Valid confidence intervals for post-model-selection predictors

Bachoc, Francois and Leeb, Hannes and Pötscher, Benedikt M.

Department of Statistics, University of Vienna, Department of Statistics, University of Vienna, Department of Statistics, University of Vienna

December 2014

Valid confidence intervals for post-model-selection predictors

François Bachoc, Hannes Leeb, and Benedikt M. Pötscher
Department of Statistics, University of Vienna

December 2014

Abstract

We consider inference post-model-selection in linear regression. In this setting, Berk et al. (2013a) recently introduced a class of confidence sets, the so-called PoSI intervals, that cover a certain non-standard quantity of interest with a user-specified minimal coverage probability, irrespective of the model selection procedure that is being used. In this paper, we generalize the PoSI intervals to post-model-selection predictors.

AMS Mathematics Subject Classification 2010: 62F25, 62J05.

Keywords: Inference post-model-selection, confidence intervals, optimal post-model-selection predictors, non-standard targets, linear regression.

1 Introduction and overview

Inference post-model-selection has proven to be a challenging problem. 'Naive' procedures, which ignore the presence of model selection, are typically invalid (e.g., in the sense that the actual coverage probability of 'naive' confidence sets can be dramatically smaller than the nominal one), and valid procedures are often hard to find; see Leeb and Pötscher (2005, 2006, 2012), Kabaila and Leeb (2006), Pötscher (2009) and references therein for an introduction to the issues involved here. Shifting the focus away from the true parameter as the target of inference, Berk et al. (2013a) recently introduced a class of confidence sets, the so-called PoSI intervals, that guarantee a user-specified minimal coverage probability after model selection in linear regression, irrespective of the model selector that is being used. In this paper, we generalize the PoSI intervals to post-model-selection predictors.

The crucial feature of the approach of Berk et al. (2013a) is that the coverage target, i.e., the quantity for which a confidence set is desired, is not the parameter in an overall model (or components thereof), but a *non-standard* quantity of interest that depends on the selected model and also on the training data. This non-standard quantity of interest is denoted by $\beta^{(n)}_{\hat{M}}$ throughout the paper (cf. Section 2 for details). Here $\hat{M}$ stands for the (data-dependent) model chosen by the model selector and n stands for sample size. The non-standard target $\beta^{(n)}_{\hat{M}}$ provides a certain vector of regression coefficients for those explanatory variables that are 'active' in the model $\hat{M}$; for a precise definition see eqs. (3) and (4) in Section 2.

For a new set of explanatory variables x_0, we first extend the PoSI-approach to obtain confidence intervals for the predictor $x_0[\hat{M}]'\beta^{(n)}_{\hat{M}}$. Here, $x_0[\hat{M}]$ denotes the set of explanatory variables from x_0 that correspond to the 'active' regressors in the model $\hat{M}$. We call $x_0[\hat{M}]'\beta^{(n)}_{\hat{M}}$ the design-dependent (non-standard) coverage target, because different design matrices in the training data typically result in different values of $x_0[\hat{M}]'\beta^{(n)}_{\hat{M}}$ even if both training data sets lead to selection of the same model $\hat{M}$. We construct PoSI confidence intervals for $x_0[\hat{M}]'\beta^{(n)}_{\hat{M}}$ that guarantee a user-specified minimal coverage probability, irrespective of the model selector that is being used. The design-dependent coverage target minimizes a certain *'in-sample'* prediction error; cf. Remark 2.6. However, when the goal is to predict a new response corresponding to a new vector x_0 of explanatory variables, this 'in-sample' optimality property may have little relevance and thus the focus on covering the design-dependent target $x_0[\hat{M}]'\beta^{(n)}_{\hat{M}}$ may be debatable.

In view of this, we next introduce an alternative coverage target that depends on the selected model but not on the training data otherwise, and that we denote by $x_0[\hat{M}]'\beta^{(\star)}_{\hat{M}}$. We call $x_0[\hat{M}]'\beta^{(\star)}_{\hat{M}}$ the design-independent (non-standard) coverage target. The design-independent coverage target minimizes the mean-squared prediction error over all (infeasible) predictors of the form $x_0[\hat{M}]'\gamma(\hat{M})$, when x_0 and the row-vectors of X are sampled from the same distribution; cf. Remark 3.2. In particular, that target does not suffer from the issues that plague the design-dependent coverage target, as discussed at the end of the preceding paragraph. For a large class of model selectors, we show that the PoSI confidence intervals constructed earlier also cover the design-independent coverage target with minimal coverage probability not below the user-specified nominal level asymptotically. In that sense, the PoSI confidence intervals for $x_0[\hat{M}]\beta^{(\star)}_{\hat{M}}$ are approximately valid, irrespective of the model selector $\hat{M}$ in that class. In simulations we find that our asymptotic result is representative of the finite-sample situation even for moderate sample sizes.

When extending the PoSI-approach to confidence intervals for both the design-dependent and the design-independent coverage target, i.e., for both $x_0[\hat{M}]'\beta^{(n)}_{\hat{M}}$ and $x_0[\hat{M}]'\beta^{(\star)}_{\hat{M}}$, we find that the resulting intervals necessarily depend not only on $x_0[\hat{M}]$ but also on those components of x_0 that are not 'active' in the model $\hat{M}$. This is problematic in situations when, after having selected a given model, only the 'active' components of x_0 are observed, e.g., in situations where observations are costly and model selection is carried out also with the goal of reducing cost by not having to observe irrelevant components of x_0. To resolve this, we also develop PoSI confidence intervals that depend on the 'active' variables $x_0[\hat{M}]$ only. These intervals are obtained by maximizing over all inactive variables and are hence larger than the intervals for the case where x_0 is known entirely. In simulations, we find that the excess width of these intervals is moderate.

One rationale for extending the PoSI-approach of Berk et al. (2013a) to problems related to prediction is that this framework seems to provide a more natural habitat for considering non-standard targets; see the discussion in Remark 2.1 of Leeb et al. (2013) as well as in Remarks 2.5, 2.8, and 3.1 given further below.

The rest of the paper is organized as follows. In Section 2, we introduce the models, the model-selection procedures, the design-dependent target, and the PoSI confidence intervals

for both the case where all explanatory variables in x_0 are observed and the case where only the 'active' explanatory variables in x_0 are available. In Section 3, we present the design-independent target and show that the PoSI confidence intervals introduced earlier also cover the design-independent target, with minimal coverage probability not below the nominal one asymptotically. In Section 4, we describe algorithms for computing the PoSI confidence intervals. Finally, the results of a simulation study are reported in Section 5. The proofs are given in the appendices.

We stress here that the 'naive' confidence interval, which ignores the data-driven model selection step and which uses the standard confidence procedure as if the selected model were correct and given a priori, is invalid also in the setting considered here. This is shown in Leeb et al. (2013) for the case where the coverage target is as in Berk et al. (2013a), and these results are easily adapted to the prediction setting considered here, and this is also confirmed by our simulation results in Section 5.

2 Confidence intervals for the design-dependent non-standard target

Consider the linear model

$$Y = X\beta + U \tag{1}$$

where X is a (real) $n \times p$ matrix of full rank $p \geq 1$, β is an unknown $p \times 1$ parameter vector and U follows an $N\left(0, \sigma^2 I_n\right)$-distribution; here σ^2, $0 < \sigma < \infty$, is the unknown error variance and I_n is the identity matrix of size n. In this section the design matrix X is treated as fixed. We refer to (1) as the overall or full linear model. The extension of our results to the case where no overall linear model is assumed, and hence all fitted models may be misspecified, is outlined in Remark 2.8.

Let $\hat{\beta}$ denote the ordinary least squares estimator for β in model (1). We further assume that, as an estimator for σ^2, we have available an (observable) random variable $\hat{\sigma}^2$ that is independent of $\hat{\beta}$ and that is distributed as σ^2/r times a chi-square distributed random variable with r degrees of freedom ($1 \leq r < \infty$). Of course, such an estimator $\hat{\sigma}^2$ always exists (provided $n > p$) and is given by the usual (unbiased) variance estimator $\hat{\sigma}^2_{OLS} = (Y - X\hat{\beta})'(Y - X\hat{\beta})/(n-p)$, in which case we have $r = n - p$, but the theory is not restricted to this case. [The extra generality of our assumption on $\hat{\sigma}^2$ provides additional flexibility and proves convenient, e.g., in the discussion of the case when no overall model is assumed, see Remark 2.8. It furthermore allows for estimators $\hat{\sigma}^2$ that not only depend on Y and X, but possibly also on other observable random variables (e.g., additional data).] The joint distribution of Y and $\hat{\sigma}^2$ depends on β and σ as well as on sample size n and will be denoted by $P_{n,\beta,\sigma}$.

We consider submodels of the full linear model (1) that are obtained by deleting columns from the matrix X. A submodel will be represented by M, a subset of $\{1, ..., p\}$, where the elements of M index the columns of X that are retained. In relation to submodels we use the following notation: For $M \subseteq \{1, ..., p\}$, we write M^c for the complement of M in $\{1, ..., p\}$. It proves useful to allow M to be the empty set. We write $|M|$ for the cardinality of M. With

$m = |M|$, let us write $M = \{j_1, ..., j_m\}$ in case $m \geq 1$. For $M \neq \varnothing$ and for an $l \times p$ matrix T, $l \geq 1$, let $T[M]$ be the matrix of size $l \times m$ obtained from T by retaining only the columns of T with indices $j \in M$ and deleting all other columns; if $M = \varnothing$ we set $T[M] = 0 \in \mathbb{R}^l$. In abuse of notation we shall, for a $p \times 1$ vector v, write $v[M]$ for $(v'[M])'$, i.e., $v[M] = (v_{j_1}, ..., v_{j_m})'$ for $m \geq 1$ and $v[M] = 0 \in \mathbb{R}$ in case $M = \varnothing$. For a given submodel M, we denote the corresponding restricted least squares estimator by $\hat{\beta}_M$, i.e.,

$$\hat{\beta}_M = \left(X'[M]X[M]\right)^{-1} X'[M]Y, \tag{2}$$

where the inverse is to be interpreted as the Moore-Penrose inverse in case $M = \varnothing$. Note that our assumption on the variance estimator implies that $\hat{\sigma}^2$ is independent of the collection $\{\hat{\beta}_M : M \subseteq \{1, ..., p\}\}$ of all restricted least squares estimators. For any given submodel M the restricted least squares estimator $\hat{\beta}_M$ is obviously an unbiased estimator of

$$\beta_M^{(n)} = \beta[M] + \left(X'[M]X[M]\right)^{-1} X'[M]X[M^c]\beta[M^c]. \tag{3}$$

Note that $\beta_M^{(n)}$ reduces to β in case $M = \{1, ..., p\}$ and to 0 in case $M = \varnothing$.

A model selection procedure $\hat{M}$ is now a (measurable) rule that associates with every $(X, Y, \hat{\sigma}^2)$ a model $\hat{M}(X, Y, \hat{\sigma}^2) \subseteq \{1, ..., p\}$. We allow for the possibility that $\hat{M}(X, Y, \hat{\sigma}^2)$ is empty. In the following we shall, in abuse of notation, often write $\hat{M}$ for $\hat{M}(X, Y, \hat{\sigma}^2)$. [In most cases $\hat{\sigma}^2$ will be a function of (X, Y), whence $\hat{M}$ will depend only on (X, Y). Allowing explicitly dependence of $\hat{M}$ on $\hat{\sigma}^2$ is thus only relevant in case $\hat{\sigma}^2$ depends on extraneous data beyond (X, Y) (and the model selection procedure actually makes use of $\hat{\sigma}^2$). We note here that in principle we could have allowed $\hat{M}$ to depend on further extraneous data, in which case $P_{n,\beta,\sigma}$ would have to be redefined as the joint distribution of Y, $\hat{\sigma}^2$, and this further extraneous data.] The post-model-selection estimator $\hat{\beta}_{\hat{M}}$ corresponding to the model selection procedure is now given by (2) with M replaced by $\hat{M}$. The preceding definition of a model selection procedure did not impose any restriction on the range of the map $\hat{M}$ except that it has to be contained in the power set of $\{1, ..., p\}$. The case where the range of the model selection procedures considered is a priori restricted to a user-specified universe $\mathcal{M}$ of admissible submodels is discussed in Remark 2.7.

The non-standard quantity of interest studied in Berk et al. (2013a) is now the random vector (with random dimension) $\beta_{\hat{M}}^{(n)}$ obtained by replacing M by $\hat{M}$ in (3). [We note that Berk et al. (2013a) do not allow for empty submodels $\hat{M}$.] The situation we shall consider in the present paper is related to Berk et al. (2013a), but is different in several aspects: Consider a fixed $x_0 \in \mathbb{R}^p$ and suppose we want to predict y_0 which is distributed as $N\left(x_0'\beta, \sigma^2\right)$, independently of Y. If one is forced to use a fixed submodel M for prediction, the predictor that would then typically be used is $x_0'[M]\hat{\beta}_M$. This predictor can be seen as estimating the infeasible predictor $x_0'[M]\beta_M^{(n)}$; cf. Remark 2.6. In the presence of model selection the predictor $x_0'[M]\hat{\beta}_M$ will then typically be replaced by the post-model-selection predictor $x_0'[\hat{M}]\hat{\beta}_{\hat{M}}$ which can be seen as a feasible counterpart to the infeasible predictor

$$x_0'[\hat{M}]\beta_{\hat{M}}^{(n)}. \tag{4}$$

The quantity in (4) will be our target for inference in this section and will be called the *design-dependent (non-standard) target* (to emphasize that it depends on the design matrix X apart from its dependence on $\hat{M}$, cf. (3), and that it is different from the standard (infeasible) predictor $x_0'\beta$ one obtains from the full model). A discussion of the merits of this target and its interpretation is postponed to Remarks 2.5 and 2.6 given below.

Let now $1-\alpha \in (0,1)$ be a nominal confidence level. In this section we are interested in confidence intervals for the design-dependent target $x_0'[\hat{M}]\beta^{(n)}_{\hat{M}}$ that are of the form

$$CI(x_0) = x_0'[\hat{M}]\hat{\beta}_{\hat{M}} \pm K(x_0, \hat{M})||s_{\hat{M}}||\hat{\sigma}, \tag{5}$$

where $\|\cdot\|$ denotes the Euclidean norm, where

$$s_M' = x_0'[M]\left(X'[M]X[M]\right)^{-1}X'[M], \tag{6}$$

and where $K(x_0, M) = K(x_0, M, r) = K(x_0, M, r, X, \alpha)$ is a non-negative constant which may depend on x_0, M, r, X, and α, but does not depend on the observations on Y and $\hat{\sigma}^2$. Here we have used the notation $a \pm b$ for the interval $[a-b, a+b]$ ($a \in \mathbb{R}$, $b \geq 0$). Note that $s_M \in \mathbb{R}^n$ for all $M \subseteq \{1, ..., p\}$. The motivation for the form of the confidence interval stems from the observation that for *fixed* M the interval $x_0'[M]\hat{\beta}_M \pm q_{S,r,1-\alpha/2}||s_M||\hat{\sigma}$ is a valid $1-\alpha$ confidence interval for $x_0'[M]\beta^{(n)}_M$, where $q_{S,r,1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of Student's t-distribution with r degrees of freedom. Note that on the event $\hat{M} = \varnothing$ the target is equal to zero and the confidence interval reduces to $\{0\}$, thus always containing the target on this event.

