Chapter 10

The de-sparsified or de-biased Lasso for confidence regions and testing

Abstract Error control based on p-values and confidence statements is a most important task in many areas of sciences. In the high-dimensional setting, assigning significance and quantifying uncertainties is challenging. While the Lasso and other sparse estimators are not tailored for this: de-sparsification is crucial and results in the so-called de-sparsified or de-biased Lasso. The low-dimensional components of the estimated parameters have asymptotic Gaussian distributions and this itself leads to the construction of tests and confidence regions which are, under additional conditions, asymptotically optimal in the established framework of semiparametric inference. A bootstrap method can be used in conjunction with the de-biased Lasso which is especially useful in presence of heteroscedastic and non-Gaussian errors and for multiple testing adjustemtn in presence of strong dependence. The de-sparsified or de-biased Lasso provides a powerful andimpotrant tool for high-dimensional statistical inference: other more generic procedures are described in Chapters 11 and 12.

10.1 Organization of the chapter

We introduce in Section 10.3 the de-sparsified or de-biased Lasso for linear models, an estimator which has been proposed by Zhang and Zhang (2014). Due to its non-sparsity, it is a regular estimator which is not exposed to the super-efficiency phenomenon. We show in Section 10.3.1 that the de-sparsified or de-biased Lasso has asymptotically a Gaussian distribution and we disucss its optimality in Section 10.3.3. The notion of optimality is here according to the framework of semiparametric inference: the estimator achieves asymptotically the Crame-Rao lower bound for the asymptotic variance. Implications and practical aspects are discussed in Section 10.3.2. We describe in Section 10.4.2 a bootstrap procedure for the de-sparsified or de-biased Lasso: it is especially useful in presence of non-Gaussian or heteroscedastic errors and for more efficient multiple testing adjustment among strongly depen-

dent tests. Finally, Section 10.5 delineates the extension to generalized linear models and alternative methods are briefly mentioned in Section 10.6.

10.2 Introduction

We consider first a high-dimensional linear model as in (2.1) while extensions are discussed in Section 10.5:

$$\mathbf{Y} = \mathbf{X}\beta^0 + \varepsilon \tag{10.1}$$

with $n \times p$ fixed or random design matrix $\mathbf{X}$, $n \times 1$ response and error vectors $\mathbf{Y}$ and ε, respectively. The errors are $\varepsilon_1, \ldots, \varepsilon_n$ are assumed to be independent with $\mathbb{E}[\varepsilon_i] = 0$, and independent of $\mathbf{X}$ (for random design). As in the previous chapters, We allow for high-dimensional settings where $p \gg n$ and we denote the active set of relevant variables as

$$S_0 = \{j;\ \beta_j^0 \neq 0,\ j = 1, \ldots, p\},$$

with cardinality $s_0 = |S_0|$. The main goals in this chapter are the construction of confidence intervals and statistical hypothesis tests for the regression parameters β_j^0 $(j = 1, \ldots, p)$ and corresponding multiple testing adjustment. The former is a highly non-standard problem in high-dimensional settings while for the latter we can use standard well-known techniques: when considering both goals simultaneously, though, one can develop more powerful multiple testing adjustments.

For assigning uncertainties in terms of confidence intervals or hypothesis testing, the standard Lasso in (2.2) seems inappropriate. It is very difficult to characterize the distribution of the estimator in the high-dimensional setting: Knight and Fu (2000) derive asymptotic results for fixed dimension as sample size $n \to \infty$ and already for such simple situations, the asymptotic distribution of the Lasso has point mass at zero. This implies, because of non-continuity of the distribution, that standard bootstrapping and subsampling schemes are delicate to apply and uniform convergence to the limit seems hard to achieve. The latter means that the estimator is exposed to undesirable super-efficiency problems. All the problems mentioned above apply not only for the Lasso but also for other sparse estimators.

10.3 Regularized projection: de-biasing or de-sparsifying the Lasso

We describe here a method, first introduced by Zhang and Zhang (2014). It is instructive to give a motivation starting with the low-dimensional setting where $p < n$ and $\mathrm{rank}(\mathbf{X}) = p$. The jth component of the ordinary least squares estimator $\hat{\beta}_{\mathrm{OLS};j}$ can be obtained as follows. Do an OLS regression of $\mathbf{X}^{(j)}$ versus all other variables $\mathbf{X}^{(-j)}$ and denote the corresponding residuals by $Z^{(j)}$. Then:

$$\hat{\beta}_{\mathrm{OLS};j} = \mathbf{Y}^T Z^{(j)} / (\mathbf{X}^{(j)})^T Z^{(j)} \tag{10.2}$$

can be obtained by a linear projection (Problem 10.1). In a high-dimensional setting with $\mathrm{rank}(\mathbf{X}) = n$, the residuals $Z^{(j)}$ would be equal to zero and the projection is ill-posed.

For the high-dimensional case with $p > n$, the idea is to pursue a regularized projection. Instead of ordinary least squares regression, we use a Lasso regression of $\mathbf{X}^{(j)}$ versus $\mathbf{X}^{(-j)}$ with corresponding residual vector $Z^{(j)}$:

$$\begin{aligned}
&\hat{\gamma}^{(j)} = \operatorname*{arg\,min}_{\gamma \in \mathbb{R}^{p-1}} \|\mathbf{X}^{(j)} - \mathbf{X}^{(-j)}\gamma\|_2^2/n + \lambda_j |\gamma|_1, \\
&Z^{(j)} = \mathbf{X}^{(j)} - \mathbf{X}^{(-j)}\hat{\gamma}^{(j)}.
\end{aligned}$$

This involves a regularization parameter λ_j for the Lasso, and hence $Z^{(j)} = Z^{(j)}(\lambda_j)$. We immediately obtain for any vector $Z^{(j)}$:

$$\begin{aligned}
&\frac{Y^T Z^{(j)}}{(\mathbf{X}^{(j)})^T Z^{(j)}} = \beta_j^0 + \sum_{k \neq j} P_{jk} \beta_k^0 + \frac{\varepsilon^T Z^{(j)}}{(\mathbf{X}^{(j)})^T Z^{(j)}}, \\
&P_{jk} = (\mathbf{X}^{(k)})^T Z^{(j)} / (\mathbf{X}^{(j)})^T Z^{(j)}.
\end{aligned} \tag{10.3}$$

We note that in the low-dimensional case with $Z^{(j)}$ being the residuals from ordinary least squares, due to orthogonality, $P_{jk} = 0$ for all $k \neq j$.

