Fridge: focused fine-tuning of ridge regression for personalized predictions

Kristoffer H. Hellton & Nils Lid Hjort

Dept. of Mathematics, University of Oslo
University of Oslo, P.O.Box 1053 Blindern, 0316 Oslo, Norway.
E-mail: kristohh@math.uio.no

January 14, 2019

Abstract

Statistical prediction methods typically require some form of fine-tuning of tuning parameter(s), with K-fold cross-validation as the canonical procedure. For ridge regression there exist numerous procedures, but common for all, including cross-validation, is that one single parameter is chosen for all future predictions. We propose instead to calculate a unique tuning parameter for each individual for which we wish to predict an outcome. This generates an individualized prediction by focusing on the vector of covariates of a specific individual. The focused ridge – fridge – procedure is introduced with a two-part contribution: 1) first we define an oracle tuning parameter minimizing the mean squared prediction error of a specific covariate vector, 2) then we propose to estimate this tuning parameter by using plug-in estimates of the regression coefficients and error variance parameter. The procedure is extended to logistic ridge regression by utilizing parametric bootstrap. For high-dimensional data, we propose to use ridge regression with cross-validation as the plug-in estimate, and simulations show that fridge gives smaller average prediction error than ridge with cross-validation for both simulated and real data. We illustrate the new concept for both linear and logistic regression models in two applications of personalized medicine: predicting individual risk and treatment response based on gene expression data. The method is implemented in the R package `fridge`.

Keywords: focused information criterion; genomics; personalized medicine; ridge regression; tuning parameters.

1 Introduction

The development of inexpensive genomic technologies has greatly contributed to the field of personalized medicine, by facilitating predictions of individualized treatment decisions and disease risks based on genetic characteristics (Hamburg and Collins, 2010). In Norway, for instance, the Norwegian Cancer Genomics Consortium (cancergenomics.no) has been founded to establish "nationwide use of individual patient genetics to guide cancer treatment". Genomic data are typically high-dimensional with the number of variables, p, greatly exceeding the number of observations, n, and this high-dimensionality is often handled by regularization, or by constructing new, low-dimensional features (Hastie et al., 2009). Penalized linear regression, the most widely used regularization technique, introduces some form of penalization of the regression coefficients in the linear model, $y_i = x_i^T\beta + \varepsilon$. Ridge regression imposes an L_2 penalty, penalizing the sum of squared regression coefficients:

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

enforcing both proportional shrinkage of all coefficients towards zero (increasing the bias, but lowering the variance) and shrinking positively correlated variables towards each other. In addition, the least informative directions in the data space are penalized more, effectively improving the predictive performance of the method (Hastie et al., 2009, p.82). Ridge regression was originally introduced as an extension of ordinary least squares (OLS) regression to handle rank-deficient data matrices (Hoerl and Kennard, 1970; van Wieringen, 2015). Ridge regression has become a standard prediction tool within genomics, and has, for instance, been shown to give the lowest error of a range of methods when predicting survival based on gene expression data (Bøvelstad et al., 2007).

As with all penalized regression methods, the L_2 penalty is controlled by a tuning parameter, λ, dictating the fit of the predictive machinery balancing between over- and under-fitting. A range of fine-tuning procedures has been proposed for ridge regression, including minimizing the mean squared error (MSE) of $\hat{\beta}(\lambda)$ (Lawless, 1981; Hemmerle, 1975), marginal maximum likelihood (Tran, 2009; Johnsen, 2011), bootstrapping (Delaney and Chatterjee, 1986), Bayesian methods (Zuliana and Perperoglou, 2016) and versions of AIC (Boonstra et al., 2015). K-fold cross-validation (CV) has nevertheless become the canonical choice of procedure (Hastie et al., 2009, p. 243). CV works by dividing the data into K parts, or folds, (typically 5 or 10) predicting each fold out-of-sample based on the remaining data, and finally averaging the squared prediction error over all folds. This is done for a range of tuning parameters, selecting the value with the lowest average error. Variations of CV include generalized cross-validation (Golub et al., 1979) and approximate cross-validation (Meijer and Goeman, 2013).

Common for all procedures is that only *one* tuning parameter value is found for all further use and future predictions. For some applications, however, it can be important to minimize the prediction error of a specific individual, rather than the average prediction error. In applications of personalized medicine, individual predictions can determine decisions with severe consequences: a predicted increase in the risk of complications could trigger a surveillance response, or a treatment with high predicted success probability can be initiated despite possible adverse side effects. Hence, our goal is to adjust the prediction model towards each specific patient by modifying the tuning parameter. When the risk or treatment response of a new patient is to be predicted, can we find a tuning parameter optimal for that particular patient's covariate vector x_0? Such targeting of a specific covariate vector is made possible when the optimal tuning parameter is based on minimizing the *expected error* of the prediction, the so-called MSE approach.

The selection of tuning parameter(s) shares parallels with the task of model selection, for which the focused information criterion (FIC) has introduced the concept of addressing a 'final outcome' of a fitted model, such as a specific prediction, instead of an overall goodness-of-fit (Claeskens and Hjort, 2008). Along the same lines, we first define an oracle personalized tuning parameter λ_{x_0}, the minimizer of the expected MSE of the ridge prediction $x_0^T\hat{\beta}(\lambda)$ as a function of λ. Second, we propose to estimate the tuning parameter by using plug-in estimates of the regression coefficients and error variance parameters in the MSE expressions. The approach aims to minimize the expected prediction error for each individual, rather than minimizing the sample prediction error over all individuals simultaneously; in other words we average over the (theoretical) distribution of y_i and not over the observed set of y_is, as done by CV. Such individualized tuning parameters require a recalculation of the ridge model for each prediction, which previously would have been a constraint. But due to current computational power, this can be viewed as one approach to tailoring predictions towards the individual.

Related work, also modifying the ridge regression tuning parameter, introduced a variable specific weight to the tuning parameter with an additional grouping determined by external data (Wiel et al., 2016). This leads to a different λ for each (group of) covariate(s), say λ_j for the jth covariate. Our approach, however, aims at producing an *individualized* tuning parameter, say $\lambda(x_0)$ for each new individual, i.e. each vector of covariates x_0. In other words, in terms of the $n \times p$ matrix X, the previous work has tailored the λ per (group of) column(s), whereas we construct one λ per row.

The outline of the paper is as follows: Section 2 introduces the framework of the procedure, the general definition of the oracle tuning parameter and our proposed approach to estimate the tuning parameter by plug-in estimates, and Section 3 gives an overview of the theoretical aspects of the procedure. Section 4 extends the fridge approach to the logistic regression model. Section 5 shows the results of simulations

comparing the focused tuning and cross-validation in real and simulated data, and Section 6 illustrates the procedure in the linear and logistic regression models with two genomic data examples.

