Lecture notes mathematical data science

Frank van der Meulen*
Delft University of Technology

March 8, 2018

Contents

*with minor modifications by Joris Bierkens

1 Introduction

Suppose we have data $\{(x_i, y_i)\}_{i=1}^n$, where $x_i = (x_{i1}, \ldots, x_{ip})$ is a vector of predictors for outcome y_i. Important questions that statistical methods seek answers to include

- Can we find a relation between y and x? This is also known as *model building*.
- Suppose we obtain a new vector of predictors x. Can we predict the corresponding value of y? Furthermore, can we quantify the uncertainty in our prediction? *Prediction & uncertainty quantification.*
- Which components of x are important for predicting y. This concerns *variable importance*.
- Closely related to the previous point is the topic of *variable selection / variable screening*. This problem concerns choosing a subset of the p predictors that are most valuable in predicting y.
- What is the effect of *setting* a predictor x to a different value on y? This concerns *causality*.

In the remainder we will assume $\{(x_i, y_i)\}_{i=1}^n$ are realisations of independent identically distributed random vectors $\{(X_i, Y_i)\}_{i=1}^n$. In case y_i is of numerical type, the most popular model is the *linear model*, where one assumes that

$$Y_i \mid X_i = \mathbf{x}_i \sim \mathcal{N}(\mu(x_i), \sigma^2).$$

Here

$$\mu(x_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}.$$

This is known as a *regression* problem. In case y_i is binary, two popular models both start from the assumption

$$Y_i \mid X_i = x_i \sim \text{Ber}\,(p_i),$$

where

$$g(p_i) = \mu(x_i).$$

The case where $g(p) = \log(p/(1-p))$ and $g(p) = \Phi^{-1}(p)$ (the inverse of the cumulative distribution function of a $N(0,1)$ random variable) are called *logistic* and *probit* regression respectively. This is known as a *classification* problem.

Note that in both models we model the distribution of Y conditional on x.

Traditionally, the number of observations n is larger than the number of predictors p. Estimators have been studied under the working assumption that p is fixed and $n \to \infty$. Nowadays, datasets where $p \gg n$ are not uncommon. Two papers with applications are Bae and Mallick (2004) and Genkin et al. (2007). If p is large various problems arise, including

- the number of models explodes as it equals 2^p;

- collinearity among predictor variables;
- variable selection by point null hypothesis testing gets troublesome because *(i)* if n is large all variables will be significant at significance level 5%, *(ii)* the overall type I error blows up.

2 Penalised linear regression and sparsity

In this section we focus on the regression setting. Without loss of generality we will assume $\beta_0 = 0$. In matrix-vector notation the model can be written as

$$Y = X\beta + \varepsilon,$$

where ε is a random vector with a $N_n(0, \sigma^2 I_n)$ distribution modelling the noise [1]. The $n \times p$ matrix X is often called the *design matrix*. The parameter $\beta = (\beta_1, \ldots, \beta_p)$ is considered unknown.

2.1 Maximum likelihood estimation

The *likelihood* is given by

$$L(\beta \mid Y) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left(-\frac{\|Y - X\beta\|^2}{2\sigma^2}\right).$$

The Maximum Likelihood Estimator (MLE) $\hat{\beta}$ equals the Least Squares Estimator as

$$\hat{\beta} = \operatorname*{argmin}_{\beta} \|Y - X\beta\|^2. \tag{1}$$

Here $\|\cdot\|$ denotes the Euclidean norm (ℓ_2-norm). The solution can be characterised using the SVD of X : suppose $X = U\Sigma V'$ where U and V are $n \times n$ and $p \times p$ orthogonal matrices respectively and Σ is an $n \times p$ matrix containing the singular values on the diagonal. Denote $\operatorname{rank}(X) = r \leq \min(n, p)$.

Lemma 1. *If $\hat{\beta}$ is defined by (1), then*

$$\hat{\beta} = X^+ Y + z.$$

Here, the Moore-Penrose inverse X^+ is given by

$$X^+ = \sum_{k=1}^{r} \sigma_k^{-1} v_k u_k'$$

and $z \in N(X)$.

[1] In numerical analysis the notation is slightly different: $Ax = b + n$.

Proof. Writing

$$X = \sum_{k=1}^{r} \sigma_k u_k v_k', \qquad Y = \sum_{k=1}^{n} \langle u_k, Y \rangle u_k$$

we have

$$\begin{aligned}
\|Y - X\beta\|^2 &= \left\| \sum_{k=1}^{n} \langle u_k, Y \rangle u_k - \sum_{k=1}^{r} \sigma_k u_k \langle v_k, \beta \rangle \right\|^2 \\
&= \left\| \sum_{k=1}^{r} \left(\langle u_k, Y \rangle - \sigma_k \langle v_k, \beta \rangle \right) u_k + \sum_{k=r+1}^{n} \langle u_k, Y \rangle u_k \right\|^2 \\
&= \sum_{k=1}^{r} \left| \langle u_k, Y \rangle - \sigma_k \langle v_k, \beta \rangle \right|^2 + \sum_{k=r+1}^{n} \left| \langle u_k, Y \rangle \right|^2 .
\end{aligned}$$

The final equality follows from Pythagoras. This is minimal if

$$\langle u_k, Y \rangle = \sigma_k \langle v_k, \beta \rangle \qquad \forall\, k = 1, \ldots r$$

which characterises the optimal β. Using $\hat{\beta} = \sum_{k=1}^{p} \langle \hat{\beta}, v_k \rangle v_k$ we have

$$\hat{\beta} = \left(\sum_{k=1}^{r} \sigma_k^{-1} v_k u_k' \right) Y + \sum_{k=r+1}^{p} \alpha_k v_k.$$

for real numbers α_k. The result follows since $\mathrm{N}(X) = \mathrm{span}\,\{v_{r+1}, \ldots, v_p\}$. □

Though the MLE is not unique, the fitted values $X\hat{\beta}$ are. We can find a simple expression for the fitted values by writing the SVD of X as $X = UDV'$, with D being the $r \times r$ matrix containing the singular values of X, U denoting the $n \times r$ matrix with orthonormal columns spanning the column space of X and V an orthogonal $r \times p$ matrix with columns spanning the row space of X:

$$\mathrm{Col}(X) = \mathrm{span}\{u_1, \ldots, u_r\} \quad \text{and} \quad \mathrm{Row}(X) = \mathrm{span}\{v_1, \ldots, v_r\}.$$

Using this we obtain

$$X\hat{\beta} = UDV'VD^{-1}U'Y = UU'Y = \sum_{k=1}^{r} u_k u_k' Y. \tag{2}$$

This shows that the vector of fitted values is the projection of Y on $\mathrm{Col}(X)$.

Now assume X is of full column rank. Then $X^+ = (X'X)^{-1}X'$ and the unique MLE is given by

$$\hat{\beta} = (X'X)^{-1}X'Y.$$

It is easy to derive that [2] *(Exercise)*

$$\mathrm{E}\left[\hat{\beta}\right] = \beta \quad \text{and} \quad \mathrm{Cov}\hat{\beta} = \sigma^2 \left(X'X\right)^{-1}.$$

The first result shows that the MLE is unbiased. However, in case $X'X$ is close to singularity, the variance becomes large. To get further insight, recall that the Mean Squared Error (MSE) of $\hat{\beta}$ for estimating β is defined by[3]

$$\mathrm{MSE}(\hat{\beta};\beta) = \mathrm{E}\left[\left\|\hat{\beta}-\beta\right\|^2\right] = \sum_{i=1}^{p}\left(\mathrm{E}\left[\hat{\beta}_i-\beta_i\right]\right)^2 + \sum_{i=1}^{p}\mathrm{Var}\left(\hat{\beta}_i\right).$$

From the SVD we know that $X'X$ has p strictly positive eigenvalues $\lambda_k = \sigma_k^2$. Therefore, $(X'X)^{-1}$ has eigenvalues $1/\sigma_k^2$. This implies

$$\sum_{i=1}^{p}\mathrm{Var}\left(\hat{\beta}_i\right) = \mathrm{tr}\left(\mathrm{Cov}\hat{\beta}\right) = \mathrm{tr}\left(\sigma^2(X'X)^{-1}\right) = \sigma^2\sum_{k=1}^{p}\frac{1}{\sigma_k^2}$$

from which we see that the variance of the MLE becomes large if one of the singular values of X is close to zero.

