

# On Various Confidence Intervals Post-Model-Selection

Hannes Leeb, Benedikt M. Pötscher and Karl Ewald

*Abstract.* We compare several confidence intervals after model selection in the setting recently studied by Berk et al. [*Ann. Statist.* **41** (2013) 802–837], where the goal is to cover not the true parameter but a certain nonstandard quantity of interest that depends on the selected model. In particular, we compare the PoSI-intervals that are proposed in that reference with the “naive” confidence interval, which is constructed as if the selected model were correct and fixed a priori (thus ignoring the presence of model selection). Overall, we find that the actual coverage probabilities of all these intervals deviate only moderately from the desired nominal coverage probability. This finding is in stark contrast to several papers in the existing literature, where the goal is to cover the true parameter.

*Key words and phrases:* Confidence intervals, model selection, nonstandard coverage target, AIC, BIC, Lasso.

## 1. INTRODUCTION AND OVERVIEW

There is ample evidence in the literature that model selection can have a detrimental impact on subsequently constructed inference procedures like confidence sets, if these are constructed in the “naive” way where the presence of model selection is ignored. Such results are reported, for example, by Brown (1967); Buehler and Feddersen (1963); Dijkstra and Veldkamp (1988); Kabaila (1998, 2009); Kabaila and Leeb (2006); Leeb (2006); Leeb and Pötscher (2003, 2005, 2006a, 2006b, 2008a, 2008b); Olshen (1973); Pötscher (1991, 2006); Pötscher and Leeb (2009); Pötscher and Schneider (2009, 2010, 2011); Sen (1979); Sen and Saleh (1987).

Recently, Berk et al. (2013) proposed a new class of confidence intervals, so-called PoSI-intervals, which

correct for the presence of model selection, in the sense that these intervals guarantee a user-specified minimal coverage probability, even if the model has been selected in a data-driven way. However, the setting of Berk et al. (2013) differs from earlier studies, in that they consider confidence intervals for *a different quantity of interest*: In the aforementioned analyses, the quantity of interest (the coverage target) is always a fixed parameter or subparameter of the data-generating model. In Berk et al. (2013), on the other hand, a different and nonstandard coverage target is considered that depends on the selected model. (Even if an overall correct model is assumed, that nonstandard coverage target does *not* coincide with a parameter in the model, except for degenerate and trivial situations.) By design, the PoSI-intervals hence do not provide a solution to the more traditional problem, where the goal is to cover a parameter in the overall model after model selection.

Berk et al. (2013) motivate the need for PoSI-intervals by the poor performance of the “naive” interval as observed in the studies mentioned in the first paragraph of this section. However, these studies do *not* deal with the performance of the “naive” procedures post-model-selection when the coverage target is as in Berk et al. (2013). This raises the question of how the “naive” interval performs when it is used to cover the coverage target considered in Berk et al. (2013).

---

Hannes Leeb is Professor, Department of Statistics, University of Vienna, Oskar-Morgenstern-Platz 1, 1090 Vienna, Austria (e-mail: [hannes.leeb@univie.ac.at](mailto:hannes.leeb@univie.ac.at)).

Benedikt M. Pötscher is Professor, Department of Statistics, University of Vienna, Oskar-Morgenstern-Platz 1, 1090 Vienna, Austria (e-mail: [benedikt.poetscher@univie.ac.at](mailto:benedikt.poetscher@univie.ac.at)).

Karl Ewald is Ph.D. Candidate, Institute for Mathematical Methods in Economics, Vienna University of Technology, Argentinierstrasse 8/E105-2, 1040 Vienna (e-mail: [karl.ewald@tuwien.ac.at](mailto:karl.ewald@tuwien.ac.at)).

The main contribution of this paper is to answer this. In particular, we compare “naive” confidence intervals and PoSI-intervals in the setting of Berk et al. (2013). [The results in the present paper are partly based on Ewald (2012), and we refer to this thesis for additional results and discussion.]

We find that the minimal coverage probability of the “naive” interval is slightly below the nominal one, while that of the various PoSI intervals is slightly above, when the coverage target is as in Berk et al. (2013) and when AIC, BIC, or the LASSO are used for model selection. In the scenarios that we consider, the coverage probabilities of all these intervals are mostly within 10% of the nominal coverage probability. In the more traditional setting where the coverage target is a parameter in the overall model, however, all these intervals generally fail to deliver the desired minimal coverage probability. (Note that the various PoSI-intervals are not designed to deal with this coverage target.) For example, consider the scenario depicted by the solid curves in Figure 1 on page 222: There, a “naive” confidence interval post-model-selection with nominal coverage probability 0.95 has a minimal coverage probability of about 0.91 and the corresponding PoSI-interval has a minimal coverage probability of about 0.96, if the coverage target is as in Berk et al. (2013). But if the coverage target is a parameter in the overall model, the minimal coverage probabilities of the “naive” interval and of the PoSI-interval drop to about 0.56 and 0.62, respectively.

The paper is organized as follows: In Section 2 we introduce the data-generating process, the model-selection procedures, the coverage targets, and various confidence procedures, including the PoSI-intervals. We consider the same assumptions and constructions as in Berk et al. (2013), as well as some additional confidence intervals. The (minimal) coverage probabilities of “naive” intervals and of PoSI-intervals are studied in Sections 3 and 4. In particular, Section 3 contains an explicit finite-sample analysis of these procedures in a simple scenario with two nested candidate models. Section 4 contains a simulation study where we compare these intervals in three more complex scenarios; the first scenario is also studied by Kabaila and Leeb (2006), and the other two scenarios are taken from Berk et al. (2013). (The code used for the computations in Section 3 and for the simulations in Section 4 is available from the first author on request.) Finally, in the Appendix we present an example with a coverage target that is similar to, but slightly different from, that considered in Berk et al. (2013). The interesting feature of this example is that the “naive” confidence interval

here is valid, in the sense that its coverage probability is never below the nominal level.

## 2. COVERAGE TARGETS AND CONFIDENCE INTERVALS

Throughout, we consider a set of  $n$  homoskedastic Gaussian observations with mean vector  $\mu \in \mathbb{R}^n$  and common variance  $\sigma^2 > 0$ , that is,

$$(2.1) \quad y = \mu + u,$$

where  $u \sim N(0, \sigma^2 I_n)$ . We further assume that we have an estimator  $\hat{\sigma}^2$  for  $\sigma^2$  that is independent of all the least-squares estimators that will be introduced shortly. See Remark 2.1(ii) for some cautionary comments regarding our assumptions on  $\hat{\sigma}^2$ . For the estimator  $\hat{\sigma}^2$ , we either assume that it is distributed as a chi-squared random variable with  $r$  degrees of freedom multiplied by  $\sigma^2/r$ , that is,  $\hat{\sigma}^2 \sim \sigma^2 \chi_r^2/r$ , for some  $r \geq 1$ , or we assume that the variance is known a priori, in which case we set  $\hat{\sigma}^2 = \sigma^2$  and  $r = \infty$ . Unless noted otherwise, all considerations that follow apply to both the known-variance case and the unknown-variance case. The joint distribution of  $y$  and  $\hat{\sigma}$  depends on the parameters  $\mu \in \mathbb{R}^n$  and  $\sigma > 0$ , and will be denoted by  $\mathbb{P}_{\mu, \sigma}$ .

The available explanatory variables are represented by the columns of a fixed  $n \times p$  matrix  $X$ , where we allow for  $p > n$ ; again, see Remark 2.1(ii). We consider models where  $y$  is regressed on a (nonempty) subset of the regressors in  $X$ : For each model  $M \subseteq \{1, \dots, p\}$  with  $M \neq \emptyset$ , write  $X_M$  for the matrix of those columns of  $X$  whose indices lie in  $M$ . Writing  $M$  as  $M = \{j_1, \dots, j_{|M|}\} \subseteq \{1, \dots, p\}$ , we thus have  $X_M = (X_{j_1}, \dots, X_{j_{|M|}})$ , where  $X_j$  denotes the  $j$ th column of  $X$  and where  $|M|$  denotes the size of  $M$ . Write  $\mathcal{M}$  for a user-specified (nonempty) collection of candidate models. Throughout, we assume that  $\mathcal{M}$  consists only of submodels of full column rank, that is, we assume that the rank of  $X_M$  equals  $|M|$  and satisfies  $1 \leq |M| \leq n$  for each  $M \in \mathcal{M}$ .