We aim at finding quantities $K(x_0, M)$ such that the confidence intervals $CI(x_0)$ satisfy

$$\inf_{\beta \in \mathbb{R}^p, \sigma > 0} P_{n,\beta,\sigma}\left(x_0'[\hat{M}]\beta^{(n)}_{\hat{M}} \in CI(x_0)\right) \geq 1-\alpha. \tag{7}$$

Note that if we replace $K(x_0, \hat{M})$ in (5) by $K_{naive} = q_{S,r,1-\alpha/2}$, then the confidence interval (5) reduces to the so-called 'naive' confidence interval which is constructed as if $\hat{M}$ were fixed a priori (thus ignoring the presence of model selection). It does not fulfill (7) as can be seen from the numerical results in Section 5, which is in line with the related results in Leeb et al. (2013).

For the construction of the quantities $K(x_0, M)$ we distinguish two cases regarding the observation on x_0: (i) The vector x_0 is observed in its entirety (regardless of which model $\hat{M}$ is selected), or (ii) only the subvector $x_0[\hat{M}]$ of x_0 is observed (note that only this subvector is needed for the computation of the post-model-selection predictor $x_0'[\hat{M}]\hat{\beta}_{\hat{M}}$). The latter case can be relevant in practical situations where the selected model is determined first and then only observations for $x_0[\hat{M}]$ (and not for the other components of x_0) are collected, e.g., out of cost considerations.

For the case (i), where x_0 is entirely observed, the following straightforward adaptation of the approach in Berk et al. (2013a) yields a constant $K_1(x_0) = K_1(x_0, r) = K_1(x_0, r, X, \alpha)$ (not depending on M) such that the resulting confidence interval (5) satisfies (7): Observe that

$$x_0'[\hat{M}]\hat{\beta}_{\hat{M}} - x_0'[\hat{M}]\beta^{(n)}_{\hat{M}} = s_{\hat{M}}'\left(Y - X\beta\right) \tag{8}$$

and define $\bar{s}_M = s_M / \|s_M\|$ if $s_M \neq 0$ and set $\bar{s}_M = 0$ if $s_M = 0$. Then obviously we have the upper bound

$$\left|\bar{s}'_{\hat{M}}\left(Y - X\beta\right)\right| / \hat{\sigma} \leq \max_{M \subseteq \{1,\ldots,p\}} \left|\bar{s}'_M\left(Y - X\beta\right)\right| / \hat{\sigma}. \tag{9}$$

Define $K_1(x_0)$ to be the smallest constant satisfying

$$P_{n,\beta,\sigma}\left(\max_{M \subseteq \{1,\ldots,p\}} \left|\bar{s}'_M\left(Y - X\beta\right)\right| / \hat{\sigma} \leq K_1(x_0)\right) \geq 1 - \alpha. \tag{10}$$

It is important to note that the probability on the left-hand side of the preceding display does neither depend on β nor σ; it also depends on the estimator $\hat{\sigma}$ only through the "degrees of freedom" parameter r: To see this note that $\bar{s}'_M\left(Y - X\beta\right) = \bar{s}'_M P_X\left(Y - X\beta\right)$, where P_X denotes the orthogonal projection on the column space of X, since $\bar{s}_M$ belongs to the column space of X. Consequently, the collection of all the quantities $\bar{s}'_M\left(Y - X\beta\right)$ is jointly distributed as $N(0, \sigma^2 C)$, independently of $\hat{\sigma}^2 \sim \left(\sigma^2/r\right)\chi^2\left(r\right)$, where the covariance matrix C depends only on x_0 and X. Hence the joint distribution of the collection of ratios $\left|\bar{s}'_M\left(Y - X\beta\right)\right| / \hat{\sigma}$ does neither depend on β nor σ and depends on the estimator $\hat{\sigma}$ only through r. It is now plain that $K_1(x_0)$ only depends on x_0, r, X, and α. Furthermore note that $K_1(x_0) = 0$ in case $x_0 = 0$; otherwise, $K_1(x_0)$ is positive, equality holds in (10), and $K_1(x_0)$ is the unique $(1-\alpha)$-quantile of the distribution of the upper bound in (9). [This follows from Lemma A.1 and from the observation that $\bar{s}'_M = 0$ for all $M \subseteq \{1, \ldots, p\}$ holds if and only if $x_0 = 0$.] Furthermore, observe that $K_1(x_0)$ coincides with a PoSI1 constant of Berk et al. (2013a) in case x_0 is one of the standard basis vectors e_i (and the universe $\mathcal{M}$ of models in Berk et al. (2013a) is the family of all non-empty subsets of $\{1, \ldots, p\}$). [This can be seen by comparison with (4.14) in Berk et al. (2013a) and noting that the maximum inside the probability in (10) effectively extends only over models satisfying $i \in M$, since $\bar{s}_M = 0$ here holds for models M with $i \notin M$.] Finally, note that $K_{naive} \leq K_1\left(x_0\right)$ clearly holds provided $x_0 \neq 0$ (since $\bar{s}'_M\left(Y - X\beta\right)/\hat{\sigma}$ follows Student's t-distribution with r degrees of freedom if $s_M \neq 0$).

As a consequence of (9) and the discussion in the preceding paragraph we thus immediately obtain the following proposition.

Proposition 2.1. *Let $\hat{M}$ be an arbitrary model selection procedure, let $x_0 \in \mathbb{R}^p$ be arbitrary, and let $K_1(x_0)$ be defined by (10). Then the confidence interval (5) with $K(x_0, \hat{M})$ replaced by $K_1(x_0)$ satisfies the coverage property (7).*

The coverage in Proposition 2.1 is guaranteed for *all* model selection procedures, and thus leads to 'universally valid post-selection inference'; cf. Berk et al. (2013a), where similar guarantees are obtained for the components of $\beta^{(n)}_{\hat{M}}$.

Consider next case (ii) where only the components of $x_0[\hat{M}]$ are observed. In this case, the confidence interval of Proposition 2.1 is not feasible in that it cannot be computed (except on the event where the selected model coincides with $\{1, ..., p\}$). A first solution is to define

$$K_2(x_0[M], M) = \sup\left\{K_1(x) : x[M] = x_0[M]\right\}, \tag{11}$$

and then to use the confidence interval (5) with $K(x_0, \hat{M})$ replaced by $K_2(x_0[\hat{M}], \hat{M})$. Note that $K_2(x_0[M], M)$, and hence the corresponding confidence interval, depends on x_0 only

via $x_0[M]$. Of course, $K_2(x_0[M], M)$ also depends on r, X, and α and we shall write $K_2(x_0[M], M, r)$ if we want to stress dependence on r. It is easy to see that $K_2(x_0[M], M)$ is finite (as it is not larger than the Scheffé constant as we shall see below). Because $K_2(x_0[M], M)$ is never smaller than $K_1(x_0)$, we have the following corollary to Proposition 2.1.

Corollary 2.2. *Let $\hat{M}$ be an arbitrary model selection procedure, let $x_0 \in \mathbb{R}^p$ be arbitrary, and let $K_2(x_0[M], M)$ be defined by (11). Then the confidence interval (5) with $K(x_0, \hat{M})$ replaced by $K_2(x_0[\hat{M}], \hat{M})$ satisfies the coverage property (7).*

The computation of $K_2(x_0[\hat{M}], \hat{M})$ is more costly than that of $K_1(x_0)$. Indeed, it requires to embed the algorithm for computing $K_1(x_0)$ in an optimization procedure. Thus, for the cases where the resulting computational cost is prohibitive, we present in the subsequent proposition larger constants $K_3(x_0[\hat{M}], \hat{M})$, K_4, and K_5 that are simpler to compute. Algorithms for computing these constants are discussed in Section 4.

For $x_0 \in \mathbb{R}^p$ and $M \subseteq \{1, ..., p\}$ define the distribution function F^*_{M,x_0} for $t \geq 0$ via

$$F^*_{M,x_0}(t) = 1 - \min \left[\begin{array}{c} 1, \Pr\left(\max_{M_* \subseteq M} \left|\vec{s}'_{M_*} V\right| > t\right) \\ + \left(2^p - 2^{|M|}\right)\left(1 - F_{Beta,1/2,(p-1)/2}\left(t^2\right)\right) \end{array} \right] \tag{12}$$

and via $F^*_{M,x_0}(t) = 0$ for $t < 0$. Here V is a random vector that is uniformly distributed on the unit sphere in the column space of X and $F_{Beta,1/2,(p-1)/2}$ denotes the $Beta\,(1/2, (p-1)/2)$-distribution function with the convention that in case $p = 1$ we use $F_{Beta,1/2,0}$ to denote the distribution function of pointmass at 1. Next define the distribution function F_{M,x_0} via

$$F_{M,x_0}(t) = \mathbb{E}_G F^*_{M,x_0}(t/G), \tag{13}$$

where G denotes a nonnegative random variable such that G^2/p follows an F-distribution with (p, r)-degrees of freedom and $\mathbb{E}_G$ represents expectation w.r.t. the distribution of G. We stress that F_{M,x_0} depends on x_0 only through $x_0[M]$, and hence the same is true for the constant $K_3(x_0[M], M)$ defined in the subsequent proposition.

Proposition 2.3. *Let $x_0 \in \mathbb{R}^p$ be arbitrary. For any $M \subseteq \{1, ..., p\}$ define $K_3(x_0[M], M)$ to be the smallest constant K such that*

$$F_{M,x_0}(K) \geq 1 - \alpha \tag{14}$$

holds. Then $K_3(x_0[M], M) = 0$ (and F_{M,x_0} is the c.d.f. of pointmass at zero) if $M = \{1, ..., p\}$ and $x_0 = 0$; otherwise, $0 < K_3(x_0[M], M) < \infty$, and equality holds in (14) if and only if $K = K_3(x_0[M], M)$. Furthermore, let $K_4 = K_3(x_0[\varnothing], \varnothing)$ and define K_5 as the $(1-\alpha)$-quantile of G. Then we have

$$K_2(x_0[M], M) \leq K_3(x_0[M], M) \leq K_4 \leq K_5 \tag{15}$$

for every M. Furthermore

$$K_2(x_0[M_2], M_2) \leq K_2(x_0[M_1], M_1), \tag{16}$$

$$K_3(x_0[M_2], M_2) \leq K_3(x_0[M_1], M_1) \tag{17}$$

hold whenever $M_1 \subseteq M_2$.

It is obvious that $K_3(x_0[M], M)$ depends, besides $x_0[M]$ and M, only on r, X, and α, whereas K_4 and K_5 only depend on r, p, and α. [Like with $K_1(x_0)$, also the other constants introduced depend on the estimator $\hat{\sigma}$ only through r.] We shall write $K_3(x_0[M], M, r)$, $K_4(r)$, and $K_5(r)$ if we want to stress dependence on r. Note that $K_1(x_0) = K_3(x_0[M_{full}], M_{full}) = K_3(x_0, M_{full})$, where $M_{full} = \{1, ..., p\}$, and that $K_3(x_0[M], M) = K_4$ holds for any M satisfying $|M| = 1$ and $\bar{s}_M \neq 0$. [In this case, the probability appearing in (12) equals $1 - F_{Beta,1/2,(p-1)/2}(t^2)$ as can be seen from the proof of Proposition 2.3.] Furthermore, K_5 is just the Scheffé constant (Scheffé (1959)); see the corresponding discussion in Section 4.8 of Berk et al. (2013a). The proof of the inequalities involving the constants K_3 and K_4 in the above proposition is an extension of an argument in Berk et al. (2013b) (not contained in the published version Berk et al. (2013a)) to find an upper-bound for their PoSI constant that does not depend on X, but only on p. [Note that K_4 is a counterpart to K_{univ} in Berk et al. (2013b).] Inequalities (16) and (17) simply reflect the fact that observing only $x_0[M]$ implies that fewer information about x_0 is provided for smaller models M. As a consequence of these inequalities it is possible that, on the event where a small model M_1 is selected, the resulting confidence interval is larger than it is on the event where a larger model M_2 is selected. Again, this simply reflects the fact that less information on x_0 is available under the smaller model. Note, however, that the just discussed phenomenon is counteracted by the fact that the length of the confidence interval also depends on $||s_M||$ and that we have $||s_{M_1}|| \leq ||s_{M_2}||$ for $M_1 \subseteq M_2$; cf. Figure 1 in Section 5.

Proposition 2.3 implies $K_2(x_0[\hat{M}], \hat{M}) \leq K_3(x_0[\hat{M}], \hat{M}) \leq K_4 \leq K_5$ and hence together with Corollary 2.2 immediately implies the following result. Note that the confidence intervals figuring in this corollary depend on x_0 only through $x_0[\hat{M}]$.

Corollary 2.4. *Let $\hat{M}$ be an arbitrary model selection procedure, and let $x_0 \in \mathbb{R}^p$ be arbitrary. Then the confidence interval (5) with $K(x_0, \hat{M})$ replaced by $K_3(x_0[\hat{M}], \hat{M})$ (K_4, or K_5, respectively) satisfies the coverage property (7).*

We conclude this section with a few remarks.

Remark 2.5. *(On the merits of $\beta^{(n)}_{\hat{M}}$ and $x_0'[\hat{M}]\beta^{(n)}_{\hat{M}}$ as targets for inference)* (i) As already noted, the (non-standard) coverage target in Berk et al. (2013a) is $\beta^{(n)}_{\hat{M}}$ (where these authors choose to represent it in what they call 'full model indexing'). While $\beta^{(n)}_{\hat{M}}$ has a clear technical meaning as the coefficient vector that provides the best approximation of $X\beta$ by elements of the form $X[\hat{M}]\gamma$ w.r.t. the Euclidean distance, the merits of this quantity as a target for statistical inference are less clear. Note that if one adopts this quantity as the target for inference, one is confronted with the fact that the target then depends on the data Y via $\hat{M}$ (implying that the target as well as its dimension are random); furthermore, different model selection procedures give rise to different targets $\beta^{(n)}_{\hat{M}}$. Also note that, e.g., the meaning of the first component of the target $\beta^{(n)}_{\hat{M}}$ depends on the selected model $\hat{M}$. The target $x_0'[\hat{M}]\beta^{(n)}_{\hat{M}}$ considered in this paper, while again being random and sharing many of the properties of $\beta^{(n)}_{\hat{M}}$ just mentioned, seems to be somewhat more amenable to interpretation as a target for inference since it is simply the random convex combination $\sum_M x_0'[M]\beta^{(n)}_M \mathbf{1}(\hat{M} = M)$ of the

(infeasible) predictors $x_0'[M]\beta_M^{(n)}$ (which one would typically use if model M is forced upon one for prediction).