2.2 Penalisation by ℓ_2-penalty: Ridge regression

The main idea behind *ridge regression* is to allow for some bias while reducing the variance. The ridge regression estimator is defined by minimising

$$\|Y - X\beta\|^2 \quad \text{subject to} \quad \|\beta\|^2 \le t.$$

By restricting β to lie in an ℓ_2-ball of radius t the variance is reduced (what would be the variance of the estimator in case $t = 0$?). Using KKT optimality theory, this minimisation problem can be shown to be equivalent to minimising

$$\|Y - X\beta\|^2 + \lambda\|\beta\|^2.$$

The parameter λ relates in a bijective way to t and is referred to as a *regularisation parameter*. Denoting the minimiser of the preceding display by $\hat{\beta}_r$ we have (for any $\lambda > 0$)

$$\hat{\beta}_r = (X'X + \lambda I)^{-1}X'Y. \tag{3}$$ *(Exercise)*

Note that the inverse appearing always exists if $\lambda > 0$, as $X'X$ is positive semi-definite. In numerical analysis estimating β in this way is known as *Tikhonov regularisation*.

[2] For a random vector $Z = (Z_1, \ldots, Z_n)$ its mean vector is defined by $\mathrm{E}[Z] = (\mathrm{E}[Z_1], \ldots, \mathrm{E}[Z_n])$ and its the covariance matrix is defined by $\mathrm{Cov}(Z) = \mathrm{E}[(Z - \mathrm{E}[Z])(Z - \mathrm{E}[Z])']$. It is easy to show that $\mathrm{Cov}(AZ) = A\mathrm{Cov}(Z)A'$.

[3] Here, we have used that for a random vector Z in $\mathbb{R}^n$ we have $\mathrm{E}\left[\|Z\|^2\right] = \mathrm{tr}(\mathrm{Cov}(Z)) + \mathrm{E}[Z']\,\mathrm{E}[Z]$.

It is instructive to compare $\hat{\beta}$ and $\hat{\beta}_r$. Comparing the estimators is most easily done using the SVD: $X = U\Sigma V'$, with Σ the $n \times p$ matrix containing the singular values. The fitted values satisfy

$$X\hat{\beta}_r = U\Sigma(\Sigma'\Sigma + \lambda I)^{-1}\Sigma' U'Y = \sum_{k=1}^{r} u_k \frac{\sigma_k^2}{\sigma_k^2 + \lambda} u_k' Y.$$

This is to be compared with (2). Hence, the ridge estimator first computes the coordinates of Y with respect to the orthonormal basis U. It then shrinks these coordinates by the factors $\sigma_k^2/(\sigma_k^2+\lambda)$. For this reason, the ridge regression estimator is an example of a *shrinkage estimator*. If the j-th predictor hardly contributes the j-th component of β will be near zero, but not exactly zero. In this sense, no variable selection is performed. The lasso is a clever variation of ridge regression that includes a variable selection property.

2.3 Penalisation by ℓ_1-penalty: the Lasso

Lasso is an abbreviation for Least Absolute Shrinkage and Selection Operator. It is defined similarly as ridge regression, replacing the ℓ_2 constraint by an ℓ_1 constraint. The lasso estimator was proposed by Tibshirani (1996)[4] and is defined by minimising

$$\|Y - X\beta\|^2 \quad \text{subject to} \quad \|\beta\|_1 \leq t.$$

which is equivalent to minimising

$$\|Y - X\beta\|^2 + \lambda\|\beta\|_1.$$

Just as in ridge regression, the parameter λ determines the bias-variance trade-off of the resulting estimator. The ℓ_1-norm has a very special property: constraining the coefficients by their ℓ_1-norm induces *sparsity* in the estimate. By sparsity we mean that the lasso solution $\hat{\beta}$ will have $\hat{\beta}_j = 0$ for many components $j \in \{1, \ldots, p\}$. The smaller the value of t (the larger value of λ), the more exactly zero coefficients in the solution. Note that this is not the case for ridge regression. If $\hat{\beta}_j = 0$, then the j-th predictor is not included in the model. Therefore, the ℓ_1-norm performs variable selection automatically, which aids in interpretability of the model. There is a "classic" picture from the book of Hastie et al. (2014) that illustrates why some coefficients are set to zero exactly. It is shown in figure 2.3. Here $\hat{\beta}$ is the MLE and the red ellipses are contours for the residual sum of squares $\|Y - X\beta\|$, centred at the MLE. For the lasso, the contours cross the light blue area in the upper corner which implies that the lasso solution will have its first component equal to zero.

You may wonder why we do not use the ℓ_0-"norm": find β to minimise

$$\|Y - X\beta\|^2 + \lambda\|\beta\|_0,$$

[4] In the signal processing literature, the lasso is also known as basis pursuit (Cf. Chen et al. (1998).

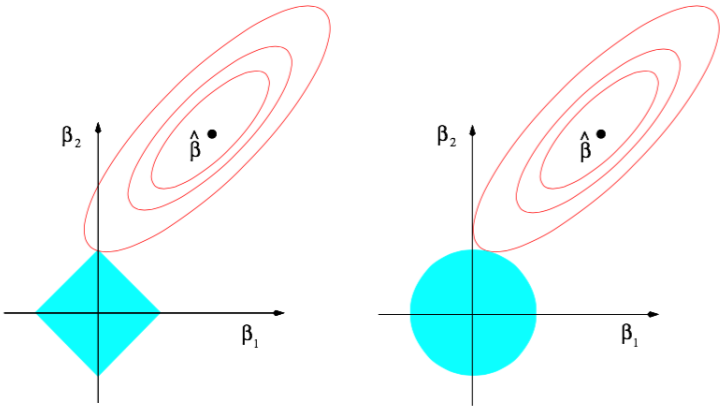

FIGURE 3.11. *Estimation picture for the lasso (left) and ridge regression (right). Shown are contours of the error and constraint functions. The solid blue areas are the constraint regions* $|\beta_1| + |\beta_2| \le t$ *and* $\beta_1^2 + \beta_2^2 \le t^2$*, respectively, while the red ellipses are the contours of the least squares error function.*

Figure 1: Illustration for ridge and lasso regression.

where $\|\beta\|_0 = \sum_{j=1}^p \mathbf{1}\{\beta_j \neq 0\}$.[5] This is indeed more natural for variable selection as we directly penalise for a large number of nonzero coefficients. The difficulty is that the resulting optimisation problem is non-convex and known to be NP-hard. Considering unit ℓ_p-balls for $p \geq 0$, the case $p = 1$ is the smallest value for which the unit ball is convex.

