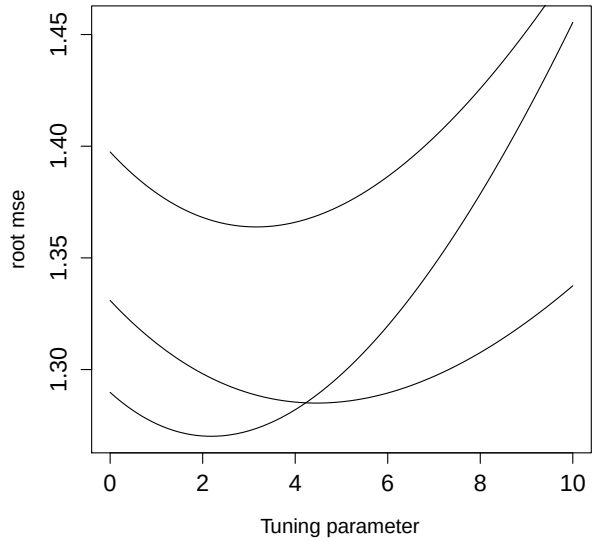

Figure 1: MSE curves for different x_0 showing minima at different values of λ.

Suppose one aims to predict the expected outcome for a specific set of covariates x_0, i.e. the focus parameter $\mu_0 = \mathbb{E}\, y_0 = x_0^T \beta$, such that the estimated ridge prediction is given

$$\hat{\mu}_0 = x_0^T \hat{\beta}(\lambda) = x_0^T (X^T X + \lambda I_p)^{-1} X^T Y.$$

If one now considers the expected MSE of the prediction, with the expectation taken with respect to the distribution of Y, the MSE will be a function of the tuning parameter λ, together with x_0 and the parameters β and σ^2:

$$\begin{aligned} \mathrm{MSE}_{\hat{\mu}}(\lambda; x_0, \beta, \sigma^2) &= \mathbb{E}_Y \left((x_0^T \hat{\beta}(\lambda) - x_0^T \beta)^2 \right) = \mathrm{Bias}^2(\hat{\mu}) + \mathrm{Var}\, \hat{\mu}, \\ &= \left\{ x_0^T ((X^T X + \lambda I_p)^{-1} X^T X - I_p) \beta \right\}^2 \\ &\quad + \sigma^2 x_0^T (X^T X + \lambda I_p)^{-1} X^T X (X^T X + \lambda I_p)^{-1} x_0. \end{aligned} \tag{2}$$

Note that we consider the error of $x_0^T \beta$ and not y_0, which simplify notation by omitting the intrinsic prediction error σ^2.

For each specific set of covariates x_0, the MSE as a function of λ will have a different minimum, as seen in Figure 1. The focused oracle tuning parameter is then defined as the minimand of the MSE curve.

Definition 1 (Oracle tuning)**.** *The oracle tuning parameter is the minimand of the mean squared prediction error*

$$\lambda_{x_0} = \arg\min_{\lambda} \mathrm{MSE}_{\hat{\mu}}(\lambda; x_0, \beta, \sigma^2), \quad \lambda \geq 0,$$

where the parameters β and σ^2 are known.

To characterize the curve of the oracle MSE as function of λ, Equation (2) is rewritten in terms of the singular value decomposition, $X = UDV^T$, as a summation over the p singular vectors $V = [v_1, \cdots, v_p]$ and values $D = \mathrm{diag}(d_1, \ldots, d_p)$ of X:

$$\mathrm{MSE}_{\hat{\mu}}(\lambda; x_0, \beta, \sigma^2) = \left\{ \lambda \sum_{i=1}^{p} \frac{x_0^T v_i v_i^T \beta}{d_i^2 + \lambda} \right\}^2 + \sigma^2 \sum_{i=1}^{p} \frac{d_i^2 (x_0^T v_i)^2}{(d_i^2 + \lambda)^2}.$$

Remark 2. The first derivative of the MSE in Equation (2) as a function of λ is given

$$\frac{\partial \, \mathrm{MSE}_{\hat{\mu}}(\lambda)}{\partial \lambda} = 2\lambda \left[\sum_{i=1}^{p} \frac{x_0^T v_i v_i^T \beta}{d_i^2 + \lambda} \right] \left[\sum_{i=1}^{p} \frac{d_i^2 x_0^T v_i v_i^T \beta}{(d_i^2 + \lambda)^2} \right] - 2\sigma^2 \sum_{i=1}^{p} \frac{d_i^2 (x_0^T v_i)^2}{(d_i^2 + \lambda)^3}. \quad (3)$$

The limit of the first derivative as $\lambda \to 0$

$$\frac{\partial}{\partial \lambda} \mathrm{MSE}_{\hat{\mu}}(\lambda; x_0, \beta, \sigma^2) \Big|_{\lambda=0} = -2\sigma^2 \sum_{i=1}^{p} \frac{(x_0^T v_i)^2}{d_i^4},$$

is negative, thus there exists a value of the tuning parameter larger than zero, $\lambda > 0$, for which $\mathrm{MSE}(\lambda)$ is smaller than $\mathrm{MSE}(0)$.

There are no explicit solutions for the minima of (2), unless all singular values are restricted to be equal (see Remark 6). The limit values of the curve are given

$$\lim_{\lambda \to 0} \mathrm{MSE}_{\hat{\mu}}(\lambda; x_0, \beta, \sigma^2) = \sigma^2 x_0^T (X^T X)^{-1} x_0, \quad \lim_{\lambda \to \infty} \mathrm{MSE}_{\hat{\mu}}(\lambda; x_0, \beta, \sigma^2) = (x_0^T \beta)^2,$$

such that if $(x_0^T \beta)^2 \geq \sigma^2 x_0^T (X^T X)^{-1} x_0$, there must exist a global minimum for $\lambda < \infty$. If the reverse is true, the global minimum can be given in the limit $\lambda \to \infty$. From Equation 3, the critical points of the MSE curves are given as the solutions of a polynomial of degree $3(p-1)+1$ giving at most $3(p-1)+1$ critical points.

