2.4.1 Practical aspects about the Lasso for prediction

From a practical perspective, prediction with the Lasso is straightforward and easy. One often uses a cross-validation (CV) scheme, e.g., 10-fold CV, to select a reasonable tuning parameter λ minimizing the cross-validated squared error risk. In addition, we can validate the accuracy of the performance by using again some cross-validation scheme. Regarding the latter, we should cross-validate the whole procedure including the selection of the tuning parameter λ. In particular, by comparing the cross-validated risk, we can roughly see whether the Lasso yields a performance which is better, equal or worse than another prediction algorithm. However, when aiming for more formal conclusions, it is not straightforward to test statistically whether the performances of two prediction algorithms are significantly different, see for example van de Wiel et al. (2009).

2.4.1.1 Binary classification of lymph node status using gene expressions

We consider a classification problem involving a binary response variable $Y \in \{0,1\}$, describing the lymph node status of a cancer patient, and we have a covariate with $p = 7129$ gene expression measurements. There are $n = 49$ breast cancer tumor samples. The data is taken from West et al. (2001). It is known that this is a difficult, high noise classification problem. The best methods achieve a cross-validated misclassification error of about 20%.

Despite the binary nature of the classification problem, we can use the Lasso as in (2.2) which yields an estimate of the conditional class probability $f(x) = \mathbf{P}[Y = 1|X = x] = \mathbb{E}[Y|X = x]$:

$$\hat{f}_\lambda(x) = x\hat{\beta}(\lambda) = \sum_{j=1}^{p} \hat{\beta}_j(\lambda)x^{(j)}.$$

Of course, we could use the Lasso also for logistic regression as described later in Chapter 3. In either case, having an estimate of the conditional class probability, denoted here by $\hat{f}_\lambda(\cdot)$, we classify as follows:

$$\hat{\mathscr{C}}_\lambda(x) = \begin{cases} 1 & \hat{f}_\lambda(x) > 1/2, \\ 0 & \hat{f}_\lambda(x) \leq 1/2. \end{cases}$$

For comparison, we consider a forward variable selection method in penalized linear logistic regression with ℓ_2-norm (Ridge-type) penalty. The optimal regularization parameter, for Lasso and forward penalized logistic regression, is chosen by 10-fold cross-validation. For evaluating the performance of the tuned algorithms, we use a cross-validation scheme for estimating the test-set misclassification error. We randomly divide the sample into 2/3 training- and 1/3 test-data and we repeat this

100 times: the average test-set misclassification error is reported in Table 2.1. Note that we run a double cross-validation: one inner level for choosing the regularization parameter and one outer level for assessing the performance of the algorithm.

Table 2.1 illustrates that the forward selection approach yields, in this example, much poorer performance than the Lasso. Forward selection methods tend to be

Lasso	forw. penalized logist. regr.
21.1%	35.25%

Table 2.1 Misclassification test set error using cross-validation

unstable (Breiman, 1996): they are of a very greedy nature striving for maximal improvement of the objective function (e.g. residual sum of squares) in every step.

Finally, we report that the Lasso selected on cross-validation average 13.12 out of $p = 7129$ variables (genes). Thus, the fitted linear model is very sparse with respect to the number of selected variables.

2.4.2 Some results from asymptotic theory

We now describe results which are developed and described in detail in Chapter 6. For simplicity, we take here an asymptotic point of view instead of finite sample results in Chapter 6. To capture high-dimensional scenarios, the asymptotics is with respect to a triangular array of observations:

$$Y_{n;i} = \sum_{j=1}^{p_n} \beta^0_{n;j} X^{(j)}_{n;i} + \varepsilon_{n;i},\ i = 1,\dots,n;\ \ n = 1,2,\dots \tag{2.6}$$

Thereby, we allow that $p = p_n \gg n$. The assumptions about $\varepsilon_{n;i}$ are as in the linear model in (2.1). A consistency result requires a sparsity assumption of the form

$$\|\beta^0\|_1 = \|\beta^0_n\|_1 = o\left(\sqrt{\frac{n}{\log(p)}}\right),$$

see Corollary 6.1 in Chapter 6. Assuming further mild regularity conditions on the error distribution, the following holds: for a suitable range of $\lambda = \lambda_n \asymp \sqrt{\log(p)/n}$, the Lasso is consistent for estimating the underlying regression function:

$$(\hat{\beta}(\lambda) - \beta^0)^T \Sigma_X (\hat{\beta}(\lambda) - \beta^0) = o_P(1)\ (n \to \infty), \tag{2.7}$$

where Σ_X equals $n^{-1}\mathbf{X}^T\mathbf{X}$ in case of a fixed design. In the case of random design, Σ_X is the covariance of the covariate X, and then (2.7) holds assuming $\|\beta^0\|_1 =$