Fridge: personalized prediction through focused fine-tuning of ridge regression

Kristoffer Hellton
Department of Mathematics, University of Oslo

May 24, 2017

Bridge

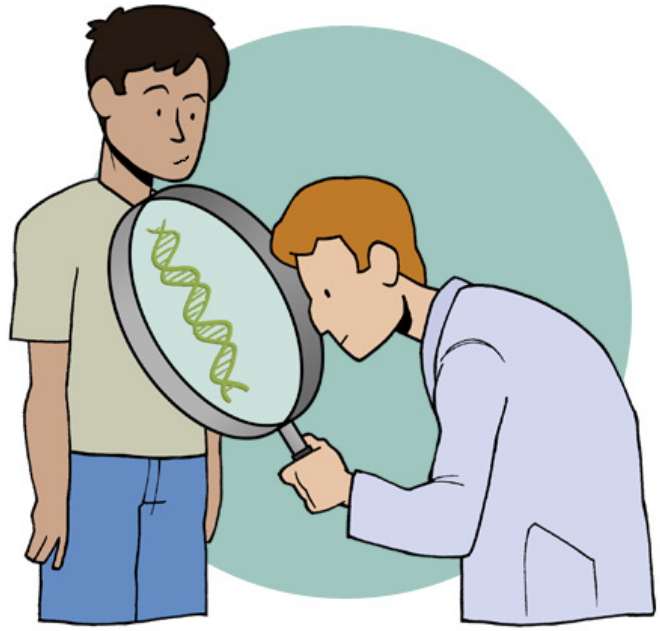
~~Bridge~~

Background: personalized medicine

Genomic data are increasingly used to achieve personalized medicine; with the prediction of individual treatment response or disease risk.

E.g. Norwegian Cancer Genomics Consortium establishes nationwide procedures for using patient genetics "to guide the adaptation of cancer treatment to the individual patient".

Scenario for the future: the doctor will first check your genetic code when visiting the doctor's office.



Ridge regression

Consider the linear regression model

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

with $\operatorname{Var} \varepsilon_i = \sigma^2$, data matrix X and outcome vector Y.

The high-dimensionality of genetic data ($p \gg n$) requires penalization; ridge regression introduces an L_2 penalty with tuning λ

$$\hat{\beta}(\lambda) = \arg\min_{\beta} \sum_{i=1}^{n} \left(y_i - x_i^T \beta \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2,$$

with the explicit solution $\hat{\beta}(\lambda) = (X^T X + \lambda I_p)^{-1} X^T Y$.

How to choose the tuning parameter

The canonical way to select the tuning parameter has become K-fold cross-validation (CV), typically 5- or 10-fold.

Or some variation: generalized cross-validation (Golub et al., 1979), approximate cross-validation (Meijer and Goeman, 2013).

For ridge regression there is a range of alternative procedures: marginal maximum likelihood, bootstrapping, Bayesian methods, different versions of AIC

Common for all, is that only *one single* tuning parameter is chosen for all future predictions.

Viewpoint of personalized medicine

After a medical study has finished and new patient enters the doctor's office, can we fine-tune the penalty towards the specific *set of covariates*, x_0, of the patient?

Assuming the linear regression model, the expected mean squared error (MSE) of the ridge prediction

$$\hat{\mu} = x_0^T \hat{\beta}(\lambda),$$

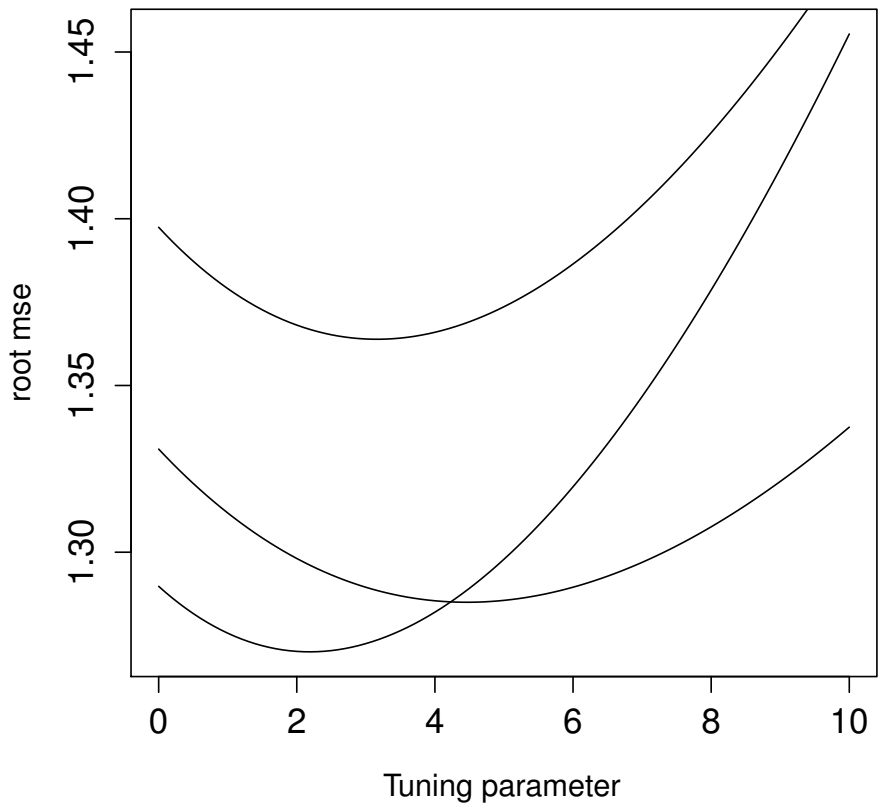
will be an *explicit* expression, over the distribution of Y:

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

Hence we can minimize $\text{MSE}_{\hat{\mu}}$ as a function of λ, to obtain an optimal oracle tuning parameter for a specific x_0:

$$\lambda_{x_0} = \arg\min_{\lambda} \text{MSE}_{\hat{\mu}}(\lambda; x_0, \beta, \sigma),$$

Each x_0 will give a different curve and an individualized tuning value, λ_{x_0}.



The oracle tuning parameter, however, requires the true β and σ^2.

We propose to estimate λ_{x_0} by using plug-in estimates of β and σ^2 in the MSE expressions to get an empirical MSE (Hellton and Hjort, 2017)

$$\begin{aligned}\hat{\lambda}_{x_0} &= \arg\min_{\lambda} \widehat{\mathrm{MSE}}_{\hat{\mu}}(\lambda; x_0, \tilde{\beta}, \tilde{\sigma}^2), \\ &= \arg\min_{\lambda} \left\{ \widehat{\mathrm{Var}}(\lambda; x_0, \tilde{\beta}, \tilde{\sigma}^2) + \widehat{\mathrm{bias}^2}(\lambda; x_0, \tilde{\beta}, \tilde{\sigma}^2) \right\}.\end{aligned}$$

We propose to use the following plug-in estimates:

for $p < n$, the standard least squares $\tilde{\beta} = (X^T X)^{-1} X^T Y$,

for $p > n$, the ridge regression estimates with cross-validation.