When using the Lasso-residuals for $Z^{(j)}$, we do not have exact orthogonality and a bias term arises. Thus, we make a bias correction in (10.3) by plugging in the Lasso estimator $\hat{\beta}$ (of the regression $\mathbf{Y}$ versus $\mathbf{X}$): the bias-corrected estimator is

$$\hat{b}_j = \frac{Y^T Z^{(j)}}{(\mathbf{X}^{(j)})^T Z^{(j)}} - \sum_{k \neq j} P_{jk} \hat{\beta}_k. \tag{10.4}$$

The estimator $\hat{b} = \{\hat{b}_j;\ j = 1, \ldots, p\}$ is not sparse with all components being different from zero. This non-sparseness happens because the first term on the right-hand side of (10.4) is non-zeroi and the second term does not cancel the first one (except for a constellation which has probability zero). Thus, the estimator in (10.4) is sometimes called the de-sparsified Lasso (van de Geer et al., 2014). We can write

(10.4) also in the following form:

$$\hat{b}_j = \hat{\beta}_j + \frac{(Z^{(j)})^T(\mathbf{Y} - \mathbf{X}\hat{\beta})}{(Z^{(j)})^T\mathbf{X}^{(j)}}. \tag{10.5}$$

Since $\hat{\beta}$ is the Lasso estimator, we see that $\hat{b}_j$ equals the Lasso estimator with an additional estimated bias correction term: therefore the name de-biased Lasso (Zhang and Zhang, 2014).

10.3.1 Limiting Gaussian distribution

We show here that the de-sparsified or de-biased Lasso has an asymptotic Gaussian distribution. Using (10.3) we obtain:

$$\sqrt{n}(\hat{b}_j - \beta_j^0) = \frac{n^{-1/2}\varepsilon^T Z^{(j)}}{n^{-1}(\mathbf{X}^{(j)})^T Z^{(j)}} + \sum_{k \neq j} \sqrt{n} P_{jk}(\hat{\beta}_k - \beta_k^0).$$

The first term on the right-hand side of the equal sign has an exact Gaussian distribution, when assuming Gaussian errors; for non-Gaussian error, one can establish an asymptotic Gaussian distribution when assuming that $\mathbb{E}|\varepsilon_i|^{2+\kappa} < \infty$ for $\kappa > 0$. We will argue below in Lemma 10.1 and Theorem 10.1 that the second term is asymptotically negligible.

Assuming such an asymptotic negligibility, we have for the variance of the leading term

$$\mathrm{Var}\Big(\frac{n^{-1/2}\varepsilon^T Z^{(j)}}{n^{-1}(\mathbf{X}^{(j)})^T Z^{(j)}}\Big) = \sigma_\varepsilon^2 \frac{\|Z^{(j)}\|_2^2/n}{|n^{-1}(\mathbf{X}^{(j)})^T Z^{(j)}|^2}.$$

Thus, when standardizing to unit variance we approximately obtain

$$\sigma_\varepsilon^{-1}\sqrt{n}\frac{n^{-1}(\mathbf{X}^{(j)})^T Z^{(j)}}{n^{-1/2}\|Z^{(j)}\|_2}(\hat{b}_j - \beta_j^0) \approx \mathcal{N}(0,1).$$

We will make this rigorous next.

Lemma 10.1. *Consider a linear model as in (10.1) with fixed design and Gaussian errors $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 I)$. Then:*

$$\sigma_\varepsilon^{-1}\sqrt{n}\mathrm{diag}\left(\frac{n^{-1}(\mathbf{X}^{(1)})^T Z^{(1)}}{n^{-1/2}\|Z^{(1)}\|_2}, \ldots, \frac{n^{-1}(\mathbf{X}^{(p)})^T Z^{(p)}}{n^{-1/2}\|Z^{(p)}\|_2}\right)(\hat{b} - \beta^0) = W + \Delta$$

where

$$W \sim \mathcal{N}_p(0,\Omega),\ \Omega_{jk} = \frac{n^{-1}(Z^{(j)})^T Z^{(k)}}{n^{-1/2}\|Z^{(j)}\|_2 n^{-1/2}\|Z^{(k)}\|_2},$$
$$|\Delta_j| \le \sigma^{-1}\sqrt{n}\lambda_j/2 \frac{1}{\|Z^{(j)}\|_2/\sqrt{n}} \|\hat{\beta} - \beta^0\|_1.$$

Proof. We write for a single component j:

$$\begin{aligned}\hat{b}_j &= \frac{(Z^{(j)})^T Y}{(Z^{(j)})^T \mathbf{X}^{(j)}} - \sum_{k\neq j} \frac{(Z^{(j)})^T \mathbf{X}^{(k)}}{(Z^{(j)})^T \mathbf{X}^{(j)}} \hat{\beta}_k + \frac{(Z^{(j)})^T \varepsilon}{(Z^{(j)})^T \mathbf{X}^{(j)}} \\ &= \beta_j^0 - \sum_{k\neq j} \frac{(Z^{(j)})^T \mathbf{X}^{(k)}}{(Z^{(j)})^T \mathbf{X}^{(j)}} (\hat{\beta}_k - \beta_k^0) + \frac{(Z^{(j)})^T \varepsilon}{(Z^{(j)})^T \mathbf{X}^{(j)}}.\end{aligned}$$

Thus, we obtain

$$\begin{aligned}&\sigma^{-1}\sqrt{n}\frac{(Z^{(j)})^T \mathbf{X}^{(j)}/n}{\|Z^{(j)}\|_2/\sqrt{n}}(\hat{b}_j - \beta_j^0) = W_j + \Delta_j, \\ &W_j = \sigma^{-1}\frac{(Z^{(j)})^T \mathbf{X}^{(j)}/n}{\|Z^{(j)}\|_2/\sqrt{n}}(Z^{(j)})^T \varepsilon/\sqrt{n} \sim \mathcal{N}(0,1), \\ &\Delta_j = \sigma^{-1}\sqrt{n}\frac{1}{\|Z^{(j)}\|_2/\sqrt{n}} \sum_{k\neq j} (Z^{(j)})^T \mathbf{X}^{(k)}/n(\beta_k^0 - \hat{\beta}_k).\end{aligned}$$

Clearly the vector $W = (W_1, \ldots, W_p)$ has a Gaussian distribution $\mathcal{N}_p(0,\Omega)$ with Ω as in the lemma.