2.4 Existence and uniqueness (optional reading)

We now turn to existence and uniqueness of the lasso. Assume $\lambda > 0$. By the KKT optimality conditions any lasso solution must satisfy

$$2X'(Y - X\hat{\beta}) = \lambda s \tag{4}$$

where s is a subgradient of the ℓ_1 norm evaluated at $\hat{\beta}$.[6] This means that the subdifferential of $\|\beta\|_1$ is the set $[-1, 1]$. So

$$\begin{cases} s_j = 1 & \text{if} \quad \hat{\beta}_j > 0 \\ s_j = -1 & \text{if} \quad \hat{\beta}_j < 0 \\ s_j \in [-1, 1] & \text{if} \quad \hat{\beta}_j = 0 \end{cases}.$$

For any subset $A \subset \{1, \ldots, p\}$ we have

$$X_A'(Y - X\hat{\beta}) = \lambda s_A$$

[5] The ℓ_0-"norm" is not a norm as we can have $\|\alpha x\|_0 \neq |\alpha|\|x\|_0$ for $\alpha \in \mathbb{R}$. Take for example $\alpha = 2$ and $x = (1, 1)$. Many authors abuse terminology by omitting the quotation marks.

[6] For a vector valued function, we say that g is a subgradient of f at x_0 if for all vector x: $f(x) - f(x_0) \geq (x - x_0)'g$. The set of all subgradients of f at x_0 is called the subdifferential of f at x_0. A point x_0 is a global minimum of a convex function f if and only if zero is contained in the subdifferential.

where X_A denotes the matrix containing all columns of X that are in the set A and s_A denotes the vector obtained from s after removing elements not in A. Now let

$$E = \{j \in \{1, \ldots, p\} \ : \ |s_j| = 1\}.$$

Applying the previous display with $A = E$ we obtain

$$X_E'(Y - X\hat{\beta}) = \lambda s_E.$$

If $j \notin E$, then $\hat{\beta}_j = 0$ which implies $X\hat{\beta} = X_E\hat{\beta}_E$. Hence

$$\begin{cases} X_E'(Y - X_E\hat{\beta}_E) = \lambda s_E \\ \hat{\beta}_{-E} = 0 \end{cases}.$$

Reordering terms in the first equation gives

$$X_E'X_E\hat{\beta}_E = X_E'Y - \lambda s_E,$$

from which we obtain that

$$\hat{\beta}_E = (X_E'X_E)^+ \left(X_E'Y - \lambda s_E\right) + z.$$

Here $z \in \mathrm{N}(X_E'X_E) = \mathrm{N}(X_E)$. So there is a unique lasso solution if X_E is of full column rank. This trivially happens when X has independent columns (in the $p < n$ setting). However, it turns out that a unique solution exists under much weaker conditions, see reference Tibshirani (2013).

The fitted values for the lasso are unique. From equation (4) it follows that s is uniquely determined by $X\hat{\beta}$ and hence unique itself. Suppose the lasso is not unique and denote two minimisers of the criterion function by $\hat{\beta}^{(1)}$ and $\hat{\beta}^{(2)}$. As s is unique we cannot have $\hat{\beta}_j^{(1)} > 0$ and $\hat{\beta}^{(2)} < 0$. This is a very beneficial for interpretation of the coefficients of the (a) lasso estimator.

2.5 Soft thresholding

A better understanding of ridge regression and the lasso can be obtained by studying the simple case where $X = I$. In that case it is easy to see that $\hat{\beta}_j = Y_j/(1 + \lambda)$. For the lasso, taking the subgradient with respect to β_j of $\beta \mapsto \|Y - \beta\|^2 + \lambda\|\beta\|_1$ gives

$$-2Y_j + 2\beta_j + \begin{cases} -\lambda & \text{if} \quad \beta_j < 0 \\ [-\lambda, \lambda] & \text{if} \quad \beta_j = 0 \\ \lambda & \text{if} \quad \beta_j > 0 \end{cases}.$$

This is solved for

$$\hat{\beta}_j = \mathrm{soft}(Y_j, \lambda/2),$$

where

$$\text{soft}(y,a) = \begin{cases} y+a & \text{if} \quad y < -a \\ 0 & \text{if} \quad -a \le y \le a \\ y-a & \text{if} \quad y > a \end{cases}$$

is called the *soft-thresholding* function. Make a sketch yourself to see the difference in shrinkage of ridge regression and the lasso. *(Exercise)*

2.6 Choosing the regularisation parameter λ

For each $\lambda > 0$, both the Lasso and ridge regression provide an estimator for β (which in turn determines the regression function). How should we choose λ? To get an understanding of this, we review the bias-variance trade off. The following is written down in a bit more generality, as the choice of regularisation parameter is important in many other estimation methods as well.

2.6.1 The bias-variance trade off

Suppose we observe (X, Y) from an unknown joint distribution and we aim to predict Y from X. [7] Suppose $f(X)$ is used to predict Y. The prediction error for f is defined by

$$PE(f) := \mathrm{E}\left[(Y - f(X))^2\right],$$

where the expectation is taken over (X, Y). The minimiser of PE is given by $f(X) = \mathrm{E}[Y \mid X]$ and this is called the true regression function. Then we get write $Y = f(X) + (Y - f(X))$ and by setting $\varepsilon = Y - f(X)$ we have

$$Y = f(X) + \varepsilon.$$

Here, $\mathrm{E}[\varepsilon] = \mathrm{E}[Y - \mathrm{E}[Y - X]] = 0$.

In a data setting, the joint distribution of (X, Y) is unknown, and hence f is unknown. We do however have data $D := \{(X_i, Y_i)\}_{i=1}^n$ to learn about this joint distribution. Suppose $\hat{f}$ is an estimator for f. That is, $\hat{f}$ is a function of D which we hope is "close" to the true regression function f. Define the prediction error (sometimes also called test error) for $\hat{f}$ is given by

$$PE(\hat{f}) = \mathrm{E}\left[(Y - \hat{f}(X))^2\right]$$

where the expectation is taken over (X, Y, D). The prediction error can be decomposed. We do this first by conditioning on $X = x$. If we define $\sigma^2 = \mathrm{Var}(\varepsilon)$, then

$$\begin{aligned} \mathrm{E}\left[(Y - \hat{f}(x))^2 \mid X = x\right] &= \sigma^2 + \mathrm{E}\left[(f(x) - \hat{f}(x))^2 \mid X = x\right] \\ &= \sigma^2 + \left(\hat{f}(x) - \mathrm{E}\left[\hat{f}(x)\right]\right)^2 + \mathrm{Var}\left(\hat{f}(x)\right). \end{aligned}$$

[7] X is a random vector here. Please do not confuse this with the design matrix, for which it is common to use the symbol X as well

For the first equality, add and subtract $f(x)$; for the second equality, add and subtract $\mathrm{E}\left[\hat{f}(x)\right]$. The first term is irreducible error, the 2nd and 3rd term are the squared bias and variance respectively. The decomposition itself is called the bias variance trade off. Unconditionally over X, the bias-variance trade off can be written as

$$PE(\hat{f}) = \mathrm{E}\left[(Y - \hat{f}(X))^2\right] = \sigma^2 + \int \left(\hat{f}(x) - \mathrm{E}\left[\hat{f}(x)\right]\right)^2 P_X(dx) + \int \mathrm{Var}\left(\hat{f}(x)\right) P_X(dx),$$

where we denote the distribution of X by P_X. Note that this quantity is unknown to us, but we can try to estimate it. In section 2.6.2 we will discuss cross-validation to estimate $PE(\hat{f})$ (there exist alternatives ways, but this is a commonly used method).

Keep in mind that for ridge regression and the Lasso, the estimator is of a specific form:

$$\hat{f}_\lambda(x) = \beta'_\lambda x.$$

Once, we have an estimator of $PE(\hat{f}_\lambda)$, say $\widehat{PE}(\hat{f}_\lambda)$, we can choose λ to minimise $\lambda \mapsto \widehat{PE}(\hat{f}_\lambda)$.