(ii) One can justly argue that the target for inference should be $x_0'\beta$ rather than $x_0'[\hat{M}]\beta_{\hat{M}}^{(n)}$ because $x_0'\beta$ is a better (infeasible) predictor in the mean-squared error sense than $x_0'[\hat{M}]\beta_{\hat{M}}^{(n)}$ provided y_0 is independent of $\hat{M}$, which will be the case if y_0 is independent of Y and $\hat{\sigma}^2$ (or if y_0 is independent of Y and $\hat{M}$ does not depend on $\hat{\sigma}^2$). [This is so since the mean-squared error of prediction of $x_0'\beta$ is not larger than the one of $x_0'[M]\beta_M^{(n)}$ for every M and since $\hat{M}$ is independent of y_0.] However, this argument does not apply if x_0 is not observed in its entirety, but only $x_0[\hat{M}]$ is observed, because then $x_0'\beta$ is not available. In this case we thus indeed have some justification for the target $x_0'[\hat{M}]\beta_{\hat{M}}^{(n)}$. If, as in Berk et al. (2013a), interest focusses on parameters rather than predictors, one can also argue that β should be the target rather than $\beta_{\hat{M}}^{(n)}$. Note, however, that the above argument justifying the target $x_0'[\hat{M}]\beta_{\hat{M}}^{(n)}$ is not applicable here. Berk et al. (2013a) rather make the case for $\beta_{\hat{M}}^{(n)}$ as a desirable target by arguing that the relevant setting is one where an overall model, and thus β, is not available. However, then the assumption in Berk et al. (2013a) on the variance estimator $\hat{\sigma}^2$ becomes problematic; see Remark 2.1(ii) in Leeb et al. (2013) as well as Remark 2.8 below.

(iii) We note the obvious fact that if the target of inference is the standard target $x_0'\beta$ then the reasoning underlying Proposition 2.1 does not apply since the difference between the post-model-selection predictor and the standard target is not independent of β. For the same reason the approach in Berk et al. (2013a)) cannot provide a solution to the problem of constructing confidence sets for the standard target β.

Remark 2.6. *(Optimality of the design-dependent target)* (i) The infeasible predictor $x_0'[M]\beta_M^{(n)}$ (for fixed M) is the best predictor in the mean-squared error sense among all predictors of the form $x_0'[M]\gamma$ in case $y_0|\, x_0 \sim N(x_0'\beta, \sigma^2)$ and x_0' is drawn from the empirical distribution of the rows of X ('in-sample prediction'). Otherwise, it does not have this optimality property (but nevertheless its feasible counterpart $x_0'[M]\hat{\beta}_M$ would typically be used if one is forced to base prediction on model M).

(ii) The optimality property in (i) carries over to the design-dependent target $x_0'[\hat{M}]\beta_{\hat{M}}^{(n)}$ provided $(y_0, x_0')'$ is independent of $\hat{M}$.

Remark 2.7. *(Restricted universe of candidate models)* (i) Suppose now that one considers only model selection procedures $\hat{M}$ that are known to take their values in a user-specified universe $\mathcal{M}$ of candidate models (i.e., $\mathcal{M}$ is a non-empty subset of the power set of $\{1, \ldots, p\}$). While the results given above apply in this situation a fortiori (by simply ignoring $\mathcal{M}$), one may want to exploit the knowledge that the model selection procedures considered only take values in $\mathcal{M}$. The results of the present section can then be generalized to this situation with no difficulty, but at the expense of more complex notation. We only sketch the most important changes necessary: In (9) and (10) the maximum now has to extend only over $M \in \mathcal{M}$, and the condition for the degenerate case $K_1(x_0) = 0$ now becomes $x_0[M_{\mathcal{M}}] = 0$, where $M_{\mathcal{M}} = \bigcup\{M : M \in \mathcal{M}\}$; note also that now $K_1(x_0)$ depends on x_0 only via $x_0[M_{\mathcal{M}}]$. Furthermore, in the definition of F^*_{M,x_0} the maximum now has to extend only over those $M_* \in \mathcal{M}$ that are subsets of M, and the constants $2^p - 2^{|M|}$ have to be replaced by the number

$c(M, \mathcal{M})$ of models $M_* \in \mathcal{M}$ that satisfy $M_* \not\subseteq M$. Note that now $K_3(x_0[M], M) = 0$ will hold if and only if $M \in \mathcal{M}$ is such that it contains any element of $\mathcal{M}$ as a subset and also $x_0[M] = 0$ holds (and K_4 will be zero if and only if $\mathcal{M} = \{\varnothing\}$). Note that now K_1 and K_3 depend on X only via $X[M_{\mathcal{M}}]$, and that all the constants K_i become dependent also on $\mathcal{M}$.

(ii) It is illustrative to consider as an example a situation where the matrix X in (1) is partitioned into $(X_1 : X_2)$ with X_1 an $n \times p_1$ matrix and $\mathcal{M}$ is the power set of $\{1, \ldots, p_1\}$ (or a subset thereof). This describes a situation where the user observes X_1 but not X_2, acts as if the model containing only the regressors in X_1 were correct, and thus selects only models belonging to $\mathcal{M}$. Note that all candidate models in $\mathcal{M}$ are misspecified (except if the true parameter happens to satisfy $\beta_i = 0$ for all $i > p_1$). While the user can then apply standard model selection procedures resulting in a selected model $\hat{M} \in \mathcal{M}$, he will typically not be able to compute the confidence intervals proposed in this section, as he will typically be lacking an estimator $\hat{\sigma}^2$ with the required properties. If subsequent to his analysis the matrix X_2 and the information that model (1) is correct is supplied to him, he could then construct $\hat{\sigma}^2$ from the correct overall model (1) and apply the generalization sketched in (i) above to set confidence intervals for his target $x_0'[\hat{M}]\beta^{(n)}_{\hat{M}}$. However, it is hard to see why – in the presence of the additional information – he would then not want to abandon his analysis and redo it within a larger universe of candidate models that also contains a true model. The preceding discussion highlights the tension between assuming that all candidate models are misspecified and assuming that an estimator $\hat{\sigma}^2$ with the required properties is available. The subsequent remark elaborates further on this point.

Remark 2.8. *(No overall model)* Berk et al. (2013a) set their results in a framework where no overall model exists (or is known to the user). That is, Y is assumed to be distributed as $N(\mu, \sigma^2 I_n)$ with no further assumption on μ. Furthermore given is an observable $n \times p$ matrix X, not necessarily of full column rank (thus allowing for the possibility that $p > n$), which gives rise to models $M \subseteq \{1, \ldots, p\}$ corresponding to the matrices $X[M]$, and a universe of models $\mathcal{M}$ such that any $M \in \mathcal{M}$ is full rank (i.e., $X[M]$ has full column rank for any $M \in \mathcal{M}$). Without loss of generality one can here assume $M_{\mathcal{M}} = \{1, \ldots, p\}$. Our results in the present section can then easily be extended to this framework. [In fact, appropriate extensions are possible even to the case where $\mathcal{M}$ contains non-full rank models, if the inverses appearing in (2) and (3) are interpreted as Moore-Penrose inverses.] The problem with such a framework, however, is that then the assumption on the estimator $\hat{\sigma}$ made above (and also in Berk et al. (2013a)) becomes problematic as it is less than clear how such a (feasible) estimator can be constructed in the absence of an overall model, except in very special cases; cf. the discussion in Remark 2.1(ii) in Leeb et al. (2013). For this reason we have abstained from formulating our results in this section in this (seemingly) more general framework.

Remark 2.9. *(Infeasible variance estimators)* (i) For later use we note that all results in this section continue to hold if $\hat{\sigma}^2$ is allowed to also depend on σ but otherwise satisfies the assumptions made earlier (e.g., if $\hat{\sigma}^2 = \sigma^2 Z/r$ where Z is an observable chi-square distributed random variable with r degrees of freedom that is independent of $P_X Y$).

(ii) In case we set $\hat{\sigma}^2 = \sigma^2$ and $r = \infty$ all of the results in this section continue to hold with obvious modifications. In particular, in Proposition 2.3 the random variable G^2 then follows a chi-squared distribution with p degrees of freedoms. We shall denote the constants

corresponding to $K_1(x_0)$, $K_2(x_0[M], M)$, $K_3(x_0[M], M)$, K_4, and K_5 obtained by setting $\hat{\sigma}^2 = \sigma^2$ and $r = \infty$ by $K_1(x_0, \infty)$, etc. We stress that these constants do *not* depend on σ.

Remark 2.10. In the proof of Proposition 2.3 union bounds are used to obtained the results for $K_3(x_0[M], M)$ and K_4. Hence, one might ask whether or not these constants as bounds for $K_2(x_0[M], M)$ are overly conservative. We provide now some evidence showing that improving $K_3(x_0[M], M)$ and K_4 will not be easy and is sometimes impossible. First, Lemma A.4 in Appendix A shows that there exist $n \times 2$ design matrices X and vectors x_0 such that $K_4 = K_1(x_0)$. Hence, in this case the union bounds used in the proof of Proposition 2.3 are all exact. Furthermore, consider the (infeasible) case where $\hat{\sigma}^2 = \sigma^2$, $r = \infty$, and consider the limiting behavior of $K_4(\infty)$ for $p \to \infty$. Inspection of the proof of Theorem 6.3 in Berk et al. (2013a) shows that $K_4(\infty)/\sqrt{p}$ then tends to $\sqrt{3}/2 \approx 0.86$. The same proof also shows that $K_1^*(x_0)/\sqrt{p}$ tends to $\sqrt{3}/2$ in probability as $p \to \infty$, where $K_1^*(x_0)$ is an analogue of $K_1(x_0, \infty)$ which is obtained from (10) (with $r = \infty$) after replacing the vectors $\bar{s}_M$ by 2^p independent random vectors, each of which is uniformly distributed on the unit sphere of the column space of X (and these vectors being independent of Y). In other words, if one ignores the particular structure of the vectors $\bar{s}_M$, then the bound $K_4(\infty)$ is close to being sharp for large values of p. Finally, using Theorem 6.2 in Berk et al. (2013a), a sequence of design matrices X can be constructed such that $\lim_{p\to\infty} K_1(x_0, \infty)/\sqrt{p} \approx 0.63$ holds for $x_0 = e_i$ (and it is stated in that reference that there are indications that other design matrices may increase this limit to 0.78). [Recall that $K_1(e_i, \infty)$ coincides with a PoSI1 constant.] This implies that for these design matrices the union bounds in the proof of Proposition 2.3 can only be improved marginally for large p.

3 Confidence intervals for the design-independent non-standard target

In this section we again consider the model (1), but now assume the $n \times p$ matrix X to be random, with X independent of U, where U again follows an $N(0, \sigma^2 I_n)$-distribution with $0 < \sigma < \infty$. We also assume that X has full column rank almost surely and that each row of X is distributed according to a common p-dimensional distribution $\mathcal{L}$ with a finite and positive definite matrix of (uncentered) second moments, which we denote by Σ. [We shall refer to the above assumptions as the maintained model assumptions of this section.] Furthermore, we assume again that we have available an estimator $\hat{\sigma}^2$ such that, conditionally on X, $\hat{\sigma}^2$ is independent of $\hat{\beta}$ and is distributed as σ^2/r times a chi-squared distributed random variable with r degrees of freedom ($1 \le r < \infty$). Thus, all the results of Section 2 remain valid in the setup of the present section if formulated conditionally on X (and if x_0 is treated as fixed). [Alternatively, if x_0 is random but independent of X, U, and $\hat{\sigma}^2$, the same is true if the results in Section 2 are then interpreted conditionally on X and x_0.] The joint distribution of Y, X, and $\hat{\sigma}^2$ (and of $\check{\sigma}$ appearing below) will be denoted by $P_{n,\beta,\sigma}$.

In this section we shall consider asymptotic results for $n \to \infty$ but where p is held constant. It is thus important to recall that all estimators, estimated models, etc. depend on sample size n. Also note that r may depend on sample size n. We shall typically suppress these

dependencies on n in the notation. Furthermore, we note that, while not explicitly shown in the notation, the rows of X and U (and thus of Y) may depend on n. [As the results in Section 2 are results for fixed n, this trivially also applies to the results in that section.]

If M_1 and M_2 are subsets of $\{1, ..., p\}$ and if Q is a $p \times p$ matrix we shall denote by $Q[M_1, M_2]$ the matrix that is obtained from Q by deleting all rows i with $i \notin M_1$ as well as all columns j with $j \notin M_2$; if M_1 is empty but M_2 is not, we define $Q[M_1, M_2]$ to be the $1 \times |M_2|$ zero vector; if M_2 is empty but M_1 is not, we define $Q[M_1, M_2]$ to be the $|M_1| \times 1$ zero vector; and if $M_1 = M_2 = \varnothing$ we set $Q[M_1, M_2] = 0 \in \mathbb{R}$.

To motivate our second target, consider now the problem of predicting a new variable $y_0 = x_0'\beta + u_0$ where x_0, u_0, X, and U are independent and $u_0 \sim N\left(0, \sigma^2\right)$. For a given model $M \subseteq \{1, ..., p\}$ we consider the (infeasible) predictor $x_0'[M]\beta_M^{(\star)}$ where

$$\beta_M^{(\star)} = \beta[M] + (\Sigma[M, M])^{-1} \Sigma[M, M^c]\beta[M^c],$$

with the convention that the inverse is to be interpreted as the Moore-Penrose inverse in case $M = \varnothing$. Note that $x_0'[M]\beta_M^{(\star)} = 0$ if $M = \varnothing$ and that $x_0'[M]\beta_M^{(\star)} = x_0'\beta$ if $M = \{1, \ldots, p\}$. A justification for considering this infeasible predictor is given in Remark 3.2 below. Given a model selection procedure $\hat{M}$ we define now the (infeasible) predictor

$$x_0'[\hat{M}]\beta_{\hat{M}}^{(\star)}$$

as our new target for inference. We call this target the *design-independent (non-standard) target* as it does not depend on the design matrix X beyond its dependence on $\hat{M}$. We discuss its merits in the subsequent remarks.

Remark 3.1. Similar to the situation in Section 2, the target $x_0'[\hat{M}]\beta_{\hat{M}}^{(\star)}$ considered in the present section is nothing else than the post-model-selection analogue to the (infeasible) predictor $x_0'[M]\beta_M^{(\star)}$, i.e., is the random convex combination $\sum_M x_0'[M]\beta_M^{(\star)}\mathbf{1}(\hat{M} = M)$ of the (infeasible) predictors $x_0'[M]\beta_M^{(\star)}$. As in Remark 2.5(ii) one can argue that the target for inference should be $x_0'\beta$ rather than $x_0'[\hat{M}]\beta_{\hat{M}}^{(\star)}$ because again $x_0'\beta$ is a better (infeasible) predictor than $x_0'[\hat{M}]\beta_{\hat{M}}^{(\star)}$ provided that (x_0', u_0) is independent of $\hat{M}$ (which, in particular, will be the case if (x_0', u_0) is independent of X, U, and $\hat{\sigma}$, or if (x_0', u_0) is independent of X, U and $\hat{M}$ does not depend on $\hat{\sigma}$). But again, this argument does not apply if x_0 is not observed in its entirety, but only $x_0[\hat{M}]$ is observed.