Under a candidate model  $M \in \mathcal{M}$ ,  $y$  is modeled as

$$y = X_M \beta_M + v_M,$$

where  $\beta_M$  corresponds to the orthogonal projection of  $\mu$  from (2.1) onto the column-space of  $X_M$ , that is,  $\beta_M = (X_M' X_M)^{-1} X_M' \mu$ . The least-squares estimator corresponding to the model  $M$  will be denoted by  $\hat{\beta}_M$ , that is,  $\hat{\beta}_M = (X_M' X_M)^{-1} X_M' y$ . The working model  $M$  is correct if  $X_M \beta_M = \mu$ ; in that case, we have  $v_M = 0$ . Otherwise, that is, if  $X_M \beta_M \neq \mu$ , the working model is incorrect, and we have  $v_M = \mu - X_M \beta_M + u$ . Irrespective of whether the working model is correct or

not, we always have  $\hat{\beta}_M \sim N(\beta_M, \sigma^2(X'_M X_M)^{-1})$ ; in particular,  $\hat{\beta}_M$  is an unbiased estimator for  $\beta_M$ , irrespective of whether or not the model  $M$  is correct. As noted earlier, we assume that the variance estimator  $\hat{\sigma}^2$  is independent of the collection of estimators  $\hat{\beta}_M$  for  $M \in \mathcal{M}$ .

To pinpoint the regression coefficient of a given regressor  $X_j$  in a model  $M$  it appears in, we write  $\beta_{j \cdot M}$  for that component of  $\beta_M$  that corresponds to the regressor  $X_j$  for each  $j \in M$ . Similarly, the components of  $\hat{\beta}_M$  are indexed as  $\hat{\beta}_{j \cdot M}$  for  $j \in M$ . This convention is called “full model indexing” in Berk et al. (2013).

Consider now a model selection procedure, that is, a data-driven rule that selects a model  $\hat{M} \in \mathcal{M}$  from the pool  $\mathcal{M}$  of candidate models and the resulting post-model-selection estimator  $\hat{\beta}_{\hat{M}}$ . The coverage target considered in Berk et al. (2013) is  $\beta_{\hat{M}}$  or components thereof. Note that this coverage target is random, because it depends on the outcome of the model selection procedure.

REMARK 2.1. (i) At least one author of the present paper believes that the merits of  $\beta_{\hat{M}}$  as a coverage target for inference are debatable: For example, the meaning of the first coefficient of  $\beta_{\hat{M}}$  depends on the selected model and hence also on the training data  $(y, X)$ ; the same applies to the dimension of  $\beta_{\hat{M}}$ . In particular, we stress that different model selection procedures (e.g., AIC, BIC, the LASSO, etc.) lead to different targets  $\beta_{\hat{M}}$ . We refer to Berk et al. (2013) for further discussion and motivation for studying  $\beta_{\hat{M}}$ . These authors make the case for  $\beta_{\hat{M}}$  by arguing that the relevant setting is one where no correct overall model is available; however, in this situation the subsequent remark becomes especially important.

(ii) While the model (2.1) is nonparametric, the distributional requirements on  $\hat{\sigma}^2$  obviously are rather restrictive. However, these are the assumptions underlying the analysis in Berk et al. (2013), and we adopt them here in order to be in line with that reference. A leading case where these requirements are fulfilled is when (2.1) is replaced by the *parametric* model  $y = X\beta + u$ , when  $X$  is as before and is assumed to be of full column rank  $p < n$ , and when  $\hat{\sigma}^2$  is the usual unbiased variance estimator in that model and  $r$  is set to  $n - p$ . In this leading case, however, the true parameter  $\beta$  in the overall model is well-defined and will then typically be the prime target of statistical inference, rather than the nonstandard coverage target introduced in Berk et al. (2013). Outside of the parametric model just discussed, the requirements on  $\hat{\sigma}^2$  made in

Berk et al. (2013), and also here, will only be satisfied in certain special cases, some of which are discussed at the end of Section 2.2 in Berk et al. (2013). [The requirements on  $\hat{\sigma}^2$  are also fulfilled (with  $r = n - q$ ), if we would maintain a true parametric model  $y = Z\theta + u$  for some observed  $n \times q$  matrix  $Z$  of rank  $q < n$  that contains  $X$  as a submatrix; however, in this case one is back to the leading case discussed above, after redefining  $\mathcal{M}$  appropriately.]

In this paper, we will mainly focus on confidence intervals for the coefficient of one particular regressor in the selected model. Without loss of generality, assume that  $X_1$  is the regressor of interest and that the coverage target is  $\beta_{1 \cdot \hat{M}}$ . To ensure that this quantity is always well-defined, we assume that the first regressor  $X_1$  is contained in all candidate models under consideration, that is, we assume that  $1 \in M$  for each  $M \in \mathcal{M}$ . We seek to construct confidence intervals for  $\beta_{1 \cdot \hat{M}}$  that are of the form

$$\hat{\beta}_{1 \cdot \hat{M}} \pm K \hat{\sigma}_{1 \cdot \hat{M}}$$

for some constant  $K > 0$ , with  $\hat{\sigma}_{1 \cdot M}^2$  defined by  $\hat{\sigma}_{1 \cdot M}^2 = \hat{\sigma}^2[(X'_M X_M)^{-1}]_{1,1}$ , where  $[\cdot \cdot \cdot]_{1,1}$  denotes the first diagonal element of the indicated matrix. Here, we abuse notation and write  $a \pm b$  for the interval  $[a - b, a + b]$ . For a given level  $1 - \alpha$  with  $0 < \alpha < 1$ , the constant  $K$  should be chosen such that the minimal coverage probability is at least  $1 - \alpha$ , that is, such that

$$(2.2) \quad \inf_{\mu, \sigma} \mathbb{P}_{\mu, \sigma}(\beta_{1 \cdot \hat{M}} \in \hat{\beta}_{1 \cdot \hat{M}} \pm K \hat{\sigma}_{1 \cdot \hat{M}}) \geq 1 - \alpha.$$

Because the distribution of  $(\hat{\beta}_{1 \cdot M} - \beta_{1 \cdot M})/\hat{\sigma}_{1 \cdot M}$  is independent of unknown parameters and also independent of  $M$ , it follows, for *fixed*  $M$ , that a confidence interval for  $\beta_{1 \cdot M}$  with minimal coverage probability  $1 - \alpha$  is given by the textbook interval  $\hat{\beta}_{1 \cdot M} \pm K_N \hat{\sigma}_{1 \cdot M}$ , where  $K_N$  is the  $(1 - \alpha/2)$ -quantile of the distribution of  $(\hat{\beta}_{1 \cdot M} - \beta_{1 \cdot M})/\hat{\sigma}_{1 \cdot M}$ —a standard normal distribution in the known-variance case and a  $t$ -distribution with  $r$  degrees of freedom in the unknown-variance case. In view of this, it is tempting to consider, as a confidence interval for  $\beta_{1 \cdot \hat{M}}$ , the interval  $\hat{\beta}_{1 \cdot \hat{M}} \pm K_N \hat{\sigma}_{1 \cdot \hat{M}}$ . Because this construction ignores the model selection step and treats the selected model  $\hat{M}$  as fixed, we will call this the “naive” confidence interval.

The PoSI-interval developed in Berk et al. (2013) is obtained by first constructing simultaneous confidence intervals for the components of  $\beta_M$  that are centered at the corresponding components of  $\hat{\beta}_M$ , for each  $M \in$

$\mathcal{M}$ , with coverage probability  $1 - \alpha$ : More formally, the PoSI-constant  $K_P$  is the unique solution to

$$(2.3) \quad \inf_{\mu, \sigma} \mathbb{P}_{\mu, \sigma}(\beta_{j \cdot M} \in \hat{\beta}_{j \cdot M} \pm K_P \hat{\sigma}_{j \cdot M} : j \in M, M \in \mathcal{M}) = 1 - \alpha,$$

where the quantities  $\hat{\sigma}_{j \cdot M}^2$  are defined like  $\hat{\sigma}_{1 \cdot M}^2$  but with  $j$  replacing 1. By construction, the PoSI-constant  $K_P$  is such that we obtain simultaneous confidence intervals for the components of  $\hat{\beta}_{\hat{M}}$  that are centered at the corresponding components of  $\hat{\beta}_{\hat{M}}$ . In other words, (2.3) implies

$$(2.4) \quad \inf_{\mu, \sigma} \mathbb{P}_{\mu, \sigma}(\beta_{j \cdot \hat{M}} \in \hat{\beta}_{j \cdot \hat{M}} \pm K_P \hat{\sigma}_{j \cdot \hat{M}} : j \in \hat{M}) \geq 1 - \alpha.$$

In particular, (2.2) holds when  $K_P$  replaces  $K$ . For computing the constant  $K_P$ , we note that the probability in (2.3) can also be written as  $\mathbb{P}_{\mu, \sigma}(|\hat{\beta}_{j \cdot M} - \beta_{j \cdot M}|/\hat{\sigma}_{j \cdot M} \leq K_P : j \in M, M \in \mathcal{M})$ . This probability is not hard to compute, because it involves only the random variables  $(\hat{\beta}_{j \cdot M} - \beta_{j \cdot M})/\hat{\sigma}_{j \cdot M}$ , which are (dependent) standard normal in the known-variance case and (dependent)  $t$ -distributed in the unknown-variance case, with an obvious dependence structure only depending on  $X$ . In particular, the probability in (2.3) does not depend on  $\mu$  or  $\sigma^2$ . Similar considerations apply, mutatis mutandis, to the constant  $K_{P1}$  that is introduced in the following paragraph.