2.6.2 Cross validation

A first estimator for $PE(\hat{f})$ that comes to mind is the Average Observed Error

$$AOE(\hat{f}) := \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{f}(X_i)^2.$$

However, this is a bad estimator: it is too optimistic in the sense hat it will underestimate $PE(\hat{f})$. Intuitively, we use the same data to derive $\hat{f}$ and to evaluate its performance. If we would have no restrictions on the form of $\hat{f}$, then we could simply define $\hat{f}(X_i) = Y_i$ to obtain $AOE(\hat{f}) = 0$! Obviously, this is an example of overfitting: one should evaluate the predictive performance on a different data set than the data that were used to find $\hat{f}$.

To deal with this problem, a popular method for setting the regularisation parameter is *cross-validation*. Choose a positive integer K. For K-fold cross-validation we split the data into K roughly equal-sized parts. One of these parts is used to assess the performance of the model (these are the "test data"), the remaining parts are used for fitting the model (these are the "training data"). In case $K = 4$ we schematically have

$$\boxed{Train}\ \boxed{Train}\ \boxed{Train}\ \boxed{Test}.$$

Using all training data we determine $\hat{f}_\lambda$. Next, we approximate $PE(\hat{f})$ using the estimator $AOE(\hat{f})$ with the test data. This is repeated for

$$\boxed{Train}\ \boxed{Train}\ \boxed{Test}\ \boxed{Train}.$$

$$\boxed{Train}\ \boxed{Test}\ \boxed{Train}\ \boxed{Train}.$$

$$\boxed{Test} \quad \boxed{Train} \quad \boxed{Train} \quad \boxed{Train}.$$

Following section 7.10 of Hastie et al. (2014) we can write this down more mathematically as follows: let $\kappa : \{1, \ldots, n\} \mapsto \{1, \ldots, K\}$ be an indexing function that indicates the partition to which observation i is allocated by the randomisation. Denote by $\hat{f}_\lambda^{-k}$ the fitted function, computed with the k-th part of the data removed. The Cross-Validation (CV) estimate of the prediction error $PE(\hat{f})$ is given by

$$\mathrm{CV}(\lambda) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{f}_\lambda^{-\kappa(i)}(X_i))^2.$$

For balancing the squared bias term and variance term, this quantity is minimised with respect to λ. In practise one usually chooses $K = 10$ as a pragmatic choice, although there is no guarantee that this is optimal. This means that we fit the model 10 times using 90% of the data. The special case of $K = n$ is known as *leave one out cross-validation*.

2.7 Computing the lasso for all values of λ: the Lasso path algorithm

Using convex optimisation it is possible to compute the lasso solution for a fixed value of λ. The collection of solutions obtained by varying $\lambda \in [0, \infty)$ is called the full solution path. There exists a fast algorithm for computing this path known as the LARS algorithm Efron et al. (2004). The `glmnet` library in R contains an implementation of this algorithm.

2.8 Theoretical justification for the lasso

Over the past 10 years an enormous amount of research on properties of the lasso has been carried out. We cannot cover all results and will only give an illustration using a result derived and discussed in a post by prof. Wasserman, see link to post. We consider a small part of the lecture notes that can be found on

`http://www.stat.cmu.edu/~larry/=sml/Assumptions.pdf`.

In this section, we assume that the regularisation parameter is fixed.

We consider estimators of the form $f_\beta(X) = \beta' X$ (both ridge regression and the lasso give estimators of this form). Such estimators are called linear estimators. Consider the setting of subsection **??**. Define

$$R(\beta) = PE(f_\beta) = \mathrm{E}\left[(Y - \beta' X)^2\right].$$

Fix $L > 0$. Let

$$\beta_* = \operatorname*{argmin}_\beta R(\beta) \quad \text{subject to} \quad \|\beta\|_1 \leq L.$$

Clearly, $f_{\beta_*}(X)$ is the optimal linear sparse regression function. Note that L corresponds bijectively to to the fixed value of the regularisation parameter λ. Define

$$R_n(\beta) = AOE(f_\beta) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \beta' X_i)^2.$$

The lasso solves

$$\hat{\beta} = \operatorname*{argmin}_{\beta} R_n(\beta) \quad \text{subject to} \quad \|\beta\|_1 \leq L.$$

Theorem 1. *Assume "all" random variables (i.e. Y_i, X_{ij}) are almost surely bounded by C. For any $\delta > 0$*

$$\mathrm{P}\left(R(\hat{\beta}) - R(\beta_*) \leq 8L^2C^2\sqrt{\frac{2}{n}\log\left(\frac{8p^2}{\delta}\right)} \right) \geq 1 - \delta.$$

For fixed L, this result shows that

$$R(\hat{\beta}) - R(\beta_*) = O_P\left(\sqrt{\frac{\log p}{n}}\right),$$

and without additional assumptions on the regression function this rate is optimal. In words: the predictive error of the Lasso comes close to predictive error of the best sparse linear predictor.

Proof. Define $Z = (Y, X)$ and

$$\gamma = (-1, \beta_1, \ldots, \beta_p)$$

so that $\beta' X - Y = \gamma' Z$. Similarly define $Z^i = (Y^i, X^i)$ so that $\beta' X^i - Y^i = \gamma' Z^i$ (we wish to reserve subscripts in the notation to denote elements in a vector or matrix). As $\mathrm{E}[\gamma' Z] = 0$

$$R(\beta) = \mathrm{Var}\big(\gamma' Z\big) = \gamma' \Sigma \gamma \quad \text{with} \quad \Sigma = \mathrm{Cov}(Z).$$

Note that $\Sigma_{k,\ell} = \mathrm{E}[Z_k Z_\ell]$. Also

$$R_n(\beta) = \frac{1}{n}\sum_{i=1}^{n}\big(\gamma' Z_i\big)^2 = \gamma' \hat{\Sigma} \gamma \quad \text{with} \quad \hat{\Sigma}_{k,\ell} = \frac{1}{n}\sum_{i=1}^{n} Z_k^i Z_\ell^i.$$

So if $\|\beta\|_1 \leq L$ then

$$\begin{aligned} |R_n(\beta) - R(\beta)| = |\gamma'(\hat{\Sigma} - \Sigma)\gamma| &\leq \sum_{k=1}^{p+1}\sum_{\ell=1}^{p+1} |\gamma_k||\gamma_\ell||\hat{\Sigma}_{k,\ell} - \Sigma_{k,\ell}| \\ &\leq (L+1)^2 \Delta_n \leq 4L^2 \Delta_n \end{aligned}$$

where

$$\Delta_n = \max_{k,\ell} |\hat{\Sigma}_{k,\ell} - \Sigma_{k,\ell}|.$$

By Hoeffding's inequality: [8]

$$\begin{aligned} \mathrm{P}(\Delta_n \geq \varepsilon) &\leq (p+1)^2 \mathrm{P}\Big(|\hat{\Sigma}_{k,\ell} - \Sigma_{k,\ell}| \geq \varepsilon\Big) \\ &\leq 2(p+1)^2 \exp\left(-\frac{n\varepsilon^2}{2C^4}\right) \\ &\leq 8p^2 \exp\left(-\frac{n\varepsilon^2}{2C^4}\right). \end{aligned}$$

Take $\varepsilon = \varepsilon_n$ with

$$\varepsilon_n = \sqrt{\frac{2C^4}{n} \log\left(\frac{8p^2}{\delta}\right)}.$$

Now $\Delta_n \leq \varepsilon_n$ with probability at least $1 - \delta$ and

$$\sup_{\beta\,:\,\|\beta\|_1 \leq L} |R_n(\beta) - R(\beta)| \leq 4L^2 \varepsilon_n.$$