Fridge for $p < n$

The focused ridge, *fridge*, with an OLS plug-in, is then defined as

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

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

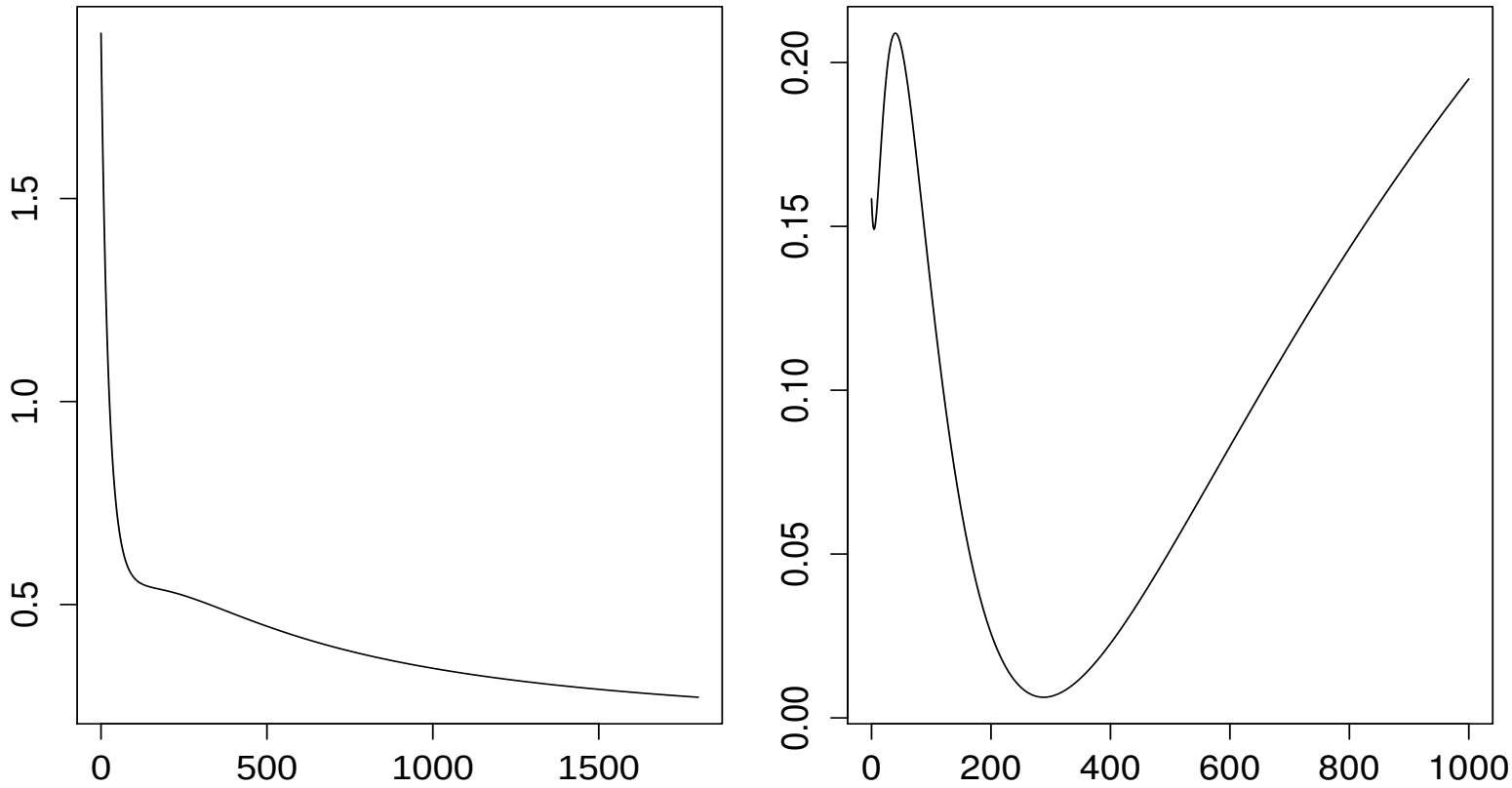
As the squared bias is estimated directly from the bias, the overestimation

$$\mathbb{E}\left(\widehat{\text{bias}}\right)^2 = \text{bias}^2 + \text{Var}\,\widehat{\text{bias}},$$

can be corrected by subtracting the variance of the bias, truncating at 0.

Some theoretical characteristics

The covariate-specific $\text{MSE}_{\hat{\mu}}(\lambda; x_0, \beta, \sigma^2)$ can have no or several minima:



If we consider an orthogonal data matrix $X^T X = MI$, the MSE is

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

with the explicit solution $\lambda_{x_0} = \frac{\sigma^2 x_0^T x_0}{M(x_0^T \beta)^2}$.

When re-expressed in terms of the geometry of x_0 and β

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

we can see that the length of x_0 does not play a role, while the length of β and α_{x_0}, the angle between x_0 and β, controls the effect of focusing.

The competitor: cross-validation

Cross-validation is the standard fine-tuning procedure, and n-fold leave-one-out CV is given by

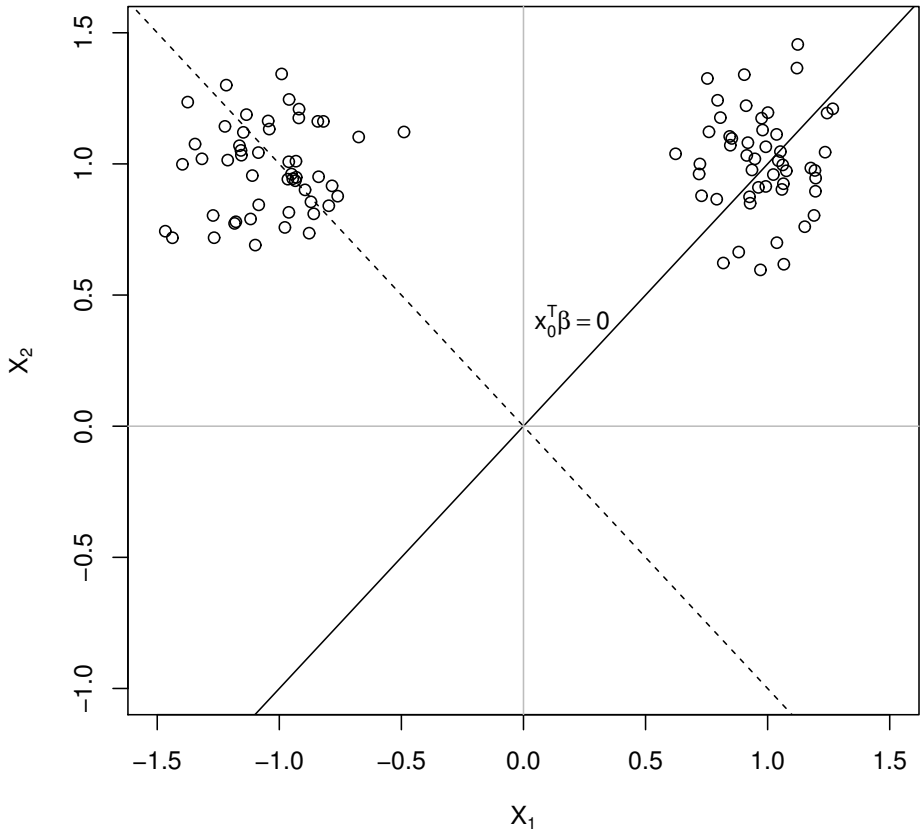
$$\hat{\lambda}_{CV} = \arg\min_{\lambda} \sum_{i=1}^{n} \left(y_i - x_i^T \hat{\beta}_{-i,\lambda} \right)^2 .$$