Remark 3.2. *(Optimality of the design-independent target)* (i) Assume that additionally $x_0' \sim \mathcal{L}$. If we are forced to use the (theoretical) predictors of the form $x_0'[M]\gamma$, then straightforward computation shows that $x_0'[M]\beta_M^{(\star)}$ provides the smallest mean-squared error of prediction among all the linear predictors $x_0'[M]\gamma$. [Note that this result corresponds to the observation made in Remark 2.6 with $\mathcal{L}$ corresponding to the empirical distribution of the rows of X.] If, furthermore, x_0 is normally distributed, then x_0 and u_0 are jointly normal and thus $x_0'[M]\beta_M^{(\star)}$ is the conditional expectation of y_0 given $x_0[M]$ and hence is also the best predictor in the class of all predictors depending only on $x_0[M]$.

(ii) Again assume that $x_0' \sim \mathcal{L}$. The discussion in (i) implies that $x_0'[\hat{M}]\beta_{\hat{M}}^{(\star)}$ has a mean-squared error of prediction not larger than the one of $x_0'[\hat{M}]\gamma(\hat{M})$ for any choice of $\gamma(\hat{M})$, provided (x_0', u_0) is independent of $\hat{M}$. If, additionally, x_0 is normally distributed, then $x_0'[\hat{M}]\beta_{\hat{M}}^{(\star)}$ is also the best predictor in the class of all predictors depending only on $x_0'[\hat{M}]$ and $\hat{M}$.

After having motivated the design-independent target, we shall, in the remainder of this section, treat x_0 as fixed (but see Remark 3.9 for the case where x_0 is random). We now proceed to show that the confidence intervals constructed in Section 2 are also valid as confidence intervals for the design-independent target $x_0'[\hat{M}]\beta_{\hat{M}}^{(\star)}$ in an asymptotic sense under some mild conditions. While the results in Section 2 apply to *any* model selection procedure whatsoever, we need here to make the following mild assumption on the model selection procedure.

Condition 3.3. *The model selection procedure satisfies: For any $M \subseteq \{1, \ldots, p\}$ with $|M| < p$ and for any $\delta > 0$,*

$$\sup \left\{ P_{n,\beta,\sigma}(\hat{M} = M|X) : \beta \in \mathbb{R}^p, \sigma > 0, \|\beta[M^c]\| / \sigma \geq \delta \right\} \to 0$$

in probability as $n \to \infty$.

Condition 3.3 is very mild and typically holds for model selection procedures such as AIC- and BIC-based procedures as well as Lasso-type procedures. In addition, we assume the following condition on the behavior of the design matrix.

Condition 3.4. *The sequence of random matrices $\sqrt{n}\,[(X'X/n) - \Sigma]$ is bounded in probability.*

Condition 3.4 holds, for example, when the rows of X are independent, or weakly dependent, and when the distribution $\mathcal{L}$ has finite fourth moments for all its components. We also introduce the following condition.

Condition 3.5. *The degrees of freedom parameters r of the sequence of estimators $\hat{\sigma}^2$ satisfy $r \to \infty$ as $n \to \infty$.*

Of course, if we choose for $\hat{\sigma}^2$ the usual variance estimator $\hat{\sigma}^2_{OLS}$ then this condition is certainly satisfied with $r = n - p$. We are now in the position to present the asymptotic coverage result. Recall that the confidence intervals corresponding to K_i with $2 \leq i \leq 5$ depend on x_0 only through $x_0[\hat{M}]$.

Theorem 3.6. *Suppose Conditions 3.3 and 3.4 hold.*

(a) Suppose also that Condition 3.5 is satisfied. Let $CI(x_0)$ be the confidence interval (5) where the constant $K(x_0, \hat{M})$ is given by the constant $K_1(x_0, r)$ defined in Section 2. Then the confidence interval $CI(x_0)$ satisfies

$$\inf_{x_0 \in \mathbb{R}^p, \beta \in \mathbb{R}^p, \sigma > 0} P_{n,\beta,\sigma}\left(x_0'[\hat{M}]\beta_{\hat{M}}^{(\star)} \in CI(x_0) \middle| X \right) \geq (1 - \alpha) + o_p(1), \tag{18}$$

where the $o_p(1)$ term above depends only on X and converges to zero in probability as $n \to \infty$. Relation (18) a fortiori holds if the confidence interval $CI(x_0)$ is based on the constants $K_2(x_0[\hat{M}], \hat{M}, r)$, $K_3(x_0[\hat{M}], \hat{M}, r)$, $K_4(r)$, or $K_5(r)$, respectively.

(b) Let $\tilde{\sigma}$ be an arbitrary estimator satisfying

$$\sup_{\beta \in \mathbb{R}^p, \sigma > 0} P_{n,\beta,\sigma}(|\tilde{\sigma}/\sigma - 1| \geq \delta \,|X\,) \xrightarrow{p} 0 \tag{19}$$

for any $\delta > 0$ as $n \to \infty$. Let further $r^* = r_n^*$ be an arbitrary sequence in $\mathbb{N} \cup \{\infty\}$ satisfying $r^* \to \infty$ for $n \to \infty$, where $\mathbb{N}$ denotes the set of positive integers. Let $CI^*(x_0)$ denote the modified confidence interval which is obtained from (5) by replacing $\hat{\sigma}$ by $\tilde{\sigma}$ and $K(x_0, \hat{M})$ by $K_1(x_0, r^*)$ ($K_2(x_0[\hat{M}], \hat{M}, r^*)$, $K_3(x_0[\hat{M}], \hat{M}, r^*)$, $K_4(r^*)$, or $K_5(r^*)$, respectively). Then $CI^*(x_0)$ satisfies the relation (18).

Theorem 3.6(a) shows that for any $x_0 \in \mathbb{R}^p$ the interval $CI(x_0)$ is an asymptotically valid confidence interval for the design-independent target and additionally that the lower bound $(1-\alpha) + o_p(1)$ for the minimal (over β and σ) coverage probability can be chosen independently of x_0. Theorem 3.6(b) extends this result to a larger class of intervals. [Note that Part (a) is in fact a special case of Part (b) obtained by setting $\tilde{\sigma} = \hat{\sigma}$ and $r^* = r$ and observing that $\hat{\sigma}$ clearly satisfies the condition on $\tilde{\sigma}$ in Part (b) under Condition 3.5.] We note that applying Theorem 3.6(b) with $\tilde{\sigma} = \hat{\sigma}$ and $r^* = \infty$ shows that Theorem 3.6(a) also continues to hold for the confidence interval that is obtained by replacing the constants $K_1(x_0, r)$ ($K_2(x_0[\hat{M}], \hat{M}, r)$, $K_3(x_0[\hat{M}], \hat{M}, r)$, $K_4(r)$, or $K_5(r)$, respectively) by the constants $K_1(x_0, \infty)$ ($K_2(x_0[\hat{M}], \hat{M}, \infty)$, $K_3(x_0[\hat{M}], \hat{M}, \infty)$, $K_4(\infty)$, or $K_5(\infty)$, respectively).

Condition (19) is a uniform consistency property. It is clearly satisfied by $\hat{\sigma}^2_{OLS}$ (and more generally by the estimator $\hat{\sigma}^2$ under Condition 3.5 as already noted above), but it is also satisfied by the post-model-selection estimator $\hat{\sigma}^2_{\hat{M}} = ||Y - X[\hat{M}]\hat{\beta}_{\hat{M}}||^2/(n - |\hat{M}|)$ provided the model selection procedure satisfies Condition 3.3, see Lemma B.2 in Appendix B for a precise result. As a consequence, Theorem 3.6(b) shows that the post-model-selection estimator $\hat{\sigma}^2_{\hat{M}}$ can be used instead of $\hat{\sigma}^2$ in the construction of the confidence interval.

Remark 3.7. *(Infeasible variance estimators)* Theorem 3.6(a) remains valid if $\hat{\sigma}^2$ is allowed to depend also on σ but otherwise satisfies the assumptions made earlier or if $\hat{\sigma}^2 = \sigma^2$ and $r = \infty$. Similarly, Theorem 3.6(b) remains valid if $\tilde{\sigma}^2$ is allowed to be infeasible.

Remark 3.8. Under the assumptions of Theorem 3.6(b) we further have that

$$\inf_{x_0 \in \mathbb{R}^p, \beta \in \mathbb{R}^p, \sigma > 0} P_{n,\beta,\sigma}\left(x_0'[\hat{M}]\beta^{(n)}_{\hat{M}} \in CI^*(x_0) \middle| X \right) \geq (1-\alpha) + o_p(1),$$

holds, where the $o_p(1)$ term above depends only on X and converges to zero in probability as $n \to \infty$. This follows easily from a repeated application of Lemma B.1 in Appendix B. [Regarding Theorem 3.6(a) recall that the finite-sample coverage result for the target $x_0'[\hat{M}]\beta^{(n)}_{\hat{M}}$ in Section 2 continues to hold in the context of the present section if interpreted conditionally on X.]

Remark 3.9. *(Random x_0)* If x_0 is random and independent of X, U, and $\hat{\sigma}^2$, Theorem 3.6 continues to hold if the result is then being interpreted as conditional on X and x_0. A particular consequence of this result conditional on X and x_0 is then that the confidence interval $CI(x_0)$ also satisfies

$$\inf_{\beta\in\mathbb{R}^p,\sigma>0} P_{n,\beta,\sigma}\left(x_0'[\hat{M}]\beta_{\hat{M}}^{(\star)} \in CI(x_0)\middle| X\right) \geq (1-\alpha) + o_p(1)$$

where again $o_p(1)$ is a function of X only (and $P_{n,\beta,\sigma}$ here represents the distribution of Y, X, $\hat{\sigma}^2$, and x_0). As noted at the beginning of this section, the results in Section 2 continue to hold if interpreted conditionally on X and x_0. As a consequence, we thus also have that

$$\inf_{\beta\in\mathbb{R}^p,\sigma>0} P_{n,\beta,\sigma}\left(x_0'[\hat{M}]\beta_{\hat{M}}^{(n)} \in CI(x_0)\middle| X\right) \geq (1-\alpha)$$

holds.

Remark 3.10. *(Relaxing the assumptions on X)* The assumption that the rows of X follow a common distribution $\mathcal{L}$ has been used only to define the matrix Σ, which in turn is used in the definition of $\beta_M^{(\star)}$. If this assumption is dropped, but instead it is assumed that Condition 3.4 holds for *some* positive matrix Σ, which is then used to define $\beta_M^{(\star)}$, Theorem 3.6 continues to hold. Note that this version of Theorem 3.6 also covers the case of nonrandom design matrices for which $n^{-1}X'X$ converges to a positive definite limit at rate $n^{-1/2}$ (or faster).

Remark 3.11. *(Restricted universe of selected models)* Theorem 3.6 can easily be generalized to the case where a universe $\mathcal{M}$ different from the power set of $\{1,\ldots,p\}$ is employed, provided the full model $\{1,\ldots,p\}$ belongs to $\mathcal{M}$.

Remark 3.12. *(Measurability issues)* Various statements concerning uncountable suprema (infima) of conditional probabilities occur in the present section and Appendix B, such as, e.g., statements that these quantities converge in probability. It is not difficult to see that – in absence of measurability – all these statements remain valid if they are properly interpreted as statements referring to outer probability. This thus relieves one from the need to establish measurability. For this reason we do not explicitly mention the measurability issues in the presentation of the results in this section as well as in Appendix B.

4 Algorithms for computing the confidence intervals

In this section we again treat X as a fixed matrix of rank p. Recall that $n \geq p$ always holds. Let Q be a $n \times p$ matrix so that the columns of Q form an orthonormal basis of the column space of X. Following Berk et al. (2013a) we define $\tilde{Y} = Q'Y$ and $\tilde{X} = Q'X$, the so-called canonical coordinates of Y and X. We then have $\tilde{Y} = \tilde{X}\beta + \tilde{U}$ with $\tilde{U} = Q'U \sim N\left(0, \sigma^2 I_p\right)$. It is now easy to see that the least squares estimator for β computed from the canonical coordinates equals $\hat{\beta}$ computed from Y and X, cf. Proposition 5.1 in Berk et al. (2013a). In particular, the independence of the given $\hat{\sigma}^2$ from the least-squares estimator holds whether we work with the original model or with the model in canonical coordinates. Furthermore,

setting $\tilde{s}'_M = x'_0[M](\tilde{X}'[M]\tilde{X}[M])^{-1}\tilde{X}'[M]$ it follows that $\|s'_M\| = \|\tilde{s}'_M\|$ and $s'_M(Y - X\beta) = \tilde{s}'_M(\tilde{Y} - \tilde{X}\beta)$. For later use define $\bar{\tilde{s}}_M = \tilde{s}_M / \|\tilde{s}_M\|$ if $\|\tilde{s}_M\| \neq 0$ and define $\bar{\tilde{s}}_M = 0$ if $\tilde{s}_M = 0$. Inspection of (10) and of the definition of F_{M,x_0} now shows that all the constants K_i remain the same whether they are computed from the original problem using the design matrix X or from the transformed problem using the canonical coordinates $\tilde{X}$. Hence, in the algorithms below we shall work with the canonical coordinates as this facilitates computation. Note that x_0 is unaffected by this transformation. The matrices Q and $\tilde{X}$ can be obtained, for example, from a SVD or a QR decomposition of X.

The following algorithm for computing $K_1(x_0)$, defined in Section 2, is similar to that of Berk et al. (2013b) for computing the PoSI constant. We present it here for completeness. From Proposition 2.3 and the observation that $K_1(x_0) = K_3(x_0[M_{full}], M_{full})$ made subsequent to this proposition we see that $K_1(x_0)$ is the solution to

$$\mathbb{E}_G \Pr\left(\max_{M_* \subseteq \{1,\ldots,p\}} \left|\bar{\tilde{s}}'_{M_*} V\right| \leq t/G\right) = 1 - \alpha.$$

The algorithm now replaces the probability in the preceding display by a Monte-Carlo estimator, analytically performs the integration w.r.t. G, and then numerically solves the resulting equation. We note that in this and the other algorithms to follow there is no need for Monte-Carlo integration w.r.t. G. We shall denote by $F^{\sharp}_{p,r}$ the c.d.f. of G ; note that then $F^{\sharp}_{p,r}(t) = F_{p,r}\left(t^2/p\right)$, where $F_{p,r}$ denotes the c.d.f. of an F-distribution with (p, r)-degrees of freedom.

Algorithm 4.1. *In case $x_0 \neq 0$, choose $I \in \mathbb{N}$ and generate independent identically distributed random vectors $V_1, \ldots, V_I$, where each V_i is uniformly distributed on the unit sphere in $\mathbb{R}^p$. Calculate the quantities $c_i = \max_{M \subseteq \{1,\ldots,p\}} \left|\bar{\tilde{s}}'_M V_i\right|$ with $\bar{\tilde{s}}_M$ as defined above. A numerical approximation to $K_1(x_0)$ is then obtained by searching for that value of K that solves*

$$\frac{1}{I}\sum_{i=1}^{I} F^{\sharp}_{p,r}\left(\frac{K}{c_i}\right) = 1 - \alpha. \tag{20}$$

In case $x_0 = 0$, set $K_1(x_0) = 0$.