These characteristics of the MSE curve can be seen for different x_0 with fixed data matrix X and β vector. For $p = 10$ and $n = 100$ there are typically zero to three critical points, and when combined with the asymptote $(x_0^T \beta)^2$ in the limit $\lambda \to \infty$ being below or above the value at $\lambda = 0$, we have the following possibilities:

- no critical points. The global minimum is in the limit $\lambda \to \infty$
- one critical point; a minimum. The global minimum is at the local minimum.
- two critical points; a minimum and a maximum. The asymptote at $\lambda \to \infty$ can be above or below the local minimum.
- three critical points; two minima and a maximum. The second minimum can be above or below the first local minimum, while the asymptote must be above the second local minimum.

Four different cases of MSE curves are shown in Figure 2.

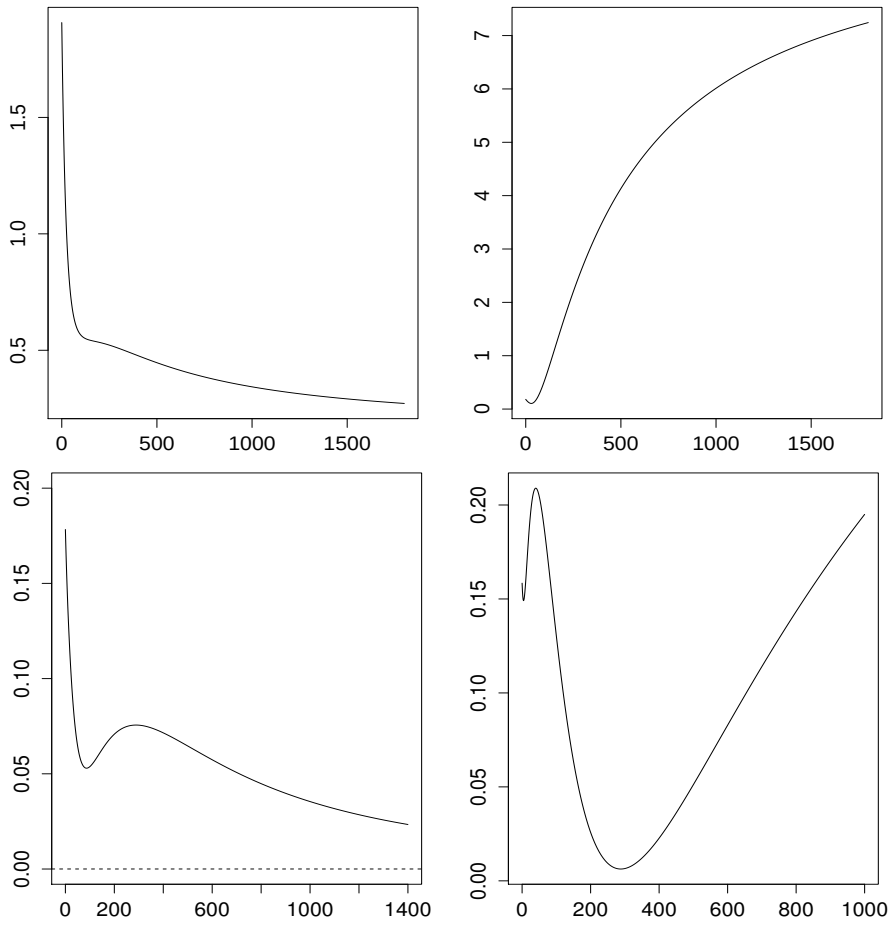

Figure 2: MSE curves as function of λ for $p = 10$. a) No critical points gives a minimum in the limit $\lambda \to \infty$. b) The classical case with one minimum and a curve increasing towards an asymptote. c) First a minimum and then a maximum, but with an asymptote $x_0^T\beta$ *below* the minimum value. d) Two local minima and a local maximum.

2.1 Estimating the tuning parameter in low dimension

The oracle value of the observation-specific tuning parameter, λ_{x_0}, will give the smallest expected prediction error, but cannot be used in practice as it requires the true value of β and σ^2. A direct way to estimate λ_{x_0} from data is to first estimate β and σ^2 by some method and plug-in the resulting estimates in Eq. (2). In low dimension ($p < n$), a natural choice of plug-in is the ordinary least squares (OLS) estimator

$$\tilde{\beta} = (X^T X)^{-1} X^T Y,$$

and the corresponding variance estimator

$$\tilde{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^{n} (y_i - x_i^T \tilde{\beta})^2,$$

assuming that $X^T X$ is of full rank.

When using the squared estimated bias directly, $(\widehat{\mathrm{Bias}})^2$, to obtain the population squared bias, Bias^2, it is necessary to subtract the variance of the bias as

$$\mathbb{E}\,\widehat{\mathrm{Bias}}^2 = \mathrm{Bias}^2 + \mathrm{Var}\,\widehat{\mathrm{Bias}},$$

see Claeskens and Hjort (2008, p. 150). These two aspects together gives the definition of our first focused tuning parameter estimator.

Definition 2 (Fridge-OLS)**.** *The fridge-OLS tuning parameter estimate is the minimand of the estimated mean squared error curve*

$$\begin{aligned}\hat{\lambda}_{x_0,OLS} &= \arg\min_{\lambda} \widehat{\mathrm{MSE}}_{\hat{\mu}}(\lambda; x_0, \tilde{\beta}, \tilde{\sigma}^2),\\ &= \arg\min_{\lambda} \left\{ \left((\widehat{\mathrm{Bias}}(\lambda))^2 - \mathrm{Var}\,\widehat{\mathrm{Bias}}(\lambda) \right)_+ + \widehat{\mathrm{Var}}(\lambda) \right\},\\ &= \arg\min_{\lambda} \Big\{ \Big((\lambda x_0^T (X^TX + \lambda I_p)^{-1} \tilde{\beta})^2 \\ &\qquad - \tilde{\sigma}^2 \lambda^2 x_0^T (X^TX + \lambda I_p)^{-1} (X^TX)^{-1} (X^TX + \lambda I_p)^{-1} x_0 \Big)_+ \\ &\qquad + \tilde{\sigma}^2 x_0^T (X^TX + \lambda I_p)^{-1} X^TX (X^TX + \lambda I_p)^{-1} x_0 \Big\}. \end{aligned} \tag{4}$$

where $\tilde{\beta}$ and $\tilde{\sigma}^2$ are the OLS estimates, and $(\cdot)_+ = \max\{\cdot, 0\}$.