For the error (or bias) term we exploit the KKT conditions of the Lasso, see Lemma 2.1, saying that

$$|Z^{(j)})^T \mathbf{X}^{(k)}/n| \le \lambda_j/2 \text{ for all } k \neq j. \tag{10.6}$$

Then, by Hölder's inequality,

$$|\Delta_j| \le 2\sigma^{-1}\sqrt{n}\frac{\lambda_j}{\|Z^{(j)}\|_2/\sqrt{n}} \|\hat{\beta} - \beta^0\|_1.$$

□

We see that the bias term Δ_j is small if $\frac{\lambda_j}{\|Z^{(j)}\|_2/\sqrt{n}}$ is small. Obviously, choosing λ_j very small leads to residuals with small $\|Z^{(j)}\|_2$ and there is a trade-off. Under some conditions, one can choose $\lambda_j \asymp \sqrt{\log(p)/n}$ such that $\|Z^{(j)}\|_2/\sqrt{n}$ is bounded away from zero: invoking the usual bound for $\|\hat{\beta} - \beta^0\|_1 = O_P(s_0\sqrt{\log(p)/n})$, see for example (2.22), this then leads to the bound

$$|\Delta_j| \le O_P(\sqrt{n}\sqrt{\log(p)/n} s_0 \sqrt{\log(p)/n}) = O_P(s_0 \log(p)/\sqrt{n}),$$

and the right-hand side is negligible if $s_0 = o(\sqrt{n}/\log(p))$. We will make this rigorous next.

Assume the following:

(B1) The design matrix $\mathbf{X}$ has compatibility constant $\phi_0^2 \geq C > 0$ bounded away from zero, and the sparsity of the regression vector is $s_0 = \|\beta^0\|_0^0 = o(\sqrt{n}/\log(p))$.

(B2,j) For $\lambda_j = C_j\sqrt{\log(p)/n}$ with $0 < L_1 \leq C_j \leq L_2 < \infty$ the residuals satisfy $\|Z^{(j)}\|_2^2/n \geq L > 0$ (where L might depend on L_1).

Theorem 10.1. *Consider a linear model as in (10.1) with fixed design and Gaussian errors $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 I)$. Assume (B1) and (B2,j), and choose $\lambda = C\sqrt{\log(p)/n}$ for $0 < M_1 \leq C \leq M_2 < \infty$ with M_1 sufficiently large. We then have that*

$$\sigma^{-1}\sqrt{n}\frac{n^{-1}(\mathbf{X}^{(j)})^T Z^{(j)}}{n^{-1/2}\|Z^{(j)}\|_2}(\hat{b}_j - \beta_j^0) \Longrightarrow \mathcal{N}(0,1).$$

If in addition (B2,j) holds for all $j = 1, \ldots, p$:

$$\sigma^{-1}\sqrt{n}\text{diag}\left(\frac{n^{-1}(\mathbf{X}^{(1)})^T Z^{(1)}}{n^{-1/2}\|Z^{(1)}\|_2}, \ldots, \frac{n^{-1}(\mathbf{X}^{(p)})^T Z^{(p)}}{n^{-1/2}\|Z^{(p)}\|_2}\right)(\hat{b} - \beta^0) = W + \Delta$$

where

$$W \sim \mathcal{N}_p(0, \Omega),\ \Omega_{jk} = \frac{n^{-1}(Z^{(j)})^T Z^{(k)}}{n^{-1/2}\|Z^{(j)}\|_2 n^{-1/2}\|Z^{(k)}\|_2},$$
$$\max_{j=1,\ldots,p} |\Delta_j| = o_P(1).$$

Proof. Assumption (B1) guarantees that when choosing $\lambda = C\sqrt{\log(p)/n}$ with C sufficiently large that

$$\|\hat{\beta} - \beta^0\|_1 = O_P(s_0\sqrt{\log(p)/n}),$$

see for example (2.22) and Chapter 6. The result then follows from Lemma 10.1 by invoking assumption (B2,j). □

The lower bound of the compatibility constant in assumption (B1) is justified in Section 6.12. Assumption (B2,j) can be justified as follows. Assume that the rows of $\mathbf{X}$ are i.i.d. from a distribution with mean zero and covariance matrix Σ (and one then conditions on $\mathbf{X}$ for a fixed design linear model). Assume that for the inverse $\Theta = \Sigma^{-1}$, $\Theta_{jj} \geq 2L > 0$. It then holds that

$$\mathbf{X}^{(j)} = \mathbf{X}^{(-j)}\gamma^{(j)} + \eta^{(j)},$$

with $\eta^{(j)} \sim \mathcal{N}_n(0, \tau_j^2 I)$, $\tau_j^2 = 1/\Theta_{jj}$ and $\eta^{(j)}$ uncorrelated from $\mathbf{X}^{(-j)}$. If the Lasso is consistent for the prediction error, saying that

$$\|\mathbf{X}^{(-j)}(\hat{\gamma}^{(j)} - \gamma^{(j)}\|_2^2/n = o_P(1), \tag{10.7}$$

we have that (B2,j) holds with probability tending to 1. Consistency of the prediction error in (10.7) holds if the sparsity is $s_j = \|\gamma^{(j)}\|_0^0 = o(\log(p)/n)$ and a restricted eigenvalue condition holds for $\mathbf{X}^{(-j)}$, see for example (2.22).