Note that for $x_0 \neq 0$ at least one of the vectors $\bar{\tilde{s}}_M$ is non-zero, implying that the quantities c_i are all non-zero with probability 1; hence the terms $F^{\sharp}_{p,r}\left(\frac{K}{c_i}\right)$ are well-defined with probability one. It is now obvious that – on the event where all c_i are non-zero – the solution K of (20) exists, is unique and positive. The costly factor in Algorithm 4.1 is the maximization involved in the computation of the quantities c_i, while searching for the value of K that solves (20), for example by bisection searches, incurs only negligible cost. In our simulations, computing $K_1(x_0)$ for $p = 10$ and $I = 10,000$ takes around one second on a personal computer. Berk et al. (2013b) found that their algorithm (which is similar to Algorithm 4.1 as noted above) is tractable for up to about $p = 20$ and $I = 1,000$, in which case the elapsed time was around one hour on 2012 desktop computer equipment.

The algorithm for computing $K_3(x_0[M], M)$ is given next. We provide this algorithm only for non-empty $M \neq \{1, ..., p\}$ since in case $M = \{1, ..., p\}$ we have $K_3(x_0[M], M) = K_1(x_0)$,

which can be computed by Algorithm 4.1, and in case M is empty we have $K_3(x_0[M], M) = K_4$, which can be computed by Algorithm 4.3 given below. We now search for the solution of the equation

$$1 - \alpha = \mathbb{E}_G \breve{F}^*_{M,x_0}(t/G)$$

where $\breve{F}^*_{M,x_0}$ is a Monte-Carlo estimator of F^*_{M,x_0} obtained by replacing the probability involving V by an empirical Monte-Carlo estimator (and where $\bar{\tilde{s}}_M$instead of $\bar{s}_M$ is being used). Observing that we need only to integrate over the range where $\breve{F}^*_{M,x_0}$ is positive (i.e., where $t/G > m_*$ defined below), the integrand can be additively decomposed into a "jump" part and a continuous part. The integral over the jump part can be expressed analytically in terms of the c.d.f. $F^\sharp_{p,r}$, whereas the integral over the continuous part is approximated by an integral over a step function, which again can be expressed in terms of the c.d.f. $F^\sharp_{p,r}$.

Algorithm 4.2. *Suppose that M satisfies $\varnothing \neq M \neq \{1, ..., p\}$. Choose $I \in \mathbb{N}$, generate independent identically distributed random vectors $V_1, \ldots, V_I$, where each V_i is uniformly distributed on the unit sphere in $\mathbb{R}^p$, and calculate the quantities $c_i = \max_{M_* \subseteq M} \left| \bar{\tilde{s}}'_{M_*} V_i \right|$ with $\bar{\tilde{s}}_{M_*}$ as defined above. Find m_* as the smallest value such that*

$$\frac{1}{I} \sum_{i=1}^{I} \mathbf{1}(c_i > t) + \left(2^p - 2^{|M|}\right) \left(1 - F_{Beta,1/2,(p-1)/2}\left(t^2\right)\right) < 1$$

holds for all $t > m_*$. Next, choose $J \in \mathbb{N}$ and find the values $m_1, ..., m_{J-1}$ so that, for $j = 1, ..., J-1$

$$\left(1 - F_{Beta,1/2,(p-1)/2}\left(m_*^2\right)\right) \frac{j}{J} = \left(1 - F_{Beta,1/2,(p-1)/2}\left(m_j^2\right)\right) \tag{21}$$

holds. Set $m_J = m_*$. A numerical approximation to $K_3(x_0[M], M)$ is then obtained by searching for that value of K that solves

$$\begin{aligned} 1 - \alpha = {} & F^\sharp_{p,r}\left(\frac{K}{m_J}\right) - \frac{1}{I} \sum_{i: c_i > m_J} \left(F^\sharp_{p,r}\left(\frac{K}{m_J}\right) - F^\sharp_{p,r}\left(\frac{K}{c_i}\right)\right) \\ & + \left(2^p - 2^{|M|}\right) \left(1 - F_{Beta,1/2,(p-1)/2}\left(m_J^2\right)\right) \frac{1}{J} \sum_{j=1}^{J-1} \left(F^\sharp_{p,r}\left(\frac{K}{m_j}\right) - F^\sharp_{p,r}\left(\frac{K}{m_J}\right)\right). \end{aligned} \tag{22}$$

Note that m_* is uniquely determined, is always positive, and satisfies $m_* \leq 1$. [In fact, $m_* < 1$ holds, except in case $c_i = 1$ for all i, which is a probability zero event.] Provided $m_* < 1$ holds, the values m_j for $j \geq 1$ are uniquely defined and satisfy $m_* < m_{J-1} < \ldots < m_1 < 1$. [In case $m_* = 1$, then any $m_j \geq 1$ would solve (21). But in this case the r.h.s. of (22) reduces to $F^\sharp_{p,r}(K)$ anyway and hence there is no need for solving equation (21).] Furthermore, note that the r.h.s. of (22) can be written as

$$\begin{aligned} & F^\sharp_{p,r}\left(\frac{K}{m_J}\right) \left[1 - \frac{1}{I} \sum_{i=1}^{I} \mathbf{1}(c_i > m_J) - \left(2^p - 2^{|M|}\right) \left(1 - F_{Beta,1/2,(p-1)/2}\left(m_J^2\right)\right) \frac{J-1}{J}\right] \\ & + \frac{1}{I} \sum_{i: c_i > m_J} F^\sharp_{p,r}\left(\frac{K}{c_i}\right) + \left(2^p - 2^{|M|}\right) \left(1 - F_{Beta,1/2,(p-1)/2}\left(m_J^2\right)\right) \frac{1}{J} \sum_{j=1}^{J-1} F^\sharp_{p,r}\left(\frac{K}{m_j}\right). \end{aligned}$$

Observing that the expression in brackets is nonnegative (in fact, positive except if $m_J = 1$) because of the definition of m_J, we see that the r.h.s. of (22) is strictly increasing in K. Furthermore, inspection of the r.h.s. of (22) shows that it is zero for $K = 0$ and converges to one for $K \to \infty$. Consequently, equation (22) has a unique solution for K, which necessarily is positive. We note that in Algorithm 4.2 the cost of searching for m_*, for the m_j's, and for K, for example by bisection searches, is negligible compared to that of computing the quantities c_i, which is again the limiting factor.

The above algorithm is based on approximating $1 - F_{Beta,1/2,(p-1)/2}\left(t^2\right)$ for $t > m_*$ by a step function from below. If we approximate by a step function from above, this results in the same algorithm except that now the second sum on the r.h.s. of equation (22) runs from $j = 0$ to $j = J - 1$ with the convention that $m_0 = 1$. A similar argument as above shows that the solution to this modification of (22) exists, is unique and is positive. Note that the solutions obtained from running both versions of the algorithm in parallel provide a lower as well as an upper bound for the solution one would obtain if the integration of the continuous part could be performed without error. These lower and upper bounds allow one to gauge whether or not J has been chosen large enough such that the effect of the numerical integration error on K is negligible. Note that running the two versions of the algorithm in parallel is not much more costly than running just one version, as only (bisection) searches are involved once the c_i's have been computed.

The following algorithm for computing K_4 is similar to the algorithm in Berk et al. (2013b), Section 7.2, for computing the universal upper-bound for the PoSI constants. The computational cost of this algorithm is negligible compared to those of Algorithms 4.1 and 4.2.

Algorithm 4.3. *In case $p > 1$, choose $J \in \mathbb{N}$ and find the values $m_1, ..., m_J$ so that, for $j = 1, ..., J$,*

$$(2^p - 1)\left(1 - F_{Beta,1/2,(p-1)/2}\left(m_j^2\right)\right) = \frac{j}{J}. \tag{23}$$

Then, K_4 is numerically approximated by the (uniquely determined and positive) constant K that solves

$$\frac{1}{J}\sum_{j=1}^{J} F_{p,r}^{\sharp}\left(\frac{K}{m_j}\right) = 1 - \alpha. \tag{24}$$

In case $p = 1$, K_4 is the (uniquely determined and positive) constant K that solves

$$F_{1,r}^{\sharp}(K) = 1 - \alpha. \tag{25}$$

Note that in case $p > 1$ the constants m_j exist, are unique, and positive; consequently, the solution K of (24) exists, is unique and positive. In case $p = 1$ the solution of (25) also exists, is unique and positive. As before, this algorithm relies on approximation by a step function from below. A version of the algorithm that uses a step function that approximates from above is obtained if equation (24) is replaced by

$$\frac{1}{J}\sum_{j=0}^{J-1} F_{p,r}^{\sharp}\left(\frac{K}{m_j}\right) = 1 - \alpha$$

with the convention that $m_0 = 1$.

Remark 4.4. (i) For the computation of the constants $K_1(x_0, \infty)$, $K_3(x_0[M], M, \infty)$, and $K_4(\infty)$ (cf. Remark 2.9) one can use the above algorithms with the only modification that the distribution function $F^{\sharp}_{p,r}$ is replaced by the distribution function of the square root of a chi-squared-distributed random variable with p degrees of freedom.

(ii) The algorithms can be appropriately extended to the generalizations described in Remarks 2.7 and 2.8.

Remark 4.5. When p is larger than 20, Algorithms 4.1 or 4.2 may not be tractable, but Algorithm 4.3 can still be. However, it is reported in Berk et al. (2013b) that, for about $p \geq 40$, it can be problematic to compute the extreme quantiles in (23). In this case, it is always possible to use the Scheffé constant K_5. In practice, it may also be reasonable (since p is large) to use the constant $K_6 = 0.86K_5$ in (5), cf. Remark 2.10. A similar advice is given in the framework of Berk et al. (2013b).

5 Simulation study

In this section, we investigate numerically the lengths and the minimal coverage probabilities of the confidence intervals introduced in Section 2. We also compare these lengths and minimal coverage probabilities with those of the 'naive' confidence interval that ignores the model selection step. As model selection procedures we consider the AIC, BIC, and LASSO model selectors, and we use design matrices that are obtained from the watershed data set of Rawlings (1998) as well as from the exchangeable and equicorrelated designs as considered in Berk et al. (2013a).

First, we consider the lengths of the confidence intervals obtained by (5) standardized by $\hat{\sigma}$, i.e., we consider $2K(x_0, \hat{M})\|s_{\hat{M}}\|$ for the six cases where $K(x_0, \hat{M})$ is replaced by either one of the five constants $K_1(x_0)$, $K_2(x_0[\hat{M}], \hat{M})$, $K_3(x_0[\hat{M}], \hat{M})$, K_4, K_5 of Section 2 or by the constant $K_{naive} = q_{S,r,1-\alpha/2}$, the $(1-\alpha/2)$-quantile of the Student's t-distribution with r degrees of freedom. We recall that the constant K_{naive} yields the 'naive' confidence interval that ignores the model selection step and that we have $K_{naive} \leq K_1(x_0) \leq ... \leq K_5$ (the first inequality holding provided $x_0 \neq 0$).

For computing the standardized length, we set $\alpha = 0.05$, $n = 29$, $p = 10$, $r = n - p$, $\sigma = 1$, and obtain X and x_0 from a data set of Rawlings (1998). This data set contains a 30×10 design matrix X_{Raw} corresponding to ten explanatory variables. These explanatory variables are a constant term (to include an intercept in the model), rainfall (inches), area of watershed (square miles), area impervious 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) and time period during which rainfall exceeded 1/4 inch/hour. Logarithms are taken of the explanatory variables except for the intercept. [In Rawlings (1998), the response corresponding to these explanatory variables is peak flow rate from watersheds.] This data set is also studied in Kabaila and Leeb (2006) and Leeb et al. (2013). We refer to it as the watershed data set, and x_0 and X are chosen such that $(x_0, X')'$ is equal to the watershed design matrix X_{Raw}.

For the so chosen values of α, n, p, r, σ, X, and x_0, we compute the standardized lengths $2K(x_0, M)\|s_M\|$ of the confidence intervals obtained by replacing $K(x_0, M)$ by K_{naive}, $K_1(x_0)$, $K_2(x_0[M], M)$, $K_3(x_0[M], M)$, K_4, and K_5, respectively. To ease the computational burden and to enable a simple presentation as in Figure 1 below, we compute the standardized lengths of the confidence intervals only for M belonging to the family $\{\{1\}, ..., \{1, ..., 10\}\}$ consisting of ten nested submodels. [This does *not* mean that we compute the constants K_i under the assumption of a restricted universe of models as described in Remark 2.7.] The computation of K_{naive}, $K_1(x_0)$, $K_3(x_0[M], M)$, K_4,and K_5 is either straightforward or is obtained from the algorithms described in Section 4. However, computing $K_2(x_0[M], M)$, for $M \neq \{1, ..., 10\}$, necessitates to compute $\sup\{K_1(x) : x[M] = x_0[M]\}$. We approximate this supremum by using a three-step Monte Carlo procedure: First, we randomly sample $100,000$ independent vectors $x \in \mathbb{R}^{10}$, so that $x[M] = x_0[M]$ and $x[M^c]$ follows a Gaussian distribution with mean vector $0 \in \mathbb{R}^{10-|M|}$ and covariance matrix $(1/n)(X'[M^c]X[M^c])$. For each of these vectors, we evaluate $K_1(x)$ with Algorithm 4.1, with $I_1 = 1,000$ Monte Carlo samples. In the second step, we keep the $1,000$ vectors x corresponding to the largest evaluations of $K_1(x)$ and we reevaluate $K_1(x)$ for them, with a number of Monte Carlo samples equal to $I_2 = 100,000$ in Algorithm 4.1. In the third step, we keep the vector x from the second step corresponding to the largest value of K_1 and we reevaluate $K_1(x)$ for this x, but this time with a number of Monte Carlo samples equal to $I_3 = 1,000,000$ in Algorithm 4.1.

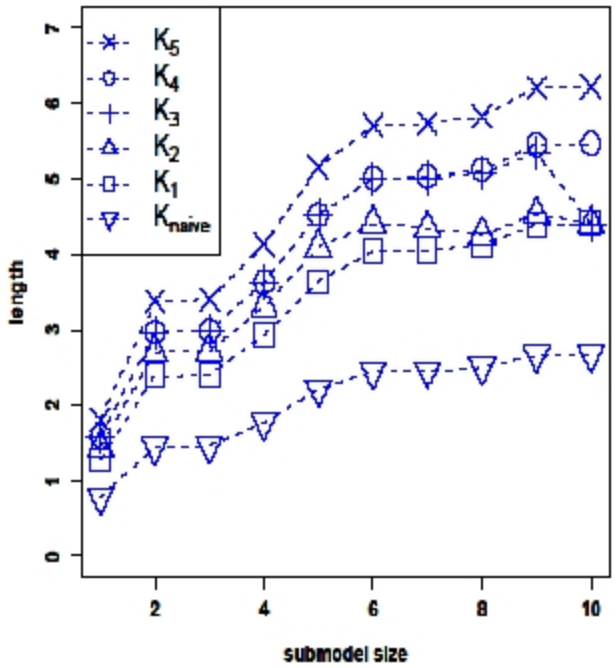

Figure 1: Standardized lengths of various confidence intervals as function of model size. Dashed lines are added to improve readability.