The results follows from the inequalities

$$\begin{aligned} R(\beta_*) &\leq R(\hat{\beta}) \\ &\leq |R(\hat{\beta}) - R_n(\hat{\beta})| + R_n(\hat{\beta}) \\ &\leq 4L^2\varepsilon_n + R_n(\beta_*) \\ &\leq 4L^2\varepsilon_n + |R_n(\beta_*) - R(\beta_*)| + R(\beta_*) \\ &\leq 8L^2\varepsilon_n + R(\beta_*). \end{aligned}$$

□

3 The Bayesian approach

3.1 Bayesian analogues of ridge regression and the lasso

In this section we use the following notation: the density of a stochastic vector X is denoted by f_X. When evaluated at x, we denote this by $f_X(x)$. At various places we use "Bayesian notation" as well : all densities are then denoted by p and the argument itself denotes both the random variable, as the value at which the density is evaluated. So $p(x) = f_X(x)$. [9]

[8] Hoeffding's inequality: suppose $X_1, \ldots, X_n$ are independent random variables with $X_i \in [-c, c]$, then for $\varepsilon > 0$

$$\mathrm{P}\big(|\bar{X}_n - \mathrm{E}\left[\bar{X}_n\right]| \geq \varepsilon\big) \leq 2\exp\left(-\frac{n\varepsilon^2}{2c^2}\right).$$

[9] Be careful: $p(x^2)$ denotes the density of X^2 evaluated in x^2, so $f_{X^2}(x^2)$. It does NOT denote $f_X(x^2)$.

Both ridge regression and the lasso have a Bayesian interpretation. If we assume σ^2 is known and fixed and moreover that $\beta \sim \mathcal{N}(0, \sigma^2\lambda^{-1}I)$ a priori, then the posterior density of β satisfies

$$p(\beta \mid D) \propto \exp\left(-\frac{1}{2\sigma^2}\|Y - X\beta\|^2 - \frac{\lambda}{2\sigma^2}\|\beta\|^2\right).$$

Here $\propto$ denotes "proportional to" and D denotes all data. The Maximum A Posterior estimator (MAP) is defined as the maximiser of the posterior density. It is easy to see that the MAP equals the ridge regression estimator. In a similar manner, if we assume $\beta \sim \mathcal{DE}(\lambda/\sigma)$ (β has the Double Exponential distribution (also known as the Laplace distribution)), meaning

$$p(\beta) = \frac{\lambda}{2\sigma}\exp\left(-\frac{\lambda}{\sigma}\|\beta\|_1\right),$$

then the MAP estimator is the lasso. A major advantage of Bayesian methods is that we can more easily assess uncertainty in parameter estimates via the posterior distribution. Various point estimators can be derived from the posterior distribution including the posterior mean and median.[10]

The value of λ is unknown and within the Bayesian paradigm it is natural to endow this parameter with a prior distribution itself. In this way we get a *hierarchical model*. For Bayesian ridge regression this takes the form:

$$\begin{aligned} Y \mid \beta, \sigma^2 &\sim \mathcal{N}(X\beta, \sigma^2 I) \\ \beta \mid \lambda, \sigma^2 &\sim \mathcal{N}(0, \sigma^2\lambda^{-1}I) \\ \sigma^2 &\sim p(\sigma^2) \\ \lambda &\sim p(\lambda). \end{aligned}$$

For the Bayesian lasso this takes the form

$$\begin{aligned} Y \mid \beta, \sigma^2 &\sim \mathcal{N}(X\beta, \sigma^2 I) \\ \beta \mid \tau^2, \sigma^2 &\sim \mathcal{N}(0, \sigma^2\tau^2 I) \\ \tau^2 \mid \lambda &\sim \mathcal{E}(\lambda^2/2) \\ \sigma^2 &\sim p(\sigma^2) \\ \lambda &\sim p(\lambda). \end{aligned}$$

Here we have used (at the second and third line of this specification) that the double exponential distribution (more specifically, the $\mathcal{DE}(\lambda/\sigma)$ distribution) is a mixture of a normal distribution with an exponential mixing density:

$$\frac{\lambda}{2\sigma}\exp\left(-\frac{\lambda}{\sigma}|z|\right) = \int_0^\infty \frac{1}{\sqrt{2\pi u\sigma^2}} e^{-\frac{z^2}{2u\sigma^2}} \frac{\lambda^2}{2} e^{-\frac{\lambda^2}{2}u} \mathrm{d}u, \qquad z \in \mathbb{R}. \tag{*(Exercise; hint: mgf)*}$$

[10] Bayesian point estimators are obtained by minimising the *expected posterior loss* which is defined by $\int L(\theta, d(x))p(\theta \mid x)d\theta$. Here $d(x)$ is an estimator for θ, $\pi(\theta \mid x)$ denotes the posterior density and L is a *loss function*. If $L(\theta, a) = (\theta - a)^2$ we get the posterior mean, if $L(\theta, a) = |\theta - a|$ we get the posterior median. If $L(\theta, a) = \mathbf{1}\{|\theta - a| \geq c\}$, the sequence of estimators approaches the MAP estimator, as $c \to 0$.

In both hierarchical schemes we have yet left the priors for both λ and σ^2 unspecified. The parameters appearing in their distributions are referred to as *hyperparameters*. Usually, these hyperparameters are chosen on two grounds:

(i) such that the variance of λ and σ^2 is high (reducing influence on the posterior of β);

(ii) such that the computations for drawing from the posterior distribution simplify. The main vehicle here is the *Gibbs sampler* and it is *partial conjugacy* that simplifies computations. More details are in section 3.3.

In the `monomvn` package in `R`, the following parametrisation is considered (See section 3.1 in Gramacy and Pantaleo (2010), we consider the case of a model without intercept)

$$\begin{aligned} Y \mid \beta, \sigma^2 &\sim \mathcal{N}(X\beta, \sigma^2 I) \\ \beta \mid \sigma^2, \tau_1^2, \ldots, \tau_p^2 &\sim \mathcal{N}(0, \sigma^2 D), \quad D = \operatorname{diag}(\tau_1^2, \ldots, \tau_p^2) \\ \tau_j^2 \mid \lambda &\overset{\text{iid}}{\sim} \mathcal{E}(\lambda^2/2) \\ \sigma^2 &\sim \mathcal{IG}(a_\sigma/2, b_\sigma/2) \\ \lambda^2 &\sim \mathcal{G}(a_\lambda, b_\lambda). \end{aligned} \tag{5}$$

$\mathcal{IG}$ and $\mathcal{G}$ are the rate- and scale-parametrised inverse gamma and gamma distribution, respectively. Note that this setup includes ridge regression by taking $\tau^2 = \tau_1^2 = \cdots = \tau_p^2$, $\tau^2 \sim \mathcal{IG}(a_\tau/2, b_\tau/2)$ and dropping λ^2. Furthermore, this setup includes the desirable property that each component of β has its own regularisation parameter, something which is completely infeasible when using cross-validation. *(Exercise)*

Whereas the lasso yields sparsity, the posterior for β will usually not be sparse, and neither the posterior mean nor the posterior median will be. From a Bayesian perspective however, to many the latter two point estimators are more natural than the posterior mode.