From a theoretical perspective, it is more elegant to use the square root Lasso (Belloni et al., 2011), described in Section 2.13, for the construction of $Z^{(j)}$: then, one can establish an analogue of the statements in the Theorem 10.1 without requiring (B2,j), see Problem 10.2. In fact, the bound for the bias term Δ_j in Lemma 10.1 becomes:

$$|\Delta_j| \leq \sigma^{-1}\sqrt{n}\lambda_j\|\hat{\beta} - \beta^0\|_1.$$

Under assumption (B1), this converges to zero in probability. In practice, it seems to make essentially no difference whether one takes the square root or plain Lasso for the construction of the $Z^{(j)}$'s (and in fact, the square root Lasso has the same solution path as the Lasso, see Peter: Reference????).

Finally, the convergence in Theorem 10.1 is uniform over the subset of the parameter space where the number of non-zero coefficients is small, e.g. over $\{\beta;\ \|\beta\|_0^0 \leq s_0\}$, where $s_0 = o(\sqrt{n}/\log(p))$ occurs in condition (B1). Therefore, we obtain *honest* confidence regions and tests, as discussed in the next subsection.

10.3.2 Confidence regions, group inference, multiple statistical testing and practical issues

Theorem 10.1 justifies the construction of confidence regions and statistical hypothesis tests. When considering a single regression parameter, we can construct a two-sided confidence interval as follows:

$$\hat{b}_j \pm \hat{\sigma}n^{-1/2}\frac{n^{-1/2}\|Z^{(j)}\|_2}{n^{-1}(\mathbf{X}^{(j)})^T Z^{(j)}}\Phi^{-1}(1-\alpha/2),$$

where $\Phi(\cdot)$ is the cumulative distribution function of $\mathcal{N}(0,1)$. Here, $\hat{\sigma}^2$ is an estimate of the error variance, as discussed below. Two-sided statistical testing for a single parameter concerns the null-hypothesis

$$H_{0,j}:\ \beta_j^0 = 0$$

versus the alternative

$$H_{A,j}:\ \beta_j^0 \neq 0.$$

Of course, one could also test other parameter values than zero.

When doing inference for a few components $G \subseteq \{1,\ldots,p\}$ of β^0, with $|G|$ of fixed cardinality (from an asymptotic perspective as $n \to \infty$), we can rely on the distribution of W_G from the second statement in Theorem 10.1. If $|G|$ is arbitrarily large, we restrict the focus to the sup-norm: in this norm, the approximation error $\|\Delta\|_\infty$ in Theorem 10.1 converges to zero. Thus, for a group $G \subseteq \{1,\ldots,p\}$, we can test a group null-hypothesis

$$H_{0,G}: \ \beta_j^0 = 0 \text{ for all } j \in G,$$

versus the logical complement $H_{A,G}: \ \beta_j^0 \neq 0$ from some $j \in G$, by considering the test-statistic

$$\max_{j\in G} \hat{\sigma}^{-1}\sqrt{n}\frac{n^{-1}(\mathbf{X}^{(j)})^T Z^{(j)}}{n^{-1/2}\|Z^{(j)}\|_2}|\hat{b}_j| \Rightarrow \max_{j\in G}|W_j|, \tag{10.8}$$

where the limit on the right hand side occurs if the null-hypothesis $H_{0,G}$ holds true. The distribution of $\max_{j\in G}|W_j|$ can be easily simulated from dependent Gaussian random variables from $\mathcal{N}_p(0,\Omega)$ where Ω is known. We also remark that sum-type statistics for large groups cannot be easily treated because $\sum_{j\in G}|\Delta_j|$ might not be reasonably upper-bounded.

We can also use the limiting distribution in (10.8) for multipe testing correction. Denote by $F_G(c) = \mathbf{P}[\max_{j\in G}|W_j| \le c]$. Then, the corrected p-values for testing all single hypotheses $H_{0,j}: \ \beta_j^0 = 0$ versus $H_{A,j}: \ \beta_j^0 \neq 0$, for $j = 1,\ldots,p$, is given by

$$P_{\mathrm{corr},j} = F_{\{1,\ldots,p\}}(\hat{\sigma}^{-1}\sqrt{n}\frac{n^{-1}(\mathbf{X}^{(j)})^T Z^{(j)}}{n^{-1/2}\|Z^{(j)}\|_2}|\hat{b}_j|), \tag{10.9}$$

which controls the familywise error rate (FWER), see Problem 10.3. The FWER is defined as follows. Denote the number of false positives by $V = V(\alpha) = \sum_{j=1}^p I(P_{\mathrm{corr},j} \le \alpha)\cdot I(H_{0,j}$ holds true$)$. Then, the familywise error rate is

$$\mathbf{P}[V > 0].$$

The correction of p-values for testing many group hypothesis H_{0,G_r} $(r = 1,\ldots,m)$ can be done similarly as in (10.9). See Problem 10.3.

Estimation of the error variance can be done by using the residual sum of squqres from Lasso fit:

$$\hat{\sigma}^2 = n^{-1}\|\mathbf{Y} - \mathbf{X}\hat{\beta}(\lambda_{\mathrm{CV}})\|_2^2, \tag{10.10}$$

where λ_{CV} is the regularization parameter optimizing a cross-validated squared error loss. Often, 10-fold cross-validation is employed. Insteaad of (10.10) one can use a more conservative estimate with the scaling factor $(n - \|\hat{\beta}(\lambda_{\mathrm{CV}}\|_0^0)^{-1}$. The use of cross-validation seems to work quite well empirically (Reid et al., 2016). . As an alternative, the so-called scaled Lasso (Sun and Zhang, 2012) can be used which

leads to a consistent estimate of the error variance: it is a fully automatic method which does not require any specification of a tuning parameter.

From a practical perspective, we need to choose the regularization parameters λ (for the Lasso regression of Y versus $\mathbf{X}$) and λ_j (for the nodewise Lasso regressions of $\mathbf{X}^{(j)}$ versus all other variables $\mathbf{X}^{(-j)}$). Regarding the former, we advocate a choice using (typically a 10-fold cross-validation; for the latter, we favor a proposal for a smaller λ_j than the one from cross-validation.

Choice of λ_j for de-sparsified Lasso. We see from the KKT conditions, see Lemma 2.1, that the numerator of the error in the bias correction term (i.e. the P_{jk}'s) is decreasing as $\lambda_j/2 \searrow 0$; for controlling the denominator, λ_j shouldn't be too small to ensure that the denominator (i.e. $n^{-1}(\mathbf{X}^{(j)})^T Z^{(j)}$) behaves reasonable (staying away from zero) for a fairly large range of λ_j.

