Confidence Intervals and Tests for High-Dimensional Models: A Compact Review

Peter Bühlmann

Abstract We present a compact review of methods for constructing tests and confidence intervals in high-dimensional models. Links to theory, finite sample performance results and software allows to obtain a "quick" but sufficiently deep overview for applying the procedures.

1 Introduction

We review some methods for assigning significance of (co-)variables or for confidence intervals of a parameter in a high-dimensional regression-type model. Our major focus is for a high-dimensional linear model

$$Y = \mathbf{X}\beta^0 + \varepsilon \tag{1}$$

with $n \times 1$ response vector Y, $n \times p$ design matrix $\mathbf{X}$, $p \times 1$ regression vector β^0 and $n \times 1$ error vector ε having i.i.d. components with $\mathbb{E}[\varepsilon_i] = 0$, $\mathrm{Var}(\varepsilon_i) = \sigma_\varepsilon^2$ and ε_i uncorrelated from $\mathbf{X}_i$. We also discuss some extensions, including generalized linear models. While there is much literature on convergence rates for parameter estimation and prediction (cf. [6]), only recent work addresses the problem of constructing confidence intervals or tests. Some recent reviews on this topic include Bühlmann et al. [5] with a focus on applications in biology, and Dezeure et al. [8] who present a much more detailed and broader treatment. The current work aims to provide a very compact and "fast to read" access to the topic, yet it still contains the main ideas and hints to software.

P. Bühlmann (✉)
Seminar for Statistics, ETH Zürich, Zürich, Switzerland
e-mail: buhlmann@stat.math.ethz.ch

© Springer International Publishing Switzerland 2015

A. Antoniadis et al. (eds.), *Modeling and Stochastic Learning for Forecasting in High Dimensions*, Lecture Notes in Statistics 217,
DOI 10.1007/978-3-319-18732-7_2

2 High-Dimensional Linear Model and Some Methods for Inference

Consider the high-dimensional linear model in (1). The goal is to test null-hypotheses $H_{0,j}: \beta_j^0 = 0$ versus $H_{A,j}: \beta_j^0 \neq 0$ (or a one-sided alternative) for individual variables with index $j \in \{1, \dots, p\}$, or to construct a confidence interval for β_j^0. In the high-dimensional setting, these tasks are non-trivial since standard least squares methodology cannot be used.

2.1 De-sparsified Lasso

Zhang and Zhang [26] propose a method based on low-dimensional regularized projection using the Lasso. A motivation can be derived from standard least squares: in the low-dimensional setting with $p < n$ and $\mathbf{X}$ having full rank, it is well-known that the ordinary least squares estimator satisfies:

$$\hat{\beta}_{\mathrm{OLS},j} \text{ is the projection of } Y \text{ onto the residuals of } Z_{\mathrm{OLS,j}},$$

where the $n \times 1$ residual vector $Z_{\mathrm{OLS,j}}$ arises from OLS regression of X_j versus all other co-variables $\mathbf{X}_{-j}$ (which is the design matrix without the jth column). In the high-dimensional setting, the projection is ill-defined since the residual vector $Z_{\mathrm{OLS,j}} \equiv 0$. The idea is to replace the residuals by a regularized version: we fit X_j versus $\mathbf{X}_{-j}$ with the Lasso and denote the corresponding residuals by Z_j (when doing this for all j's, this is the nodewise Lasso from Meinshausen and Bühlmann [18]). We then look at the projection

$$Z_j^T Y / Z_j^T X_j = \beta_j^0 + \sum_{k \neq j} \beta_k^0 Z_j^T X_k / Z_j^T X_j + Z_j^T \varepsilon / Z_j^T X_j.$$

The first term on the right-hand side is what we aim for, the second one is a bias, and the third one is the noise component with mean zero. To get rid of the bias, we employ a bias correction using (again) the Lasso: this leads to a new estimator

$$\hat{b}_j = Z_j^T Y / Z_j^T X_j - \sum_{k \neq j} \hat{\beta}_k Z_j^T X_k / Z_j^T X_j \quad (j = 1, \dots, p), \tag{2}$$

where $\hat{\beta}$ denotes the Lasso estimator for the regression of Y versus $\mathbf{X}$. A typical choice for the regularization parameter involved in Z_j and for $\hat{\beta}$ is based on cross-validation of the corresponding Lasso estimations. The estimator $\hat{b}$ is not sparse and hence the name "de-sparsified Lasso". One can show that the error in bias estimation is asymptotically negligible [10, 24, 26] on the $1/\sqrt{n}$-scale, and one then obtains

$$\sqrt{n}(\hat{b}_j - \beta_j^0) \Rightarrow \mathcal{N}(0, \sigma_\varepsilon^2 \Omega_{jj}) \ (n \to \infty), \ \ \Omega_{jj} = \frac{\|Z_j\|_2^2/n}{(Z_j^T X_j/n)^2}. \tag{3}$$