3.2 Other shrinkage estimators

Over the past 10 years many more shrinkage estimators for estimating sparse signals have been proposed. Here we consider one example: the *Horseshoe estimator* by Carvalho et al. (2010).

$$\begin{aligned} Y \mid \beta, \sigma^2 &\sim \mathcal{N}(X\beta, \sigma^2 I) \\ \beta \mid \sigma^2, \tau_1^2, \ldots, \tau_p^2 &\sim \mathcal{N}(0, \sigma^2 D), \quad D = \operatorname{diag}(\tau_1^2, \ldots, \tau_p^2) \\ \tau_j &\overset{\text{iid}}{\sim} \mathcal{C}^+(0, 1) \\ \sigma^2 &\sim p(\sigma^2). \end{aligned} \tag{6}$$

Here, $C^+(0,1)$ denotes the standard half-Cauchy distribution: $X \sim \mathcal{C}^+(0,1)$ if the density $f_X(x) = (2/\pi)(1 + x^2)^{-1}$. The motivation for this choice and the name of the prior

will be part of the exercises. As an appetiser: if $\beta \mid \tau^2 \sim \mathcal{N}(0, \tau^2)$ and $\tau \sim \mathcal{C}^+(0, 1)$, then it can be shown that for $\beta \neq 0$

$$\frac{K}{2} \log\left(1 + \frac{4}{\beta^2}\right) \leq p(\beta) \leq K \log\left(1 + \frac{2}{\beta^2}\right),$$

where $K = 1/\sqrt{2\pi^3}$. As the tails decrease to zero slowly, large values of β are expected a priori. At the same time, the density tends to infinity near zero and the mass near zero models the expectation that $\beta \approx 0$.

3.3 Computational methods

For complex models it is usually impossible to draw directly from the posterior. Fortunately, there exists a versatile class of randomised algorithms known as *Markov Chain Monte Carlo* methods. Probably the best known algorithm is the Gibbs sampler, see for instance section 3.7 of Young and Smith (2010). In a simple setting this algorithm is easily explained: suppose we wish to sample from a bivariate (intractable) density $f_{(X,Y)}(x, y)$, which we know up to a proportionality constant. Now assume we know how to sample from the conditional distributions $f_{X|Y}(x \mid y)$ and $f_{Y|X}(y \mid x)$. The Gibbs sampler is an iterative algorithm:

(i) Set $i = 1$ and initialise x^0.

(ii) Draw y^i from $f_{Y|X}(\cdot \mid x^{i-1})$.

(iii) Draw x^i from $f_{X|Y}(\cdot \mid y^i)$.

(iv) Set $i := i + 1$ and return to step (ii).

Under weak conditions, the Markov chain $\{(x^i, y^i)\}_i$ will converge to its invariant distribution. To get a bit more feeling for this, the Markov kernel for moving from y to y' is given by

$$K(y' \mid y) = \int f_{X|Y}(x \mid y) f_{Y|X}(y' \mid x) dx.$$

Note that $\int K(y' \mid y) dy' = 1$. The product $K(y' \mid y) f(y)$ is symmetric in (y, y'):

$$\begin{aligned} K(y' \mid y) f_Y(y) &= \int f_{X|Y}(x \mid y) f_{Y|X}(y' \mid x) f_Y(y) dy \\ &= \int f_{(X,Y)}(x, y) \frac{f_{(X,Y)}(x, y')}{f_X(x)} dy. \end{aligned}$$

This implies the Markov chain on y is reversible with respect of $f(y)$:

$$K(y' \mid y) f_Y(y) = K(y \mid y') f_Y(y').$$

This relation is also referred to as *detailed balance*. Integrating over y reveals that $f(y)$ is invariant for K:

$$\int K(y' \mid y) f_Y(y) dy = f_Y(y').$$

Hence, y and y' have the same marginal distribution

The above Gibbs sampler is a two-component Gibbs sampler. The sampler easily generalises to the case with n components and can be summarised by "iteratively sampling from the full conditionals of the target density". Whereas the above derivations only concern the chain for y, we can also show that $f_{(X,Y)}$ is invariant for the chain on (x, y).

Lemma 2. *If $(X, Y) \sim f_{(X,Y)}$ and (X', Y') are generated by one cycle of the Gibbs sampler, then $(X', Y') \sim f_{(X,Y)}$.*

Proof. Note that the Markov kernel for the Gibbs sampler is given by

$$L(x', y' \mid x, y) = f_{X|Y}(x' \mid y) f_{Y|X}(y' \mid x').$$

Hence for a measurable set A we have

$$\begin{aligned}
\mathrm{P}\big((X', Y') \in A\big) &= \int \mathbf{1}_A((x', y')) L(x', y' \mid x, y) f_{(X,Y)}(x, y) dx dy dx' dy' \\
&= \int \mathbf{1}_A((x', y')) f_{X|Y}(x' \mid y) f_{Y|X}(y' \mid x') f_{(X,Y)}(x, y) dx dy dx' dy' \\
&= \int \mathbf{1}_A((x', y')) f_{X|Y}(x' \mid y) f_{Y|X}(y' \mid x') f_Y(y) dy dx' dy' \\
&= \int \mathbf{1}_A((x', y')) f_{(X,Y)}(x', y) f_{Y|X}(y' \mid x') dy dx' dy' \\
&= \int \mathbf{1}_A((x', y')) f_X(x') f_{Y|X}(y' \mid x') dx' dy' \\
&= \int \mathbf{1}_A((x', y')) f_{(X,Y))}(x', y') dx' dy',
\end{aligned}$$

showing that $f_{(X,Y)}$ is invariant. □

We illustrate the Gibbs sampler for a simplified version of the ridge regression problem

$$\begin{aligned}
Y \mid \beta &\sim \mathcal{N}(X\beta, I) \\
\beta \mid \lambda &\sim \mathcal{N}(0, \lambda I) \\
\lambda &\sim \mathcal{IG}(a/2, b/2).
\end{aligned}$$

Now

$$\begin{aligned}
p(Y, \beta, \lambda) &\propto p(Y \mid \beta) p(\beta \mid \lambda) p(\lambda) \\
&\propto \exp\left(-\frac{1}{2}\|Y - X\beta\|^2\right) \lambda^{-p/2} \exp\left(-\frac{1}{2\lambda}\|\beta\|^2\right) \lambda^{-(a/2+1)} e^{-b/(2\lambda)}.
\end{aligned}$$

Some straightforward calculations reveal that

$$\begin{aligned}
\beta \mid Y, \lambda &\sim \mathcal{N}(\Sigma_\lambda X'Y, \Sigma_\lambda) \qquad \Sigma_\lambda = (X'X + \lambda I)^{-1} \\
\lambda \mid \beta &\sim \mathcal{IG}((a+p)/2, (b + \|\beta\|^2)/2)
\end{aligned}$$

from which we see that the chosen priors are partially conjugate. Note that the posterior mean β conditional on (Y, λ) equals the ridge regression estimator using regularisation parameter λ.

For the hierarchical model in (5) we have :

$$\begin{array}{rcl} \beta \mid \sigma^2, Y, \{\tau_j^2\}_{j=1}^p & \sim & \mathcal{N}(\hat{\beta}, \sigma^2\Sigma^{-1}), \qquad \Sigma = X'X + D_\tau^{-1}, \quad \hat{\beta} = \Sigma^{-1}X'Y \\ \sigma^2 \mid \beta, Y, \{\tau_j^2\}_{j=1}^p & \sim & \mathcal{IG}((p+n+a_\sigma)/2, (b_\sigma + \psi)/2) \\ & & \psi = \|Y - X\beta\|^2 + \beta' D_\tau^{-1}\beta \\ \tau_j^2 \mid \beta_j, \sigma^2, \lambda & \sim & \mathcal{GIG}(1/2, \lambda^2, \beta_i^2/\sigma^2) \\ \lambda^2 \mid \tau_1^2, \ldots, \tau_p^2 & \sim & \mathcal{G}\left(a_\lambda + p, b_\lambda + \frac{1}{2}\sum_{j=1}^p \tau_j^2\right) \end{array} \tag{7}$$

Here $Z \sim \mathcal{GIG}(p, a, b)$, for the definition of the generalised inverse gaussian distribution: see section 7.