Therefore, the strategy is as follows.

1. Compute a Lasso regression of $\mathbf{X}^{(j)}$ versus all other variables $\mathbf{X}^{(-j)}$ using (typically a 10-fold) cross-validation, and the corresponding residual vector is denoted by $Z^{(j)}$.
2. Compute $\|Z^{(j)}\|_2^2/((\mathbf{X}^{(j)})^T Z^{(j)})^2$ which is the asymptotic variance of $\hat{b}_j/\sigma_\varepsilon$, assuming that the error in the bias correction is negligible.
3. Increase the variance by 25%, i.e., $V_j = 1.25\|Z^{(j)}\|_2^2/((\mathbf{X}^{(j)})^T Z^{(j)})^2$.
4. Search for the smallest λ_j such that the corresponding residual vector $Z^{(j)}(\lambda_j)$ satisfies:

$$\|Z^{(j)}(\lambda_j)\|_2^2/((\mathbf{X}^{(j)})^T Z^{(j)}(\lambda_j))^2 \leq V_j.$$

This procedure is similar to the choice of λ_j advocated in Zhang and Zhang (2014).

For a description about computational implementation of the de-biased or de-sparsified Lasso and its comparison to other methods, see also Section 10.4, we refer to Dezeure et al. (2015).

10.3.3 Asymptotic efficiency

We look at the question whether the de-sparsified estimator $\hat{b}_j$ for a single component is asymptotically efficient, reaching the smallest possible asymptotic variance, that is, the semiparametric efficiency bound (Bickel et al., 1998, cf.). To do so, we assume that $\mathbf{X}$ is random with rows being i.i.d. from a distribution with mean zero and covariance matrix Σ. Assume that

(C,j) $\Theta = \Sigma^{-1}$ exists and $0 < L \leq \Theta_{jj} \leq U\infty$ for some constants $0 < L < U < \infty$.

In the low-dimensional classical case with $p < n$ and $\text{rank}(\mathbf{X}) = p$, the semiparametric efficiency bound describing the smallest asymptotic variance with the class of regular estimators is $\Theta_{jj} = (\Sigma^{-1})_{jj}$. This is also known as the Cramér-Rao bound.

Denote the ℓ_0 and the scaled ℓ_2 sparsity of the jth row of Θ by

$$s_j = \sum_{k \neq j} I(\Theta_{jk} \neq 0),$$
$$t_j^2 = \sum_{k \neq j} \Theta_{jk}^2 / \Theta_{jj}^2.$$

The following then holds.

Theorem 10.2. *Consider $j \in \{1, \ldots, p\}$. Assume (B1), (C,j), $s_j = o(n/\log(p))$, $t_j^2 \leq C < \infty$ for some constant $C < \infty$ and that the restricted minimal eigenvalue for $(\mathbf{X}^{(-j)})^T \mathbf{X}^{(-j)}/n$ is bounded away from zero. Furthermore, assume that the distribution of the rows of $\mathbf{X}$ (each being the same) is sub-Gaussian. Then,*

$$\sqrt{n}(\hat{b}_j - \beta_j^0) = U_j + \Gamma_j,$$
$$U_j \sim \mathcal{N}(0, \sigma^2 \Theta_{jj}),\ \Gamma_j = o_P(1)\ (n \to \infty).$$

Proof. When conditioning on $\mathbf{X}$, we can invoke Theorem 10.1. The asymptotic variance equals

$$\sigma^2 \lim_{n \to \infty} \frac{\|Z^{(j)}\|_2^2/n}{|(Z^{(j)})^T \mathbf{X}^{(j)}/n|^2}.$$

The following holds:

$$(Z^{(j)})^T \mathbf{X}^{(j)}/n = \|Z^{(j)}\|_2^2/n + n^{-1}(Z^{(j)})^T \mathbf{X}^{(-j)} \gamma^{(j)} + n^{-1}(Z^{(j)})^T \mathbf{X}^{(-j)}(\hat{\gamma}^{(j)} - \gamma^{(j)}).$$

Using the KKT conditions as in (10.6) and $\|\gamma^{(j)}\|_1 \leq \sqrt{s_j}\|\gamma^{(j)}\|_2$, we have that

$$|n^{-1}(Z^{(j)})^T \mathbf{X}^{(-j)} \gamma^{(j)}| \leq 2\lambda_j \sqrt{s_j} \|\gamma^{(j)}\|_2,$$

and

$$|n^{-1}(Z^{(j)})^T \mathbf{X}^{(-j)}(\hat{\gamma}^{(j)} - \gamma^{(j)})| \leq 2\lambda_j \|\hat{\gamma}^{(j)} - \gamma^{(j)}\|_1.$$

Since $t_j = \|\gamma^{(j)}\|_2^2$, see (15.5), by assumption we have $\|gamma^{(j)}\|_2 \leq \sqrt{C} < \infty$ and that the compatibility constant for $\mathbf{X}^{(-j)}$ is bounded away from zero: due to sub-Gaussianity of $\mathbf{X}$ we then obtain $\|\hat{\gamma}^{(j)} - \gamma^{(j)}\|_1 \leq O_P(s_j \sqrt{\log(p)/n})$. Therefore, since $s_j = o(n/\log(p))$, we have that the asymptotic variance is behaving like

$$\sigma^2/(n^{-1}\|Z^{(j)}\|_2^2). \tag{10.11}$$

Now, again invoking the compatibility condition for $\mathbf{X}^{(-j)}$, the sparsity assumption for s_j and sub-Gaussianity of $\mathbf{X}$ we have that $\|Z^{(j)}\|_2^2/n = 1/\Theta_{jj} + o_P(1)$ and due to boundedness of Θ_{jj} from (C,j), this completes the proof. □

The conclusion $\|Z^{(j)}\|_2^2/n = 1/\Theta_{jj} + o_P(1)$ is satisfied whenever the Lasso is consistent for the prediction error: as discussed in the proof above, this holds assuming a restricted eigenvalue condition for $\mathbf{X}^{(-j)}$ and that the regression of $\mathbf{X}^{(j)} = \mathbf{X}^{(-j)}\gamma^{(j)} + \text{error}$ is sparse with $s_j = \|\gamma^{(j)}\|_0^0 = o(n/\log(p))$.