The convergence as $n \to \infty$ encompasses that the dimension $p = p(n) \gg n$ tends to infinity as well, at a potentially much faster rate than sample size. We thus have an asymptotic pivot and we can then construct p-values for $H_{0,j}$ or confidence intervals by plugging in an estimate for σ_ε^2, see Sect. 2.3. In fact, the asymptotic variance is the smallest possible (among regular estimators) and it reaches the Cramér-Rao lower bound [24]: thus, statistical tests and confidence intervals derived from (3) are asymptotically optimal. Furthermore, the convergence in (3) to a Gaussian limit is uniform for a large part of the parameter space and thus, we obtain honest confidence intervals [11].

It is important to outline the assumptions which are used to establish the result in (3). Assume that the design $\mathbf{X}$ consists of (possibly fixed realizations of) i.i.d. rows whose distribution has a $p \times p$ covariance matrix Σ. The main conditions are as follows:

(A1) The rows of $\mathbf{X}$ have a (sub-)Gaussian distribution and the smallest eigenvalue of Σ is bounded away from zero.
(A2) The matrix Σ^{-1} is row-sparse: the maximal number of non-zero entries in each row is bounded by $o(\sqrt{n/\log(p)})$.
(A3) The linear model is sparse: the number of non-zero entries of β^0 is $o(\sqrt{n}/\log(p))$.
(A4) The error ε has a (sub-) Gaussian distribution.

We note that these assumptions imply the ones in van de Geer et al. [24]. The most restrictive conditions are (A2) regarding the design and (A3) saying that the linear model needs to be rather sparse.

2.2 *Ridge Projection*

The estimator in (2) is has a linear part and a non-linear bias correction. A similar construction can be made based on the Ridge estimator:

$$\hat{\beta}_{\text{Ridge}} = (n^{-1}\mathbf{X}^T\mathbf{X} + \lambda I)^{-1} n^{-1} \mathbf{X}^T Y. \tag{4}$$

A main message is that the Ridge estimator has substantial bias when $p \gg n$: in fact, it estimates a projected parameter

$$\theta^0 = P\beta^0, \ \ P = \mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-}\mathbf{X},$$

where $(\mathbf{X}\mathbf{X}^T)^{-}$ denotes a generalized inverse of $\mathbf{X}\mathbf{X}^T$ [22].

The bias for θ^0 can be made arbitrarily small by choosing λ sufficiently small, and a quantitative bound is given in Bühlmann [3]. A potentially substantial bias occurs, however, due to the difference between θ^0 and the target β^0. Since

$$\frac{\theta^0}{P_{jj}} = \beta_j^0 + \sum_{k \neq j} \frac{P_{jk}}{P_{jj}} \beta_k^0,$$

this bias can be estimated and corrected with

$$\sum_{k \neq j} \frac{P_{jk}}{P_{jj}} \hat{\beta}_k,$$

where $\hat{\beta}$ is the Lasso estimator. Thus, we construct a bias corrected Ridge estimator

$$\hat{b}_{R;j} = \frac{\hat{\beta}_{\text{Ridge};j}}{P_{jj}} - \sum_{k \neq j} \frac{P_{jk}}{P_{jj}} \hat{\beta}_k,\ j = 1, \ldots, p. \tag{5}$$

A typical choice of the regularization parameter in (4) for $\hat{\beta}_{\text{Ridge}}$ is $\lambda = \lambda_n = n^{-1}$ and we can use cross-validation for the regularization parameter in the Lasso $\hat{\beta}$. This estimator has the following property [3]:

$$\begin{aligned}
&\sigma_\varepsilon^{-1} \Omega_{R;jj}^{-1/2} (\hat{b}_{R;j} - \beta_j^0) \approx Z + \Delta_j,\ \ Z \sim \mathcal{N}(0,1),\\
&\Omega_R = (\hat{\Sigma} + \lambda)^{-1} \hat{\Sigma} (\hat{\Sigma} + \lambda)^{-1},\ \hat{\Sigma} = n^{-1} \mathbf{X}^T \mathbf{X},\\
&|\Delta_j| \leq \sigma_\varepsilon^{-1} \max_{k \neq j} \Omega_{R;jj}^{-1/2} \left| \frac{P_{jk}}{P_{jj}} \right| \|\hat{\beta} - \beta^0\|_1.
\end{aligned} \tag{6}$$

Here, the "$\approx$" symbol represents an approximation which becomes exact as $\lambda \searrow 0^+$. The problem here is that the behavior of $|P_{jk}/P_{jj}|$ and of the diagonal elements $\Omega_{R;jj}$ are not easily under control, but they are observed for fixed design $\mathbf{X}$ so that it is possible to construct an upper bound as discussed next.