2 Method

Consider the data $\{y_i, x_i\}$, consisting of n observations of a continuous outcome y_i and p-dimensional vector of covariates x_i, following the linear regression model

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

with unknown p-dimensional regression coefficients, β, and *iid* noise, ε_i, with zero mean and variance σ^2. We denote the outcome vector Y, and the $n \times p$ data matrix X. For $p < n$ and a data matrix of full rank, the OLS estimate of β is given by $\tilde{\beta} = (X^T X)^{-1} X^T Y$.

In the high-dimensional situation with $p > n$, the least squares criterion requires a penalty to give a unique solution, and ridge regression introduces an L_2 penalty (Hoerl and Kennard, 1970), also known as Tikhonov regularization,

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

with the *explicit* solution

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

Suppose we aim to predict the expected outcome of y_0 for the vector of covariates of a specific individual x_0, the focus parameter $\mu_0 = E y_0 = x_0^T \beta$, 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 we then consider the expected MSE of the prediction, where the expectation is 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 vector of covariates x_0, i.e. representing a new individual or patient, the MSE will have a different minimum as a function of λ, as seen in Figure 1. We will aim to estimate these MSE curves separately for each x_0 to lower the expected prediction error of each individual, and as this optimal value is given for known parameters β and σ^2, it is termed the oracle tuning parameter.

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.

The oracle value of the personalized 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 other method and plug-in the resulting estimates in Eq. (2).

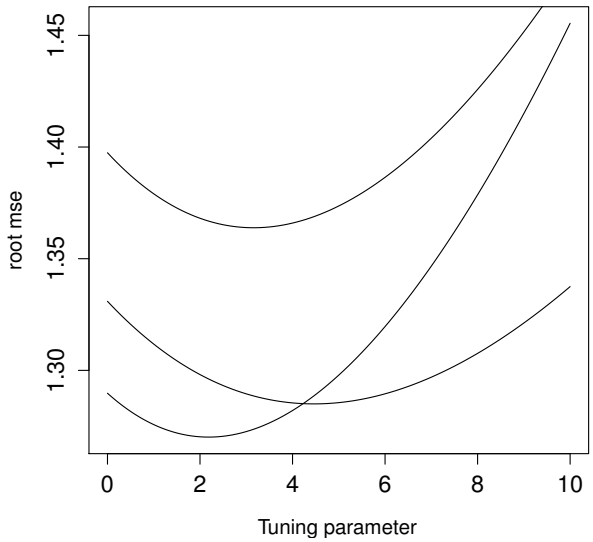

Figure 1: MSE curves for different x_0 demonstrating how the minima occur at different values of λ.

In low dimension ($p < n$) the simplest choice of plug-in estimate is the ordinary least squares (OLS) estimator (assuming X is of full rank), and corresponding variance estimator

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

In the high-dimensional situation ($p \gg n$), one choice of plug-in estimate is ridge regression tuned by standard CV. Alternative plug-in estimators include lasso (L_1 penalty) and principal component regression both fine-tuned by CV, or the OLS estimate combined with the Moore-Penrose pseudo-inverse. However, when ridge regression is already considered an appropriate prediction method for the problem at hand, it is natural to also use ridge as a plug-in estimate.

The MSE in Eq. (2) relies on the squared bias, Bias^2, estimated by squaring the estimated bias, $(\widehat{\text{Bias}})^2$, directly. The resulting estimate will be biased as $\mathbb{E}((\widehat{\text{Bias}})^2) = \text{Bias}^2 + \text{Var}\,\widehat{\text{Bias}}$, and a correction of the overestimation is necessary (Claeskens and Hjort, 2008, p. 150). We can correct the squared bias, for instance by subtracting the variance of the bias and truncate at zero

$$\max\{(\widehat{\text{Bias}})^2 - \text{Var}\,\widehat{\text{Bias}}, 0\},$$

or using a smooth correction

$$(\widehat{\text{Bias}})^2 - \frac{(\widehat{\text{Bias}})^2}{(\widehat{\text{Bias}})^2 + 1} \text{Var}\,\widehat{\text{Bias}}.$$

We will in our proposed procedure consider the first option.

In the low-dimensional case, the OLS estimates can be used as plug-in estimates:

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

mean squared error curve

$$\begin{aligned}
\hat{\lambda}_{x_0,OLS} &= \arg\min_{\lambda} \widehat{\mathrm{MSE}}_{\hat{\mu}}(\lambda; x_0, \tilde{\beta}, \tilde{\sigma}^2), \\
&= \arg\min_{\lambda} \left\{ \left((\widehat{\mathrm{Bias}}(\lambda))^2 - \mathrm{Var}\,\widehat{\mathrm{Bias}}(\lambda) \right)_+ + \widehat{\mathrm{Var}}(\lambda) \right\}, \\
&= \arg\min_{\lambda} \Big\{ \Big((\lambda x_0^T (X^T X + \lambda I_p)^{-1} \tilde{\beta})^2 \\
&\qquad - \tilde{\sigma}^2 \lambda^2 x_0^T (X^T X + \lambda I_p)^{-1} (X^T X)^{-1} (X^T X + \lambda I_p)^{-1} x_0 \Big)_+ \\
&\qquad + \tilde{\sigma}^2 x_0^T (X^T X + \lambda I_p)^{-1} X^T X (X^T X + \lambda I_p)^{-1} x_0 \Big\}. \qquad (3)
\end{aligned}$$

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

We can construct a simplified version of fridge by omitting the bias correction

$$\begin{aligned}
\hat{\lambda}^*_{x_0,OLS} = \arg\min_{\lambda} \Big\{ & (x_0^T ((X^T X + \lambda I_p)^{-1} X^T X - I_p) \tilde{\beta})^2 \qquad (4) \\
& + \tilde{\sigma}^2 x_0^T (X^T X + \lambda I_p)^{-1} X^T X (X^T X + \lambda I_p)^{-1} x_0 \Big\}.
\end{aligned}$$

which ensures a continuous first derivative.

In the high-dimensional case, we propose to use ridge regression with the tuning parameter found by CV as the plug-in estimate:

Definition 3 (Fridge-ridge)**.** *The fridge-ridge tuning parameter estimate is the minimand of the estimated mean squared error curve*

$$\begin{aligned}
\hat{\lambda}_{x_0,ridge} = \arg\min_{\lambda} \Big\{ & \Big((\lambda x_0^T (X^T X + \lambda I_p)^{-1} \hat{\beta}(\hat{\lambda}_{\mathrm{CV}}))^2 \\
& - \hat{\sigma}^2 \lambda^2 x_0^T (X^T X + \lambda I_p)^{-1} (X^T X + \hat{\lambda}_{\mathrm{CV}} I_p)^{-1} X^T X (X^T X + \hat{\lambda}_{\mathrm{CV}} I_p)^{-1} (X^T X + \lambda I_p)^{-1} x_0 \Big)_+ \\
& + \hat{\sigma}^2 x_0^T (X^T X + \lambda I_p)^{-1} X^T X (X^T X + \lambda I_p)^{-1} x_0 \Big\},
\end{aligned}$$

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

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

with the effective degrees of freedom $df(\hat{\lambda}_{CV}) = \mathrm{tr}(X(X^T X + \hat{\lambda}_{CV})^{-1} X^T)$.