Remark 3. A simplified version of fridge can be defined by omitting the bias correction, which insures a continuous first derivative giving faster convergence for numerical optimizers locating the minima:

$$\begin{aligned}\hat{\lambda}^*_{x_0,OLS} &= \arg\min_{\lambda} \left\{ (\widehat{\mathrm{Bias}}(\lambda))^2 + \widehat{\mathrm{Var}}(\lambda) \right\}, \\ &= \arg\min_{\lambda} \Big\{ (x_0^T ((X^TX + \lambda I_p)^{-1} X^TX - I_p) \tilde{\beta})^2 \\ &\qquad + \tilde{\sigma}^2 x_0^T (X^TX + \lambda I_p)^{-1} X^TX (X^TX + \lambda I_p)^{-1} x_0 \Big\}. \end{aligned} \tag{5}$$

The estimated MSE curves with the pilot estimates $\tilde{\beta}$ and $\tilde{\sigma}^2$ largely exhibit the same behavior as the oracle curves: the limit of the first derivative as $\lambda \to 0$

$$\frac{\partial}{\partial \lambda} \widehat{\mathrm{MSE}}(\lambda; x_0, \tilde{\beta}, \tilde{\sigma}^2) \Big|_{\lambda=0} = -2\tilde{\sigma}^2 \sum_{i=1}^{p} \frac{(x_0^T v_i)^2}{d_i^4} < 0,$$

is also negative, such that there exists a tuning parameter, $\lambda > 0$, for which $\widehat{\mathrm{MSE}}(\lambda) < \widehat{\mathrm{MSE}}(0)$. Also the limit of the estimated MSE as $\lambda \to 0$ remains the same:

$$\lim_{\lambda \to 0} \widehat{\mathrm{MSE}}(\lambda; x_0, \tilde{\beta}, \tilde{\sigma}^2) = \tilde{\sigma}^2 x_0^T (X^TX)^{-1} x_0,$$

but the asymtotic limit as $\lambda \to \infty$ changes to

$$\lim_{\lambda \to \infty} \widehat{\mathrm{MSE}}(\lambda; \tilde{\beta}, x_0, \tilde{\sigma}^2) = \max \left\{ 0, (x_0^T \tilde{\beta})^2 - \tilde{\sigma}^2 x_0^T (X^TX)^{-1} x_0 \right\}.$$

As a consequence, the global minimum is more often found in the limit $\lambda \to \infty$.

2.2 Estimating the tuning parameter: high dimension

In the high-dimensional situation with $p \gg n$, one needs an alternative to the OLS estimate. One possibility is to substitute the non-invertible X^TX with the Moore-Penrose pseudo-inverse:

$$\tilde{\beta}^+ = (X^TX)^+ X^TY,$$

and use the estimate directly to estimate the noise variance

$$\tilde{\sigma}^2_+ = \frac{1}{n - df(\tilde{\beta}^+)} \sum_{i=1}^{n} \left(y_i - x_i^T \tilde{\beta}^+ \right)^2, \quad df(\tilde{\beta}^+) = \mathrm{tr}(X(X^TX)^+ X^T).$$

Definition 3 (Fridge-OLS+)**.** *The fridge-OLS+ estimator is the minimand of the estimated mean squared error curve using the Moore-Penrose pseudo-inverse in the OLS estimate:*

$$\hat{\lambda}_{x_0,OLS+} = \arg\min_{\lambda}\Bigg\{\Big((\lambda x_0^T(X^TX+\lambda I_p)^{-1}\tilde{\beta}^+)^2 - \tilde{\sigma}_+^2\lambda^2 x_0^T(X^TX+\lambda I_p)^{-1}(X^TX)^+(X^TX+\lambda I_p)^{-1}x_0\Big)_+ + \tilde{\sigma}_+^2 x_0^T(X^TX+\lambda I_p)^{-1}X^TX(X^TX+\lambda I_p)^{-1}x_0\Bigg\}.$$

Remark 4. The estimate $\hat{\lambda}_{x_0,OLS+}$ is equivalent to the standard OLS based estimate $\hat{\lambda}_{x_0}$, when $p < n$ and X^TX is of full rank.

A second alternative is to use ridge regression with cross-validation as a pilot estimate, which would give a less variable estimate of β than OLS for higher p:

Definition 4 (Fridge-ridge)**.** *The fridge-ridge tuning parameter estimate is the minimand of the estimated mean squared error curve*

$$\begin{aligned}\hat{\lambda}_{x_0,ridge} &= \arg\min_{\lambda}\widehat{\mathrm{MSE}}_{\hat{\mu}}(\lambda; x_0, \hat{\beta}(\hat{\lambda}_{\mathrm{CV}}), \hat{\sigma}^2),\\ &= \arg\min_{\lambda}\left\{\Big((\widehat{\mathrm{Bias}}(\lambda))^2 - \mathrm{Var}\,\widehat{\mathrm{Bias}}(\lambda)\Big)_+ + \widehat{\mathrm{Var}}(\lambda)\right\},\\ &= \arg\min_{\lambda}\Bigg\{\Big((\lambda x_0^T(X^TX+\lambda I_p)^{-1}\hat{\beta}(\hat{\lambda}_{\mathrm{CV}}))^2\\ &\quad -\tilde{\sigma}^2\lambda^2 x_0^T(X^TX+\lambda I_p)^{-1}(X^TX+\hat{\lambda}_{\mathrm{CV}}I_p)^{-1}X^TX(X^TX+\hat{\lambda}_{\mathrm{CV}}I_p)^{-1}(X^TX+\lambda I_p)^{-1}x_0\Big)_+\\ &\quad + \tilde{\sigma}^2 x_0^T(X^TX+\lambda I_p)^{-1}X^TX(X^TX+\lambda I_p)^{-1}x_0\Bigg\},\end{aligned}$$

where $\hat{\lambda}_{\mathrm{CV}}$ *is found by cross-validation, giving the standard ridge estimates*

$$\hat{\beta}(\hat{\lambda}_{\mathrm{CV}}) = (X^TX+\lambda I_p)^{-1}X^TY, \quad \hat{\sigma}^2 = \frac{1}{n - df(\hat{\lambda}_{CV})}\sum_{i=1}^{n}\Big(y_i - x_i^T\hat{\beta}(\hat{\lambda}_{CV})\Big)^2$$

with $df(\hat{\lambda}_{CV}) = \mathrm{tr}(X(X^TX+\hat{\lambda}_{CV})^{-1}X^T)$.