2.2.1 Inference Based on an Upper Bound

Assuming the so-called compatibility condition on the design $\mathbf{X}$ [6, Ch.6.2], we obtain that

$$|\Delta_j| \leq \Omega_{R;jj}^{-1/2} \max_{k \neq j} \left| \frac{P_{jk}}{P_{jj}} \right| O_P(s_0 \sqrt{\log(p)/n}),$$

and in practice, we use an upper bound of the form

$$\Delta_j^{\text{bound}} := \Omega_{R;jj}^{-1/2} \max_{k \neq j} \left| \frac{P_{jk}}{P_{jj}} \right| (\log(p)/n)^{1/2 - \xi}, \tag{7}$$

for some small $0 < \xi < 1/2$, typically $\xi = 0.05$; this bound is motivated via an implicit assumption that $s_0 \leq (n/\log(p))^{\xi}$.

Inference can then be based on (6) with the upper bound in (7). For example, for testing $H_{0,j} : \beta_j^0 = 0$ against the two-sided alternative $H_{A,j} : \beta_j^0 \neq 0$ we use the upper bound for the p-value

$$2(1 - \Phi((\sigma_\varepsilon^{-1}\Omega_{R;jj}^{-1/2}|\hat{b}_{R;j}| - \Delta_j^{\text{bound}})_+)),$$

and an analogous construction can be used for a two-sided $1-\alpha$ confidence interval for β_j^0:

$$[\hat{b}_{R;j} - a, \hat{b}_{R;j} + a],$$
$$a = (\Phi^{-1}(1-\alpha/2) + \Delta_j^{\text{bound}})\sigma_\varepsilon\Omega_{R;jj}^{1/2}.$$

The main conditions used for proving consistency of the Ridge-based inference method are as follows:

(B1) As assumption (A1).
(B2) The linear model is sparse: for $0 < \xi < 1/2$ which is used in (7), the number of non-zero entries of β^0 is $O((n/\log(p))^{\xi})$.
(B3) The error ε has a Gaussian distribution.

It is expected that assumption (B3) could be relaxed to sub-Gaussian distributions as in (A4). No condition is required in terms of sparsity of Σ^{-1} as in (A2), but typically the method does not lead to optimality as with the de-sparsified Lasso estimator from Sect. 2.1.

2.3 *Estimation of the Error Variance*

The de-sparsified Lasso and the Ridge projection method in Sects. 2.1 and 2.2 require an estimate of σ_ε for construction of tests or confidence intervals.

The scaled Lasso [23] leads to a consistent estimate of the error variance: it is a fully automatic method which does not need a user-specific choice of a tuning parameter. Reid et al. [21] present an empirical comparison of various estimators which suggests that the alternative scheme of residual sum of squares of a cross-validated Lasso solution exhibits has good finite-sample performance.

2.4 *Multi Sample Splitting*

Sample splitting is a generic method for construction of p-values. The sample is randomly split in two halves with corresponding indices from disjoint sets $I_1, I_2 \subset$

$\{1,\ldots,n\}$, $I_1 \cup I_2 = \{1,\ldots,n\}$ with $|I_1| = \lfloor n/2 \rfloor$ and $|I_2| = n - \lfloor n/2 \rfloor$. A variable selection technique $\hat{S} \subseteq \{1,\ldots,p\}$ is used on the first half I_1, denoted by $\hat{S}(I_1)$: a prime example is the Lasso where $\hat{S} = \{j;\ \hat{\beta}_j \neq 0\}$, and other selectors $\hat{S}$ can be derived from a sparse estimator in the same way. With the fewer variables from $\hat{S}$, we can obtain p-values based on the second half I_2 and using classical t-tests from ordinary least squares: that is, we only use the subsample $(Y_{I_2}, \mathbf{X}_{I_2,\hat{S}})$ of the data, with obvious notational meaning of the sub-indices. Such a procedure is implicitly contained in Wasserman and Roeder [25]. Sample splitting avoids that we would use the data twice for selection and inference which would lead to over-optimistic p-values.

