

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 $\mathbb{E}[\varepsilon_i] = 0$, $\text{Var}(\varepsilon_i) = \sigma_\varepsilon^2$ 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

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:

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

where the $n \times 1$ residual vector $Z_{\text{OLS},j}$ arises from OLS regression of X_j versus all other co-variables \mathbf{X}_{-j} (which is the design matrix without the j th column). In the high-dimensional setting, the projection is ill-defined since the residual vector $Z_{\text{OLS},j} \equiv 0$. 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

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

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

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

The convergence as $n \rightarrow \infty$ 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:

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

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

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

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

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

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]:

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

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

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

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

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

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

$$\begin{aligned} & [\hat{b}_{R_{:j}} - a, \hat{b}_{R_{:j}} + a], \\ & a = (\Phi^{-1}(1 - \alpha/2) + \Delta_j^{\text{bound}}) \sigma_\varepsilon \Omega_{R_{:j}}^{1/2}. \end{aligned}$$

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 $I_1, I_2 \subset$

$\{1, \dots, n\}$, $I_1 \cup I_2 = \{1, \dots, n\}$ with $|I_1| = \lfloor n/2 \rfloor$ and $|I_2| = n - \lfloor n/2 \rfloor$. A variable selection technique $\hat{S} \subseteq \{1, \dots, p\}$ 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 $j = 1, \dots, p$ (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

Modeling and Stochastic Learning for Forecasting in
High Dimensions

Antoniadis, A.; Poggi, J.-M.; Brossat, X. (Eds.)

2015, X, 339 p. 105 illus., 49 illus. in color., Softcover

ISBN: 978-3-319-18731-0