Other estimators for σ^2 in the high-dimensional setting have also been proposed (Dicker, 2014).

For the MSE of the ridge estimate, $\mathrm{MSE}\left(\hat{\beta}(\lambda)\right)$, it has been shown that there always exists a value, $\lambda > 0$, for which the MSE of ridge regression will be smaller than the MSE of OLS (Hoerl and Kennard, 1970; van Wieringen, 2015). This is also true for the MSE of the prediction $x_0^T \beta$; there always exists a tuning parameter value, $\lambda > 0$, with smaller mean squared prediction error. If Eq. (2) is rewritten in terms of the singular value decomposition, $X = UDV^T$, as a summation over the singular vectors, $v_1, \cdots, v_p$, and singular values, $d_1, \ldots, d_p$

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

the first derivative of $\text{MSE}_{\hat{\mu}}$ with respect to λ

$$\frac{\partial}{\partial\lambda}\,\text{MSE}_{\hat{\mu}}(\lambda; x_0, \beta, \sigma^2) = 2\lambda\left[\sum_{i=1}^{p}\frac{x_0^T v_i v_i^T \beta}{d_i^2+\lambda}\right]\left[\sum_{i=1}^{p}\frac{d_i^2 x_0^T v_i v_i^T \beta}{(d_i^2+\lambda)^2}\right] - 2\sigma^2\sum_{i=1}^{p}\frac{d_i^2(x_0^T v_i)^2}{(d_i^2+\lambda)^3}, \tag{5}$$

is always negative in the limit, $\lambda \to 0$:

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

Thus there always exists a tuning parameter value larger than zero, $\lambda > 0$, for which $\text{MSE}(\lambda)$ is smaller than $\text{MSE}(0)$.

The MSE of $\beta^T\beta$ has a single global minimum, which is not the case for the MSE of the prediction, $x_0^T\hat{\beta}(\lambda)$, as the MSE curves in Eq. (2) can have several local minima. Extra care therefore needs to be taken when using numerical optimizers to locate the global minimum. There are no explicit solutions for the minima, except for the case of all singular values being equal, and the limit values of the curve are given

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

Thus if $(x_0^T\beta)^2 \geq \sigma^2 x_0^T(X^TX)^{-1}x_0$, there must exist a global minimum for $\lambda < \infty$. In the reverse case, however, the global minimum can be given in the limit $\lambda \to \infty$.

We illustrate these characteristics in the case of $p = 10, n = 100$ with a fixed x_0 and normally distributed data and regression coefficients, $x_j \sim N(0, I_p), \beta \sim N(0, I_p)$. For this setup we typically see zero to three critical points, as shown in Figures 2 and 3. There can be no critical points (Figure 2a), giving the global minimum in the limit, $\lambda \to \infty$, or one critical point (Figure 2), a minimum, giving the global minimum at the local minimum. Further, one can have two critical points (Figure 3a), a minimum and a maximum, or three critical points (Figure 3b), two minima and a maximum, where simulations suggest that the second minimum is always below the first local minimum.

The MSE curves with the plug-in estimates exhibit the same behavior as the oracle curves. As $\lambda \to 0$ the first derivative will always be negative, such that there exists an optimal tuning parameter larger than zero. However, the asymptotic limit of the estimated MSE as $\lambda \to \infty$ changes to

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

The correction of the squared bias by truncating at zero will more often produce a global minimum in the limit $\lambda \to \infty$.

3 The benefits of focusing

The goal of fridge is to lower the expected prediction error of a specific covariate vector x_0, and to what degree this can be achieved depends on the x_0, X and β in question. In this section, we explore the theoretical characteristics of fridge in simplified examples, to showcase how the relation between x_0 and β affects the procedure.

We first consider the oracle setting, where β and σ^2 are known, to demonstrate that the effect of focusing acts through the inner product between the vector of covariates and the regression coefficients, $x_0^T\beta$. Suppose the covariates are transformed to give a diagonal covariance matrix with equal entries,

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

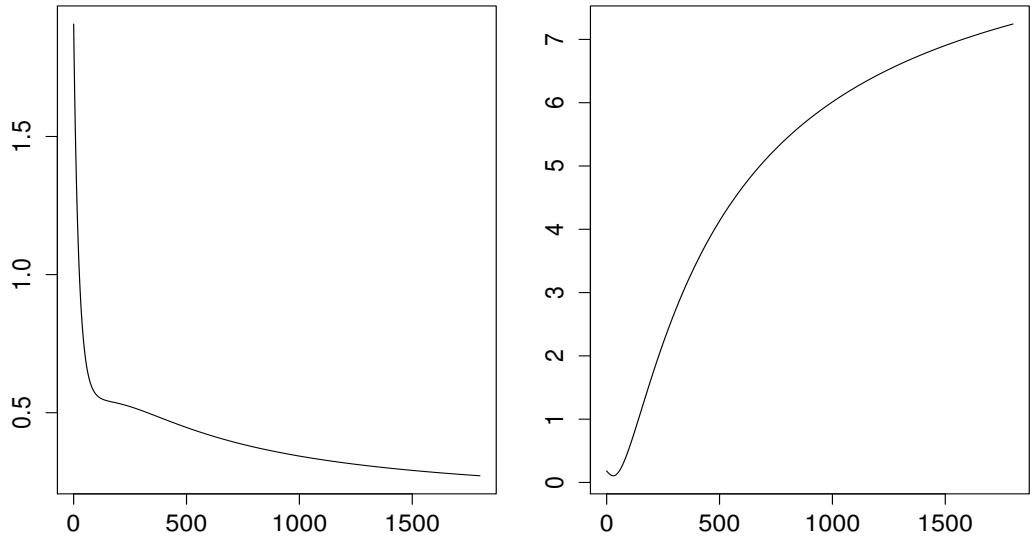

Figure 2: MSE curves as a function of λ for $p = 10$. a) With no critical points the minimum is in the limit $\lambda \to \infty$. b) The classical case with one minimum and a curve increasing towards an asymptote.