LOOCV has an explicit expression for ridge regression, and is therefore particularly easy to calculate:

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

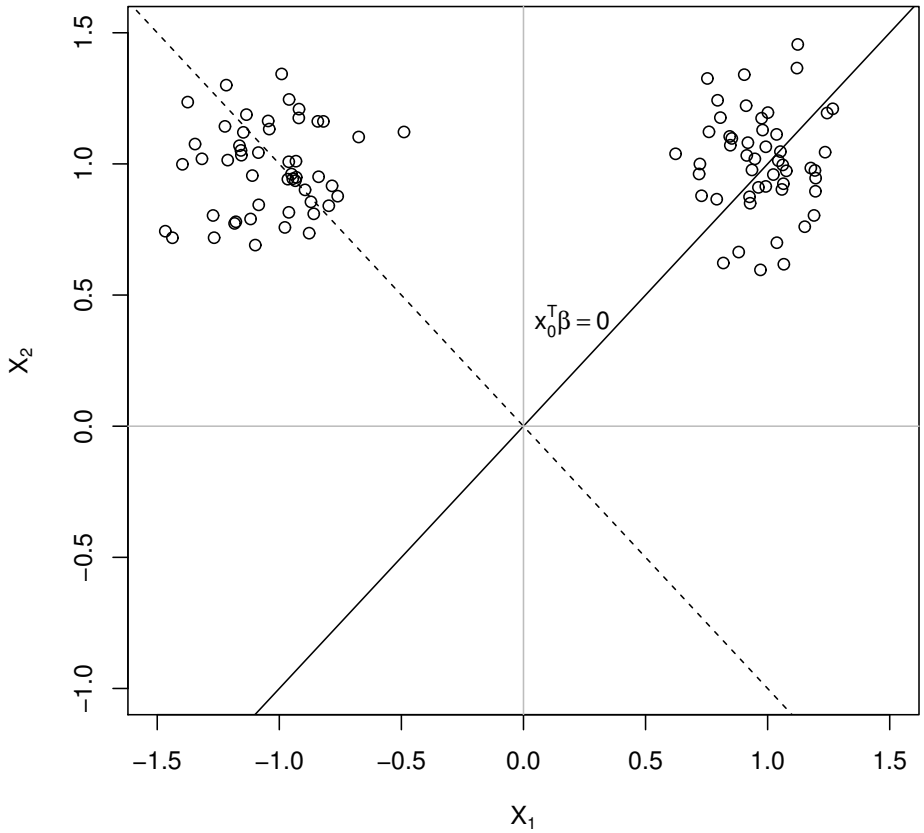
But what is the difference between cross-validation and the fridge?