A modification of the preceding procedure, which is also proposed in Berk et al. (2013), is useful when inference is focused on a particular component of  $\hat{\beta}_{\hat{M}}$ , instead of on all components. Recall that the coverage target in (2.2) is the first component of  $\hat{\beta}_{\hat{M}}$ , that is,  $\beta_{1 \cdot \hat{M}}$ . The PoSI-constant  $K_{P1}$  provides simultaneous confidence intervals for  $\beta_{1 \cdot M}$  centered at  $\hat{\beta}_{1 \cdot M}$  for each  $M \in \mathcal{M}$ . In particular,  $K_{P1}$  is the unique solution to

$$(2.5) \quad \inf_{\mu, \sigma} \mathbb{P}_{\mu, \sigma}(\beta_{1 \cdot M} \in \hat{\beta}_{1 \cdot M} \pm K_{P1} \hat{\sigma}_{1 \cdot M} : M \in \mathcal{M}) = 1 - \alpha.$$

Again, by construction, (2.2) holds when  $K_{P1}$  replaces  $K$ .

Like the PoSI-constants discussed so far, other procedures for controlling the family-wise error rate can be used. Consider, for example, Scheffé's method: Recall that  $X$  denotes the matrix of all available explanatory variables, and note that  $(\hat{\beta}_{j \cdot M} - \beta_{j \cdot M})$  is a linear function of  $Y - \mu$ , that is, a function of the form

$v'(Y - \mu)$ , for a certain vector  $v \neq 0$  in the span of  $X$ . The Scheffé constant  $K_S$  is chosen such that

$$\mathbb{P}_{\mu, \sigma} \left( \sup_{\substack{v \in \text{span}(X) \\ v \neq 0}} \frac{v'(Y - \mu)}{\hat{\sigma} \|v\|} \leq K_S \right) = 1 - \alpha.$$

Then the relations (2.4) and, in particular, (2.2) hold when  $K_S$  replaces both  $K$  and  $K_P$ . Note that the probability in the preceding display does not depend on  $\mu$  and  $\sigma$ , and that the constant  $K_S$  is easily computed as follows: Let  $s$  denote the rank of  $X$ . In the known-variance case,  $K_S$  is the square root of the  $(1 - \alpha)$ -quantile of a chi-square distribution with  $s$  degrees of freedom. In the unknown-variance case,  $K_S$  is the square root of the product of  $s$  and the  $(1 - \alpha)$ -quantile of an  $F$ -distribution with  $s$  and  $r$  degrees of freedom.

Using the constants  $K_P$ ,  $K_{P1}$ , or  $K_S$  gives valid confidence intervals post-model-selection, that is, intervals that satisfy (2.2), because these constants give simultaneous confidence intervals for all quantities of interest that can occur; for example, (2.4) follows from (2.3), which in turn guarantees that (2.2) holds when  $K_P$  replaces  $K$ . One advantage of this is that a coverage probability of at least  $1 - \alpha$  is guaranteed, *irrespective* of the model selection procedure  $\hat{M}$  (as long as it takes values in  $\mathcal{M}$ ). In particular, this is guaranteed even if the model is selected by statistically inane methods like the SPAR-procedure mentioned in Section 4.9 of Berk et al. (2013). The price for this is that the PoSI constants  $K_P$  and  $K_{P1}$  may be overly conservative for a *particular* model selection procedure  $\hat{M}$ . [In this context, we note that equality holds in (2.4) for the SPAR-procedure, and that equality holds in (2.2) for a variant of the SPAR-procedure which selects that model  $\hat{M}$  which maximizes  $|\hat{\beta}_{1 \cdot M}|/\hat{\sigma}_{1 \cdot M}$  over  $M \in \mathcal{M}$ . Because such model selection procedures are hard to justify from a statistical perspective, we will not further consider SPAR and its variant here.]

Last, we will also consider the obvious approach where one chooses the smallest constant  $K$  such that (2.2) is satisfied. We will denote this constant by  $K_*$  (provided it exists). This is, of course, a well-known standard construction; see Bickel and Doksum (1977), page 170, for example. By definition, the interval in (2.2) with  $K_*$  replacing  $K$  is the shortest interval of that form whose minimal coverage probability is  $1 - \alpha$ . Note that  $K_*$  depends on the model selection procedure in question, and that computation of this quantity can be cumbersome as it requires computation of the finite-sample distribution of  $\hat{\beta}_{1 \cdot \hat{M}}/\hat{\sigma}_{1 \cdot \hat{M}}$ . However, explicit computation of this constant is feasible in some

cases [cf. the results in Section 3 and also the more general results of Leeb and Pötscher (2003)], and this constant can also be computed or approximated in a variety of other scenarios [e.g., by adapting the results of Pötscher and Schneider (2010) or the procedures of Andrews and Guggenberger (2009)]. Also note that we have  $K_* \leq K_{p1} \leq K_p \leq K_S$  by construction.

The procedures discussed so far are concerned with coverage targets like  $\beta_{\hat{M}}$  that depend on the selected model. This should be compared to the more classical parametric setting where the coverage target is the underlying true parameter: Assume that the data is generated by an overall linear model, that is, assume that the parameter  $\mu$  in (2.1) satisfies  $\mu = X\beta$  for the overall regressor matrix  $X$  introduced earlier, and that  $\text{rank}(X) = p < n$  holds. And assume that inference is focused on (components of) the parameter  $\beta$ . In this setting, the effect of model selection on subsequently constructed confidence intervals can be dramatic. For example, Kabaila and Leeb (2006) show that the minimal coverage probability of the “naive” confidence interval for  $\beta_1$ , that is, the quantity

$$\inf_{\beta, \sigma} \mathbb{P}_{X\beta, \sigma}(\beta_1 \in \hat{\beta}_{1, \hat{M}} \pm K_N \hat{\sigma}_{1, \hat{M}}),$$

can be much smaller than the nominal coverage probability  $1 - \alpha$ ; in fact, this minimal coverage probability can, for example, be smaller than 0.5, depending on the regressor matrix  $X$  in the overall model  $y = X\beta + u$ . The main reason for this more dramatic effect is that  $\hat{\beta}_{1, M}$  is a biased estimator for  $\beta_1$  whenever the model  $M$  is incorrect, whereas  $\hat{\beta}_{1, M}$  is always unbiased for  $\beta_{1, M}$ . Of course, valid confidence intervals post-model-selection can also be constructed when the coverage target is  $\beta_1$ , namely, by replacing  $K_N$  in the preceding display by the smallest constant  $K$  such that the resulting minimal coverage probability equals  $1 - \alpha$  (provided it exists). For the computation or approximations of this constant in particular situations, we refer to the papers cited in the preceding paragraph.