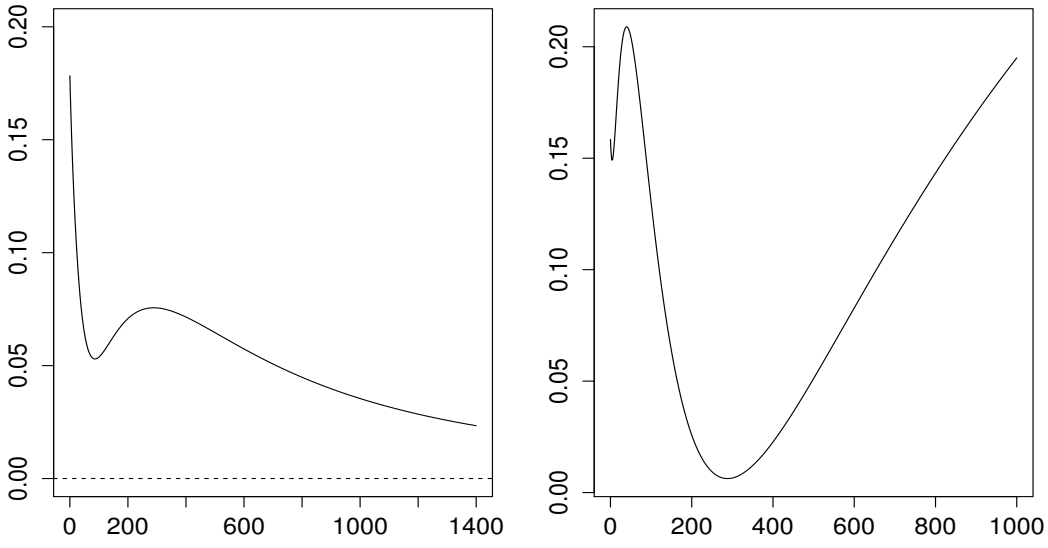

Figure 3: MSE curves as a function of λ for $p = 10$. a) First a minimum and then a maximum, but with an asymptote $x_0^T\beta$ *below* the minimum value. b) Two local minima and a local maximum.

such that the columns of X are orthogonal. This is an artificial data matrix allowing for an explicit solution with the crucial aspect being equal diagonal entries. The oracle MSE from Eq. (2) can then be written out with an explicit solution

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

The oracle tuning parameter is thus controlled by $x_0^T\beta$, the inner product between the focus covariate vector and regression coefficients, and the tuning parameter can be re-expressed in terms of the geometry of x_0 and β, specifically the length of β, $\|\beta\|$, and the angle between the vectors x_0 and β, α_{x_0}:

$$\lambda_{x_0} = \frac{\sigma^2}{\|\beta\|^2 \cos^2 \alpha_{x_0}}, \tag{6}$$

It is evident that the length of x_0, $\|x_0\|$, does not influence the value of the oracle tuning in the orthogonal case. In addition, if $p = 1$, the x_0 in Eq. (6) cancels out, such that λ_{x_0} does not depend on x_0 at all and the estimator has in some sense *lost* its focus.

The covariate vector x_0 influences the oracle tuning through its relation to β as the angle α measures how close the prediction is to the mean response. When the true outcome is close to the mean of Y, $\cos \alpha_{x_0}$ will be close to zero (meaning x_0 and β are close to being orthogonal). This causes the oracle tuning parameter to blow up, $\lambda \to \infty$, shrinking the estimated prediction towards the mean. The length of β on the other hand acts as a measure of the signal strength, such that larger values of β, i.e. a stronger signal, warrants a stronger penalization in the optimal case, while weaker signal requires less penalization. As previously stated, the resulting prediction error of fridge in the oracle case will be uniformly smaller than the OLS prediction error, corresponding to $\lambda = 0$;

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

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

$$\text{MSE}_{\hat{\mu}}(\lambda; x_0, \beta, \sigma^2) = \left\{\lambda x_0^T V (D^2 + \lambda I_p)^{-1} V^T \beta\right\}^2 + \sigma^2 x_0^T V D (D^2 + \lambda I_p)^{-1} D V^T x_0, \tag{7}$$

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 up-weighting the vectors associated with large singular values. The data matrix, therefore, gives the premise for which directions in the covariate space that are considered more important. Hence, how x_0 and β are spanned by the first singular vectors of X will together determine 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, and $V^T x_0$ and $V^T\beta$ can be viewed as *new* a covariate vector and regression coefficients oriented along the singular vectors:

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

If instead some of the singular values are substantially larger than the rest, the first singular vectors of the data matrix will dominate, and it is crucial how x_0 and β are spanned by the first singular vectors. If all variables are equally correlated

$$X^T X = \begin{bmatrix} M & R & \dots & R \\ R & M & & R \\ \vdots & & \ddots & \vdots \\ R & R & R & M \end{bmatrix},$$

the first singular vector, $v_1 = [1, 1, \ldots, 1]^T$, has a singular value substantially larger than the rest: $d_1 \gg d_2 = \cdots = d_p$. For such a data matrix, the part of x_0 and β spanned by v_1 will be heavily emphasized. If β is spanned by v_1 alone and x_0 is orthogonal to v_1, the bias part becomes zero, forcing $\lambda \to \infty$ and the prediction towards the mean. In the reverse situation, if β is not properly spanned by v_1, the same shrinkage towards the mean will occur. But when β and x_0 are spanned by v_1 alone, the tuning parameter will approximately be given by the length of β, scaled by the largest singular value, $\lambda_{x_0} = \frac{\sigma^2}{d_1^2 \|\beta\|^2}$.

It is also a question how much of the oracle optimality is lost by estimating the tuning parameter through the plug-in approach. In the orthogonal case, the distribution of the estimated bias of fridge-OLS is given

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

such that the estimated fridge-OLS tuning parameter is explicitly given

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

When combining the estimated tuning parameter with the ridge prediction

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

the risk of fridge-OLS is given

$$\text{risk}\left(x_0^T \hat{\beta}(\hat{\lambda}_{x_0,OLS})\right) = \begin{cases} (x_0^T \beta)^2 & \text{if } |x_0^T \tilde{\beta}| \leq \tilde{\sigma}\sqrt{x_0^T x_0/M}, \\ \mathbb{E}\left(\frac{(x_0^T \tilde{\beta})^2 - \tilde{\sigma}^2 x_0^T x_0/M}{(x_0^T \tilde{\beta})^2}\, x_0^T \tilde{\beta} - x_0^T \tilde{\beta}\right)^2 & \text{if } |x_0^T \tilde{\beta}| > \tilde{\sigma}\sqrt{x_0^T x_0/M}. \end{cases}$$

If the residuals are assumed to be normally distributed, $\varepsilon_i \sim N(0, \sigma^2)$, the risk can be visualized in two dimensions, for either fixed x_0 or β. Figure 4 displays contour plots of the risk of fridge-OLS for an orthogonal data matrix, scaled to give OLS risk equal to 1. The left panel shows a contour plot of the risk as a function of β for fixed covariates, $x_0 = [-5, 2]$, illustrating that the risk of fridge-OLS is constant along lines parallel to $x_0^T \beta = 0$. Thus fridge-OLS will have lower risk than OLS within a trench following the line $x_0^T \beta = 0$. The right panel shows the risk as a function of x_0 for fixed regression coefficients, $\beta = [-5, 2]$, and then the risk will be constant along radial lines from the origin. This illustrates that it is the angle between x_0 and β, and not the length of x_0, which determines the benefit of the focused approach. The fridge-OLS will have lower risk than OLS within a *cone*, oriented along the line $x_0^T \beta = 0$.