Depending on p, n and in particular the structure of β, other pilot estimates of β can preform better than ridge regression. Additional options include lasso (L_1 penalty) or principal component regression (PCR) using cross-validation to decide additional tuning parameters. As the variance of the bias based on these pilot estimates can not be expressed explicitly, we defined the simplified versions only.

Definition 5 (Simplified Fridge-PCR and Fridge-lasso)**.** *The simplified fridge-PCR and fridge-lasso estimates are the minimizers of the estimated MSE curves:*

$$\hat{\lambda}^*_{x_0,PCR} = \arg\min_{\lambda}\Bigg\{\Big((\lambda x_0^T(X^TX+\lambda I_p)^{-1}\hat{\beta}_{PCR})^2 + \tilde{\sigma}^2 x_0^T(X^TX+\lambda I_p)^{-1}X^TX(X^TX+\lambda I_p)^{-1}x_0\Bigg\},$$

$$\hat{\lambda}^*_{x_0,lasso} = \arg\min_{\lambda}\Bigg\{\Big((\lambda x_0^T(X^TX+\lambda I_p)^{-1}\hat{\beta}_{lasso})^2 + \tilde{\sigma}^2 x_0^T(X^TX+\lambda I_p)^{-1}X^TX(X^TX+\lambda I_p)^{-1}x_0\Bigg\},$$

where $\hat{\beta}_{PCR}$ is the principal component regression estimate and $\hat{\beta}_{lasso}$ is the lasso estimates both with cross-validation tuning parameters, and $\tilde{\sigma}^2$ is a suitable variance estimate.

Reid et al. (2013) and Dicker (2014) have proposed several estimators for σ^2 in the case of lasso regression.

3. The benefits of fridging

The benefit of fridge is governed by several aspects: the focus x_0, the structure of β, the structure of the data matrix X, and the outcomes Y. The first part of this section will establish $x_0^T\beta$ as the *key quantity* to answer which x_0 benefits from focusing and the role of the data matrix as determining the important directions in the data spaces.

3.1 The role of the focus

The key question from a practical point of view: is it possible to identify a priori which covariates x_0 will benefit from the focused approach? In the oracle setting where β is known this can be answered straightforward.

Remark 5 (One dimensional case). Let $p = 1$, then the ridge prediction is

$$x_0\hat{\beta}(\lambda) = x_0 \sum_{i=1}^{n} x_i y_i \left(\sum_{i=1}^{n} x_i^2 + \lambda \right)^{-1} = \frac{M}{M+\lambda} x_0 \tilde{\beta}, \quad M = \sum_{i=1}^{n} x_i^2,$$

giving the oracle mean squared error as

$$\mathrm{MSE}_{\hat{\mu}}(\lambda; x_0, \beta, \sigma^2) = x_0^2 \left(-\frac{\lambda}{M+\lambda}\beta \right)^2 + x_0^2 \frac{\sigma^2 M}{(M+\lambda)^2}.$$

As x_0^2 is a common factor, the oracle tuning minimizing the MSE will be independent of x_0:

$$\lambda_{x_0} = \frac{\sigma^2}{\beta^2},$$

and *the estimator has lost its focus!* This is, however, not the case in higher dimension, $p \geq 2$.

To demonstrate that the key role of the focus is through the relation between x_0^T and β, we consider an artificial data matrix that allow for an explicit expression of the tuning parameter.It is worth noticing that crucial aspect are the equal entries of the diagonal.

Remark 6 (Orthogonal case). Suppose the covariates are transformed to give a diagonal covariance matrix with equal entries,

$$X^T X = diag(M, \ldots, M) = MI, \quad M = \sum_{i=1}^{n} x_{i,j}^2,$$

such that the columns of X are orthogonal. Then the oracle MSE is given

$$\begin{aligned} \mathrm{MSE}_{\hat{\mu}}(\lambda; x_0, \beta, \sigma^2) &= \left\{ x_0^T (M(MI_p + \lambda I_p)^{-1} - I_p)\beta \right\}^2 + \sigma^2 M x_0^T (MI_p + \lambda I_p)^{-2} x_0, \\ &= (x_0^T\beta)^2 \frac{\lambda^2}{(M+\lambda)^2} + \sigma^2 x_0^T x_0 \frac{M}{(M+\lambda)^2}, \end{aligned}$$

with a minimand with the explicit expression

$$\lambda_{x_0} = \frac{\sigma^2 x_0^T x_0}{(x_0^T\beta)^2}. \tag{6}$$

Result 1. *In the orthogonal case, the oracle tuning parameter is controlled by $x_0^T\beta$, the inner product between the focus covariate and regression coefficients, and the tuning parameter in* (6) *can be expressed by the geometry of x_0 and β:*

$$\lambda_{x_0} = \frac{\sigma^2}{\|\beta\|^2 \cos^2 \alpha_{x_0}},$$

where $\|\beta\|$ is the length of β, and α_{x_0} the angle between the vectors x_0 and β in the variable space.

Remark 7. Importantly, the length of x_0, $\|x_0\|$ *does not influence* the value of the oracle tuning in the orthogonal case.

The oracle tuning is therefore given by the relation between x_0 and β, in particular the angle α, measuring how close the prediction is to the mean response, and the length of β, a measure of the signal strength. When the prediction is close to zero (for centered variables), such that $\cos \alpha_{x_0}$ approaches zero and x_0 and β becomes orthogonal, oracle tuning parameter blows up, $\lambda \to \infty$ and shrinks the prediction towards zero. A zero prediction $x_0^T\beta = 0$ implies that the vectors x_0 and β does not contain mutual information. Regarding the length of β as a measure of signal strength, the larger β values can handle a harder penalization in the oracle case, while small β requires less penalization.

The resulting prediction error with the fridge tuning parameter

$$\text{MSE}_{\hat{\mu}}(\lambda_0; x_0, \beta, \sigma^2) = \frac{\sigma^2 x_0^T x_0 (x_0^T\beta)^2}{\sigma^2 x_0^T x_0 + (x_0^T\beta)^2} < \text{MSE}_{\hat{\mu}}(0; x_0, \beta, \sigma^2) = \sigma^2 x_0^T x_0,$$

will be uniformly smaller than the OLS prediction error, corresponding to $\lambda = 0$. This expresses the fact that with negative first derivative at $\lambda = 0$, there always exits a $\lambda > 0$ for which the prediction error is smaller the error of OLS. The decrease in prediction error of fridge compared to OLS depends on how close $(x_0^T\beta)^2$ is to zero.