### 3. SOME FINITE-SAMPLE RESULTS

In this section we give a finite-sample analysis of the confidence intervals discussed so far, where we consider a simple model selection procedure that selects among two nested models using a likelihood ratio test. More precisely, maintaining the setting of Section 2, let  $X$  now be an  $n \times 2$  matrix of rank 2, and assume that  $\mathcal{M} = \{M_1, M_2\}$  with  $M_1 = \{1\}$  and  $M_2 = \{1, 2\}$  throughout this section. For the model selector, we set  $\hat{M} = M_2$  if  $|\hat{\beta}_{2, M_2}|/\hat{\sigma}_{2, M_2}$  is larger than  $C$ , and

$\hat{M} = M_1$  otherwise, where  $C > 0$  is a user-specified constant. Arguably, any reasonable model selection procedure in this setting must be equivalent to a likelihood ratio test, at least asymptotically; cf. Kabaila and Leeb (2006). In the numerical examples that follow, we will consider  $C = \sqrt{2}$ , such that the resulting model selector  $\hat{M}$  corresponds to selection by the classical Akaike information criterion (AIC); this model selector is asymptotically equivalent to several other model selectors, including the GCV model selection criterion of Craven and Wahba (1978/79) and the  $S_p$  criterion of Tukey (1967); cf. Leeb (2008). Furthermore, we will also consider  $C = \sqrt{\log(n)}$ , corresponding to the BIC model selection criterion. Throughout this section, let  $\phi(\cdot)$  and  $\Phi(\cdot)$  denote the density and the cumulative distribution function (c.d.f.) of the univariate standard Gaussian distribution, and set  $\Delta(x, c) = \Phi(x + c) - \Phi(x - c)$ . And, last, we will write  $\rho$  for the correlation coefficient between the two components of  $\hat{\beta}_{M_2}$ , that is,  $\rho = -[(X'_{M_2} X_{M_2})^{-1}]_{1,2} / ([X'_{M_2} X_{M_2})^{-1}]_{1,1} [X'_{M_2} X_{M_2})^{-1}]_{2,2}^{-1/2}$ .

The following result describes the coverage probability of the interval  $\hat{\beta}_{1, \hat{M}} \pm K \hat{\sigma}_{1, \hat{M}}$  in two scenarios, namely, when the coverage target is  $\beta_{1, \hat{M}}$  and when the coverage target is  $\beta_{1, M_2}$ . Note that, in case the model  $M_2$  is correct, that is, if we have  $\mu = X\beta$  for some  $\beta \in \mathbb{R}^2$ , and hence also  $y = X\beta + u$ , then this second scenario reduces to the classical parametric setting described at the end of Section 2; in particular, we then have  $\beta_{M_2} = \beta$  and, thus,  $\beta_{1, M_2} = \beta_1$ .

**PROPOSITION 3.1.** *In the setting of this section, we have*

$$\begin{aligned} & \mathbb{P}_{\mu, \sigma}(\beta_{1, \hat{M}} \in \hat{\beta}_{1, \hat{M}} \pm K \hat{\sigma}_{1, \hat{M}}) \\ &= \mathbb{E} \left[ \Delta \left( 0, \frac{\hat{\sigma}}{\sigma} K \right) \Delta \left( \zeta, \frac{\hat{\sigma}}{\sigma} C \right) \right. \\ & \quad \left. + \int_{-(\hat{\sigma}/\sigma)K}^{(\hat{\sigma}/\sigma)K} \left( 1 - \Delta \left( \frac{\zeta + \rho z}{\sqrt{1 - \rho^2}}, \frac{(\hat{\sigma}/\sigma)C}{\sqrt{1 - \rho^2}} \right) \right) \right. \\ & \quad \left. \cdot \phi(z) dz \right] \end{aligned}$$

and

$$\begin{aligned} & \mathbb{P}_{\mu, \sigma}(\beta_{1, M_2} \in \hat{\beta}_{1, \hat{M}} \pm K \hat{\sigma}_{1, \hat{M}}) \\ &= \mathbb{P}_{\mu, \sigma}(\beta_{1, \hat{M}} \in \hat{\beta}_{1, \hat{M}} \pm K \hat{\sigma}_{1, \hat{M}}) \\ & \quad + \mathbb{E} \left[ \left( \Delta \left( \frac{\rho \zeta}{\sqrt{1 - \rho^2}}, \frac{\hat{\sigma}}{\sigma} K \right) \right. \right. \\ & \quad \left. \left. - \Delta \left( 0, \frac{\hat{\sigma}}{\sigma} K \right) \right) \Delta \left( \zeta, \frac{\hat{\sigma}}{\sigma} C \right) \right], \end{aligned}$$

with  $\zeta = \beta_{2 \cdot M_2} / \text{SD}(\hat{\beta}_{2 \cdot M_2})$ , where  $\text{SD}(\cdot)$  denotes the standard deviation. The expectations on the right-hand sides are taken with respect to  $\hat{\sigma}/\sigma$ . In the known-variance case,  $\hat{\sigma}/\sigma$  is constant equal to one and the expectations are trivial; in the unknown-variance case,  $\hat{\sigma}/\sigma$  is distributed like the square root of a chi-squared distributed random variable with  $r$  degrees of freedom divided by  $r$ , that is,  $\hat{\sigma}/\sigma \sim \sqrt{\chi_r^2/r}$ .

PROOF. The statements for the known-variance case are simple adaptations of the finite-sample statements of Proposition 3 in Kabaila and Leeb (2006). For the unknown-variance case, it suffices to note that  $\hat{\sigma}/\sigma$  is independent of  $\{\hat{\beta}_{M_1}, \hat{\beta}_{M_2}\}$ . With this, the statements are then obtained by conditioning on  $\hat{\sigma}/\sigma$  and by using the formulae for the known-variance case derived earlier.  $\square$

Proposition 3.1 provides explicit formulas that also allow us to compute (minimal) coverage probabilities numerically. For the following discussion, fix the values of  $C$  and  $K$ , that is, the critical value  $C$  of the hypothesis test that is used for model selection and the value  $K$  that governs the length of the confidence interval post-model-selection. We first note that  $\mathbb{P}_{\mu, \sigma}(\beta_{1 \cdot M_2} \in \hat{\beta}_{1 \cdot \hat{M}} \pm K \hat{\sigma}_{1 \cdot \hat{M}})$  is strictly smaller than  $\mathbb{P}_{\mu, \sigma}(\beta_{1 \cdot \hat{M}} \in \hat{\beta}_{1 \cdot \hat{M}} \pm K \hat{\sigma}_{1 \cdot \hat{M}})$  whenever  $\rho \zeta \neq 0$ , because the two probabilities differ by a correction term (namely, the expected value on the right-hand side of the second display in Proposition 3.1) which is negative whenever  $\rho \zeta \neq 0$ . If  $\rho \zeta = 0$ , the two probabilities are equal. And if  $\rho = 0$ , it is easy to see that both probabilities are equal to  $\mathbb{E}[\Delta(0, K \hat{\sigma}/\sigma)] = F(K) - F(-K)$ , irrespective of  $\zeta$ , where  $F$  denotes the c.d.f. of a  $t$ -distribution with  $r$  degrees of freedom in the unknown-variance case and the standard Gaussian c.d.f. in the known-variance case. Next, we note that the coverage probabilities depend only on  $r$ ,  $\zeta$ , and  $\rho$ . (Recall that  $r$  denotes the degrees of freedom of  $\hat{\sigma}^2$  in the unknown-variance case, and that we have set  $r = \infty$  in the known-variance case.) Note that  $\zeta$  is a function of the regressor matrix  $X_{M_2}$  and of the unknown parameters  $\mu$  and  $\sigma^2$ , while  $\rho$  is a function of  $X_{M_2}$  only. Moreover, it is easy to see that the coverage probabilities are symmetric both in  $\zeta$  and in  $\rho$  around the origin. Concerning the influence of  $r$ , it can be shown that the coverage probabilities for the known-variance case provide a uniform approximation to those in the unknown-variance case, uniformly in the unknown parameters, where the approximation error goes to zero as  $r \rightarrow \infty$ ; this follows from the results of

Leeb and Pötscher (2003) using standard arguments. In the examples that follow, we found that the results for the known-variance case and for the unknown-variance case are similar, and that these results are visually hard to distinguish from each other, unless  $r$  is extremely small like, for example, 3. We therefore focus on the known-variance case in the following because it provides a good approximation to the unknown-variance case as long as  $r$  is not too small.

We proceed to comparing the case where the coverage target is  $\beta_{1 \cdot \hat{M}}$  as in Berk et al. (2013) with the more standard case where the coverage target is the parameter  $\beta_{1 \cdot M_2}$ , in terms of the coverage probabilities of confidence intervals post-model-selection. Recall that the nonstandard target depends on the training data as well as on the model selection procedure employed, whereas the standard target does not. Consider first the case where  $C = \sqrt{2}$ , corresponding to the AIC model selector. For several of the confidence intervals introduced in the preceding section, the results are visualized in Figure 1, for the case where the coverage target is  $\beta_{1 \cdot \hat{M}}$  (top panel) and for the case where the coverage target is  $\beta_{1 \cdot M_2}$  (bottom panel). Note that the range of the vertical axes (displaying coverage probability) in the two panels is quite different.