4 Fridge for logistic ridge regression

The focused approach to ridge regression can also be extended to generalized linear models by utilizing parametric bootstrap to obtain expressions for the variance and squared bias. Consider logistic ridge regression: independent responses, $y_i \in \{0, 1\}$ for $i = 1, \ldots, n$, are distributed as

$$y_i \sim \text{Bernoulli}(\text{logit}^{-1}(x_i^T \beta)),$$

for covariate vector x_i with the link function $\text{logit}^{-1}(x_i^T \beta) = \exp(x_i^T \beta)/(1 + \exp(x_i^T \beta))$, where the regression coefficients, $\hat{\beta}$, are estimated by maximizing the penalized log-likelihood with an L_2 penalty

$$\sum_{i=1}^{n} [y_i \log(p_i) + (1 - Y_i) \log(1 - p_i)] - \lambda \sum_{j=1}^{p} \beta_j^2, \quad p_i = \text{logit}^{-1}(x_i^T \beta),$$

typically using the Newton-Raphson algorithm (Le Cessie and Van Houwelingen, 1992).

We then propose the following procedure for extending the fridge concept to logistic regression:

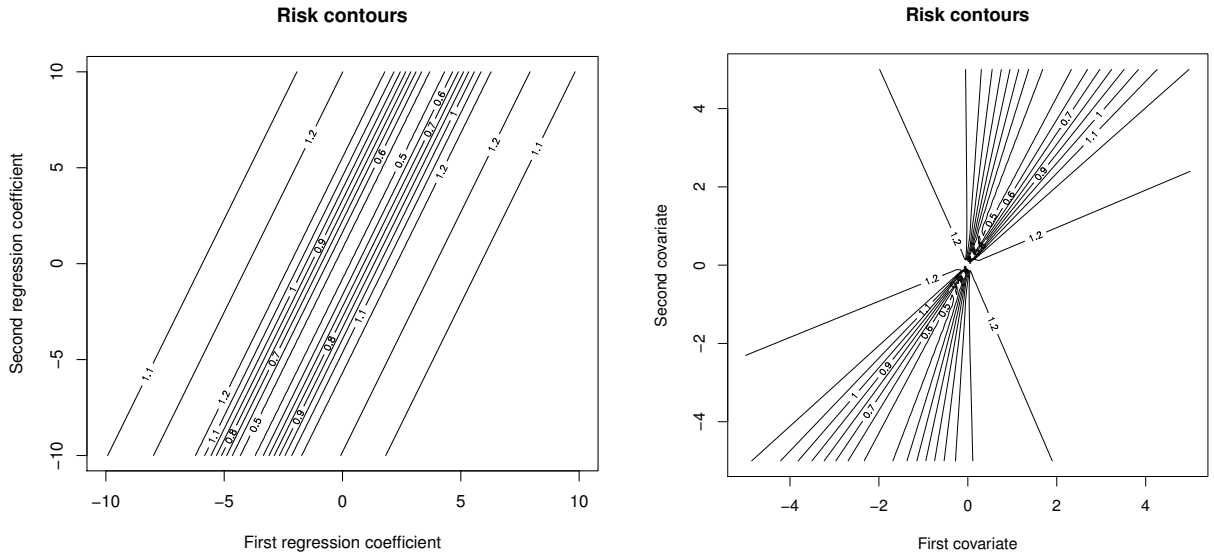

Figure 4: Contour plots of the risk of fridge-OLS for an orthogonal data matrix scaled relative to the OLS risk a) as a function of β_1 and β_2 for a fixed covariate vector $x_0 = (-5, 2)$, and b) as a function of $x_{0,1}$ and $x_{0,2}$ for fixed regression coefficients $\beta = (-5, 2)$.

1. Use parametric bootstrap with plug-in $\tilde{\beta}$ to simulate $r = 1, \ldots, M$ bootstrap samples of n observations $Y^{(r)}$
$$y_i^{(r)} \sim \text{Bernoulli}(\text{logit}^{-1}(x_i^T \tilde{\beta})), \qquad i = 1, \ldots, n.$$
2. Over a suitable grid of λ, holding the tuning parameter value fixed:
   - calculate $\hat{\beta}_\lambda^{(r)}$ for each bootstrap sample,
   - find the squared bias and variance of $x_0^T \hat{\beta}_\lambda^{(r)}$, compared to $x_0^T \tilde{\beta}$ as the population parameter,
   - add the terms, yielding the MSE.
3. Set the estimate $\hat{\lambda}_{x_0}$ to the tuning parameter value with the smallest MSE over the grid of λ.

With the use of the `glmnet` `R` package, the estimate $\hat{\beta}_\lambda^{(r)}$ can be calculated for a long sequence of λ within each bootstrap iteration, greatly increasing calculation speed. Similar fridge procedures of Cox's proportional hazard regression may also be established.

5 Simulation: comparison with cross-validation

To compare the predictive performance of fridge and ridge with CV, we perform simulations with both simulated and real data. K-fold cross-validation is the most widely used fine-tuning procedure, probably due to its conceptual simplicity: The data is divided into 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 (Stone, 1974; Allen, 1974). Currently 10- or 5-fold cross-validation has become the default approach in modern statistics and machine learning (Hastie et al., 2009). In the case of ridge regression, leave-one-out cross-validation (LOOCV) error has an explicit

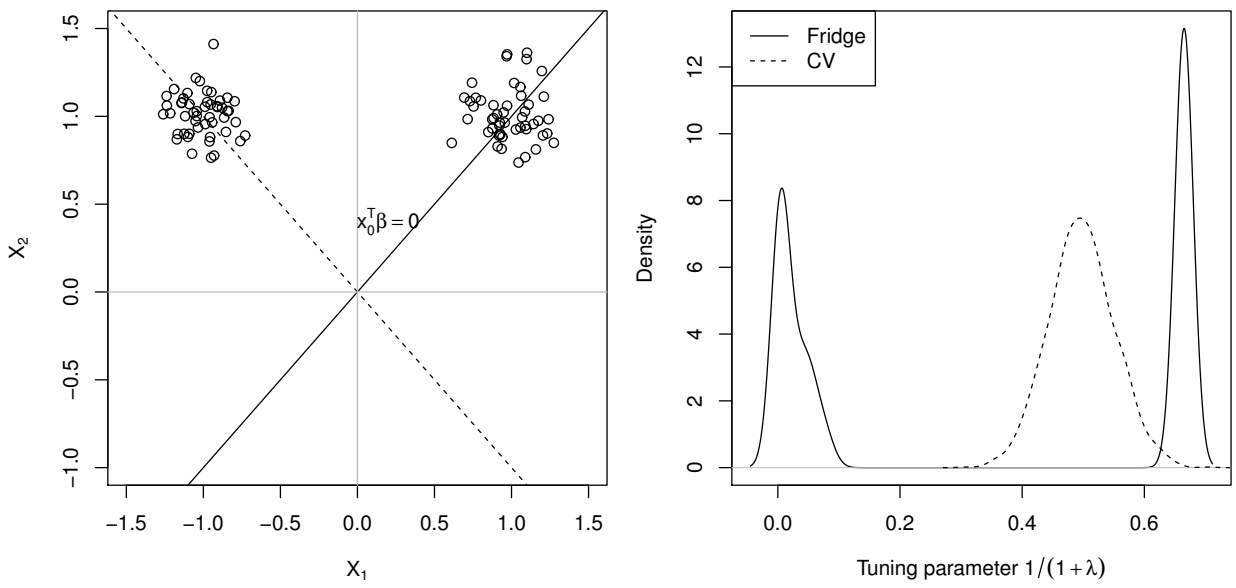

Figure 5: a) Data concentrated in two distinct clusters centered at $(1,1)$ and $(-1,1)$ b) Density plot of $1/(\lambda+1)$ for fridge (black) and cross-validation (dashed) for $\beta = [-1,1]$ and $\sigma^2 = 1$.

expression (Golub et al., 1979), thus being particularly easy to calculate

$$\hat{\lambda}_{\text{CV}} = \arg\min_{\lambda} \sum_{i=1}^{n} \left(\frac{y_i - x_i^T \hat{\beta}(\lambda)}{1 - w_{ii}} \right)^2, \quad w_{ii} = x_i^T (X^T X + \lambda I)^{-1} x_i,$$

and LOOCV will be used for the remainder of the paper.