If this condition does not hold, for example if the the regression of $\mathbf{X}^{(j)}$ versus $\mathbf{X}^{(-j)}$ is not sparse, one can actually obtain a smaller variance of the de-sparsified Lasso. The following holds unders some conditions:

$$\sqrt{n}(\hat{b}_j - \beta_j^0) \Longrightarrow \mathcal{N}\left(0, \sigma^2 \frac{\Theta_{jj}}{1 + \Theta_{jj}(\mathbb{E}[|X^{(-j)}(\gamma^{(j)} - \gamma^{*,(j)})|^2])}\right),$$

where $\gamma^{*,j}$ is a specific sparse approximation of $\gamma^{(j)}$; and thus, if $\gamma^{(j)}$ is sparse, $\gamma^{*,(j)} = \gamma^{(j)}$ and the asymptotic variance equals $\sigma^2\Theta_{jj}$ as in Theorem 10.1. For further details we refer to van de Geer (2017). More general results for efficiency in high-dimensional model are discussed in Janková and van de Geer (2016).

10.4 Heteroscedastic errors, the bootstrap and some empirical results

We discuss here some extensions and conclude with a small empirical study.

10.4.1 Heteroscedastic errors

Theorem 10.1 does not hold for heteroscedastic errors, where $\varepsilon_1, \ldots, \varepsilon_n$ independent with $\text{Var}(\varepsilon_i) = \sigma_i^2$. In such a situation, the variance of the de-sparsified or de-biased Lasso in (10.4) asymptotically behaves as

$$\text{Var}(\sqrt{n}\frac{(Z^{(j)})^T\mathbf{X}^{(j)}}{n}\hat{b}_j) \asymp \text{Var}(n^{-1/2}\sum_{i=1}^n Z_i^{(j)}\varepsilon_i) = n^{-1}\sum_{i=1}^n (Z_i^{(j)})^2\sigma_i^2.$$

The quantity can be consistently estimated by

$$\hat{\omega}_j^2 := n^{-1}\sum_{i=1}^n (Z_i^{(j)}\hat{\varepsilon}_i - n^{-1}\sum_{r=1}^n Z_r^{(j)}\hat{\varepsilon}_r)^2$$

assuming e.g. that $\max_i |Z_i^{(j)}| \leq C < \infty$ and $\|\mathbf{X}(\hat{\beta} - \beta^0)\|_2^2/n = o_P(1)$, see also Bühlmann and van de Geer (2015). This is a version of the robust sandwich formula in presence of heterogeneous errors Huber (1967); White (1980). Similarly, we can estimate the covariance between $\sqrt{n}\frac{(Z^{(j)})^T\mathbf{X}^{(j)}}{n}\hat{b}_j$ and $\sqrt{n}\frac{(Z^{(k)})^T\mathbf{X}^{(j)}}{n}\hat{b}_j$ by the empirical covariance of $Z^{(j)} \circ \hat{\varepsilon}$ and $Z^{(k)} \circ \hat{\varepsilon}$, where "$\circ$" denotes the Hadamard product, i.e., $(a \circ b)_i = a_i b_i$ for two vectors a, b of the same dimension.

Theorem 10.1 can then be extended in a straightforward way for the case where the errors are independent Gaussian $\varepsilon \sim \mathcal{N}(0, \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2))$. We still invoke assumption (B1) and (B2,j) but assume in addition that the heteroscedasticity is such that $\|\hat{\beta} - \beta^0\|_1 = O_P(s_0\sqrt{\log(p)/n})$. The latter is a small extension of Corollary by simply assuming that $\sigma_i^2 \leq C < \infty$. One then obtains that

$$\hat{\omega}_j^{-1}\sqrt{n}\frac{(Z^{(j)})^T\mathbf{X}^{(j)}}{n}(\hat{b}_j - \beta_j^0) \Longrightarrow \mathcal{N}(0,1).$$

Analogously, for the multivariate case, the asymptotic covariance can estimated as indicated above.

10.4.2 The residual bootstrap

As described in Section 10.2, bootstrapping the Lasso does not lead to a consistent estimate of the underlying sampling distribution which in turn could be used for constructing confidence statements. The reason is, here discussed from another view point, that the bootstrap essentially only works for estimators having an asymptotic Gaussian distribution (Giné and Zinn, 1989, 1990), but the Lasso as a sparse estimaytor has also asymptoticaly point mass at zero (Knight and Fu, 2000). The de-sparsified or de-biased Lasso, however, has an asymptotic Gaussian distribution as discussed in Theorem 10.1. Therefore, the bootstrap is expected to consistently estimate its normalized sampling distribution. And indeed, this is the case as described next.

We consider a residual bootstrap. We use the Lasso for computing residuals $\hat{\varepsilon} = \mathbf{Y} - \mathbf{X}\hat{\beta}$ and centered residuals $\hat{\varepsilon}_{\mathrm{cent},i} = \hat{\varepsilon}_i - \overline{\hat{\varepsilon}}$ $(i = 1, \ldots, n)$, where $\overline{\hat{\varepsilon}} = n^{-1}\sum_{i=1}^n \hat{\varepsilon}_i$. The bootstrapped errors are then constructed as

$$\varepsilon_1^*, \ldots, \varepsilon_n^* \text{ i.i.d. (re-)sampled from the centered residuals } \hat{\varepsilon}_{\mathrm{cent},i} \ (i = 1, \ldots, n).$$

The bootstrapped response variables are constructed as

$$Y^* = \mathbf{X}\hat{\beta} + \varepsilon^*. \tag{10.12}$$

and the bootstrap sample is $\{(\mathbf{X}_i, Y_i^*)\}_{i=1}^n$, reflecting the fact of fixed (non-random) design. Here and in the sequel $\mathbf{X}_i$ denotes the $p \times 1$ row vectors of $\mathbf{X}$ $(i = 1, \ldots, n)$.