It is rather straightforward to see that such a principle works if

$$\begin{aligned} &\hat{S}(I_1) \supseteq S_0 = \{j;\ \beta_j^0 \neq 0\}, \\ &|\hat{S}(I_1)| < n/2, \end{aligned} \tag{8}$$

where $\hat{S}(I_1)$ denotes the selector based on the subsample with indices I_1. Furthermore, multiple testing adjustment over all components $j = 1,\ldots,p$ (see Sect. 3.2) can be done in a powerful way, e.g., Bonferroni correction only needs an adjustment with a factor $|\hat{S}(I_1)|$ which is often much smaller than p. A drawback of the method is its severe sensitivity of how the sample is split: Meinshausen et al. [20] propose repeated splitting of the sample (multi sample splitting) and show how to combine the corresponding dependent p-values. The latter is of independent interest and the procedure is described below in Sect. 2.4.1.

Such a multi sample splitting method leads to p-values which are already adjusted for multiple testing, either for the familywise error rate or the false discovery rate. The main conditions which are required for the method are (8): when using the Lasso as a screening method (typically with either a cross-validated choice of λ or taking a fixed fraction of the variables entering the Lasso path first), they are implied by the following:

(C1) As assumption (A1).

(C2) beta-min assumption:

$$\min_{j \in S_0} |\beta_j^0| \gg \sqrt{s_0 \log(p)/n},$$

and $s_0 = o(n/\log(p))$ where $s_0 = |S_0|$ denotes the number of non-zero entries of β^0.

(C3) As assumption (A4).

The beta-min assumption in (C2) is rather unpleasant since, for example, we would like to find out with significance testing whether a regression coefficient is large or smallish (or zero): thus, an a-priori assumption excluding smallish coefficients is unpleasant. The condition can be somewhat relaxed to "zonal assumptions" which still require that there is a gap between large and smallish coefficients and restrict the number of smallish coefficients [4].

2.4.1 Aggregation of p-Values

With the multi sample splitting approach described above we obtain the following: for testing the null-hypothesis $H_{0,j}: \ \beta_j^0 \neq 0$, when repeating the sample splitting B times, we get p-values

$$P_j^{(1)}, \dots, P_j^{(B)}.$$

The problem, in general, is how to aggregate many p-values which can be arbitrarily dependent to a single p-value P_j. The following Lemma is very general and might be of interest in other problems.

Lemma 1 ([20]) *Assume that we have B p-values* $P^{(1)}, \dots, P^{(B)}$ *for testing a null-hypothesis* H_0*, i.e., for every* $b \in \{1, \dots, B\}$ *and any* $0 < \alpha < 1$, $\mathbb{P}_{H_0}[P^{(b)} \leq \alpha] \leq \alpha$*. Consider for any* $0 < \gamma < 1$ *the empirical* γ*-quantile of the values* $\{P^{(b)}/\gamma;\ b = 1, \dots, B\}$*:*

$$Q(\gamma) = \ \min\left(\textit{empirical } \gamma\textit{-quantile } \{P^{(1)}/\gamma, \dots, P^{(B)}/\gamma\}, 1\right).$$

Furthermore, consider a suitably corrected minimum value of $Q(\gamma)$ *over a range which is lower bounded by a positive constant* $\gamma_{\min}$*:*

$$P = \min\left((1 - \log(\gamma_{\min})) \min_{\gamma \in (\gamma_{\min}, 1)} Q(\gamma), 1 \right). \tag{9}$$

Then, both $Q(\gamma)$ *(for any* $\gamma \in (0, 1)$*) and P are conservative p-values satisfying for any* $0 < \alpha < 1$*:* $\mathbb{P}_{H_0}[Q(\gamma) \leq \alpha] \leq \alpha$ *or* $\mathbb{P}_{H_0}[P \leq \alpha] \leq \alpha$*, respectively.*

A simple generic aggregation rule is with $\gamma = 1/2$: multiply the raw p-values by the factor 2 and take the sample median. Potential power improvement is possible with an adaptive version searching for the best γ as in (9) but paying a price in terms of the factor $(1 - \log(\gamma_{\min}))$ (which e.g. is ≈ 3.996 for $\gamma_{\min} = 0.05$).