The standardized lengths of the confidence intervals corresponding to the constants K_{naive}, $K_1,\ldots,$ K_5 are reported in Figure 1 for the ten nested submodels mentioned before. We first see that for each of the constants K_{naive}, K_1, K_4, and K_5 the standardized length of the confidence interval increases with submodel size, which must hold since these constants do not depend on the submodel M and since the term $||s_M||$ increases with submodel size (for nested submodels as considered here). However, as discussed after Proposition 2.3, the values of K_2 and K_3 decrease with submodel size for nested submodels. Figure 1 shows that the combined effect of the increase of $||s_M||$ and the decrease of K_2 and K_3 with submodel size can be an increase or a decrease of the standardized lengths of the confidence intervals. Indeed, the standardized lengths increase globally (i.e., from submodel size 1 to 10), but can decrease locally (for example, the standardized length of the confidence interval obtained from K_2 decreases from submodel size 6 to submodel size 8; for the interval obtained from K_3 the standardized length decreases from submodel size 9 to submodel size 10). In Figure 1 the decreases of the standardized lengths occur only between submodel sizes for which $||s_M||$ is almost constant with M (which can be seen from the standardized lengths obtained from, say, K_5, since they are proportional to $||s_M||$). We also see from Figure 1 that the 'naive' interval is much shorter than the other intervals (at the price of not having the correct minimal coverage probability). The difference in standardized length between the intervals based on K_1 and K_2, respectively, is noticeable but not dramatic. A larger increase in standardized length is noted when comparing the interval based on the costly-to-compute constant K_2 with the one obtained from K_3, especially for submodel sizes 6 to 9. Furthermore, the standardized lengths of the confidence intervals obtained from K_3 are very close to those obtained from K_4 for model size 1 to 8. Finally, in Figure 1 we also see that the confidence intervals obtained from K_1, K_2, and K_3 have the same standardized length when the model size is 10, and that the same is true for the confidence intervals obtained from K_3 and K_4 when the model size is 1. This, of course, is not a coincidence, but holds necessarily as has been noted in the discussion of Proposition 2.3.

Additional computations of confidence interval lengths, with X and x_0 now randomly generated, yield results very similar to those in Figure 1. For the sake of brevity, these results are not shown here. We find, in particular, that the standardized length of the confidence interval obtained from K_3 always increase with submodel size when they are averaged with respect to X and x_0, but as in Figure 1 can decrease locally when not averaged. [In these additional numerical studies we did not consider the constant K_2 due to the high computational cost involved in its evaluation.]

We next investigate the minimal coverage probabilities of the intervals obtained from the constants K_{naive}, K_1, K_3, and K_4 when used as confidence intervals for the target $x_0[\hat{M}]'\beta^{(n)}_{\hat{M}}$ on the one hand as well as for the target $x_0[\hat{M}]'\beta^{(\star)}_{\hat{M}}$ on the other hand. We do not report results with confidence intervals obtained from K_2, since the computation of K_2 is too costly for the study we present below. The results for confidence intervals obtained from K_5 would be qualitatively similar to these for confidence intervals obtained from K_4, so we do not report them for the sake of brevity.

We consider minimal coverage probabilities in the setting where $\alpha = 0.05$, $p = 10$, $n = 20$ or $n = 100$, and the variance parameter is estimated by the standard unbiased estimator

using the full model, so that $r = n - p$. For model selection we consider, similar as in the simulation study of Leeb et al. (2013), AIC-, BIC-procedures, and the LASSO. For these three procedures, we always protect the first explanatory variable (which corresponds to an intercept term) from selection. However, the information that the first variable is protected is not used in computing the constants K_i, i.e., we do not use a restricted universe of models as described in Remark 2.7. [Additional simulations with no intercept term and no protected explanatory variable lead to results very similar to the ones given in Table 1 below.] For the AIC- and BIC-procedures we use the `step()` function in `R`, with penalty parameter `k` equal to 2 for AIC and $\log(n)$ for BIC. The AIC and BIC objective functions are minimized through a greedy general-to-specific search over the resulting 2^{p-1} candidate models (recall that the intercept is protected). For the LASSO, the selected model corresponds to the explanatory variables for which the LASSO estimator has non-zero coefficients. More precisely, we use the `lars` package in `R` and follow suggestions outlined in Efron et al. (2004): To protect the first regressor, 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 design 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`, but otherwise use the default settings. The selected model is comprised of those regressors in $\tilde{X}$ for which the corresponding LASSO coefficients are non-zero, plus the first column of X.

The 10×10 matrix Σ of (uncentered) second moments is chosen to be of the form

$$\Sigma = \begin{pmatrix} 1 & 0 \cdots 0 \\ 0 & \\ \vdots & \tilde{\Sigma} \\ 0 & \end{pmatrix},$$

where we consider three choices for the 9×9 matrix $\tilde{\Sigma}$. For the first case, $\tilde{\Sigma}$ is obtained by removing the first row and column of the 10×10 empirical covariance matrix (standardized by $n - 1 = 29$) of the variables in the watershed design matrix X_{Raw} . For the second case, we define $\mathbf{X}^{(\tilde{p})}(a)$ as in Section 6.1 of Berk et al. (2013a), with $\tilde{p} = 9$ and $a = 10$ and we set $\tilde{\Sigma} = (\mathbf{X}^{(\tilde{p})}(a))'(\mathbf{X}^{(\tilde{p})}(a))$. For the third case, we define $\mathbf{X}^{(\tilde{p})}(c)$ as in Section 6.2 of Berk et al. (2013a), with $\tilde{p} = 9$ and $c = \sqrt{0.8/(\tilde{p} - 1)}$ and we set $\tilde{\Sigma} = (\mathbf{X}^{(\tilde{p})}(c))'(\mathbf{X}^{(\tilde{p})}(c))$. Similar as in Berk et al. (2013a) and Leeb et al. (2013), we refer to the data set obtained in the second case as the exchangeable data set, and to the one obtained in the third case as the equicorrelated data set. For a given configuration of n and Σ, we then sample independently $n+1$ vectors of dimension 10×1 such that for each of these vectors the first component is 1 and the remaining nine components are jointly normally distributed with mean zero and covariance matrix $\tilde{\Sigma}$. The transposes of the first n of theses vectors now form the rows of the $n \times p$ design matrix X, while the $n + 1$-th of these vectors is used for the p-dimensional vector x_0.

Consider now a given configuration of n, Σ, the model selection procedure, the target (either the design-dependent or the design-independent target), as well as of a matrix X and a vector x_0 that have been obtained in the manner just described. Then, we estimate the minimal (over β and σ^2) coverage probabilities (conditional on X and x_0) of the confidence intervals obtained from the constants K_{naive}, K_1, K_3, and K_4 for the given target under

investigation. The minimal coverage probabilities are estimated by a three-step Monte Carlo procedure similar to that of Leeb et al. (2013), where for each of the constants K_{naive}, K_1, K_3, and K_4 we do the following: We first sample independently $m_1 = 1,000$ parameters β from a p-dimensional random vector b where Xb follows a standard Gaussian distribution within the column-space of X. Then, for each of these vectors β, we draw $I_1 = 1000$ Monte Carlo samples from the overall model using β and $\sigma = 1$ as the true parameters. For each Monte Carlo sample, we use the standard unbiased estimator $\hat{\sigma}^2$ of the error variance (under the full linear model), we carry out the model-selection procedure $\hat{M}$, and we record whether or not the target currently under investigation is covered by the confidence interval obtained from (5) with $K(x_0, \hat{M})$ replaced by the constant K under investigation. For each β, the I_1 recorded results are then averaged, resulting in m_1 Monte Carlo estimates of the coverage probabilities for the sampled vectors β. Then for the $m_2 = 100$ vectors β corresponding to the smallest estimated coverage probabilities from the first step, we repeat the Monte Carlo procedures, but this time with $I_2 = 10,000$ Monte Carlo samples, and we record the vector β that yields the smallest estimate for the coverage probability in this second step. Performing these two steps for each of the four constants K_{naive}, K_1, K_3, and K_4 results in four vectors $\beta(1)$, $\beta(2)$, $\beta(2)$, and $\beta(4)$. In a third step, we now reevaluate the coverage probability of any of the four confidence intervals at each of the vectors $\beta(j)$, $j = 1, \ldots, 4$, this time now with $I_3 = 100,000$ Monte Carlo samples, and record, for each of the confidence intervals, the minimum of these four estimates of the coverage probabilities. This is then used as the final estimate of the minimal coverage probability of the confidence interval under consideration. We stress here that the minimal coverage probabilities found by this Monte Carlo 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.

Table 1 summarizes the estimated minimal coverage probabilities for the various confidence sets and targets, and for the model-selection procedures and data sets considered in the study. The conclusions are pretty much the same for the three data sets. First, we observe that, for $n = 20$, the differences of minimal coverage probabilities between the design-dependent and independent targets can be significant, especially for the 'naive' intervals and for the other intervals in case the LASSO model selector is used. However, for $n = 100$, these differences are very small for all the configurations. This is in line with Lemma B.1 in Appendix B, stating that for a large family of model selection procedures, the difference of coverage probabilities between the two targets vanishes, uniformly in β and σ^2, when n increases. For $n = 100$, the results are thus almost identical for the two targets: For the three model selection procedures, the confidence intervals obtained from the constants K_1, K_3, and K_4 are valid, while the 'naive' confidence intervals are moderately too short, so that their minimal coverage probabilities are below the nominal level, with a minimum of 0.84.

For $n = 20$ and when AIC or BIC is used, the 'naive' confidence intervals fail to have the right coverage probabilities to a somewhat larger extent than in case $n = 100$. Their minimal coverage probabilities can be as small as 0.81 for the design-dependent target and 0.74 for the design-independent target. Furthermore, the confidence intervals obtained from the constants K_1, K_3, and K_4 remain valid here for both targets. However, when $n = 20$ and the LASSO model selector is used, the results for the design-independent target are drastically different from those obtained with the AIC- or BIC-procedures: All four confidence intervals have

Data set	n	Model selector	Target							
			design-dependent $x_0[\hat{M}]'\beta^{(n)}_{\hat{M}}$				design-independent $x_0[\hat{M}]'\beta^{(\star)}_{\hat{M}}$			
			K_{naive}	K_1	K_3	K_4	K_{naive}	K_1	K_3	K_4
Watershed	20	AIC	0.84	0.99	1.00	1.00	0.79	0.97	0.99	0.99
	20	BIC	0.84	0.99	1.00	1.00	0.74	0.96	0.98	0.98
	20	LASSO	0.90	1.00	1.00	1.00	0.18	0.48	0.61	0.61
	100	AIC	0.87	0.99	1.00	1.00	0.88	0.99	1.00	1.00
	100	BIC	0.88	0.99	1.00	1.00	0.87	0.99	1.00	1.00
	100	LASSO	0.88	0.99	1.00	1.00	0.87	0.99	1.00	1.00
Exchangeable	20	AIC	0.83	0.99	1.00	1.00	0.80	0.98	0.99	0.99
	20	BIC	0.84	0.99	1.00	1.00	0.76	0.97	0.99	0.99
	20	LASSO	0.90	1.00	1.00	1.00	0.46	0.86	0.93	0.92
	100	AIC	0.89	0.99	1.00	1.00	0.90	0.99	1.00	1.00
	100	BIC	0.90	0.99	1.00	1.00	0.90	0.99	1.00	1.00
	100	LASSO	0.90	0.99	1.00	1.00	0.90	0.99	1.00	1.00
Equicorrelated	20	AIC	0.83	0.99	1.00	1.00	0.79	0.98	0.99	0.99
	20	BIC	0.81	0.99	1.00	1.00	0.74	0.98	0.99	0.99
	20	LASSO	0.88	1.00	1.00	1.00	0.39	0.71	0.79	0.79
	100	AIC	0.84	0.99	1.00	1.00	0.84	0.99	1.00	1.00
	100	BIC	0.86	0.99	1.00	1.00	0.86	0.99	1.00	1.00
	100	LASSO	0.88	1.00	1.00	1.00	0.88	1.00	1.00	1.00

Table 1: Monte Carlo estimates of the minimal coverage probabilities (w.r.t. β and σ^2) of various confidence intervals. The nominal coverage probability is $1 - \alpha = 0.95$ and $p = 10$.

minimal coverage probabilities for the design-independent target that are below, and in most cases significantly below, the nominal level. The failure of all the confidence intervals is here often more pronounced than the failure of the 'naive' confidence intervals when other model selectors are used. Especially for the watershed data set, the estimated minimal coverage probability is 0.18 for the 'naive' interval and 0.48 for the confidence interval based on K_1. The reason for this phenomenon can be traced to the observation that the LASSO model selector, as implemented here and for the parameters used in the stochastic search for the smallest coverage probability, selects models that are significantly smaller than those AIC and BIC select. In particular, the LASSO procedure often excludes regressors for which the corresponding regression coefficients are not small. In our simulation study, selecting a small model, that excludes regressors with significant coefficients, makes the difference between the design-dependent and independent targets larger. Since the confidence intervals are designed to cover the former target, they hence have a hard time to cover the latter when the two targets are significantly different. In other words, for $n = 20$, the left-hand side of the display in Condition 3.3 is not small for the LASSO procedure, so that Theorem 3.6 does not provide a good approximation of the finite-sample situation. [We stress here that this conclusion holds for the LASSO procedure as implemented here, that selects the penalty by cross-validation. Other implementations of the LASSO may of course give different results.]

The results in Table 1 concern the coverage probabilities conditional on the design matrix X and on x_0, and thus can depend on the values of X and x_0 used. In additional (non-exhaustive) simulations we repeated the above analysis for other values of X and x_0 and found similar results.