3.2 The role of the data matrix

The effect of the data matrix X is best understood as a modification of x_0 and β, relative to orthogonal case. Consider the general case where the singular value decomposition of the data matrix is $X = UDV^T$, giving the mean square error

$$\text{MSE}_{\hat{\mu}}(\lambda; x_0, \beta, \sigma^2) = \left\{\lambda x_0^T V (D^2 + \lambda I_p)^{-1} V^T \beta\right\}^2 + \sigma^2 x_0^T V D (D^2 + \lambda I_p)^{-1} D V^T x_0,$$

where the matrix of singular vectors V rotate the original x_0 and β. The data matrix determines the value of λ by projecting the x_0 and β along the singular vectors and upweighting the vectors associated with large singular values. The data matrix therefore gives the premise for which directions in the covariate space are considered more important. In consequence, how x_0 and β are spanned by the first singular vectors of X determines the optimal value of the tuning parameter. If all singular values are equal, such that all directions carry the same weights, the data matrix $X = U(MI_p)V^T$ works as a rotation matrix through the singular vectors, $V^T x_0$ and $V^T\beta$ could be viewed as new covariates and regression coefficients, oriented along the singular vectors:

$$\text{MSE}_{\hat{\mu}}(\lambda; x_0, \beta, \sigma^2) = (x_0^T V V^T \beta)^2 \frac{\lambda^2}{(M+\lambda)^2} + \sigma^2 x_0^T V V^T x_0 \frac{M}{(M+\lambda)^2},$$

with an explicit expression $\lambda_{x_0} = \frac{\sigma^2 x_0^T V V^T x_0}{(x_0^T V V^T \beta)^2}$.

3.3 The effect of estimation

It is also a question how much of the oracle optimality is lost by estimating the tuning parameter through the plug-in approach. Consider first the case where effect of the data matrix is negligible, such that there are no dominating singular values, for instance in the case when uncorrelated variables. Then the key quantity, $x_0^T\beta$, will determine the observation-specific tuning parameter, as seen for the orthogonal design matrix.

Remark 8 (One dimensional case). For $p = 1$, the estimated bias of the OLS pilot is given

$$\widehat{\text{Bias}} = -\frac{\lambda}{M+\lambda}x_0\tilde{\beta}, \quad \text{Var}\,\widehat{\text{Bias}} = \frac{\sigma^2\lambda^2 x_0^2}{M(M+\lambda)^2},$$

such that the estimated mean square error becomes

$$\widehat{\text{MSE}}(\lambda;\tilde{\beta},x_0) = x_0^2\left\{\left(\tilde{\beta}^2 - \frac{\sigma^2}{M}\right)_+\left(\frac{\lambda}{M+\lambda}\right)^2 + \frac{\sigma^2}{M}\left(1-\frac{\lambda}{M+\lambda}\right)^2\right\}.$$

The tuning parameter estimate, still independent of x_0 and without a focus, is given

$$\hat{\lambda}_{x_0,OLS} = \frac{M\sigma^2}{\left(M\tilde{\beta}^2-\sigma^2\right)_+}.$$

When $M\tilde{\beta}^2$ approaches or becomes smaller than σ^2, the tuning parameter explodes, $\hat{\lambda}_{x_0} \to \infty$, making the fridge prediction exactly zero

$$x_0\hat{\beta}(\hat{\lambda}_{x_0,OLS}) = \frac{(M\tilde{\beta}^2-\sigma^2)_+}{\sigma^2+(M\tilde{\beta}^2-\sigma^2)_+}x_0\tilde{\beta} = \begin{cases} 0 & \text{if } M\tilde{\beta}^2 \le \sigma^2, \\ \dfrac{M\tilde{\beta}^2-\sigma^2}{M\tilde{\beta}^2}x_0\tilde{\beta} & \text{if } M\tilde{\beta}^2 > \sigma^2. \end{cases} \tag{7}$$

Thus the risk function of the fridge-OLS, scaled by the OLS risk, is

$$\text{risk}\left(x_0\hat{\beta}(\hat{\lambda}_{x_0,OLS})\right) = \begin{cases} \dfrac{M(x_0\beta)^2}{\sigma^2} & \text{if } V^2 \le \sigma^2, \\ \dfrac{1}{\sigma^2}\,\mathbb{E}\left(\dfrac{V^2-\sigma^2}{V^2}V - \sqrt{M}x_0\beta\right)^2 & \text{if } V^2 > \sigma^2. \end{cases}$$

with the notation $V = \sqrt{M}\tilde{\beta} \sim N(\sqrt{M}x_0\beta, 1)$. But as there is no focusing effect in one dimension, we first extend to the p-dimensional case before evaluating the risk functions.

Result 2 (Orthogonal case)**.** *In the orthogonal case, the fridge-OLS gives the bias*

$$\widehat{\text{Bias}} = -\frac{\lambda}{M+\lambda}x_0^T\tilde{\beta}, \quad \text{Var}\,\widehat{\text{Bias}} = \frac{\sigma^2\lambda^2 x_0^T x_0}{M(M+\lambda)^2},$$

with estimated MSE as

$$\widehat{\text{MSE}}(\beta,x_0,a) = \left(x_0^T\beta - \frac{\sigma^2 x_0^T x_0}{M}\right)_+\left(\frac{\lambda}{M+\lambda}\right)^2 + \frac{\sigma^2 x_0^T x_0}{M}\left(1-\frac{\lambda}{M+\lambda}\right)^2,$$

The fridge-OLS tuning estimate is then given

$$\hat{\lambda}_{x_0,OLS} = \frac{\sigma^2 M x_0^T x_0}{\left(M(x_0^T\tilde{\beta})^2 - \sigma^2 x_0^T x_0\right)_+},$$

with the prediction

$$x_0^T\hat{\beta}(\hat{\lambda}_{x_0,OLS}) = \frac{\left(M(x_0^T\tilde{\beta})^2 - \sigma^2 x_0^T x_0\right)_+}{\sigma^2 M x_0^T x_0 + \left(M(x_0^T\tilde{\beta})^2 - \sigma^2 x_0^T x_0\right)_+}\, x_0^T\tilde{\beta}, \tag{8}$$

$$= \begin{cases} 0 & \text{if } |x_0^T\tilde{\beta}| \leq \sigma\sqrt{x_0^T x_0/M}, \\ \dfrac{(x_0^T\tilde{\beta})^2 - \sigma^2 x_0^T x_0/M}{(x_0^T\tilde{\beta})^2}\, x_0^T\tilde{\beta} & \text{if } |x_0^T\tilde{\beta}| > \sigma\sqrt{x_0^T x_0/M}. \end{cases} \tag{9}$$