2.5 Stability Selection

Stability Selection [19] is an even (much) more generic method than the multi sample splitting from Sect. 2.4. It can be applied to any structure estimation problem such as edges in a graph: variable selection in a regression problem is a special case thereof which we discuss now a bit further.

As with multi sample splitting, we randomly split the sample in two halves with indices I_1 and I_2, respectively, and we consider a variable selection method $\hat{S} \subseteq \{1,\ldots,p\}$. The idea is to analyze the stability of $\hat{S}(I_1)$, based on the half-sample I_1, when subsampling the data, and in fact, we do not make any use of the other half of the sample I_2. Thus, denote by I^* a random subsample of size $\lfloor n/2 \rfloor$. We consider the event that a single variable j is selected by $\hat{S}(I^*)$ based on the subsample I^*, $j \in \hat{S}(I^*)$, and we compute its probability

$$\pi(j) = \mathbb{P}^*[j \in \hat{S}(I^*)].$$

In practice, this probability is computed based on $B \approx 100$ random subsamples and calculating empirical relative frequencies.

The main problem is to determine a threshold $1/2 < \tau_{\mathrm{thr}} \leq 1$ such that $\pi(j) \geq \tau$ implies that variable j is selected in a "stable way". This can be formalized as follows: denote by $V = |\cup_{j \in S_0^c} \{\pi(j) \geq \tau\}|$, that is, the number of false positive selections. Then, assuming some conditions as outlined below, the following formula holds [19]:

$$\mathbb{E}[V] \leq \frac{1}{2\tau_{\mathrm{thr}} - 1} \frac{q^2}{p}, \tag{10}$$

where $q \geq |\hat{S}(I^*)|$ (almost surely). For example, q can be specified as the top q variables of a ranking (or selection) scheme, e.g., the q variables having largest regression coefficients in absolute value (if there are fewer than q coefficients with non-zero values, just take all of them). For the Lasso based on the first half-sample, since it selects at most $\lfloor n/2 \rfloor$ variables, a good value of q might be in the range of $n/10$ to $n/3$.

The formula (10) can then be inverted to determine a threshold τ_{thr} for a given bound of $\mathbb{E}[V]$ and a given q (which specifies the selection method $\hat{S}$). For example, by tolerating $\mathbb{E}[V] \leq 5$, a specified $q = 30$ and $p = 1{,}000$ we choose

$$\tau_{\mathrm{thr}} = (1 + \frac{q^2}{p}\frac{1}{5})/2 = (1 + \frac{30^2}{1{,}000}\frac{1}{5})/2 = 0.59$$

and such a choice then satisfies $\mathbb{E}[V] \leq 5$. When using the tolerance bound $\mathbb{E}[V] \leq \alpha$, the corresponding threshold τ_{thr} leads to a procedure where

$$\mathbb{P}[V > 0] \leq \mathbb{E}[V] \leq \alpha,$$

and hence, with control of the familywise error rate.

The main assumptions for validity of (10) are here sketched only:

(D1) The selector $\hat{S}$ is performing better than random guessing.

(D2) An exchangeability condition holds implying that it is equally likely that a noise variable is selected by $\hat{S}$.

The formal assumptions are given in Meinshausen and Bühlmann [19]. In fact, assumption (D1) is a mild condition while (D2) is rather restrictive: however, it was shown empirically that formula (10) approximately holds even for scenarios where (D2) does not hold. Interestingly, a beta-min assumption as in (C2) is not required for Stability Selection.

2.6 A Summary of an Empirical Study

We briefly summarize the results from a fairly large empirical study in Dezeure et al. [8]. An overall conclusion is that the multi sample splitting and the Ridge projection method are often somewhat more reliable for familywise error control (type I error control) than the de-sparsified Lasso procedure; on the other hand, the de-sparsified Lasso has often (a bit) more power in comparison to multi sample splitting and Ridge projection. However, these findings depend on the particular case and they are not consistent among all considered settings. Figure 1 illustrates the familywise error control and power of various methods for 96 different scenarios, varying over different covariate designs, sparsity degrees and structure of active sets, and signal to noise ratios.

From a practical point of view, if one is primarily concerned about false positive statements, the multi sample splitting method might be preferable: especially for logistic linear models (see Sect. 3.1), the adapted version of multi sample splitting was found to be most "robust" for reliable error control.