The bootstrapped estimator of the de-sparsified Lasso $\hat{b}_j$ and its corresponding (after appropriate scaling) robustly estimated standard deviation $\hat{\omega}$, or it s non-robust analogue $\hat{\sigma}$ are defined by the plug-in rule, where the estimators are computed from the bootstrap sample. We denote them by $\hat{b}_j^*, \hat{\omega}^*$ or $\hat{\sigma}^*$, respectively. We aim to estimate the distribution of the asymptotic pivot

$$T_j = \frac{\hat{b}_j - \beta_j^0}{\widehat{s.e.}_j},$$

where

$$\widehat{s.e.}_j = n^{-1/2}\hat{\omega}\frac{1}{|(Z^{(j)})^T\mathbf{X}^{(j)}/n|} \text{ for the robust version,} \tag{10.13}$$

$$\widehat{s.e.}_j = n^{-1/2}\hat{\sigma}\frac{\|Z^{(j)}\|_2/\sqrt{n}}{|(Z^{(j)})^T\mathbf{X}^{(j)}/n|} \text{ for the standard version.} \tag{10.14}$$

The bootstrap approximation is

$$T_j^* = (\hat{b}_j^* - \hat{\beta}_j)/\widehat{s.e.}_j^*,$$

and the quantiles of T_j^* will converge to the quantiles of T_j. Denote by $q_{j;\nu}^*$ the ν-quantile of the bootstrap distribution of T_j^*. We then construct two-sided $100(1-\alpha)\%$ confidence intervals for the jth coefficient β_j^0 as

$$\mathrm{CI}_j = [\hat{b}_j - q_{j;1-\alpha/2}^*\,\widehat{s.e.}_j, \hat{b}_j - q_{j;\alpha/2}^*\,\widehat{s.e.}_j]. \tag{10.15}$$

Bootstrapping pivots in classical low-dimensional settings is known to improve the level of accuracy of confidence intervals and hypothesis tests (Hall, 1992). For the high-dimensional case as discussed here, higher-order accuracy have not been worked out. Nevertheless, empirical results suggest that the bootstrap approach has an advantage over the normal approximation in Theorem 10.1, especially in presence of non-Gaussian errors ε. In addition, the bootstrap can also be used for approximating the distribution of $\max_{j\in G} T_j/\widehat{s.e.}_j$ which is useful for inference over large groups $G \subseteq \{1,\ldots,p\}$ and multiple testing adjustment. We refer to Dezeure et al. (2017) for further details.

Figure 10.1 displays some finite sample results from a simulation study with $n = 100$ and $p = 500$. The design matrix is generated as i.i.d. rows from a $\mathcal{N}_p(0,\Sigma)$ distribution with a Toeplitz covariance matrix where $\Sigma_{j,k} = 0.9^{|j-k|}$. The sparsity is chosen as $s_0 = 3$ and the active set is randomly sampled from $\{1,\ldots,p\}$. The non-zero regression coefficients are i.i.d. sampled from a Uniform$([-2,2])$ distribution. Finally, the errors are i.i.d. non-Gaussian from a scaled and centere χ_1^2 dsitribution, that is, $\varepsilon_i = \frac{\zeta_i - 1}{\sqrt{2}}$ with $\zeta_1,\ldots,\zeta_n$ i.i.d. $\sim \chi_1^2$. The bootstrap improves over the cases with the worst under-coverage. In addition, because the errors are i.i.d. and thus homoscedastic, there is not much by using the robust standard error. On the other hand (not shown here), for heteroscedastic errors, the robust version performs much better

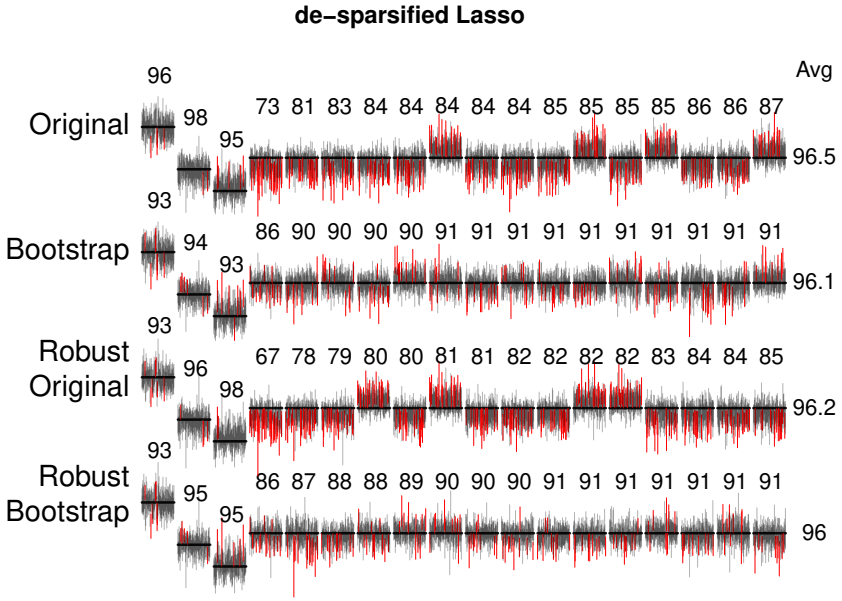

Fig. 10.1 Two-sided 95% confidence intervals for the de-sparsified Lasso estimator. From left to right 18 coefficients are shown with a black horizontal bar of a certain height illustrating the value of the coefficient. Only the first three coefficients differ from zero. The other 15 coefficients presented are those with the lowest confidence interval coverage for that particular method (in increasing order from left to right). 100 response vectors were generated for different realizations of the errors but with fixed design matrix $\mathbf{X}$. Each of these realizations leads to a confidence interval for each coefficient in the model. The 100 confidence intervals are drawn as vertical lines and ordered from left to right in the column corresponding to the particular underlying coefficient whose value is indicated by the horizontal bar. The line segments are colored black if they cover the true coefficient and colored red otherwise. The number above each coefficient corresponds to the number of confidence intervals, out of 100, which end up covering the truth. The average coverage probability over all coefficients is provided in a column to the right of all coefficients. The first two rows correspond to the case with the standard error as in (10.14, and the thrid and fourth row to the case with the robust standard error (10.13). The figure is taken from Dezeure et al. (2017).

in comparison to the one with the wrong (non-robust) standard error. In view of this, one should always use the robust standard error as it also works for heteroscedastic errors. One could also use the wild bootstrap to deal with heteroscedastic errors. Details can be found in Dezeure et al. (2017).