Alternatively, the simplified fridge-OLS prediction is given

$$x_0^T\hat{\beta}(\hat{\lambda}^*_{x_0,OLS}) = \frac{M(x_0^T\tilde{\beta})^2}{M(x_0^T\tilde{\beta})^2 + \sigma^2 x_0^T x_0} x_0^T\tilde{\beta}.$$

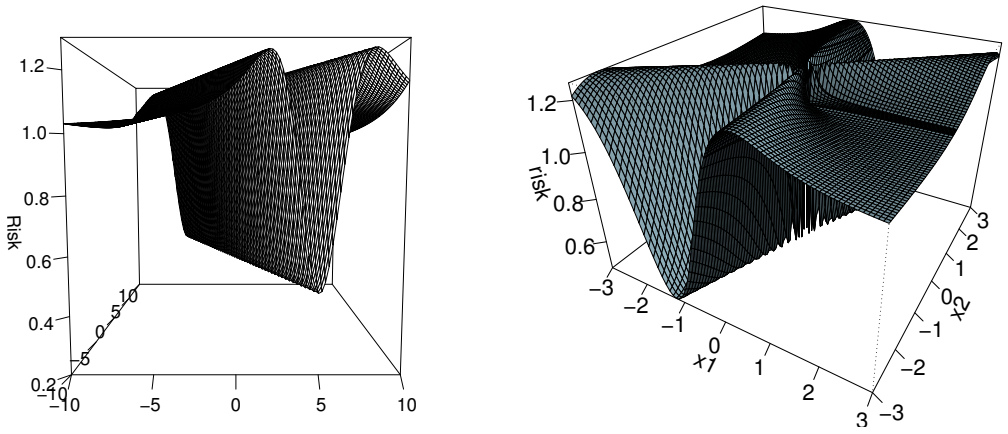

Figure 3: Panel a) The risk surface for the original FIC estimator in two dimensions for varying β and fixed $x_0 = (-5, 2)$. Panel b) The risk surface in two dimensions for varying x_0 and fixed $\beta = (-5, 2)$

To visualize how the risk varies with x_0, it is necessary with more than one dimension and for $p = 2$ the risk can be shown as a function of β and x_0 separately. Figure 3a) shows the risk as function of β for fixed x_0, with risk surface is scaled by the risk of the OLS estimator, $\text{risk}(x_0\tilde{\beta}) = \frac{\sigma^2 x_0^T x_0}{M}$. It is seen that the fridge-OLS has a lower risk than OLS in a trench following the line $x_0^T\beta = 0$. Figure 3b) shows the risk as function of x_0 when the true β is fixed, also scaled by the risk of the OLS estimator. The important observation is that fridge-OLS has lower risk than OLS within *cone* centered on the line $x_0^T\beta = 0$. Within this section of the covariate space, the focused shrinking towards zero is particularly beneficial in prediction.

4. Comparison with cross-validation

The most widely used fine-tuning procedure for tuning parameters is undoubtedly K-fold cross-validation, probably due its conceptual simplicity: the data is divided in K folds with each part held out and predicted by fitting a model on the remaining folds. A range of tuning parameters can then be tested and one will choose the value with the lowest error, averaged over all folds. The procedure was introduced by Stone (1974) and Allen (1974), proposing the prediction residual error sum of squares (PRESS) criterion, equivalent to n-fold or

leave-one-out cross-validation. Currently 10- and 5-fold cross-validation have become the default approach in modern statistics and machine learning Hastie et al. (2009).

In leave-one-out cross-validation, each outcome is predicted using all *other* observations: If $X_{[i]}$ and $Y_{[i]}$ are the X and Y with the ith row deleted, such that the regression model for each observation is

$$\hat{\beta}_{[i]}(\lambda) = (X_{[i]}^T X_{[i]} + \lambda I_p)^{-1} X_{[i]}^T Y_{[i]},$$

the LOOCV criterion is given

$$\mathrm{CV}(\lambda) = \sum_{i=1}^{n} e_{[i]}^2 = \sum_{i=1}^{n} \left(y_i - x_i^T \hat{\beta}_{[i]} \right)^2.$$

For ridge regression in particular, the leave-one-out prediction error has an explicit expression (see Golub et al. (1979)) avoiding the standard iterative procedure. Due to a special relation for matrix inverses, the leave-one-out prediction error can be expressed as a weighted version prediction error from a model using all the data, and the cross-validation tuning parameter is thus given

$$\hat{\lambda}_{\mathrm{CV}} = \arg\min_{\lambda} \sum_{i=1}^{n} \left(\frac{y_i - x_i^T \hat{\beta}(\lambda)}{1 - x_i^T (X^T X + \lambda I)^{-1} x_i} \right)^2. \tag{10}$$

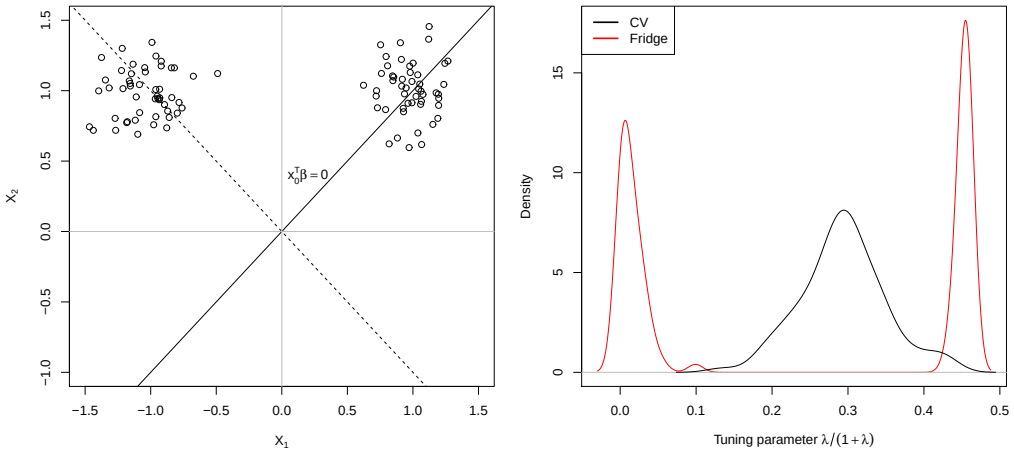

Figure 4: a) Data concentrated in two distinct clusters centered at $(1, 1)$ and $(-1, 1)$ b) Density plot of $\lambda/(\lambda + 1)$ for fridge (red) and cross-validation (black.)