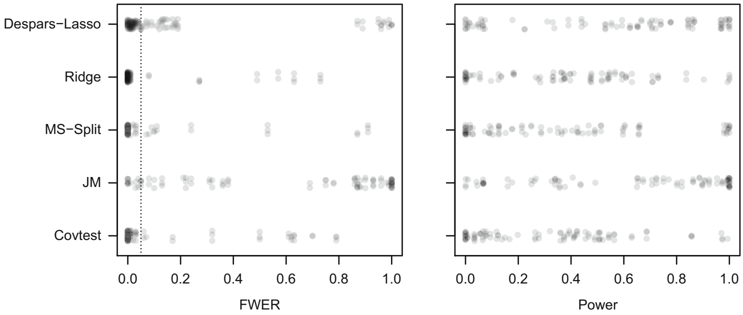

Fig. 1 Ninety-six different simulation scenarios, all with $p = 500$ and $n = 100$, with varying covariate design, sparsity and structure of the active set, and signal to noise ratio. Each *dot* represents a scenario, shown with jittered plotting. Five methods: De-sparsified Lasso (Despars-Lasso, as in Sect. 2.1), Ridge projection (Ridge, as in Sect. 2.2), Multi sample splitting (MS-Split, as in Sect. 2.4), a method from Javanmard and Montanari [10] (JM), covariance test from Lockhart et al. [13] (Covtest). *Left panel*: familywise error rate (FWER) with nominal level at 0.05 indicated by the *dotted line*; *right panel*: power (Power) representing the fraction of correctly identified active variables with non-zero regression coefficients. The figure is similar to some graphical representations in Dezeure et al. [8]

2.6.1 Supporting Theoretical Evidence and Discussion of Various Assumptions

Supporting evidence from theory, for the performance results in the empirical study, can be based by discussing the main assumptions underlying the different methods. The de-sparsified Lasso method is expected to work well and is most powerful if the design matrix is sparse in terms of its corresponding row-sparsity of Σ^{-1} (assumption (A2)) and if the linear model is rather sparse as well (assumption (A3)). The Ridge projection method does allow for designs with non-sparse rows of Σ^{-1}; however, the less restrictive assumption come with a price in that there is no optimality results in terms of power. The multi sample splitting method, which performs empirically quite reliably, has a theoretical drawback as it requires a zonal or the stronger beta-min assumption for the underlying regression coefficients (assumption (C2)); in terms of sparsity for the linear model, the multi sample splitting method is justified for a broader regime, allowing for $s_0 = o(n/\log(p))$ (assumption (C2)), than the required $s_0 = o(\sqrt{n}/\log(p))$ in assumption (A2) for the de-sparsified Lasso.

Stability Selection is controlling the number of false positives $\mathbb{E}[V]$ and not e.g. the familywise error rate (except when controlling $\mathbb{E}[V]$ at a very low level α which implies familywise error control at level α). The restrictive theoretical assumption is the exchangeability condition (D2): however, it seems that this condition is far from necessary. Stability Selection does not require a beta-min assumption as in (C2).

2.7 Other Methods

Very much related to the de-sparsified Lasso in Sect. 2.1 is a proposal by Javanmard and Montanari [10]. Their method is proved to be asymptotically optimal without requiring sparsity of the design as in condition (A2). Empirical evidence suggests though that the error control is not very reliable, see Fig. 1.

Bootstrap methods have been suggested to construct confidence intervals and p-values [7, 12]. They seem to work well for the components where the true parameter value equals zero, but they are often poor for the other components with non-zero parameters. Furthermore, multiple testing adjustment often requires a huge number of bootstrap replicates for reasonable computational approximation of tail events.

The covariance test [13] has been recently proposed as an "adaptive" method for assigning significance for the Lasso. Asymptotic validity of the test was shown under rather restrictive assumptions, in particular a restrictive beta-min assumption in the spirit of condition (C2). Empirical results of the covariance test are illustrated in Fig. 1, indicating that its power is comparably poor and error control is less reliable than for example for the Ridge projection or multi sample split method.

Another interesting proposal is due to Meinshausen [17]: we outline more details in Sect. 3.4.

3 Extensions and Further Topics

We briefly discuss here important extensions and additional issues.

3.1 *Generalized Linear Models*

Generalized linear models can be immediately treated with the multi sample splitting method or Stability Selection. Instead of e.g. the Lasso, we use ℓ_1-norm regularized maximum likelihood estimation for the selector $\hat{S}$, and low-dimensional inference (for the multi sample splitting method) is then based on maximum likelihood methodology.