10.5 Extensions for generalized linear models

For generalized linear models, introduced in Chapter 3, one can de-sparsify or de-bias the Lasso estimator by considering the KKT conditions and using a regularized inversion of these: this approach has been described in van de Geer et al. (2014).

A simpler version of obtaining inferential statements is to use weighted regression. We restrict ourselves to the specific case of logistic regression: a more general treatment is given in Dezeure et al. (2015). Logistic regression is usually fitted by applying the iteratively reweighted least squares (IRLS) algorithm where at every iteration one solves a weighted least squares problem (Hastie et al., 2001). The idea is now to apply the Lasso for the logistic regression model, compute corresponding weights and then use the de-sparsified or de-biased Lasso on the transformed response and covariates.

Denote by $\hat{\pi}_i, i = 1, \ldots, n$ the estimated conditional probabilities for the binary responses, and $\hat{\pi}$ denotes the vector of these probabilities.

From Hastie et al. (2001), the adjusted response variable becomes

$$\mathbf{Y}_{\text{adj}} = \mathbf{X}\hat{\beta} + W^{-1}(Y - \hat{\pi}),$$

and the weighted least squares problem is

$$\hat{\beta}_{\text{new}} = \arg\min_{\beta}(\mathbf{Y}_{adj} - \mathbf{X}\beta)^T W(\mathbf{Y}_{\text{adj}} - \mathbf{X}\beta),$$

with diagonal weight matrix

$$W = (\hat{\pi}(1 - \hat{\pi})).$$

We rewrite, $Y_W = \sqrt{W}\mathbf{Y}_{adj}$ and $X_W = \sqrt{W}\mathbf{X}$. Note that one then obtains the parameter estimate

$$\hat{\beta}_{\text{new}} = \arg\min_{\beta} \|\mathbf{Y}_w - \mathbf{X}_w\beta\|_2^2/n$$

from regression of Y_W versus X_W. This motivates to use the de-sparsified or de-biased Lasso for the response Y_W versus X_W and then to obtain p-values and confidence regions as described before (but based on the weighted data $\mathbf{Y}_W$ and $\mathbf{X}_W$). This approach is detailed in Dezeure et al. (2015).

10.6 Related and alternative methods

The idea of de-biasing the Lasso in linear models has also been proposed by using a vector $Z^{(j)}$ in (10.5) which arises from Ridge regression, that is $Z^{(j)}$ is essentially the jth row vector of $(\mathbf{X}^T\mathbf{X}+\lambda_X I)^{-1}\mathbf{X}^T$ (Bühlmann, 2013). There is no such elegant theory as for the de-biased Lasso in Theorem 10.1 and 10.2. Empirically, the method seems rather reliable for type I error control over a variety of design matrices while it pays a price in terms of efficiency (Dezeure et al., 2015).

Javanmard and Montanari (2014) propose to choose the vector $Z^{(j)}$ in (10.5) from optimizing the variance of the estimator under a newr orthogonality constraint. We see from (10.11) that the asymptotic variance of the de-sparsified estimator $\hat{b}_j$ behaves as

$$\sigma^2/(n^{-1}\|Z^{(j)}\|_2^2).$$

Futhermore, the proof of Lemma 10.1 reveals that the de-sparsified Lasso satisfies a near-orthogonality constraint, due to the KKT conditions from the Lasso:

$$|(Z^{(j)})^T\mathbf{X}^{(k)}/n| \leq 2\lambda_X. \tag{10.16}$$

From this perspective, one can proceed to find a vector $Z^{(j)}$ which maximizes $\|Z^{(j)}\|_2^2/n$ under the constraint (10.16). This can be done using a convex program, as advocated by Javanmard and Montanari (2014), and following the argumentaion above, the estimator should exibit good efficiency. In numerical studies, however, the procedure seems to be "over-optimized" and does not reliably control the type I error (Dezeure et al., 2015). In fact, the idea of choosing a reasonable λ_X for the de-sparsified Lasso as described at the end of Section 10.3.2 is going against the idea of optimizing the variance: instead, it takes the view point that a somewhat larger variance of the estimator leads to more reliable type I error control.

For Gaussian graphical models, the idea and "philosophy" of the de-biased Lasso for linear models has been adapted and worked out in Ren et al. (2015) and Janková and van de Geer (2017), thereby relying on nodewise regression as discussed in Section 15.4.2 in Chapter 15.

Alternative methods rely on subsampling with corresponding inferential statements such as p-values or confidence regions, see Chapter 11, or geared towards stability with controling the expected number of false positive selections as dicusssed in Chapter 12). These methods can be used for a broad variety of models: the price for this generality includes a decrease in efficiency (power) and theoretical justifications which assume stronger (sufficient) conditions.

Problems

10.1. Prove that formula (10.2) holds.

10.2. Prove the analogue of the first statement in Theorem 10.1 by using the square root Lasso (Belloni et al., 2011), described in Section 2.13, for the construction of $Z^{(j)}$ without assuming condition (B2,j).

10.3. (i) Prove that the correction of p-values in (10.9) asymptotically controls the familywise error rate, saying that $\limsup_{n\to\infty} \mathbf{P}[V > 0] \le \alpha$.

(ii) Consider the situation of testing $H_{0,G_1}, \ldots H_{0,G_m}$ for various groups $G_r \subseteq \{1,\ldots,p\}$ $(r = 1,\ldots,m)$ and denote by $G := \cup_{r=1}^{m} G_r$. Assume that the statistical test for H_{0,G_r} is based on the statistic

$$\max_{j\in G_r} \hat{\sigma}^{-1} \sqrt{n} \frac{n^{-1}(\mathbf{X}^{(j)})^T Z^{(j)}}{n^{-1/2}\|Z^{(j)}\|_2} |\hat{b}_j|,$$

with resulting p-values P_{G_r}. Denote the corrected p-values by

$$P_{\mathrm{corr},G_r} = 1 - F_G(P_{G_r}).$$

Show that this correction asymptotically controls the familywise error rate.