To illustrate the difference between the focused tuning and cross-validation, consider an example where the data matrix consists of different clusters. Figure 4a) shows a data example $(p = 2)$ with two distinct clusters centered at $(1, 1)$ and $(-1, 1)$, respectively. If the regression coefficients are given $\beta = [-1, 1]$, the outcome for the right cluster will be close to zero, $y_i \simeq 0$, while the outcome for the left cluster will be close to two, $y_i \simeq 2$, given a small noise variance. The line implied by $x^T\beta = 0$ is marked in black. The clusters will then require a very different level of penalization to produce an optimal prediction; the right cluster requires a stronger penalization and the left cluster requires a weaker penalty. Figure 4b) shows the distribution of the observation-specific tuning, $\lambda_{x_0}/(1 + \lambda_{x_0})$ in red, for each of the observations seen in Figure 4a). The corresponding distribution of the tuning parameter estimated by leave-one-out cross-validation, $\hat{\lambda}_{CV}/(1 + \hat{\lambda}_{CV})$ over multiple sets

of simulated y_i is shown in black. Figure 4b) displays clearly that the difference in optimal tuning parameter for the two clusters are captured by the fridge procedure, with the right cluster corresponding to a high tuning parameter and the left cluster to a smaller tuning value. Cross-validation on the other hand estimates an overall tuning parameter, averaging over all covariates, and thus selects a tuning parameter value inappropriate for both two clusters.

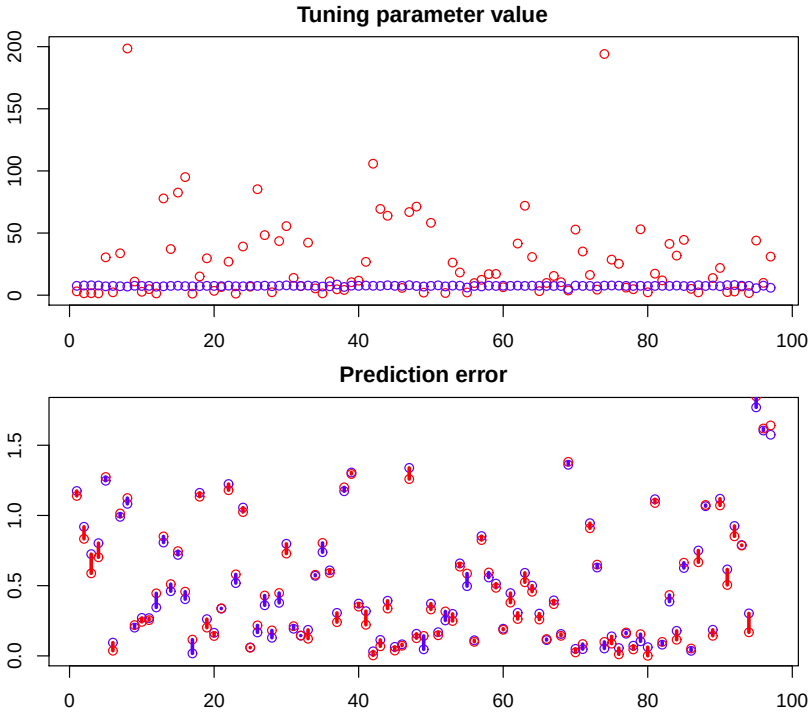

Figure 5: Top; tuning parameter estimates for cross-validation (blue) and fridge-OLS (red) for each out-of-sample prediction. Bottom; prediction error of log-PSA values for cross-validation (blue) and the fridge (red) with the difference colored according to the method with the lowest error.

5. Data examples

In the following section, we illustrate the fridge procedure using real data. Stamey et al. (1989) examined the relation between prostate specific antigen (PSA) and 7 clinical measurement, such as cancer volume and prostate weight, in 97 prostate cancer patients. The goal was to predict the log PSA values based on the different clinical measurements. Even though the data are low-dimensional ($p < n$), the sample size is not very large compared to the number of variables and we used fridge with a ridge pilot to estimate the focused tuning parameter. To imitate the case of new unknown sets of covariates x_0, each observation is predicted out-of-sample based on all other data points, and we recorded the mean squared prediction error relative to the corresponding outcome for comparison with cross-validation:

$$\frac{1}{n}\sum_{i=1}^{n}(x_i^T\hat{\beta}_{-i}(\lambda_{x_0}) - y_i)^2.$$

The top panel of Figure 5 shows the estimated tuning parameters for each observation (out-of-sample) in the prostate cancer data set. The estimated cross-validation tuning pa-