A Appendix: Proofs for Section 2

Lemma A.1. *Suppose W is a random $m \times 1$ vector that has a density that is positive almost everywhere. Let $a_1, \ldots, a_L$, for some $L \in \mathbb{N}$, be elements of $\mathbb{R}^m$, not all of which are zero. Define $h(w) = \max_{l=1,\ldots,L} |a_l' w|$, and set $H(t) = \Pr(h(W) \leq t)$ for $t \in \mathbb{R}$. Then H is continuous on $\mathbb{R}$, satisfies $H(t) = 0$ for $t \leq 0$, and is strictly increasing on $[0, \infty)$.*

Proof: For $t < 0$ the event $\{h(W) \leq t\}$ is empty; for $t = 0$ this event is an intersection of the sets $\{a_l' W = 0\}$ where at least one of these sets has probability zero because W possesses a density and not all a_l are zero. Consequently, $H(t) = 0$ for $t \leq 0$ follows. Because H is a distribution function, continuity of H on $\mathbb{R}$ will follow if we can establish continuity on $(0, \infty)$. Now, for every $t > 0$ the event $\{h(W) = t\}$ is contained in the union of the events $\{|a_l' W| = t\}$ for which $a_l \neq 0$ holds. Since any of these events has probability zero, it follows that $\Pr(\{h(W) = t\}) = 0$ and consequently H is continuous on $(0, \infty)$. It remains to establish the claim regarding strict monotonicity: For $t > 0$ the set $A(t) = \{w : h(w) \leq t\}$ contains a sufficiently small ball centered at the origin because $h(0) = 0$ and h is continuous, and consequently $H(t) > 0$ follows by the assumption on the density of W. It hence suffices to show that $0 < t_1 < t_2$ implies $H(t_1) < H(t_2)$. Because not all a_l are zero and h is positively homogeneous of degree one, we can find an element $w_1 \in A(t_1)$ such that $h(w_1) = t_1$ holds. But then there exists an l_1 such that $|a_{l_1}' w_1| = t_1$ and $|a_l' w_1| \leq t_1$ for all l hold. In fact, we may assume that $a_{l_1}' w_1 = t_1$ holds (otherwise we change the sign of w_1). Consider the set B

consisting of all $w \in \mathbb{R}^m$ such that $a'_{l_1}(w - w_1) > 0$ and such that $|a'_l(w - w_1)| < (t_2 - t_1)/2$ for every l. Then $B \subseteq A(t_2) \backslash A(t_1)$ holds, since for $w \in B$

$$h(w) \leq \max_{l=1,\ldots,L} \left|a'_l(w - w_1)\right| + h(w_1) < (t_2 - t_1)/2 + t_1 = (t_1 + t_2)/2 < t_2,$$

$$a'_{l_1} w > a'_{l_1} w_1 = t_1 > 0,$$

and hence also $h(w) \geq \left|a'_{l_1} w\right| > t_1$ hold. But B obviously has positive Lebesgue measure, implying that $H(t_2) - H(t_1) = \Pr(A(t_2) \backslash A(t_1)) > 0$. ■

Remark A.2. In the special case where $W = W_1 W_2$ with W_1 a random $m \times 1$ vector having a density that is positive almost everywhere, with W_2 a random variable that is independent of W_1, is positive almost surely, and has a density that is almost everywhere positive on $(0, \infty)$, an alternative, and perhaps simpler, proof is as follows: Set $H^*(t) = \Pr(h(W_1) \leq t)$. We conclude that H^* is continuous on $\mathbb{R}$ and satisfies $H^*(t) = 0$ for $t \leq 0$ by repeating the corresponding arguments in the preceding proof. The same properties for $H(t) = \mathbb{E}_{W_2} H^*(t/W_2)$ then follow immediately. To establish strict monotonicity of H on $[0, \infty)$ consider $0 \leq t_1 < t_2$. It is not difficult to see that we can then find $w_2 > 0$ such that $H^*(t_1/w_2) < H^*(t_2/w_2)$ holds since otherwise H^* would have to be constant on $[0, \infty)$ which is impossible since $H^*(0) = 0$ and H^* is a distribution function. By continuity of H^* then also $H^*(t_1/w'_2) < H^*(t_2/w'_2)$ must hold for every w'_2 in a sufficiently small neighborhood of w_2. Since H^* is nondecreasing and since the distribution of W_2 puts positive mass on the aforementioned neighborhood, we can conclude that $\mathbb{E}_{W_2} H^*(t_1/W_2) < \mathbb{E}_{W_2} H^*(t_2/W_2)$, i.e., that $H(t_1) < H(t_2)$ holds.

The following lemma will be used in the proof of Proposition 2.3 below.

Lemma A.3. *Suppose F^* is a distribution function on $\mathbb{R}$ that is continuous at zero. Let S be a random variable that is positive with probability one and has a continuous distribution function. Then $F(t) = \mathbb{E}_S F^*(t/S)$ is continuous on $\mathbb{R}$.*

Proof: Let S^* be a random variable which is independent of S and which has distribution function F^*. Then $F(t) = \mathbb{E}_S \mathbb{E}_{S^*} \mathbf{1}(S^* \leq t/S) = \Pr(SS^* \leq t)$. Because $S^* \neq 0$ holds almost surely by the assumption on F^*, we have $\Pr(SS^* = t) = \mathbb{E}_{S^*} \mathbb{E}_S \mathbf{1}(S = t/S^*)$. Since S has a continuous distribution function, we have $\mathbb{E}_S \mathbf{1}(S = t/S^*) = 0$ almost surely, implying that $\Pr(SS^* = t) = 0$. ■

Proof of Proposition 2.3: To prove the first claim observe that in case $M = \{1, \ldots, p\}$ and $x_0 = 0$ it is obvious that F^*_{M,x_0}, and hence also F_{M,x_0}, is the indicator function of $[0, \infty)$, which then implies $K_3(x_0[M], M) = 0$. If $M = \{1, \ldots, p\}$ and $x_0 \neq 0$, then F_{M,x_0} is continuous on $\mathbb{R}$, satisfies $F_{M,x_0}(t) = 0$ for $t \leq 0$, and is strictly increasing on $[0, \infty)$ in view of Lemma A.1 since $F_{M,x_0}(t)$ here reduces to $P_{n,\beta,\sigma}\left(\max_{M_* \subseteq \{1,\ldots,p\}} \left|\bar{s}'_{M_*}(Y - X\beta)\right|/\hat{\sigma} \leq t\right)$ (see (26) and (27) below) and since not all $\bar{s}_{M_*}$ are zero.

In case M is a proper subset of $\{1, \ldots, p\}$ and $p > 1$ holds we argue as follows: Note that then $F^*_{M,x_0}(0) = 0$ holds since $2^p - 2^{|M|} > 1$ and since $F_{Beta,1/2,(p-1)/2}(0) = 0$. Hence F^*_{M,x_0} is continuous at $t = 0$. We may apply Lemma A.3 to conclude that F_{M,x_0} is continuous on $\mathbb{R}$ and thus satisfies $F_{M,x_0}(0) = 0$ (since $F_{M,x_0}(t) = 0$ for $t < 0$ by its definition). Next let $0 \leq t_1 < t_2$. Because $F^*_{M,x_0}(0) = 0$ as noted before and because

$F^*_{M,x_0}(1) = 1$ (since $\Pr\left(\max_{M_*\subseteq M}\left|\bar{s}'_{M_*}V\right| > 1\right) = 0$ and $F_{Beta,1/2,(p-1)/2}(1) = 1$) we thus can find a positive g_0 such that $F^*_{M,x_0}(t_1/g_0) < F^*_{M,x_0}(t_2/g_0)$ holds (if not, constancy of F^*_{M,x_0} on $[0,\infty)$ would have to follow). Because of continuity from the right at t_1/g_0 it follows that $F^*_{M,x_0}(t_1/g) < F^*_{M,x_0}(t_2/g)$ also holds for all $g < g_0$ in a sufficiently small neighborhood of g_0 that is contained in $(0,\infty)$. Because $F^*_{M,x_0}(t_1/g) \leq F^*_{M,x_0}(t_2/g)$ holds for every $g > 0$ and because G has a density that is positive everywhere on $(0,\infty)$, the strict inequality $\mathbb{E}_G F^*_{M,x_0}(t_1/G) < \mathbb{E}_G F^*_{M,x_0}(t_2/G)$ follows. This establishes strict monotonicity of F_{M,x_0} on $[0,\infty)$ also in this case.

Finally, if M is a proper subset of $\{1,\ldots,p\}$ and $p = 1$ holds, then M is empty and F^*_{M,x_0} reduces to the indicator function of $[1,\infty)$. But then $F_{M,x_0}(t) = \Pr(G \leq t)$ which obviously is continuous on $\mathbb{R}$, takes the value zero at $t = 0$, and is strictly increasing on $[0,\infty)$. This completes the proof of the first claim.

To prove the remaining claims observe that $\bar{s}_{M_*}$ belongs to the column space of X for every $M_* \subseteq \{1,\ldots,p\}$ and hence we have

$$\begin{aligned} & P_{n,\beta,\sigma}\left(\max_{M_*\subseteq\{1,\ldots,p\}}\left|\bar{s}'_{M_*}(Y - X\beta)\right|/\hat{\sigma} > t\right) \\ = \; & P_{n,\beta,\sigma}\left(\max_{M_*\subseteq\{1,\ldots,p\}}\left|\bar{s}'_{M_*}P_X(Y - X\beta)/\left\|P_X(Y - X\beta)\right\|\right| > \left(\hat{\sigma}/\left\|P_X(Y - X\beta)\right\|\right)t\right), \quad (26) \end{aligned}$$

where $P_X(Y - X\beta)/\left\|P_X(Y - X\beta)\right\|$ and $\left\|P_X(Y - X\beta)\right\|/\hat{\sigma}$ are independent since the random variables $P_X(Y - X\beta)/\left\|P_X(Y - X\beta)\right\|$, $\left\|P_X(Y - X\beta)\right\|$, and $\hat{\sigma}$ are mutually independent. Consequently, the probability given above can be represented as

$$\Pr\left(\max_{M_*\subseteq\{1,\ldots,p\}}\left|\bar{s}'_{M_*}V\right| > t/G\right) \quad (27)$$

where V and G are independent and otherwise are as in the definition of F^*_{M,x_0} and F_{M,x_0}. Now, using first independence of V and G and then a union bound twice we have for $t \geq 0$

$$\begin{aligned} \Pr\left(\max_{M_*\subseteq\{1,\ldots,p\}}\left|\bar{s}'_{M_*}V\right| > t/G\right) &= \int \Pr\left(\max_{M_*\subseteq\{1,\ldots,p\}}\left|\bar{s}'_{M_*}V\right| > t/g\right) dF_G(g) \\ \leq \; & \int \min\left[1, \Pr\left(\max_{M_*\subseteq M}\left|\bar{s}'_{M_*}V\right| > t/g\right) + \Pr\left(\max_{M_*\not\subseteq M}\left|\bar{s}'_{M_*}V\right| > t/g\right)\right] dF_G(g) \\ \leq \; & \int \min\left[1, \Pr\left(\max_{M_*\subseteq M}\left|\bar{s}'_{M_*}V\right| > t/g\right) + \sum_{M_*\not\subseteq M}\Pr\left(\left|\bar{s}'_{M_*}V\right| > t/g\right)\right] dF_G(g) \\ = \; & \int \min\left[1, \Pr\left(\max_{M_*\subseteq M}\left|\bar{s}'_{M_*}V\right| > t/g\right) + \sum_{M_*\not\subseteq M}\Pr\left(\left(\bar{s}'_{M_*}V\right)^2 > t^2/g^2\right)\right] dF_G(g) \\ \leq \; & \int \left(1 - F^*_{M,x_0}(t/g)\right) dF_G(g) = \mathbb{E}_G\left(1 - F^*_{M,x_0}(t/G)\right) = 1 - F_{M,x_0}(t), \quad (28) \end{aligned}$$

where F_G here denotes the c.d.f. of G. The last inequality follows from the fact that $\Pr\left(\left(\bar{s}'_{M_*}V\right)^2 > t^2/g^2\right)$ is either equal to zero (if $\bar{s}_{M_*} = 0$) or is equal to $1 - F_{Beta,1/2,(p-1)/2}\left(t^2/g^2\right)$

(if $\bar{s}_{M_*} \neq 0$) as is easy to see; for the case where M is the empty set also observe that $\Pr\left(\max_{M_* \subseteq M} \left|\bar{s}'_{M_*} V\right| > t/g\right) = 0$ for $t \geq 0$ because $\bar{s}_{\varnothing} = 0$. In view of (10) the chain of inequalities in (26), (28) establishes $K_1(x_0) \leq K_3(x_0[M], M)$. It follows that $K_1(x) \leq K_3(x[M], M) = K_3(x_0[M], M)$ for every x satisfying $x[M] = x_0[M]$, implying $K_2(x_0[M], M) \leq K_3(x_0[M], M)$. The inequality (16) is obvious and inequality (17) follows since for $t \geq 0$ we have (again noting that expressions like $\Pr\left(\max_{M_* \subseteq M_1} \left|\bar{s}'_{M_*} V\right| > t\right)$ for $t \geq 0$ are equal to zero if M_1 is empty)

$$
\begin{aligned}
& \Pr\left(\max_{M_* \subseteq M_2} \left|\bar{s}'_{M_*} V\right| > t\right) + \left(2^p - 2^{|M_2|}\right)\left(1 - F_{Beta,1/2,(p-1)/2}\left(t^2\right)\right) \\
\leq\ & \Pr\left(\max_{M_* \subseteq M_1} \left|\bar{s}'_{M_*} V\right| > t\right) + \Pr\left(\max_{M_* \subseteq M_2, M_* \not\subseteq M_1} \left|\bar{s}'_{M_*} V\right| > t\right) \\
& + \left(2^p - 2^{|M_2|}\right)\left(1 - F_{Beta,1/2,(p-1)/2}\left(t^2\right)\right) \\
\leq\ & \Pr\left(\max_{M_* \subseteq M_1} \left|\bar{s}'_{M_*} V\right| > t\right) + \sum_{M_* \subseteq M_2, M_* \not\subseteq M_1} \Pr\left(\left|\bar{s}'_{M_*} V\right| > t\right) \\
& + \left(2^p - 2^{|M_2|}\right)\left(1 - F_{Beta,1/2,(p-1)/2}\left(t^2\right)\right) \\
\leq\ & \Pr\left(\max_{M_* \subseteq M_1} \left|\bar{s}'_{M_*} V\right| > t\right) + \left(2^p - 2^{|M_1|}\right)\left(1 - F_{Beta,1/2,(p-1)/2}\left(t^2\right)\right).
\end{aligned}
$$

The relation $K_3(x_0[M], M) \leq K_4$ is now immediate. Finally, $1 - F^*_{\varnothing,x_0}(t) \leq 1$ for all $t \in \mathbb{R}$ and $1 - F^*_{\varnothing,x_0}(t) = 0$ for $t > 1$ lead to

$$
\begin{aligned}
1 - F_{\varnothing,x_0}(t) &= \mathbb{E}_G\left(1 - F^*_{\varnothing,x_0}(t/G)\right) = \mathbb{E}_G\left(\left(1 - F^*_{\varnothing,x_0}(t/G)\right)\mathbf{1}(t \leq G)\right) \\
&\leq \mathbb{E}_G \mathbf{1}(t \leq G) = 1 - \Pr(G \leq t),
\end{aligned}
$$

which proves $K_4 \leq K_5$. ■

Lemma A.4. *Assume $p = 2$ and $n \geq 2$. Then there exists a design matrix X and a vector x_0 such that $K_4 = K_1(x_0)$.*

Proof: Assume first that $n = 2$. In view of the definition of $K_4 = K_3(x_0[\varnothing], \varnothing)$ it suffices to exhibit a 2×2 matrix X and a 2×1 vector x_0 such that equality holds between the far l.h.s. and the far r.h.s. of (28) for $M = \varnothing$ and all $t \geq 0$. Inspection of (28) shows that for this it suffices to find X and x_0 such that

$$
\Pr\left(\max_{\varnothing \neq M_* \subseteq \{1,2\}} \left|\bar{s}'_{M_*} V\right| > c\right) = \min\left(1, \sum_{\varnothing \neq M_* \subseteq \{1,2\}} \Pr\left(\left|\bar{s}'_{M_*} V\right| > c\right)\right)
$$

holds for every $c \geq 0$ and that $\bar{s}'_{M_*} \neq 0$ for every $\varnothing \neq M_* \subseteq \{1, 2\}$. This is achieved for

$$
X = X^{(2)} = \begin{bmatrix} 1 & \cos(2\pi/3) \\ 0 & \sin(2\pi/3) \end{bmatrix}
$$

and $x_0^{(2)\prime} = (\cos(4\pi/3), \sin(4\pi/3))\, X^{(2)}$: Then $\bar{s}'_{\{1\}} = -(1,0)$, $\bar{s}'_{\{2\}} = -(\cos(2\pi/3), \sin(2\pi/3))$, and $\bar{s}'_{\{1,2\}} = (\cos(4\pi/3), \sin(4\pi/3))$. Consequently, the event $\{\max_{\varnothing \neq M_* \subseteq \{1,2\}} |\bar{s}'_{M_*} V| > c\}$ is either the entire space or is the disjoint union of the events $\{\left|\bar{s}'_{\{1\}} V\right| > c\}$, $\{\left|\bar{s}'_{\{2\}} V\right| > c\}$ and $\{\left|\bar{s}'_{\{1,2\}} V\right| > c\}$. In the case $n > 2$ simply set

$$X = \left(X^{(2)\prime}, 0, \ldots, 0\right)'$$

and $x_0' = (\cos(4\pi/3), \sin(4\pi/3), 0, \ldots, 0)\, X$. ■

Remark A.5. Further examples of pairs X, x_0 satisfying the above lemma can be generated from the matrix X constructed in the proof by premultiplying X by an orthogonal matrix and leaving x_0 unchanged.