In each panel of Figure 1, we see that the effect of model selection on the resulting coverage probabilities depends on the correlation coefficient  $\rho$ , with larger values of  $\rho$  corresponding to smaller minimal coverage probabilities. But the strength of the effect varies greatly with the scenario, that is, on whether the coverage target is  $\beta_{1 \cdot \hat{M}}$  or  $\beta_{1 \cdot M_2}$ . When the coverage target is  $\beta_{1 \cdot \hat{M}}$  (top panel in Figure 1), we see that the effect of model selection is comparatively minor: The smallest coverage probabilities are always obtained for the “naive” interval, whose coverage probability here can be smaller as well as larger than the nominal 0.95. Irrespective of the true parameters, the actual coverage probability of the “naive” interval is quite close to the nominal one here. The other intervals, that is, the PoSI-, the PoSI-, and the Scheffé-interval, all have coverage probabilities larger than 0.95. (The minimal coverage probabilities here are obtained for  $\zeta = 0$ , but we found this not to be the case for other model selection procedures, i.e., for other values of  $C$ .) When the coverage target is  $\beta_{1 \cdot M_2}$  (bottom panel in Figure 1), however, we get a very different picture: For  $\rho = 0.9$ , the minimal coverage probability of all the intervals considered there is much smaller than 0.95, with minima between 0.55 (“naive”) and 0.65 (Scheffé). For  $\rho = 0.5$ , the minimal coverage probabilities of the

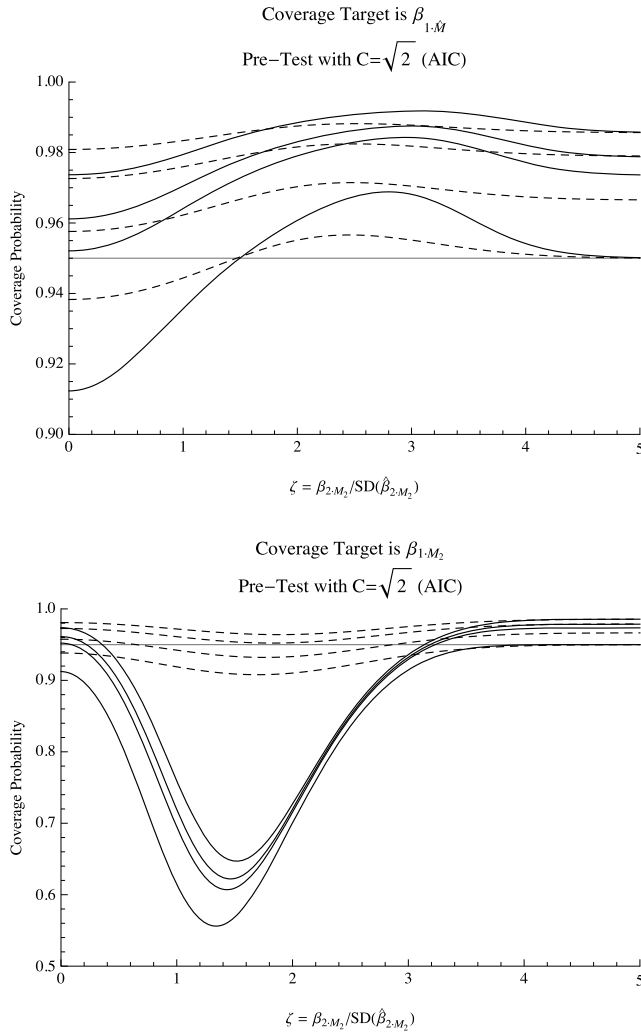


FIG. 1. Coverage probability of several confidence intervals in the known-variance case, as a function of the scaled parameter  $\zeta = \beta_{2.M_2}/SD(\hat{\beta}_{2.M_2})$ , using the model selection procedure with  $C = \sqrt{2}$ , that is, AIC. The nominal coverage probability is  $1 - \alpha = 0.95$ , indicated by a gray horizontal line. The coverage target is  $\beta_{1.\hat{M}}$  (top panel) and  $\beta_{1.M_2}$  (bottom panel). In each panel, the four solid curves are computed for  $\rho = 0.9$ , and the four dashed curves are for  $\rho = 0.5$ . The curves in each group of four are ordered: Starting from the top, the curves show the coverage probabilities for  $K_S$  (Scheffé),  $K_P$  (PoSI),  $K_{P1}$  (PoSI1), and  $K_N$  (naive).

“naive” interval and of the PoSI1-interval are below, while those of the other intervals are above, the nominal 0.95. For very small values of  $\rho$ , the coverage probabilities of all the intervals considered in Figure 1 are visually indistinguishable from horizontal lines as a function of  $\zeta$  (and hence are not shown here), irrespective of the coverage target. For  $\rho = 0.1$ , for example, the coverage probability of the “naive” interval is about 0.95, while that of the other intervals is above 0.95, ordered by their length. (This should not come

as a surprise since in case  $\rho = 0$  model selection has no effect on estimating the regression coefficients; furthermore, the two targets are identical in this case.)

Figure 1 illustrates that the coverage probability of confidence intervals post-model-selection depends crucially on whether the coverage target is  $\beta_{1.\hat{M}}$  as in Berk et al. (2013) or the more classical coverage target  $\beta_{1.M_2}$ . We stress here again that the PoSI-intervals and the Scheffé-interval have not been designed to deal with the case where the coverage target is  $\beta_{1.M_2}$ . For a more detailed analysis of the “naive” interval in the case where the coverage target is  $\beta_{1.M_2}$ , we refer to Kabaila and Leeb (2006).

For the other values of  $C$  that we consider, that is, for  $C = \sqrt{\log(n)}$  for various values of  $n$ , we found the following: When the coverage target is  $\beta_{1.\hat{M}}$ , the results are very similar to those shown in the top panel of Figure 1. To conserve space, we do not show these results here. When the target is  $\beta_{1.M_2}$ , the resulting curves are of the same shape but steeper, with coverage probabilities decreasing as  $C$  increases. This is so because larger values of  $C$  lead to more frequent selection of the smaller model  $M_1$ , causing more bias in the resulting post-model-selection estimator; we refer to Leeb and Pötscher (2005) and, in particular, Figure 3 in that reference, for further discussion and analysis of this phenomenon.

We next compare the confidence intervals for  $\beta_{1.\hat{M}}$  introduced in Section 2 through their minimal coverage probability as a function of the correlation coefficient  $\rho$ . In particular, for various values of  $C$ , we compute the quantity on the left-hand side of (2.2) for specific  $K$ 's, namely, for  $K_N$  (“naive”), for  $K_P$  (PoSI), for  $K_{P1}$  (PoSI1), for  $K_S$  (Scheffé), and for  $K_*$  (the smallest valid  $K$ ). By construction, we have  $K_* \leq K_{P1} \leq K_P \leq K_S$ , so that the resulting curves of minimal coverage probabilities are also arranged in increasing order.

All the minimal coverage probabilities shown in Figure 2 are within 5% of the nominal level 0.95. For the “naive” intervals corresponding to  $K_N$  (the first four curves from the bottom), the minimal coverage probability is below 0.95 (except for the trivial case where  $\rho = 0$ ), but not by much. The intervals with  $K_*$  have minimal coverage probabilities of exactly 0.95, for every value of  $C$ , by construction (but note that  $K_*$  depends on  $C$ , whereas  $K_S, K_P, K_{P1}$ , and  $K_N$  do not). Hence, the curves corresponding to the  $K_*$ 's for the four values of  $C$  considered here are constant and sit on top of each other. And, again by construction, all other intervals are slightly too large in the sense that

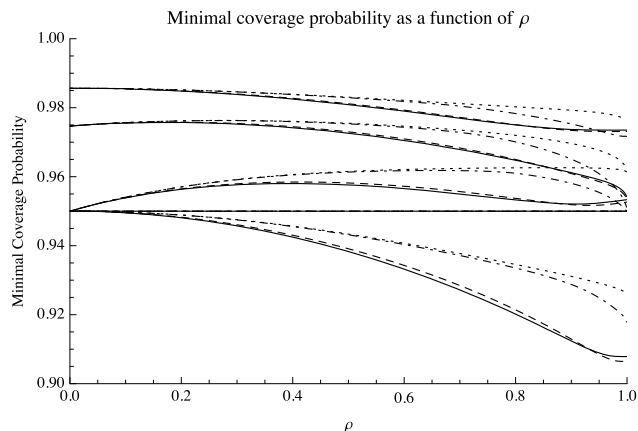


FIG. 2. Minimal coverage probabilities of the confidence intervals for  $\beta_{1,\hat{M}}$  as a function of  $\rho$  in the known-variance case, for  $C = \sqrt{2}$  (solid curves),  $C = \sqrt{\log(10)}$  (dashed curves),  $C = \sqrt{\log(100)}$  (dot-dashed curves), and  $C = \sqrt{\log(1000)}$  (dotted curves). The nominal coverage probability is  $1 - \alpha = 0.95$ . For each value of  $C$ , the corresponding five curves are ordered: Starting from the top, the curves correspond to the intervals with  $K_S$ ,  $K_P$ ,  $K_{P1}$ ,  $K_*$ , and  $K_N$ .

their minimal coverage probability exceeds the nominal level 0.95. Concerning the influence of  $C$ , we see that larger values of  $C$  correspond to slightly larger minimal coverage probabilities for the intervals corresponding to  $K_N$ ,  $K_{P1}$ ,  $K_P$ , and  $K_S$ , and for most values of  $\rho$ ; it should be noted, however, that—in contrast to the case of the standard target—here the target changes with  $C$ . Overall, the difference between the coverage probabilities of all these intervals is not dramatic.