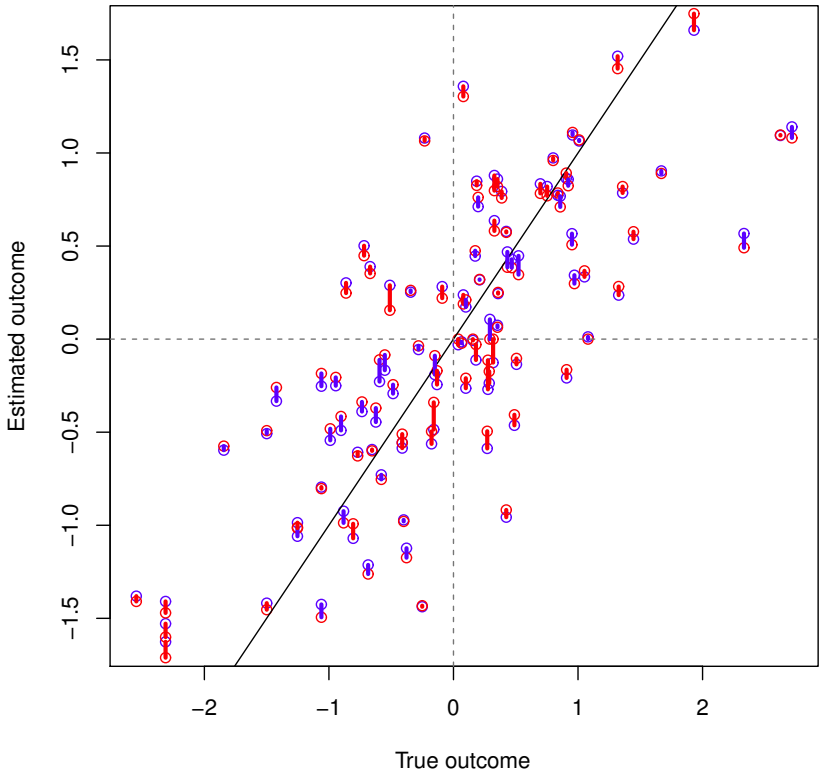

Figure 6: The out-of-sample predictions of log PSA plotted against the true value, with ridge with cross-validation in blue and fridge-ridge in red. The difference is colored according to the method with the least squared error.

rameter (in blue) is quite stable for each observation due to the large sample size, while the fridge-ridge estimate (in red) varies considerably across the different x_i. The bottom panel shows the squared prediction error of each log PSA observation for cross-validation (in blue) and the fridge (in red) with the difference colored according to the method with the lowest error. There are small difference between the methods and both can do better or worse than the other for specific x_0s. However, fridge-ridge gives a smaller error then cross-validation in 56.3 % of the cases and on average the squared prediction error is 1.1 % lower for fridge-ridge than for ridge regression with CV.

Figure 6 shows the true y_i plotted against the out-of-sample prediction $\hat{y}_i = x_i^T \hat{\beta}(\hat{\lambda}_{x_i})$. It is seen that fridge penalizes more the observations with y_i close to zero than CV, forcing the prediction towards zero.

6. Discussion

The increasing availability of data allows for individualized prediction procedures, for instance by focusing the tuning parameter towards specific covariate sets. With our definition of the optimal tuning parameter as the minimand of the mean squared error prediction, one can also consider other loss functions or risk measures. The current formulation requires not distributional assumption for the noise, apart from the linear model. We present one way of estimating this tuning parameter; by plug-in estimates in the risk expressions. Different plug-in or pilot estimates are considered.

In the high-dimensional case, the so-called projection bias, as explored by Shao and Deng (2012), poses an additional problem. Any linear estimator is in fact only consistent

for a projection of the true β unto the row space of the observed data matrix. If the row space of X does not properly spanned the true coefficient vector, the bias can be substantial, and the projection bias cannot be quantified from the observations. One approach to avoid the bias is to use a non-linear estimator such as the lasso, requiring the assumption that β is sparse (Meier, 2016).

The focused tuning parameter is also connected to random effect models, when viewed in the Bayesian context. As ridge regression corresponds to a Gaussian prior on the regression coefficients, β, (Lindley and Smith, 1972)

$$y_i = N(X\beta, \sigma^2), \quad \beta \sim N(0, \tau^2),$$

the variance, τ^2 will be inversely proportional to the tuning parameter. In such a framework, the observation-specific λ_{x_i} can be formulated as Gaussian prior with a observation-specific variance. This can further be view as an individual scaling of the β:

$$y_i = x_i^T \lambda_{x_i}^{-1/2} \beta + \varepsilon_i = x_i^T \beta_i + \varepsilon_i, \quad \beta_i = \lambda_{x_i}^{-1/2} \beta, \quad i = 1, \ldots, n,$$

giving observation-specific regression coefficients, similar to a random effects model. Such random effects in a mixed model framework are typically assumed to follow a multivariate normal distribution and estimated using empirical Bayes methods, an approach which could also be suitable for fridge.

References

Allen, D. M. (1974). The relationship between variable selection and data agumentation and a method for prediction. *Technometrics*, 16(1):125–127.

Boonstra, P. S., Mukherjee, B., and Taylor, J. M. (2015). A small-sample choice of the tuning parameter in ridge regression. *Statistica Sinica*, 25(3):1185.

Bøvelstad, H. M., Nygård, S., Størvold, H. L., Aldrin, M., Borgan, Ø., Frigessi, A., and Lingjærde, O. C. (2007). Predicting survival from microarray dataa comparative study. *Bioinformatics*, 23(16):2080–2087.

Claeskens, G. and Hjort, N. L. (2008). *Model selection and model averaging*. Cambridge University Press.

Delaney, N. J. and Chatterjee, S. (1986). Use of the bootstrap and cross-validation in ridge regression. *Journal of Business & Economic Statistics*, 4(2):255–262.

Dicker, L. H. (2014). Variance estimation in high-dimensional linear models. *Biometrika*, 101(2):269–284.

Frank, I. E. and Friedman, J. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135.

Golub, G. H., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223.

Hastie, T., Friedman, J., and Tibshirani, R. (2009). *The elements of statistical learning*. Springer, 2 edition.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Lindley, D. V. and Smith, A. F. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–41.

Meier, L. (2016). High-dimensional regression and inference. *Handbook of Big Data*, page 305.

Meijer, R. J. and Goeman, J. J. (2013). Efficient approximate k-fold and leave-one-out cross-validation for ridge regression. *Biometrical Journal*, 55(2):141–155.

Reid, S., Tibshirani, R., and Friedman, J. (2013). A study of error variance estimation in lasso regression. *arXiv preprint arXiv:1311.5274*.

Shao, J. and Deng, X. (2012). Estimation in high-dimensional linear models with deterministic design matrices. *The Annals of Statistics*, 40(2):812–831.

Stamey, T. A., Kabalin, J. N., McNeal, J. E., Johnstone, I. M., Freiha, F., Redwine, E. A., and Yang, N. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate. ii. radical prostatectomy treated patients. *The Journal of urology*, 141(5):1076–1083.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal statistical society. Series B (Methodological)*, 36(1):111–147.

Tran, M. N. (2009). Penalized maximum likelihood principle for choosing ridge parameter. *Communications in Statistics-Simulation and Computation*, 38(8):1610–1624.

Zuliana, S. U. and Perperoglou, A. (2016). The weight of penalty optimization for ridge regression. In *Analysis of Large and Complex Data*, pages 231–239. Springer.