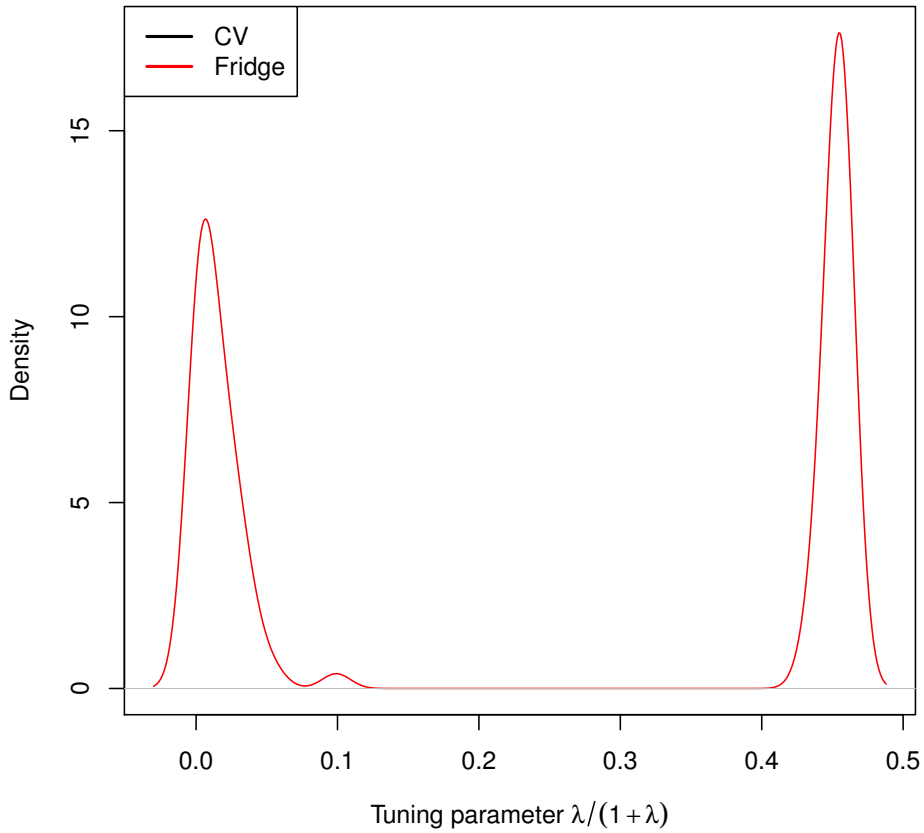
Consider known $\beta = [1, -1]$ and $\sigma^2 = 0.16$, and a data matrix X with two clusters of observations:



The right cluster will have y_i close to 0, and the left cluster y_i close to 2. This is equivalent to α_{x_0} being either close to 0 or $\pi/2$.

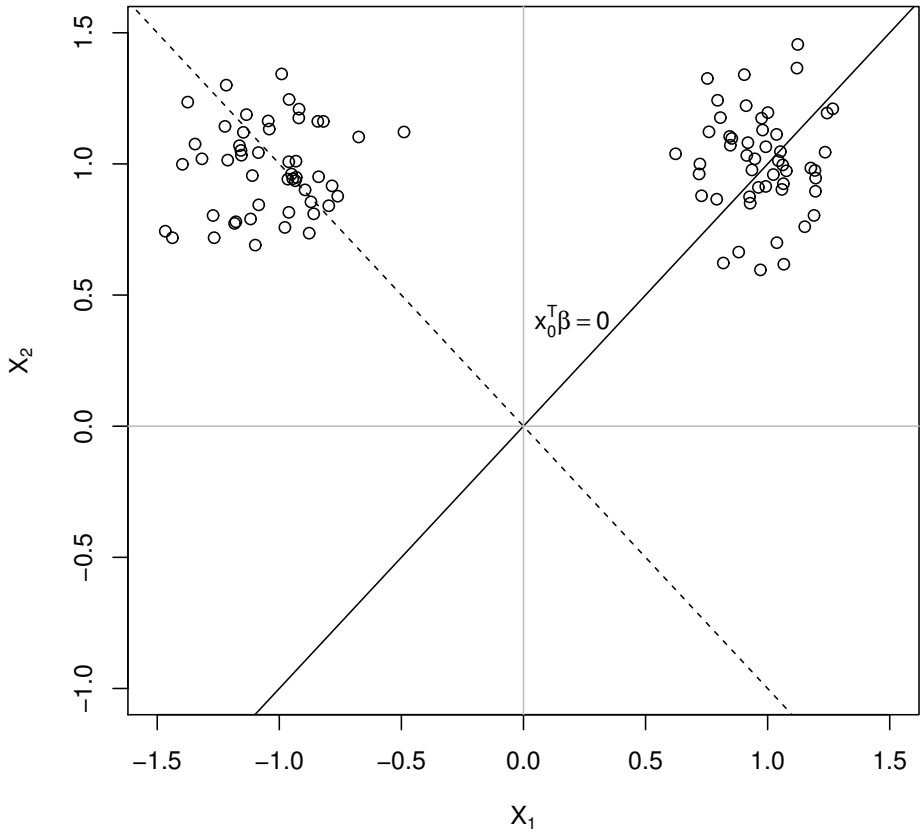
Consider known $\beta = [1, -1]$ and $\sigma^2 = 0.16$, and a data matrix X with two clusters of observations:

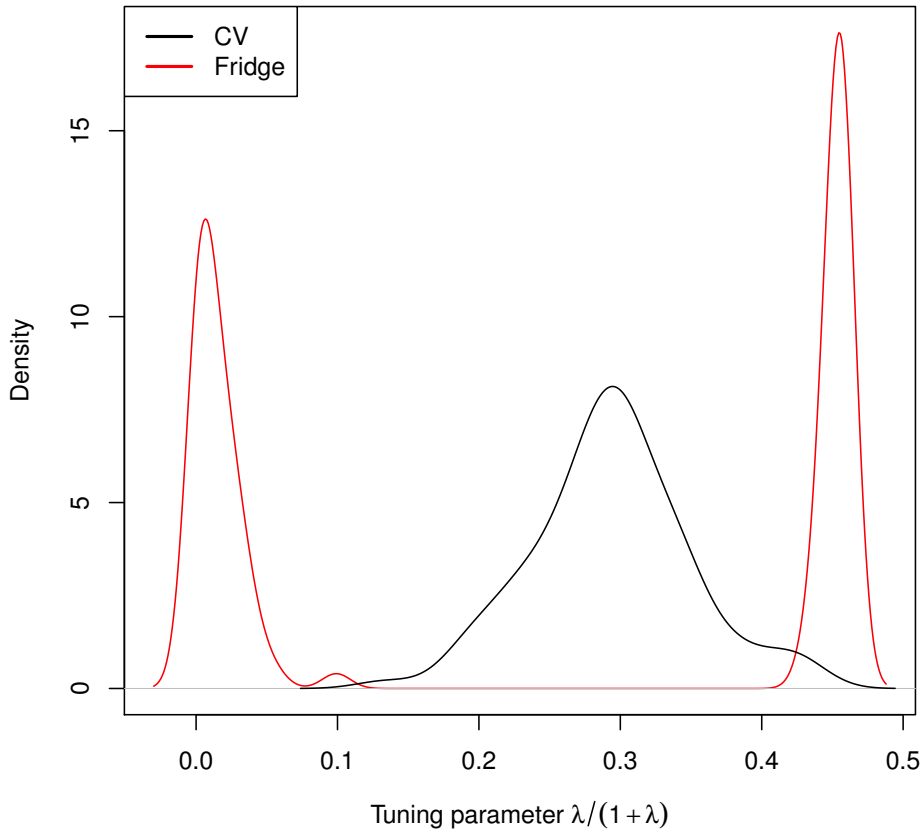




The right cluster will have y_i close to 0, and the left cluster y_i close to 2. This is equivalent to α_{x_0} being either close to 0 or $\pi/2$.

Consider known $\beta = [1, -1]$ and $\sigma^2 = 0.16$, and a data matrix X with two clusters of observations:





The right cluster will have y_i close to 0, and the left cluster y_i close to 2. This is equivalent to α_{x_0} being either close to 0 or $\pi/2$.