Last, we compare the confidence intervals for  $\beta_{1,\hat{M}}$  through the values of the constants  $K$  that correspond to the intervals in question. By construction,  $K_S$  and  $K_N$  are constant as a function of  $\rho$ . Note that the constants  $K_N$ ,  $K_P$ ,  $K_{P1}$ , and  $K_S$  do not depend on the model selection procedure that is being used (and thus not on  $C$ ), while the constant  $K_*$  does depend on the model selection procedure (and thus on  $C$ ). For a given model selection procedure, the constant  $K_*$  is the smallest number  $K$  for which (2.2) holds; in particular, the interval corresponding to  $K$  has minimal coverage probability smaller/equal/larger than  $1 - \alpha$  if and only if  $K$  is smaller/equal/larger than  $K_*$ .

The interpretation of Figure 3 is similar to that of Figure 2, the main difference being that the lengths considered here are somewhat more distorted than the minimal coverage probabilities considered earlier. The “naive” interval is up to about 10% too short, while the intervals corresponding to  $K_{P1}$ ,  $K_P$ , and  $K_S$  are too

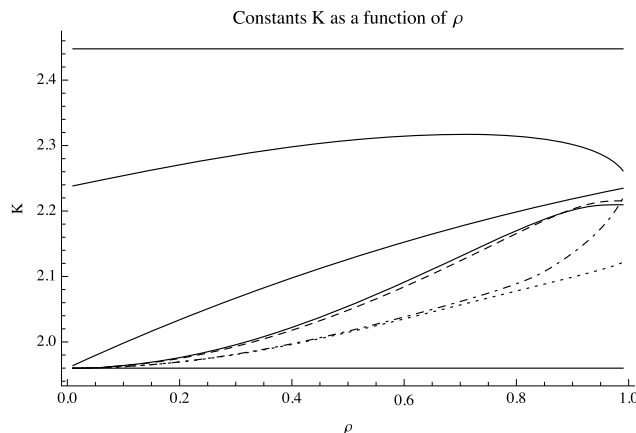


FIG. 3. The constants  $K$  that govern the width of the confidence intervals as a function of  $\rho$  in the known-variance case, using the model selection procedure with critical value  $C$ . The nominal coverage probability is  $1 - \alpha = 0.95$ . Starting from the top, the five solid curves show  $K_S$ ,  $K_P$ ,  $K_{P1}$ ,  $K_*$  for  $C = \sqrt{2}$  (AIC), and  $K_N$ . The remaining curves show  $K_*$  for  $C = \sqrt{\log(10)}$  (dashed curve), for  $C = \sqrt{\log(100)}$  (dash-dotted curve), and for  $C = \sqrt{\log(1000)}$  (dotted curve).

long, namely, by up to about 5%, 15%, 25%, respectively. We also see that  $K_*$  decreases as  $C$  increases for most values of  $\rho$ , which is consistent with the observations made in the second-to-last paragraph.

#### 4. SIMULATION STUDY

We now compare the “naive” interval, the PoSI interval, and (a variant of) the PoSI interval for  $\beta_{1,\hat{M}}$  by their respective minimal coverage probabilities in a simulation study where the data are generated from a Gaussian overall linear model  $M_{\text{full}}$ , say, of the form  $Y = X\beta + u$  with 30 observations, 10 explanatory variables, and i.i.d. standard normal errors. Moreover, we also study these intervals when the coverage target is  $\beta_1 = \beta_{1,M_{\text{full}}}$  (instead of  $\beta_{1,\hat{M}}$ ). For the estimator  $\hat{\sigma}^2$ , we use the usual unbiased variance estimator obtained by fitting the overall model; hence, we have  $r = n - p = 20$  here. [To be precise, while the constants  $K_N$  as well as  $K_{P1}$  are computed as detailed in Section 2, we consider instead of  $K_P$  defined by (2.3) the larger constant  $K_{P'}$  which is obtained from (2.3) when  $\mathcal{M}$  is replaced by the collection of *all* nonempty subsets of  $\{1, \dots, p\}$ . We shall refer to the resulting interval also as a PoSI-interval in this section. The reason for this choice is that code for computing  $K_{P'}$  is publicly available from the authors of Berk et al. (2013), so that  $K_{P'}$  is the PoSI-constant likely to be used by practitioners. Note that  $K_{P1} \leq K_P \leq K_{P'}$  holds, and



hence the performance of the interval based on  $K_P$  can be easily deduced from Table 1.]

As model selectors, we consider AIC, BIC, and the LASSO: For AIC we use the `step()` function in R with its default settings, subject to the constraint that the regressor of interest, that is, the first one, is always included; this corresponds to minimizing the AIC objective function through a greedy general-to-specific search over the  $2^9$  candidate models (i.e.,  $\mathcal{M}$  consists of all submodels of the overall model that contain the first regressor). Similarly, for BIC we use the `step()` function with the penalty parameter equal to  $\log(30)$ . And for the LASSO, we basically select those regressors for which the LASSO-estimator has nonzero coefficients. [More precisely, we use the `lars` package in R and follow suggestions outlined in Efron et al. (2004), Section 3.4: To protect the regressor of interest (the first one), we first compute the residual of the orthogonal projection of  $y$  on the first regressor; write  $\tilde{y}$  for this residual vector, and write  $\tilde{X}$  for the regressor matrix  $X$  with the first column removed. We then compute the LASSO-estimator for a regression of  $\tilde{y}$  on  $\tilde{X}$  using the `lars()` function; the LASSO-penalty is chosen by 10-fold cross-validation using the `cv.lars()` function (in both functions, we set the `intercept` parameter to `FALSE`, and otherwise use the default settings). The selected model is comprised of those regressors in  $\tilde{X}$  for which the corresponding LASSO coefficients are nonzero, plus the first column of  $X$ .]

Three designs are considered for the design matrix  $X$ : For design 1, we take the regressor matrix from the data-example from Section 3 of Kabaila and Leeb (2006) (for which the minimal coverage probability of a “naive” nominal 95% interval for  $\beta_1$ , based on a different variance estimator, was found to be no more than 0.63 in that paper). For design 2 and 3, respectively, we consider the exchangeable design and the equicorrelated design studied in Sections 6.1 and 6.2 of Berk et al. (2013). The exchangeable design is such that the corresponding PoSI-constant is small asymptotically, and the equicorrelated design corresponds to a large PoSI-constant asymptotically; cf. Theorem 6.1 and Theorem 6.2 in Berk et al. (2013). For the equicorrelated design (design 3), the difference between the PoSI-interval and the “naive” interval is thus expected to be most pronounced.