To illustrate the difference between the focused tuning and cross-validation, consider an example where the data matrix consists of different clusters. Figure 5a) 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, $x_i^T\beta \simeq 0$, while the outcome for the left cluster will be close to two, $x_i^T\beta \simeq 2$. 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 5b) shows the distribution of the covariate vector specific tuning, $1/(1+\lambda_{x_0})$ in red, for each of the observations seen in Figure 5a). The corresponding distribution of the tuning parameter estimated by leave-one-out cross-validation, $1/(1+\hat{\lambda}_{CV})$ over multiple sets of simulated y_i is shown in black. Figure 5b) 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 large tuning parameter value and the left cluster to a small tuning parameter value. Cross-validation, on the other hand, estimates an overall tuning parameter, averaging over all individuals, and thus selects a tuning parameter value inappropriate for both clusters.

In low dimensions, the risk of fridge-OLS and fridge-ridge can be compared to ridge with cross-validation for varying β. In Figure 6, the average squared prediction error is shown in the case of $p = 2$ and $n = 50$, when $\beta = [b,b]$ and the data matrix X and focus x_0 are drawn from a uniform distribution $X, x_0 \sim U(-1,1)$. It is seen that the risk of fridge-OLS is slightly higher than cross-validation for β close to zero, but lower for medium β. The fridge-ridge has the same risk as ridge with cross-validation for β close to zero, but a higher risk for medium β.

Lastly, we compare fridge to standard ridge in a simulation study based on real high-dimensional covariates, with simulated y_i and known β coefficients. The data consist of gene expression profiles measured

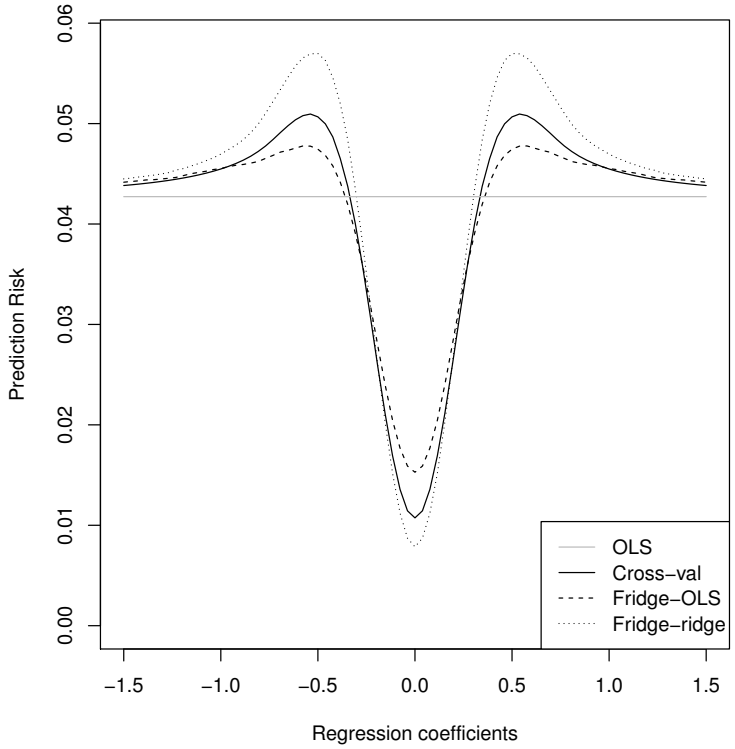

Figure 6: Risk functions for OLS, ridge with cross-validation, fridge-OLS, fridge-ridge for varying β.

in 40 glioma patients (Moeckel et al., 2014), and the data are described and analyzed in detail in Section 6. The simulation was done on the 1000 genes with the largest variance, and was carried out by drawing standard normally distributed residuals, $\varepsilon_i \sim N(0,1)$, and regression coefficients from the normal distribution, $\beta \sim N(0, 0.05)$, to ensure a suitable signal-to-noise ratio. We then simulated 200 sets of outcomes Y and calculated the average squared prediction error of fridge-ridge and ridge regression with cross-validation for each individual. The bottom panel of Figure 7 shows the relative difference in the MSE of fridge-ridge compared to ridge with cross-validation, ordered from best to worst, and fridge gives a lower MSE for a majority of the observations. The upper panel shows the out-of-sample estimated tuning parameter on an inverse scale for CV, fridge-ridge and the oracle fridge, and demonstrates that fridge-ridge estimates well the large oracle tuning parameters, but is less precise when estimating the smallest oracle values.

6 Data examples

We demonstrate the proposed fridge-ridge procure in two examples where accurate prediction of individual patients can be considered more important than overall accuracy; specifically the prediction of complication risk and the response of treatment.

6.1 Prediction of weight gain

We demonstrate the fridge procedure in a study investigating whether gene expression can be used to predict weight gain (Cashion et al., 2013), available in the EMBL-EBI ArrayExpress database (www.ebi.ac.uk/arrayexpress) under accession number E-GEOD-33070. Kidney transplant recipients are known to gain substantial weight during the first year after transplantation, with a reported average increase of 12 kg (Patel, 1998). Such large weight gain over a short time period results in increased risk for adverse health effects, e.g. cardiovascular

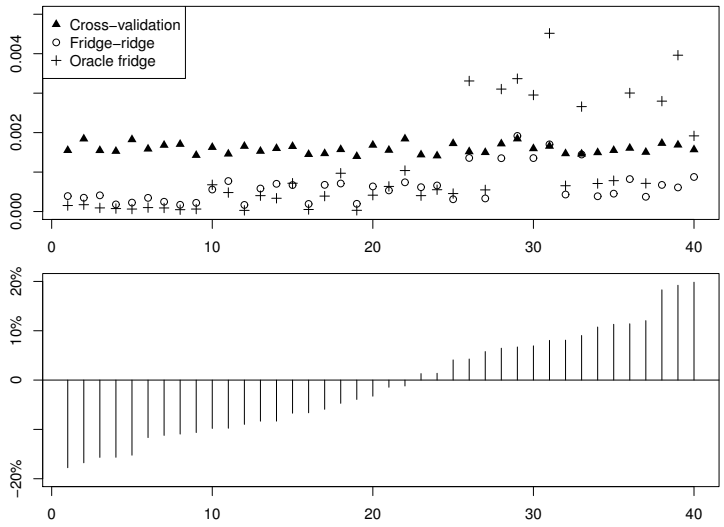

Figure 7: Top panel: the inverse of the estimated tuning parameter for each out-of-sample observation given by cross-validation (triangle), oracle fridge (cross) and fridge-ridge (circle). Bottom panel: the corresponding relative difference in MSE for fridge-ridge compared to ridge with cross-validation.

disease, contributing to an overall worse outcome for patients. The weight gain has been partly explained by patients experiencing better appetite, due to the use of prescribed steroids, and by patients having less restrictive diets after transplantation. However, steroid-free protocols do not alone reduce the risk of obesity, suggesting that other causes also contribute (Elster et al., 2008).