The de-sparsified Lasso or the Ridge projection method are most easily adapted via additional weights as in iteratively reweighted least squares estimation [15]. The weights $w_i = w_i(\beta^0)$ $(i = 1, \ldots n)$ can be estimated by plugging in the ℓ_1-norm regularized maximum likelihood estimate; we can then proceed with new weighted data

$$\tilde{Y} = WY, \;\; \tilde{\mathbf{X}} = W\mathbf{X}, \;\; W = \mathrm{diag}(w_1, \ldots, w_n),$$

and apply the procedures from Sects. 2.1 and 2.2.

3.2 *Multiple Testing Correction*

Adjustment to multiple testing can be based using standard procedures which require valid p-values for individual tests as input: even under arbitrary dependence among the p-values, we can use e.g. the Bonferroni-Holm method for controlling the familywise error rate or the procedure from Benjamini and Yekutieli [1] to control the false discovery rate.

For the de-sparsified Lasso or Ridge projection method, one can use a simulation-based method which is less conservative than Bonferroni-Holm in presence of dependence: the details are given in Bühlmann [3].

We note that the multi sample splitting method from Sect. 2.4 as in the software package `hdi` (see Sect. 3.3) yields p-values which are adjusted for controlling the familywise error or false discovery rate.

3.3 R-Package `hdi`

The R-package `hdi` [16] contains implementations of various methods, namely the de-sparsified Lasso, the Ridge projection, the multi sample splitting method and of Stability Selection. We refer to to Dezeure et al. [8] how to use the procedures and what the various R-functions can do.

3.4 Testing Groups of Parameters

There might be considerable interest in testing the null-hypothesis $H_{0,G}: \ \beta_j^0 = 0$ for all $j \in G$, where $G \subseteq \{1, \ldots, p\}$ corresponds to a group of variables. The alternative is $H_{A,G}:$ there exists $j \in G$ with $\beta_j^0 \neq 0$.

Based on the de-sparsified Lasso or Ridge projection method, one can use a simulation-based procedure to obtain an approximate distribution of $\max_{j \in G} |\hat{b}_j|$ under the null-hypothesis $H_{0,G}$. We refer to Bühlmann [3] for the details. The multi sample splitting method can be modified for testing $H_{0,G}$, as described in Mandozzi and Bühlmann [14].

An interesting and very different proposal is given by Meinshausen [17] which can be used for testing individual but also groups of variables (and the latter is the main motivation in that work): the procedure does not even require an identifiability condition in terms of the design matrix **X** as it automatically determines whether a parameter or a group of parameters is identifiable.

3.5 Selective Inference

Especially with confidence intervals, one would typically report only for a few selected variables. An interesting approach to account for the selection effect, in terms of the false coverage rate, is presented in Benjamini and Yekutieli [2]. Their procedure can be applied for confidence intervals from e.g. the de-sparsified Lasso or the Ridge projection method from Sects. 2.1 or 2.2.

3.6 Some Thoughts on Bayesian Methods

For expository simplicity, consider a Gaussian linear model with Gaussian prior for the regression coefficients $\beta = (\beta_1, \ldots, \beta_p)$:

$$\begin{aligned} &\beta_1, \ldots, \beta_p \text{ i.i.d. } \sim \mathcal{N}(0, \tau^2), \\ &Y|\beta \sim \mathcal{N}_n(\mathbf{X}\beta, \sigma^2). \end{aligned} \tag{11}$$

The maximum a-posteriori estimator is then the Ridge estimator

$$\hat{\beta}_{\mathrm{MAP}} = \mathrm{argmin}_{\beta} \|Y - \mathbf{X}\beta\|_2^2/n + \frac{\sigma^2}{\tau^2 n}\|\beta\|_2^2.$$

For τ^2 large, this is the Ridge estimator with small regularization parameter λ as in Sect. 2.2.

Denote by β^* a realization from the prior distribution, and we are interested in constructing an interval which contains β^* with high probability. Alternatively, when adopting the frequentist Bayesian viewpoint (cf. [9]), we assume that the data is generated from a true parameter β^0, and we are interested to construct an interval which covers β^0 with high probability, based on a Bayesian model in (11). As discussed in Sect. 2.2, we know that for τ^2 large or σ^2 very small, $\hat{\beta}_{\mathrm{MAP}}$ is essentially unbiased for $\theta^* = P\beta^*$ (or $\theta^0 = P\beta^0$), where P is as in Sect. 2.2, but it can be severely *biased* for β^* (or β^0) in the high-dimensional scenario with $p \gg n$. Thus, the standard (Gaussian prior) Bayesian credible region centered around $\hat{\beta}_{\mathrm{MAP}}$ seems rather flawed for covering β^* or β^0 in the frequentist Bayesian paradigm.