More precisely, for the first design, we take the regressor matrix from a data set from Rawlings, Pantula and Dickey (1998) (page 179), where the response

is peak flow rate from watersheds, and where the explanatory variables are rainfall (inches), which is the regressor of interest here, that is, the first column of  $X$ , as well as area of watershed (square miles), area impermeable to water (square miles), average slope of watershed (percent), longest stream flow in watershed (thousands of feet), surface absorbency index (0 = complete absorbency; 100 = no absorbency), estimated soil storage capacity (inches of water), infiltration rate of water into soil (inches/hour), time period during which rainfall exceeded 1/4 inch/hour, and a constant term to include an intercept in the model. Logarithms are taken of the response and of all explanatory variables except for the intercept. For the second design, we define  $\mathbf{X}^{(p)}(a)$  as in Section 6.1 in Berk et al. (2013) with  $p = 10$  and we choose  $a = 10$  here, and we set  $X = U\mathbf{X}^{(p)}(a)$ , where  $U$  is a collection of  $p$  orthonormal  $n$ -vectors obtained by first drawing a set of  $p$  i.i.d. standard Gaussian  $n$ -vectors and then applying the Gram–Schmidt procedure. And for the third design, we define  $\mathbf{X}^{(p)}(c)$  as in Section 6.2 in Berk et al. (2013), but such that the regressor of interest is the first one, where we choose  $c = \sqrt{0.8/(p-1)}$ , and we set  $X = V\mathbf{X}^{(p)}(c)$ , where  $V$  is obtained by drawing an independent observation from the same distribution as  $U$  before. (Because we consider only orthogonally invariant methods here, the coverage probabilities under study are invariant under orthogonal transformations of the columns of the design matrix. In particular, the coverage probabilities for the second and for the third design actually do not depend on the matrices  $U$  and  $V$ .)

For each of the three design matrices, we simulate coverage probabilities under the model  $Y = X\beta + u$  for randomly selected values of the parameter  $\beta$ , we identify those  $\beta$ 's for which the simulated coverage probability gets small, and we correct for bias as explained in detail shortly. For example, consider the case where the coverage target is  $\beta_1$  and where the “naive” confidence interval is used with AIC as the model selector. We first select 10,000 parameters  $\beta$  by drawing i.i.d. samples from a random  $p$ -vector  $b$  such that  $Xb$  follows a standard Gaussian distribution within the column-space of  $X$ . For each of these  $\beta$ 's, we approximate the corresponding coverage probability by the coverage rate obtained from 100 Monte Carlo samples. In particular, we draw 100 Monte Carlo samples from the overall model using  $\beta$  as the true parameter. For each Monte Carlo sample, we compute the model selector  $\hat{M}$  and the resulting “naive” confidence interval, and we record whether  $\beta_1$  is covered or not. The 100 recorded results

TABLE 1

*Smallest coverage probabilities (rounded to two digits of accuracy after the comma) found in MC study for the coverage targets  $\beta_{1,\hat{M}}$ , and  $\beta_1$ , using AIC, BIC, and the LASSO for model selection, for the PoSI-interval, the PoSII-interval, and the “naive” interval, each with nominal coverage probability 0.95*

Coverage target	Model selector	Confidence interval	Design 1 (watershed)	Design 2 (exchangeable)	Design 3 (equicorr.)
$\beta_{1,\hat{M}}$	AIC	PoSI	1.00	1.00	0.99
		PoSII	0.99	0.99	0.98
		Naive	0.89	0.92	0.81
	BIC	PoSI	1.00	1.00	0.99
		PoSII	0.98	0.99	0.98
		Naive	0.89	0.86	0.84
	LASSO	PoSI	1.00	1.00	1.00
		PoSII	1.00	1.00	1.00
		Naive	0.95	0.95	0.93
$\beta_1$	AIC	PoSI	0.85	0.91	0.83
		PoSII	0.76	0.91	0.77
		Naive	0.62	0.82	0.54
	BIC	PoSI	0.62	0.65	0.48
		PoSII	0.51	0.66	0.43
		Naive	0.43	0.51	0.26
	LASSO	PoSI	0.09	0.12	0.05
		PoSII	0.08	0.12	0.03
		Naive	0.07	0.10	0.01

are then averaged, resulting in a coverage rate that provides an estimator for the coverage probability of the interval if the true parameter is  $\beta$ . After repeating this for each of the 10,000  $\beta$ 's, we compute the resulting smallest coverage rate as an estimator for the minimal coverage probability of the confidence interval. The smallest coverage rate, as an estimator for the smallest coverage probability (over the 10,000 selected  $\beta$ 's), is clearly biased downward. To correct for that, we then take those 1000 parameters  $\beta$  that gave the smallest coverage rates and re-estimate the corresponding coverage probabilities as explained earlier, but now using 1000 Monte Carlo samples. For that parameter  $\beta$  that gives the smallest coverage rate in this second run, we run the simulation again but now with 500,000 Monte Carlo samples, to get a reliable estimate of the corresponding coverage probability. This procedure is also used, mutatis mutandis, to evaluate the performance of the PoSII-interval and of the PoSI-interval (with constant  $K_{p'}$ ), with AIC, BIC, and the LASSO as model selectors, and also in the case where the coverage target is  $\beta_{1,\hat{M}}$ . We stress here that the smallest coverage rates found by this procedure are simulation-based results obtained by a stochastic search over a 10-dimensional

parameter space, and thus only provide approximate upper bounds for the true minimal coverage probabilities (cf., e.g., the results for the PoSI-interval and the PoSII-interval, when the coverage target is  $\beta_1$ , when BIC is used for model selection, and when the second design matrix is used for  $X$ ). Table 1 summarizes the results.

For AIC and BIC, the results of the simulation study reinforce the impression already gained in the theoretical analysis in Section 3: When the coverage target is  $\beta_{1,\hat{M}}$ , the PoSII-interval as well as the PoSI-interval are somewhat too long and the “naive” interval is somewhat too short, resulting in moderate over- and under-coverage, respectively. Both over- and under-coverage are more pronounced than in the simple model studied in Section 3. In contrast, when the coverage target is  $\beta_1$ , then the actual coverage probability of all intervals can again be far below the nominal level. As expected, the difference between the “naive” interval and the PoSII-interval (resp., PoSI-interval) is most pronounced for design 3. The results for BIC are quite similar to those for AIC, when the coverage target is  $\beta_{1,\hat{M}}$ ; but when the target is  $\beta_1$ , all intervals based on BIC have poorer coverage properties

compared to the intervals based on AIC, with minima close to, or below, 0.5 in some cases. This is because BIC selects smaller models than AIC, typically causing more bias in the resulting post-model-selection estimator [that phenomenon is analyzed in greater detail in Leeb and Pötscher (2005) and Pötscher (2009)]. The results for the LASSO stand out: When the coverage target is  $\beta_{1,\hat{M}}$ , the PoSI-interval (resp., PoSI-interval) gives smallest probabilities very close to one, while the smallest coverage probability of the naive interval is very close to the nominal level (0.95). But when the coverage target is  $\beta_1$ , all intervals have smallest coverage probabilities of around 0.1 and below. The reason for this is that the LASSO model selector, as implemented here and for the parameters used in the stochastic search for the smallest coverage probability, selects the smallest possible model in most cases, that is, the model containing only the first regressor. In other words, the model selected by the LASSO is “nearly nonrandom.” When the target is  $\beta_{1,\hat{M}}$ , this entails that the naive interval is approximately valid and that both PoSI intervals are too large. [Indeed, the naive interval is valid if the underlying model selector always chooses a fixed (nonrandom) model; cf. the discussion following (2.2).] But when the target is  $\beta_1$ , the model selected by the LASSO typically suffers from severe bias, resulting in very small coverage probabilities for all intervals.

Other model selectors can, of course, give results different from those in Table 1. The model selectors chosen here represent a selection of popular methods from the contemporary literature that exhibit an interesting range of possible scenarios for the minimal coverage probabilities of confidence intervals post-model-selection.