Weight gain is fundamentally caused by a too high intake of calories, relative to the energy expenditure, but the individual response is seen to be substantial. Hence genetic variations have been thought to be a contributing factor, and several genes have already been linked to obesity and weight gain. To investigate the predictive power of genomic data regarding weight gain, gene expression profiles were measured in adipose tissue taken from kidney transplant patients (Cashion et al., 2013). Subcutaneous adipose tissue was considered particularly well-suited as it is involved in appetite regulation and can be easily obtained from the patients during surgery. Tissue samples from 25 transplant patient were collected at the time of surgery, and mRNA levels were measured using Affymetrix Human Gene 1.0 ST arrays, obtaining gene expression profiles for 28 869 genes. Patients were weighed at transplantation and at a follow-up time of one year, resulting in a one-year recorded weight gain. Additional covariates, such as race and gender, were collected to adjust for possible confounding, but gene expression variability was not associated with neither characteristics.

As excessive weight gain can have severe consequences for the patients, the goal is to predict the future weight increase based on the available gene expression profiles. When a large increase in weight is predicted, additional measures such as diet restrictions or physiotherapy could be set in effect. It is thus important to predict the weight gain of each transplant recipient as accurate as possible. In such a setting, a focused tuning parameter tailored to the covariate vector of each patient could give an advantage.

We will predict each of the 25 measured patients out-of-sample based on the remaining 24 observations, considering the hold-out observation as the new covariate vector x_0. The plug-in estimates for $\tilde{\beta}$ and $\tilde{\sigma}^2$ are given by ridge regression with the LOOCV tuning parameter, following Definition 3.

Figure 8 shows the out-of-sample predicted change in weight estimated by fridge (in black) and ridge with cross-validation (in gray), plotted against the true change in weight. The difference between the predictions from the two methods is colored black when fridge gives the lowest error, and grey otherwise. In general, it is seen that both methods achieve a good out-of-sample predictive performance. The observations in the upper right corner of Figure 8 are penalized less by the focused approach, compared to cross-validation,

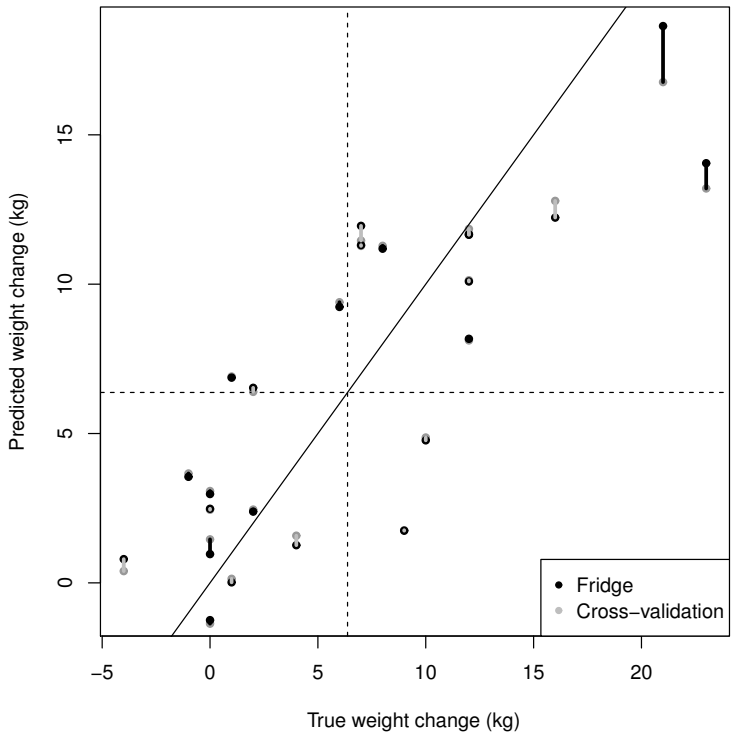

Figure 8: The out-of-sample predicted change in weight for fridge (in black) and ridge with cross-validation (in gray) against the true change in weight. The difference between the fridge and ridge with CV is colored according to the method with the lowest squared prediction error.

which results in smaller prediction error. This demonstrates how fridge aims to give a smaller penalty to predictions far from the outcome mean and a larger penalty to the predictions close to the outcome mean.

Fridge gives a smaller prediction error compared to ridge with cross-validation for around 44% of the observations. But the overall improvement in the individual predictions gives an average squared prediction error of 15.60 for fridge and 16.24 for standard ridge, which yields a 4.0 % decrease when using the focused approach. The average parameter estimate over the leave-one-out models was, $\hat{\sigma}^2 = 6.7$, for the variance and, $\hat{\lambda}_{cv} = 12.2$, for the cross-validation tuning parameter.

6.2 Prediction of treatment response

The prediction of treatment response, in particular within cancer treatment, is an important application in the field of personalized medicine. We illustrate the logistic fridge procedure with treatment response data in glioma tumor samples.

High-grade gliomas, cancer tumors in the brain, are amongst the most deadly human tumors, and in treatment trials only around 20% of patients respond to therapy. It has been investigated whether gene expression can be used to identify glioma patients that will profit from cancer therapy (Moeckel et al., 2014), and the data are publicly available in the EMBL-EBI ArrayExpress database (www.ebi.ac.uk/arrayexpress) under accession number E-GEOD-76900. Decisions regarding treatment for glioma patients are currently based on age and performance status (Weller et al., 2012), and increased use of molecular markers would be beneficial. As a follow-up validation study, 18 samples of high-grade gliomas were treated with Sunitinib (tyrosine kinase inhibitor) to measure treatment response in terms of decrease in proliferation rate, or cell growth, over a 6 hour time period, compared to a control sample of the same tissue (Moeckel et al., 2014). To illustrate logistic fridge, we dichotomize the outcome in a high and low treatment response group, relative to the median value. Before treatment, genetic expression profiles of 19 410 genes were measured using Affymetrix Human gene 1.1 arrays, and for analysis, we used the 3000 genes most correlated with the treatment response.