A disadvantage of the Gibbs sampler is that it is computationally intensive and convergence can be painfully slow. For really large datasets one often resorts to computing the MAP estimator. In this case one does not use the Gibbs sampler, but other specialised algorithms, such as the *EM (Expectation Maximisation) algorithm*. We do not go into details here. Speeding up computations for Markov Chain Monte Carlo methods is presently a very active field of research.

4 Classification

For the probit model we have

$$L(\beta \mid D) = \prod_{i=1}^n \Phi(x_i'\beta)^{Y_i}(1 - \Phi(x_i'\beta))^{1-Y_i}.$$

For regression the ℓ_p-penalty ($p \in \{1, 2\}$) are added to the negative loglikelihood. This can be generalised

$$\text{minimise} \quad -\sum_{i=1}^n \left[Y_i \log \Phi(x_i'\beta) + (1 - Y_i)\log(1 - \Phi(x_i'\beta))\right] + \lambda\|\beta\|_p^p.$$

Within the Bayesian setup, there is a nice trick to change the problem to a new problem closely related to regression. The trick consists of introducing auxiliary variables $Z_1, \ldots, Z_n$ such that

$$\begin{array}{rcl} Y_i & = & \mathbf{1}\{Z_i \geq 0\} \\ Z_i & \overset{\text{ind}}{\sim} & \mathcal{N}(x_i'\beta, 1). \end{array} \tag{8}$$

It is easy to verify that

$$\mathrm{P}(Y_i = 1) = \mathrm{P}(Z_i \geq 0) = \mathrm{P}\big(Z_i - x_i'\beta \geq -x_i'\beta\big) = \Phi(x_i'\beta).$$

Let $Z = (Z_1, \ldots, Z_n)$. We devise a Gibbs sampler to sample from the joint distribution of (Z, β), conditional on Y. Note that

$$\begin{aligned}
&\Phi(x'\beta)^y \left(1 - \Phi(x'\beta)\right)^{1-y} \\
&\quad = \Phi(x'\beta)y + \big(1 - \Phi(x'\beta)\big)(1 - y) \\
&\quad = y \int \mathbf{1}_{\{z \geq 0\}} \varphi(z; x'\beta, 1) dz + (1 - y) \int \mathbf{1}_{\{z < 0\}} \varphi(z; x'\beta, 1) dz.
\end{aligned}$$

Hence the density is

$$y\mathbf{1}_{\{z \geq 0\}} \varphi(z; x'\beta, 1) + (1 - y)\mathbf{1}_{\{z < 0\}} \varphi(z; x'\beta, 1)$$

with respect to the product measure of the measure that puts mass 1 on both $\{0\}$ and $\{1\}$ times Lebesgue measure. This implies

$$p(Z, \beta \mid Y) \propto p(\beta) \prod_{i=1}^{n} \big(Y_i \mathbf{1}_{\{Z_i \geq 0\}} + (1 - Y_i)\mathbf{1}_{\{Z_i < 0\}}\big) \varphi(Z_i; x_i'\beta, 1)$$

and Gibbs sampling is performed by iteratively sampling from the full conditionals of β and Z. Details are part of the exercises.

For *logistic regression*, a similar trick as in (8) can be applied by considering the hierarchical model

$$\begin{aligned}
Y_i &= \mathbf{1}\{Z_i \geq 0\} \\
Z_i &= X_i\beta + \varepsilon_i \\
\varepsilon_i &\overset{\text{iid}}{\sim} L.
\end{aligned}$$

Here L denotes the logistic distribution: $\varepsilon \sim L$ if $\mathrm{P}(\varepsilon \leq x) = (1 + e^{-x})^{-1}$. An application of Bayesian logistic regression for text categorisation is given in Genkin et al. (2007).

5 Example: relevance vector machines

Here we briefly explain a method for regression and sparsity introduced in Tipping (2001) (see also section 7.2 in Bishop (2006)). The method has an analogue to classification as well. We now explain the method. If we define the *kernel function* K by $K(x, y) = \exp\left(-\frac{1}{2\sigma^2}\|x - y\|^2\right)$, then the following hierarchical model is assumed for the data $\mathcal{D} := \{(x_i, y_i)\}_{i=1}^n$:

$$\begin{aligned}
y_i \mid x_i, \beta, w &\overset{\text{ind}}{\sim} \mathcal{N}\left(w_0 + \sum_{j=1}^{n} w_j K(x_j, x_i), \beta^{-1}\right) \\
w \mid \alpha &\sim \mathcal{N}\left(0, A^{-1}\right), \qquad \text{where} \quad A = \operatorname{diag}(\alpha_0, \ldots, \alpha_n).
\end{aligned}$$

Here, $\alpha_0, \ldots, \alpha_n, \beta$ are hyperparameters. This model implies that the prior on the regression function is given by the series expansion

$$r(x) = w_0 + \sum_{j=1}^{n} w_j K(x, x_j)$$

where the weights are assigned a probability distribution [11]. Using Bayes' theorem it is not difficult to derive that

$$w \mid \mathcal{D}, \alpha, \beta \sim \mathcal{N}\left(\beta \Sigma K' y, \Sigma = (A + \beta K' K)^{-1}\right),$$

where K is the $(n+1) \times (n+1)$ symmetric matrix with elements $K(x_i, x_j)$.

A Bayesian approach would imply that $\alpha_0, \ldots, \alpha_n, \beta$ need to be fixed at some value. Alternatively, these can be provided with a prior distribution, albeit at the cost of additional computations. To obtain sparsity, an *empirical Bayes approach is taken.* Here, the distributions on α_i and β are Dirac measures at the corresponding maximisers of the marginal likelihood which is given by

$$p(y_1, \ldots, y_n \mid x_1, \ldots, x_n, \alpha, \beta) = \int p(y_1, \ldots, y_n \mid x_1, \ldots, x_n, \beta, w) p(w \mid \alpha) dw.$$

This is sometimes called the *evidence approximation* or *type-2 maximum likelihood.* The density in the preceding display is in fact the multivariate normal density $\mathcal{N}(0, C)$, evaluated at $(y_1, \ldots, y_n)$. Here

$$C = \beta^{-1} I + K A^{-1} K'.$$

The availability of a closed form expression for the marginal likelihood facilitates the empirical Bayes approach. For details about the optimisation we refer to section 7.2 in Bishop (2006).

It turns out that a proportion of the α_i is infinite in this maximisation, corresponding to weights $w_i \to 0$. Therefore, effectively the number of terms in the regression function can be way smaller than the sample size n (yielding sparsity). Those components $x \mapsto K(x, x_i)$ for which $w_i \neq 0$ are called the *relevance vectors.*

6 Further reading

The presented methods are by now classical and serve as a starting point. Many variations have been proposed with fancy names such as *support vector machines* and *Boltzmann machines.* Depending on the application, these can perform better or worse (ultimately depending on a chosen performance measure).

Likelihood penalisation is certainly not the only way to deal with a large number of predictors. The "machine learning" community has introduced many algorithmic

[11] As the basis functions in the expansion depend on the x_i this does not entirely fit into the Bayesian framework. The machine learning community appears to be quite pragmatic on these kind of things.

methods for obtaining good predictions without using any kind of statistical model. Well known algorithms include *decision trees* (CART), *adaptive boosting* (adaboost) and *random forests*, to name a few. In the projects it may well be worthwhile to investigate whether these algorithms can give useful insights. You could for example have 2 students in your group trying out one of these methods.

Useful references are Hastie et al. (2014) (mostly frequentist), Bishop (2006) (mainly written from the Bayesian point of view), Murphy (2012) and Kuhn and Johnson (2013) (contains lot's of practical advice and pointers to relevant `R` libraries).