Individualized risk predictions

Personalized medicine: genomics data are used to predict risk of complications

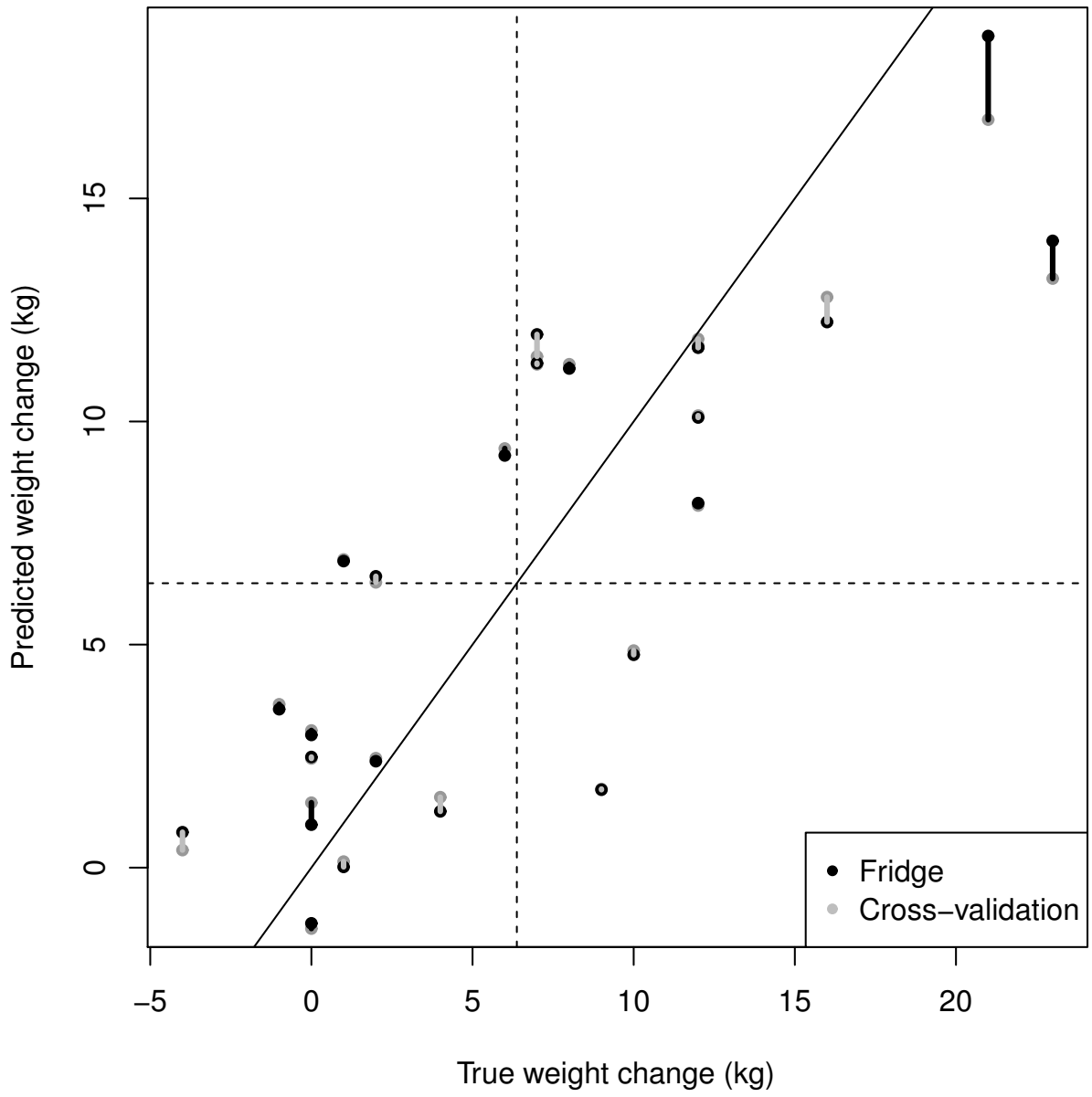
Cashion et al. (2013) investigated whether gene expression from adipose tissue can be used to predict future weight gain:

- measuring 28 869 genes,
- in 25 kidney transplant patients.

To illustrate the focused approach, we predict each observation out-of-sample using fridge and ridge regression with standard LOOCV. For fridge, we use the ridge-regression with CV as the plug-in estimate implemented in the R package `fridge`.

The difference in prediction error for each observation between fridge (black) and CV (grey) is colored according to the method with the lowest error.

Fridge gives a 4 % lower averaged squared prediction error compared to cross-validation.