Figure 9 shows the ROC curves for the logistic out-of-sample prediction of the 18 samples, with the predictions based on fridge (black line) and the ridge with cross-validation (dashed line). It is seen that the ROC curves are identical for most of the predictions probably due to the low number of samples, but the fridge preforms better in predicting the zero values. Fridge performs better in terms of area under the curve (AUC) with an AUC value of 0.914, compared to an AUC value of 0.877 for ridge regression with CV.

7 Discussion

The development of personalized medicine will increase the demand for new prediction methodologies targeting the individual, where one possible approach is to allow tuning parameter(s) to vary with the covariate values for which the prediction is to be given. A covariate vector specific tuning parameter, λ_{x_0}, can be defined as the minimizer of the expected prediction error, and we have proposed to estimate this tuning parameter by plugging in separately obtain regression coefficients and noise variance estimates in the theoretical MSE expressions. As the MSE approach minimizes the error on a *population level*, it allows for the added focus on specific individuals. Standard cross-validation, on the other hand, cannot be focused in the same way, as it relies on averaging over all observed outcomes to minimize the *sample* prediction error. Where fridge tries to minimize the expected squared prediction error for each individual, cross-validation minimizes the observed squared prediction error over all individuals. The difference lies in averaging over the theoretical distribution of y_i, instead of averaging over the observed outcomes. Our simulations demonstrate, however, the adaptability and robustness of cross-validation, and that the benefits of focusing can be lost if the utilized plug-in estimates are not precise enough. Future works should explore whether one can predetermine for which covariate values the fridge approach will out-perform CV. Further, also alternative loss functions and more direct approaches to estimation should be explored.

The use of ridge regression combined with cross-validation as the plug-in estimate can also be set in the following conceptual framework: in an initial step, cross-validation is used to establish an average or

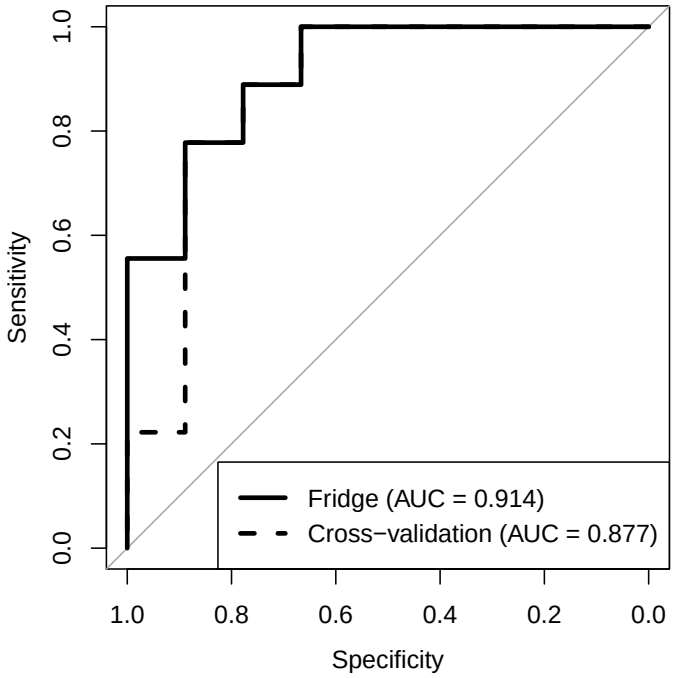

Figure 9: The ROC curve for fridge-ridge (black) and ridge with cross-validation (dashed).

overall suitable level of penalization, followed by a focusing step where the average value is tweaked to a necessary stronger or weaker penalization, depending on the individual predictions. 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 with a variance inversely proportional to the tuning parameter, a covariate vector specific λ_{x_i} can be formulated as a Gaussian prior with a variance specific to the covariate vector. This can further be viewed as an individual-specific scaling of 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 individual-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.

Based on the plug-in approach to estimate the tuning parameter, a possible extension of fridge is to utilize external data to form the plug-in estimate. The integration of external data, so-call co-data, such as p-values from earlier studies, gene annotations or other prior knowledge (Tai and Pan, 2007; Bergersen et al., 2011), is typically achieved by up- and down-weighting variables according to some measure of importance and the approach has also been studied in the setting of ridge regression (Wiel et al., 2016). Basing the plug-in estimates in the fridge framework additionally, or solely, on external data opens for new approach to utilizing co-data in integrative analyses.

References

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

Bergersen, L. C., I. K. Glad, and H. Lyng (2011). Weighted lasso with data integration. *Statistical applications in genetics and molecular biology 10*(1).

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

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

Cashion, A., A. Stanfill, F. Thomas, L. Xu, T. Sutter, J. Eason, M. Ensell, and R. Homayouni (2013). Expression levels of obesity-related genes are associated with weight change in kidney transplant recipients. *PloS one 8*(3), e59962.

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

Delaney, N. J. and S. Chatterjee (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.

Elster, E. A., D. B. Leeser, C. Morrissette, J. M. Pepek, A. Quiko, D. A. Hale, C. Chamberlain, C. Salaita, A. D. Kirk, and R. B. Mannon (2008). Obesity following kidney transplantation and steroid avoidance immunosuppression. *Clinical transplantation 22*(3), 354–359.

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

Hamburg, M. A. and F. S. Collins (2010). The path to personalized medicine. *New England Journal of Medicine 363*(4), 301–304.

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

Hemmerle, W. J. (1975). An explicit solution for generalized ridge regression. *Technometrics 17*(3), 309–314.

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

Johnsen, H. G. (2011). Regularized parameter estimation and the choice of tuning parameters. Master's thesis, University of Oslo.

Lawless, J. (1981). Mean squared error properties of generalized ridge estimators. *Journal of the American Statistical Association 76*(374), 462–466.

Le Cessie, S. and J. C. Van Houwelingen (1992). Ridge estimators in logistic regression. *Applied statistics 41*(1), 191–201.

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

Moeckel, S., K. Meyer, P. Leukel, F. Heudorfer, C. Seliger, C. Stangl, U. Bogdahn, M. Proescholdt, A. Brawanski, and A. Vollmann-Zwerenz (2014). Response-predictive gene expression profiling of glioma progenitor cells in vitro. *PloS one 9*(9), e108632.

Patel, M. G. (1998). The effect of dietary intervention on weight gains after renal transplantation. *Journal of Renal Nutrition 8*(3), 137–141.

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

Tai, F. and W. Pan (2007). Incorporating prior knowledge of predictors into penalized classifiers with multiple penalty terms. *Bioinformatics 23*(14), 1775–1782.

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

van Wieringen, W. N. (2015). Lecture notes on ridge regression. *arXiv preprint arXiv:1509.09169*.

Weller, M., R. Stupp, M. Hegi, and W. Wick (2012). Individualized targeted therapy for glioblastoma: fact or fiction? *The Cancer Journal 18*(1), 40–44.

Wiel, M. A., T. G. Lien, W. Verlaat, W. N. Wieringen, and S. M. Wilting (2016). Better prediction by use of co-data: adaptive group-regularized ridge regression. *Statistics in Medicine 35*(3), 368–381.

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