7 Appendix: distributions

Parametrisations of certain distributions differ among texts, so we summarise some frequently used distributions

1. $\mathcal{E}(\lambda)$. Exponential distribution with intensity (rate) λ:
$$p(x) = \lambda e^{-\lambda x} \mathbf{1}\{x \geq 0\}.$$

2. $\mathcal{G}(a, b)$. Gamma distribution with shape parameter $a > 0$ and rate parameter $b > 0$:
$$p(x) \propto x^{a-1} e^{-bx} \mathbf{1}\{x \geq 0\}.$$

3. $\mathcal{DE}(\lambda)$. Double exponential distribution with rate λ:
$$p(x) = \frac{\lambda}{2} e^{-\lambda |x|}.$$

4. $\mathcal{GIG}(p, a, b)$ Generalised inverse gaussian distribution ($p \in \mathbb{R}$, $a > 0$, $b > 0$):
$$p(x) \propto x^{p-1} \exp\left(-\frac{1}{2}\left(ax + b/x\right)\right) \mathbf{1}\{x \geq 0\}.$$

5. $\mathcal{IG}(a, b)$ Inverse gamma distribution:
$$p(x) \propto x^{-a-1} e^{-b/x} \mathbf{1}\{x \geq 0\}.$$

6. $\mathcal{N}_k(\mu, \Sigma)$. Multivariate normal distribution with mean vector μ and covariance matrix Σ.
$$p(x) = (2\pi)^{-k/2} |\Sigma|^{-1/2} \exp\left((x - mu)' \Sigma^{-1} (x - \mu)\right).$$
This distributions is obtained by starting with a sequence of independent random variables $Z_1, \ldots Z_k$ with common standard normal distributions. Denote $Z = (Z_1, \ldots, Z_k)'$. If we define
$$X = \mu + LZ,$$
then $X \sim N_k(\mu, \Sigma)$ with $\Sigma = LL'$.

References

Bae, K. and Mallick, B.K. (2004) *Gene selection using a two-level hierarchical Bayesian model*, Bioinformatics **20**(18), 3423–3430

Bishop, C.M. (2006) *Pattern Recognition and Machine Learning*, Springer.

Carvalho, C.M., Polson, N.G. and Scott, J.G. (2010) *The horseshoe estimator for sparse signals*, Biometrika **97**(2), 465–480

Chen, S., Donoho, D.L. and Saunders, M. (1998) *Atomic decomposition for basis pursuit*, SIAM Journal on Scientific Computing **20**(1), 33–61

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004) *Least angle regression*, Annals of Statistics **32**(2), 407–499

Genkin, A., Lewis, D.D. and Madigan, D. (2007) *Large-Scale Bayesian Logistic Regression for Text Categorization*, Technometrics **49**(3), 291–304.

Gramacy, R.B. and Pantaleo, E. (2010) *Shrinkage Regression for Multivariate Inference with Missing Data, and an Application to Portfolio Balancing*, Bayesian Analysis **5**(2), 237–262

Hastie, T., Tibshirani, R. and Friedman, J. (2014) *The Elements of Statistical Learning*, 2nd edition, Springer (the pdf is freely available)

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

Kuhn, M. and Johnson, K. (2013) *Applied Predictive Modeling*, Springer.

Murphy, K.P. (2012) *Machine Learning, A Probabilistic Perspective*, MIT Press.

Tibshirani, R. (1996) *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society: Series B **58**(1), 267–288.

Tibshirani, R.J. (2013) *The lasso problem and uniqueness* Electronic Journal of Statistics **7**, 1456–1490.

Tipping, M.E. (2001) *Sparse Bayesian Learning and the Relevance Vector Machine*, Journal of Machine Learning Research **1**, 211–244.

Young, G.A. and Smith, R.L. (2010) *Essentials of Statistical Inference*, Cambridge University Press

8 Exercises

Non-starred exercises are to be handed in as homework exercises.

1. Show that in case $r = p$

$$\hat{\beta}_r = \left(I + \lambda(X'X)^{-1}\right)^{-1} \hat{\beta}$$

and conclude that $\hat{\beta}_r$ is biased for estimating β. Here, we condition on X, which henceforth can be considered deterministic.

2. * Show that

$$\hat{\beta}_r = X'(\lambda I + XX')^{-1}Y.$$

When is this formula computationally cheaper than equation (3) for computing $\hat{\beta}_r$?
Hint: Write $\beta = X'\alpha$ with $\alpha = \lambda^{-1}(Y - X\beta)$.

3. Instead of either an ℓ_1 or ℓ_2 penalty, one can also consider a convex combination of these two penalties. This is called the "elastic net" because it is claimed to be "like a stretchable fishing net that retains all the big fish". Actually, the elastic net is implemented in the `glmnet` package in R. For $\alpha \in [0, 1]$, the objective function to be minimised is

$$\|Y - X\beta\|^2 + \lambda \left(\alpha\|\beta\|_2^2 + (1 - \alpha)\|\beta\|_1\right).$$

Show that this optimisation problem can be reduced to a lasso problem on modified data.

4. * Look up the proof on the theoretical justification of the lasso at http://www.stat.cmu.edu/~larry/=sml/Assumptions.pdf. List all typo's in the proof.

5. In this exercise we will compare schemes (5) and (6). Suppose $X = I$, $\sigma^2 = 1$ and $\lambda = 1$. Note that the crucial difference between the Horseshoe estimator and the Bayesian lasso lies in the choice of distribution for τ_j $(j = 1, \ldots, p)$.

   (a) Show that

$$\mathrm{E}[\beta_i \mid Y_i, \tau_i] = (1 - u_i)Y_i,$$

   with $u_i = (1 + \tau_i^2)^{-1}$.

   (b) Show that for the Horseshoe prior u_i has a Beta distribution with parameters 1/2 and 1/2.

   (c) Derive an expression for the density of u_i in case $\tau_i^2 \sim Exp(1/2)$.

   (d) Plot the densities derived in parts (b) and (c) in one figure. Using this figure explain why the Horseshoe prior is suitable for discovering a few strong signals in a sparse setting.

(e) Now consider the Horseshoe prior with an additional factor $\alpha > 0$, where the second line of (6) is replaced with

$$\beta \mid \sigma^2, \tau_1^2, \ldots, \tau_p^2 \quad \sim \quad N(0, \sigma^2 \alpha^2 D), \quad D = \text{diag}(\tau_1^2, \ldots, \tau_p^2).$$

Investigate the effect of α by computing the density of u_i for this setting. Make a small simulation experiment that illustrates this effect. Please explain carefully.

6. * Prove the relations in (7).

7. Consider the Bayesian probit model. Suppose that we take a $N(0, \Sigma)$ prior for β. When using the Maximum A Posteriori estimator this corresponds to ridge probit regression.

   (a) Show that sampling from β conditional on (Y, Z) boils down to sampling from the $N(A^{-1}X'Z, A^{-1})$ distribution, where X is the $n \times p$ design matrix containing all predictor varialbes and $A = X'X + \Sigma^{-1}$.

   (b) Denote by x_i the i-th row of the matrix X (considered as a column vector). Show that sampling $Z_1, \ldots, Z_n \mid \beta$ conditional on (Y, β) boils down to sampling $Z_i \sim TN(x_i'\beta, 1, Y_i)$. Here, $TN(\mu, \sigma^2, u)$ denotes the normal distribution with mean μ and variance σ^2 that is truncated to be positive if $u = 1$ and negative if $u = 0$.

   (c) Suppose we wish to replace te ℓ_2 penalty by an ℓ_1 penalty. Adjust your prior specification and give the steps for the corresponding Gibbs sampler. Hint: read the first three pages of Bae and Mallick (2004).