Of course, in the classical Bayesian inference paradigm, such a bias does not occur, even when $p \gg n$, since the distribution of $\beta|Y$ is Gaussian with mean $\mathbb{E}[\beta|Y] = \hat{\beta}_{\mathrm{MAP}}$.

4 Conclusions

We provide a compact review of some methods for constructing tests and confidence intervals in high-dimensional models. The main assumptions underlying each method as well as a summary of empirical results are presented: this helps to understand, also from a comparative perspective, the strengths and weaknesses of the different approaches. Furthermore, a link to the R-package hdi is made. Thus, the user and practitioner obtains a "quick" but sufficiently deep overview for applying the procedures.

References

1. Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics, 29*, 1165–1188.
2. Benjamini, Y., & Yekutieli, D. (2005). False discovery rate-adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association, 100*, 71–81.
3. Bühlmann, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli, 19*, 1212–1242.
4. Bühlmann, P., & Mandozzi, J. (2013). High-dimensional variable screening and bias in subsequent inference, with an empirical comparison. *Computational Statistics*. Published online doi:10.1007/s00180-013-0436-3.

5. Bühlmann, P., Meier, L., & Kalisch, M. (2014). High-dimensional statistics with a view towards applications in biology. *Annual Review of Statistics and Its Applications, 1*, 255–278.
6. Bühlmann, P., & van de Geer, S. (2011). *Statistics for high-dimensional data: Methods, theory and applications*. Heidelberg/New York: Springer.
7. Chatterjee, A., & Lahiri, S. (2013). Rates of convergence of the adaptive LASSO estimators to the oracle distribution and higher order refinements by the bootstrap. *Annals of Statistics, 41*, 1232–1259.
8. Dezeure, R., Bühlmann, P., Meier, L., & Meinshausen, N. (2014). High-dimensional inference: Confidence intervals, p-values and software `hdi`. Preprint arXiv:1408.4026.
9. Diaconis, P., & Freedman, D. (1986). On the consistency of Bayes estimates (with discussion). *Annals of Statistics, 14*, 1–63.
10. Javanmard, A., & Montanari, A. (2013). Confidence intervals and hypothesis testing for high-dimensional regression. arXiv:1306.3171.
11. Li, K.-C. (1989). Honest confidence regions for nonparametric regression. *Annals of Statistics, 17*, 1001–1008.
12. Liu, H., & Yu, B. (2013). Asymptotic properties of lasso+mls and lasso+ridge in sparse high-dimensional linear regression. arXiv:1306.5505.
13. Lockhart, R., Taylor, J., Tibshirani, R., & Tibshirani, R. (2014). A significance test for the Lasso. *Annals of Statistics, 42*(2), 413–468.
14. Mandozzi, J., & Bühlmann, P. (2013). Hierarchical testing in the high-dimensional setting with correlated variables. arXiv:1312.5556.
15. McCullagh, P., & Nelder, J. (1989). *Generalized linear models* (2nd ed.). London: Chapman & Hall.
16. Meier, L. (2013). *hdi: High-dimensional inference*. R package version 0.0-1/r2.
17. Meinshausen, N. (2013). Assumption-free confidence intervals for groups of variables in sparse high-dimensional regression. arXiv:1309.3489.
18. Meinshausen, N., & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics, 34*, 1436–1462.
19. Meinshausen, N., & Bühlmann, P. (2010). Stability selection (with discussion). *Journal of the Royal Statistical Society Series B, 72*, 417–473.
20. Meinshausen, N., Meier, L., & Bühlmann, P. (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association, 104*, 1671–1681.
21. Reid, S., Tibshirani, R., & Friedman, J. (2013). A study of error variance estimation in Lasso regression. arXiv:1311.5274.
22. Shao, J., & Deng, X. (2012). Estimation in high-dimensional linear models with deterministic design matrices. *Annals of Statistics, 40*, 812–831.
23. Sun, T., & Zhang, C.-H. (2012). Scaled sparse linear regression. *Biometrika, 99*, 879–898.
24. van de Geer, S., Bühlmann, P., Ritov, Y., & Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics, 42*(3), 1166–1202.
25. Wasserman, L., & Roeder, K. (2009). High dimensional variable selection. *Annals of Statistics, 37*, 2178–2201.
26. Zhang, C.-H., & Zhang, S. (2014). Confidence intervals for low-dimensional parameters in high-dimensional linear models. *Journal of the Royal Statistical Society Series B, 76*, 217–242.