Logistic fridge

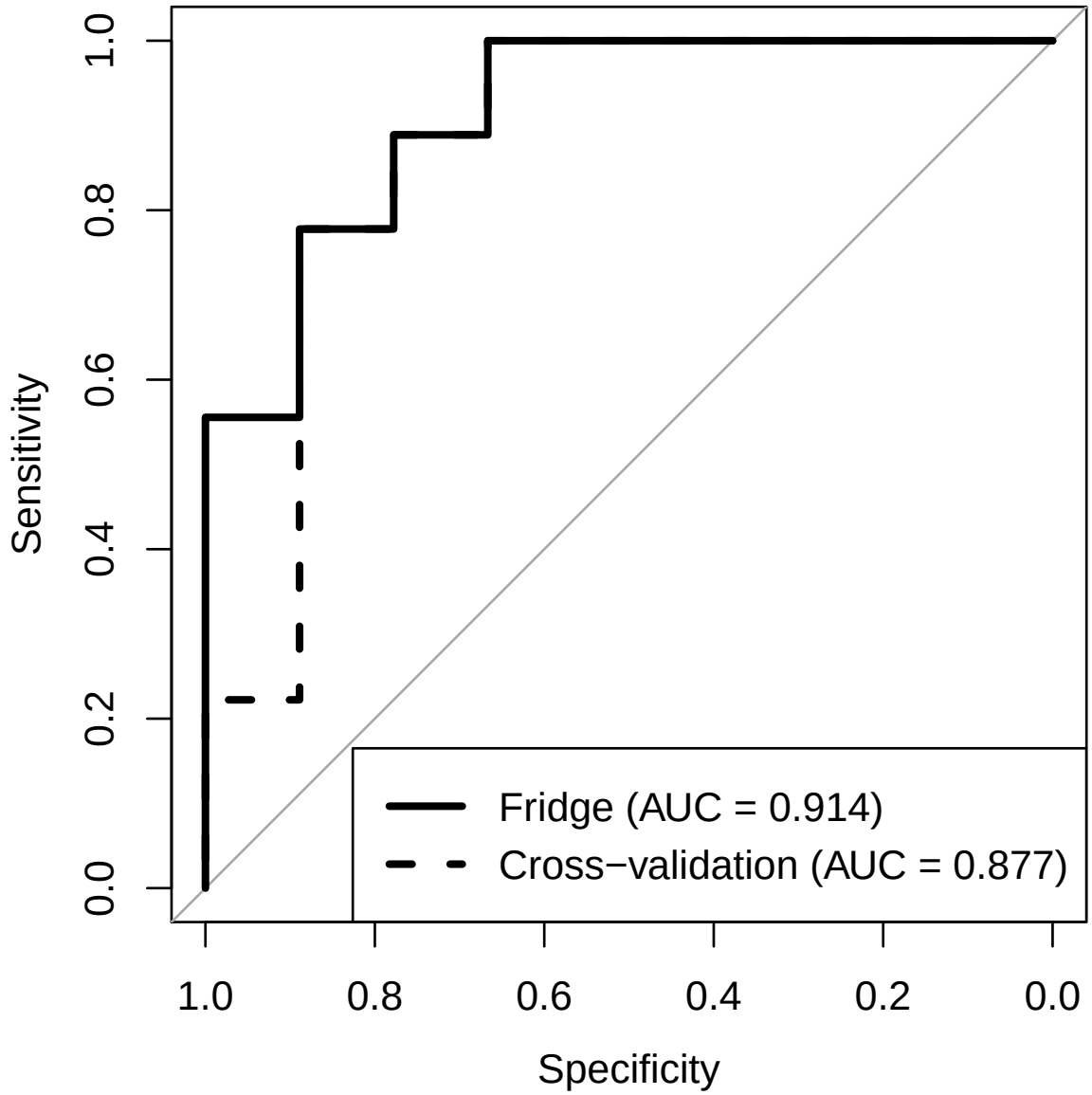
Fridge can be extended to logistic regression by using parametric bootstrap to estimate the squared bias and variance expressions.

Möckel et al. (2014) investigated whether gene expression can be used to predict the response of a specific treatment in glioma tumors, with the outcome being either a satisfactory response or no response. We have measurements

- of the top 3000 genes associated with the outcome,
- for 18 patient samples.

Each observation is predicted out-of-sample using logistic fridge and logistic ridge regression with standard LOOCV, as implemented by `glmnet`

The ROC curves for the two methods show that fridge gives a slightly higher Area Under the Curve (AUC) compared to cross-validation.



A ridge-bridge; personalized covariate selection

For a covariate of a specific patients

$$x_0 = (x_{0,1}, \ldots, x_{0,p}),$$

can an additional subset selection, $x_{0,S}$ be put on top of fridge?

We look for the optimal $x_{0,S}^T \beta_{S,\lambda}$; simultaneously searching for both the best selection of covariates and the best ridge tuning parameter.

Consider a subset $S \in \{1, \ldots, p\}$ where

$$\hat{\beta}_S(\lambda) = (X_S^T X_S + \lambda I_{|S|})^{-1} X_S^T Y,$$

then the bias and variance of the prediction can be estimated separately:

$$\text{MSE}(X_0, S, \lambda) = \text{Var}(X_0, S, \lambda) + \text{bias}^2(X_0, S, \lambda).$$

A ridge-bridge

We need to optimize over both parameters λ and S; for instance by

1. minimizing λ for each subset S to obtain an estimated risk,
2. then selecting the subset $\hat{S}$ with the lowest risk,
3. yielding the estimated tuning $\hat{\lambda}(\hat{S})$.

The subset-specific bias has a general expression using the projection matrix π_S:

$$\text{bias}(x_0, S, \lambda) = x_0^T A_S^T(\lambda)\beta, \qquad A_S(\lambda) = X^T X(X^T X + \lambda I_{|S|})^{-1}\pi_S - I.$$

Some issues: For large p, a search over all 2^p models is infeasible.
The connection to Lasso?

Hellton, K. H. and Hjort, N. L. (2017). Personalized predictions through focused fine-tuning of ridge regression. *Statistics in Medicine, in revision*.

The R package `fridge` is available at GitHub under `khellton/fridge`

Thank you!