#### APPENDIX: CONFIDENCE SETS UNDER ZERO-RESTRICTIONS POST-MODEL-SELECTION

Let  $y$  and  $\hat{\sigma}^2$  be as in Section 2, and consider  $\mathcal{M} = \{M_0, M_1\}$ , where each of the two candidate models  $M_i$  is full rank. Suppose we are interested in the coefficient of the first regressor  $X_1$ , that is assumed present in  $M_1$  but absent in  $M_0$ . In the notation introduced in Section 2, we thus have  $1 \in M_1$  and  $1 \notin M_0$ . Let  $\hat{M}$  be any model selection procedure that chooses only between  $M_0$  and  $M_1$ . As the model-dependent coverage target, we consider the coefficient of  $X_1$ , which is not restricted under  $M_1$ , and which is restricted to zero under  $M_0$ . More precisely, set  $b_{M_1} = \beta_{1,M_1}$ , set  $b_{M_0} = 0$ ,

and let the target be  $b_{\hat{M}}$ . We consider a “naive” confidence interval for  $b_{\hat{M}}$  that is defined as

$$I_{\hat{M}} = \begin{cases} \hat{\beta}_{1,M_1} \pm k_N \hat{\sigma}_{1,M_1}, & \text{if } \hat{M} = M_1, \\ \{0\}, & \text{if } \hat{M} = M_0, \end{cases}$$

where  $k_N$  is chosen so that  $\mathbb{P}_{\mu,\sigma}(\beta_{1,M_1} \in \hat{\beta}_{1,M_1} \pm k_N \hat{\sigma}_{1,M_1}) = 1 - \alpha$ . [The constant  $k_N$  is the  $(1 - \alpha/2)$ -quantile of a standard normal distribution in the known-variance case and the  $(1 - \alpha/2)$ -quantile of a  $t$ -distribution with  $r$  degrees of freedom in the unknown-variance case.] The actual coverage probability of  $I_{\hat{M}}$ , as a confidence interval for  $b_{\hat{M}}$ , is at least equal to the nominal coverage probability  $1 - \alpha$ , because

$$\begin{aligned} \mathbb{P}_{\mu,\sigma}(b_{\hat{M}} \in I_{\hat{M}}) &= \mathbb{P}_{\mu,\sigma}(\beta_{1,M_1} \in I_{M_1} \text{ and } \hat{M} = M_1) \\ &\quad + \mathbb{P}_{\mu,\sigma}(0 \in \{0\}, \hat{M} = M_0) \\ &= \mathbb{P}_{\mu,\sigma}(\beta_{1,M_1} \in I_{M_1} \text{ and } \hat{M} = M_1) \\ &\quad + \mathbb{P}_{\mu,\sigma}(\hat{M} \neq M_1) \\ &= \mathbb{P}_{\mu,\sigma}(\beta_{1,M_1} \in I_{M_1} \text{ or } \hat{M} \neq M_1) \geq 1 - \alpha, \end{aligned}$$

where the inequality in the last step holds in view of the choice of  $k_N$ .

#### ACKNOWLEDGMENTS

We thank the anonymous referee and the Editor for helpful comments and feedback. We also thank Richard Berk, Lawrence Brown, Andreas Buja, Kai Zhang, and Linda Zhao for providing us with the code to compute the PoSI-constant  $K_{P'}$  used in Section 4; the entire “PoSI-group” at the University of Pennsylvania for inspiring discussions during Hannes Leeb’s visit; and Francois Bachoc for constructive feedback.

Karl Ewald supported in part by Deutsche Forschungsgemeinschaft (DFG) Grant FOR916, and Hannes Leeb supported in part by FWF Grant P26354.

#### REFERENCES

- ANDREWS, D. W. K. and GUGGENBERGER, P. (2009). Hybrid and size-corrected subsampling methods. *Econometrica* **77** 721–762. [MR2531360](#)
- BERK, R., BROWN, L., BUJA, A., ZHANG, K. and ZHAO, L. (2013). Valid post-selection inference. *Ann. Statist.* **41** 802–837. [MR3099122](#)
- BICKEL, P. J. and DOKSUM, K. A. (1977). *Mathematical Statistics: Basic Ideas and Selected Topics*. Holden-Day, Oakland, CA. [MR0443141](#)
- BROWN, L. (1967). The conditional level of Student’s  $t$  test. *Ann. Math. Stat.* **38** 1068–1071. [MR0214210](#)

- BUEHLER, R. J. and FEDDERSEN, A. P. (1963). Note on a conditional property of Student's  $t$ . *Ann. Math. Stat.* **34** 1098–1100. [MR0150864](#)
- CRAVEN, P. and WAHBA, G. (1978/79). Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.* **31** 377–403. [MR0516581](#)
- DIJKSTRA, T. K. and VELDKAMP, J. H. (1988). Data-driven selection of regressors and the bootstrap. In *Lecture Notes in Econom. and Math. Systems* **307** 17–38. Springer, New York.
- EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–499. [MR2060166](#)
- EWALD, K. (2012). On the influence of model selection on confidence regions for marginal associations in the linear model. Master's thesis, Univ. Vienna.
- KABAILA, P. (1998). Valid confidence intervals in regression after variable selection. *Econometric Theory* **14** 463–482. [MR1650037](#)
- KABAILA, P. (2009). The coverage properties of confidence regions after model selection. *Int. Stat. Rev.* **77** 405–414.
- KABAILA, P. and LEEB, H. (2006). On the large-sample minimal coverage probability of confidence intervals after model selection. *J. Amer. Statist. Assoc.* **101** 619–629. [MR2256178](#)
- LEEB, H. (2006). The distribution of a linear predictor after model selection: Unconditional finite-sample distributions and asymptotic approximations. In *Optimality. Institute of Mathematical Statistics Lecture Notes—Monograph Series* **49** 291–311. IMS, Beachwood, OH. [MR2338549](#)
- LEEB, H. (2008). Evaluation and selection of models for out-of-sample prediction when the sample size is small relative to the complexity of the data-generating process. *Bernoulli* **14** 661–690. [MR2537807](#)
- LEEB, H. and PÖTSCHER, B. M. (2003). The finite-sample distribution of post-model-selection estimators and uniform versus nonuniform approximations. *Econometric Theory* **19** 100–142. [MR1965844](#)
- LEEB, H. and PÖTSCHER, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory* **21** 21–59. [MR2153856](#)
- LEEB, H. and PÖTSCHER, B. M. (2006a). Can one estimate the conditional distribution of post-model-selection estimators? *Ann. Statist.* **34** 2554–2591. [MR2291510](#)
- LEEB, H. and PÖTSCHER, B. M. (2006b). Performance limits for estimators of the risk or distribution of shrinkage-type estimators, and some general lower risk-bound results. *Econometric Theory* **22** 69–97. [MR2212693](#)
- LEEB, H. and PÖTSCHER, B. M. (2008a). Can one estimate the unconditional distribution of post-model-selection estimators? *Econometric Theory* **24** 338–376. [MR2422862](#)
- LEEB, H. and PÖTSCHER, B. M. (2008b). Model selection. In *Handbook of Financial Time Series* (T. G. Andersen, R. A. Davis, J.-P. Kreiß and Th. Mikosch, eds.) 785–821. Springer, New York.
- OLSHEN, R. A. (1973). The conditional level of the  $F$ -test. *J. Amer. Statist. Assoc.* **68** 692–698. [MR0359198](#)
- PÖTSCHER, B. M. (1991). Effects of model selection on inference. *Econometric Theory* **7** 163–185. [MR1128410](#)
- PÖTSCHER, B. M. (2006). The distribution of model averaging estimators and an impossibility result regarding its estimation. In *Time Series and Related Topics. Institute of Mathematical Statistics Lecture Notes—Monograph Series* **52** 113–129. IMS, Beachwood, OH. [MR2427842](#)
- PÖTSCHER, B. M. (2009). Confidence sets based on sparse estimators are necessarily large. *Sankhyā* **71** 1–18. [MR2579644](#)
- PÖTSCHER, B. M. and LEEB, H. (2009). On the distribution of penalized maximum likelihood estimators: The LASSO, SCAD, and thresholding. *J. Multivariate Anal.* **100** 2065–2082. [MR2543087](#)
- PÖTSCHER, B. M. and SCHNEIDER, U. (2009). On the distribution of the adaptive LASSO estimator. *J. Statist. Plann. Inference* **139** 2775–2790. [MR2523666](#)
- PÖTSCHER, B. M. and SCHNEIDER, U. (2010). Confidence sets based on penalized maximum likelihood estimators in Gaussian regression. *Electron. J. Stat.* **4** 334–360. [MR2645488](#)
- PÖTSCHER, B. M. and SCHNEIDER, U. (2011). Distributional results for thresholding estimators in high-dimensional Gaussian regression models. *Electron. J. Stat.* **5** 1876–1934. [MR2970179](#)
- RAWLINGS, J. O., PANTULA, S. G. and DICKEY, D. A. (1998). *Applied Regression Analysis: A Research Tool*, 2nd ed. Springer, New York. [MR1631919](#)
- SEN, P. K. (1979). Asymptotic properties of maximum likelihood estimators based on conditional specification. *Ann. Statist.* **7** 1019–1033. [MR0536504](#)
- SEN, P. K. and SALEH, A. K. M. E. (1987). On preliminary test and shrinkage  $M$ -estimation in linear models. *Ann. Statist.* **15** 1580–1592. [MR0913575](#)
- TUKEY, J. W. (1967). Discussion of “Topics in the investigation of linear relations fitted by the method of least squares” by F. J. Anscombe. *J. Roy. Statist. Soc. Ser. B* **29** 47–48.