B Appendix: Proofs for Section 3

In the subsequent lemma we assume that $\tilde{\sigma}_1$ and $\tilde{\sigma}_2$ are defined on the same probability space as are Y, X, and $\hat{\sigma}^2$. In slight abuse of notation, we shall then denote by $P_{n,\beta,\sigma}$ the joint distribution of Y, X, $\hat{\sigma}^2$, $\tilde{\sigma}_1$, and $\tilde{\sigma}_2$. We note that an argument corresponding to a special case of this lemma has been used in Ewald (2012).

Lemma B.1. *Suppose that the maintained model assumptions of Section 3 are satisfied. Assume further that Conditions 3.3 and 3.4 hold. Let $\mathcal{W}$ be the set of all measurable non-negative functions of the form $W(x_0, X, M)$. Then, for any two sequences of random variables $\tilde{\sigma}_1 = \tilde{\sigma}_{1,n}$ and $\tilde{\sigma}_2 = \tilde{\sigma}_{1,n}$ (which may be functions of σ) satisfying*

$$\sup_{\beta \in \mathbb{R}^p, \sigma > 0} P_{n,\beta,\sigma}\left(|(\tilde{\sigma}_i/\sigma) - 1| > \delta \middle| X\right) \to 0 \tag{29}$$

in probability as $n \to \infty$ for every $\delta > 0$ and for $i = 1, 2$, we have that

$$\begin{aligned} &\sup_{x_0 \in \mathbb{R}^p, \beta \in \mathbb{R}^p, \sigma > 0, W \in \mathcal{W}} \left| P_{n,\beta,\sigma}\left(\left|x_0'[\hat{M}]\hat{\beta}_{\hat{M}} - x_0'[\hat{M}]\beta^{(\star)}_{\hat{M}}\right| \leq W(x_0, X, \hat{M})\tilde{\sigma}_1 \middle| X\right)\right. \\ &\left. - P_{n,\beta,\sigma}\left(\left|x_0'[\hat{M}]\hat{\beta}_{\hat{M}} - x_0'[\hat{M}]\beta^{(n)}_{\hat{M}}\right| \leq W(x_0, X, \hat{M})\tilde{\sigma}_2 \middle| X\right)\right| \end{aligned}$$

converges to 0 in probability as $n \to \infty$.

Proof: Because the number of variables p is fixed, it suffices to show for arbitrary but fixed $M \subseteq \{1, \ldots, p\}$ that

$$\begin{aligned} Q_n \;=\; &\sup_{x_0 \in \mathbb{R}^p, \beta \in \mathbb{R}^p, \sigma > 0, W \in \mathcal{W}} \left| P_{n,\beta,\sigma}\left(\left|x_0'[M]\hat{\beta}_M - x_0'[M]\beta^{(\star)}_M\right| \leq W_M\tilde{\sigma}_1; \hat{M} = M \middle| X\right)\right. \\ &\left. - P_{n,\beta,\sigma}\left(\left|x_0'[M]\hat{\beta}_M - x_0'[M]\beta^{(n)}_M\right| \leq W_M\tilde{\sigma}_2; \hat{M} = M \middle| X\right)\right| \end{aligned}$$

goes to 0 in probability, where we have used the abbreviation $W_M = W(x_0, X, M)$. We may assume in what follows that $M \neq \varnothing$ since otherwise Q_n is zero. Furthermore, Q_n does not

change its value if the supremum is restricted to those x_0 which have $\|x_0[M]\| = 1$ (since the expression inside the supremum is identically zero if x_0 satisfies $x_0[M] = 0$ and since otherwise the norm of $x_0[M]$ can be absorbed into W_M). Hence we have

$$\begin{aligned} Q_n &= \sup_{x_0 \in S(M), \beta \in \mathbb{R}^p, \sigma > 0, W \in \mathcal{W}} \Big| P_{n,\beta,\sigma}\left(|\sigma e_1 + \sigma e_2| \le W_M \tilde{\sigma}_1; \hat{M} = M \Big| X \right) \\ &\quad - P_{n,\beta,\sigma}\left(|\sigma e_2| \le W_M \tilde{\sigma}_2; \hat{M} = M \Big| X \right) \Big| \end{aligned} \tag{30}$$

where we have used the abbreviations $S(M) = \{x_0 \in \mathbb{R}^p : \|x_0[M]\| = 1\}$,

$$e_1 = \sigma^{-1} n^{1/2} x_0'[M] \left(\left(X'[M]X[M] \right)^{-1} X'[M]X[M^c] - \left(\Sigma[M,M] \right)^{-1} \Sigma[M, M^c] \right) \beta[M^c]$$

and

$$e_2 = \sigma^{-1} n^{1/2} x_0'[M] \left(X'[M]X[M] \right)^{-1} X'[M] \left(Y - X\beta \right).$$

Note that we have also absorbed a factor $n^{1/2}$ into W_M, which is possible because of the supremum operation w.r.t. W_M. Using the inequality $|\Pr(A \cap C) - \Pr(B \cap C)| \le \Pr(A^c \cap B \cap C) + \Pr(A \cap B^c \cap C)$ we can bound the absolute value inside the supremum in (30) by

$$\begin{aligned} &P_{n,\beta,\sigma}\left(|\sigma e_1 + \sigma e_2| > W_M \tilde{\sigma}_1; |\sigma e_2| \le W_M \tilde{\sigma}_2; \hat{M} = M \Big| X \right) \\ &+ P_{n,\beta,\sigma}\left(|\sigma e_1 + \sigma e_2| \le W_M \tilde{\sigma}_1; |\sigma e_2| > W_M \tilde{\sigma}_2; \hat{M} = M \Big| X \right). \end{aligned} \tag{31}$$

Let now $\delta_{n,1}$ be an arbitrary sequence of positive numbers converging to zero. Then we can further bound the above expression by

$$\begin{aligned} &P_{n,\beta,\sigma}\left(\tilde{\sigma}_1 \left(W_M - \delta_{n,1} \right)_+ \le |\sigma e_2| \le \tilde{\sigma}_2 W_M \Big| X \right) \\ &+ P_{n,\beta,\sigma}\left(\tilde{\sigma}_2 W_M \le |\sigma e_2| \le \tilde{\sigma}_1 \left(W_M + \delta_{n,1} \right) \Big| X \right) \\ &+ 2P_{n,\beta,\sigma}\left(|\sigma e_1| \ge \tilde{\sigma}_1 \delta_{n,1}; \hat{M} = M \Big| X \right). \end{aligned} \tag{32}$$

By the assumption on the estimators $\tilde{\sigma}_1$ and $\tilde{\sigma}_2$ we can find a sequence $\delta_{n,2} < 1$ of positive numbers converging to zero such that

$$\sup_{\beta,\sigma} P_{n,\beta,\sigma}\left(\min_{i=1,2} |(\tilde{\sigma}_i/\sigma) - 1| > \delta_{n,2} \Big| X \right) \to 0 \tag{33}$$

in probability as $n \to \infty$. This can easily be seen from a diagonal sequence argument. Now,

using (30), (31), (32), and (33), we have

$$\begin{aligned}
Q_n &\leq \sup_{x_0\in S(M),\beta\in\mathbb{R}^p,\sigma>0,W\in\mathcal{W}} P_{n,\beta,\sigma}\left(\left(1-\delta_{n,2}\right)\left(W_M-\delta_{n,1}\right)_+ \leq |e_2| \leq \left(1+\delta_{n,2}\right)W_M\middle| X\right) \\
&\quad + \sup_{x_0\in S(M),\beta\in\mathbb{R}^p,\sigma>0,W\in\mathcal{W}} P_{n,\beta,\sigma}\left(\left(1-\delta_{n,2}\right)W_M \leq |e_2| \leq \left(1+\delta_{n,2}\right)\left(W_M+\delta_{n,1}\right)|\, X\right) \\
&\quad +2\sup_{x_0\in S(M),\beta\in\mathbb{R}^p,\sigma>0,W\in\mathcal{W}} P_{n,\beta,\sigma}\left(|e_1| \geq \left(1-\delta_{n,2}\right)\delta_{n,1}; \hat{M}=M\middle| X\right) + o_p(1) \\
&\leq 2\sup_{x_0\in S(M),\beta\in\mathbb{R}^p,\sigma>0,W\in\mathcal{W}} P_{n,\beta,\sigma}\left(\left(1-\delta_{n,2}\right)\left(W_M-\delta_{n,1}\right)_+ \leq |e_2| \leq \left(1+\delta_{n,2}\right)\left(W_M+\delta_{n,1}\right)\middle| X\right) \\
&\quad +2\sup_{x_0\in S(M),\beta\in\mathbb{R}^p,\sigma>0} P_{n,\beta,\sigma}\left(|e_1| \geq \left(1-\delta_{n,2}\right)\delta_{n,1}; \hat{M}=M\middle| X\right) + o_p(1) \\
&= 2Q_{n,1} + 2Q_{n,2} + o_p(1).
\end{aligned}$$

We first bound $Q_{n,1}$ as follows: Observe that, conditionally on X, the quantity e_2 is normally distributed with mean zero and variance given by $c_n\left(x_0, X\right) = x_0'[M]\left(n^{-1}X'[M]X[M]\right)^{-1}x_0[M]$. By Condition 3.4 the variance $c_n\left(x_0, X\right)$ converges to $c(x_0) = x_0'[M]\left(\Sigma[M,M]\right)^{-1}x_0[M] > 0$ in probability, and in fact even uniformly in $x_0 \in S(M)$. Since $\Sigma[M,M]$ is obviously positive definite, $0 < c_* \leq c(x_0) \leq c^* < \infty$ must hold for all $x_0 \in S(M)$. Consequently,

$$\sup_{x_0\in S(M)}\left|\left(c_n^{1/2}\left(x_0, X\right)/c^{1/2}\left(x_0\right)\right) - 1\right| \tag{34}$$

converges to zero in probability. Therefore we can find a sequence $\delta_{n,3} \in (0,1)$ converging to zero for $n \to \infty$ such that the event D_n where (34) is less than $\delta_{n,3}$ has probability converging to 1. On this event $c_n\left(x_0, X\right)$ is then positive for sufficiently large n and we have on D_n and

for sufficiently large n

$$\begin{aligned}
Q_{n,1} &= 2 \sup_{x_0 \in S(M), W \in \mathcal{W}} \left\{ \Phi\left((1+\delta_{n,2}) (W_M + \delta_{n,1}) / c_n^{1/2}(x_0, X) \right) \right. \\
&\quad \left. - \Phi\left((1-\delta_{n,2}) (W_M - \delta_{n,1})_+ / c_n^{1/2}(x_0, X) \right) \right\} \\
&\leq 2 \sup_{x_0 \in S(M), W \in \mathcal{W}} \left\{ \Phi\left(\frac{1+\delta_{n,2}}{1-\delta_{n,3}} (W_M + \delta_{n,1}) / c^{1/2}(x_0) \right) \right. \\
&\quad \left. - \Phi\left(\frac{1-\delta_{n,2}}{1+\delta_{n,3}} (W_M - \delta_{n,1})_+ / c^{1/2}(x_0) \right) \right\} \\
&\leq 2 \sup_{x_0 \in S(M), W \in \mathcal{W}} \left\{ \Phi\left(\frac{1+\delta_{n,2}}{1-\delta_{n,3}} (W_M + \delta_{n,1}) / c^{1/2}(x_0) \right) \right. \\
&\quad \left. - \Phi\left(\frac{1-\delta_{n,2}}{1+\delta_{n,3}} (W_M - \delta_{n,1}) / c^{1/2}(x_0) \right) \right\} \\
&\leq 2 \sup_{x_0 \in S(M), z \geq 0} \left\{ \Phi\left(\frac{1+\delta_{n,2}}{1-\delta_{n,3}} \left(z + \delta_{n,1} / c^{1/2}(x_0) \right) \right) \right. \\
&\quad \left. - \Phi\left(\frac{1-\delta_{n,2}}{1+\delta_{n,3}} \left(z - \delta_{n,1} / c^{1/2}(x_0) \right) \right) \right\} \\
&\leq 2 \sup_{z \geq 0} \left\{ \Phi\left(\frac{1+\delta_{n,2}}{1-\delta_{n,3}} \left(z + \delta_{n,1} / c_*^{1/2} \right) \right) - \Phi\left(\frac{1-\delta_{n,2}}{1+\delta_{n,3}} \left(z - \delta_{n,1} / c_*^{1/2} \right) \right) \right\},
\end{aligned}$$

where Φ denotes the standard normal cdf. But the far right-hand side in the above display obviously converges to zero for $n \to \infty$ since $\delta_{n,1}$, $\delta_{n,2}$, as well as $\delta_{n,3}$ converge to zero. We have thus established that $Q_{n,1}$ converges to zero in probability as $n \to \infty$.

We next turn to $Q_{n,2}$. In case $M = \{1, \ldots, p\}$, we have that $e_1 = 0$, and hence $Q_{n,2} = 0$. Otherwise, from Condition 3.3 we can conclude (from a diagonal sequence argument) the existence of a sequence of positive numbers $\delta_{n,4}$ that converge to zero for $n \to \infty$ such that

$$\sup \left\{ P_{n,\beta,\sigma}(\hat{M} = M | X) : \beta \in \mathbb{R}^p, \sigma > 0, \| \beta[M^c] \| / \sigma \geq \delta_{n,4} \right\} \to 0$$

in probability as $n \to \infty$. Then

$$\begin{aligned}
Q_{n,2} &\leq \sup_{x_0 \in S(M), \| \beta[M^c] \| / \sigma \geq \delta_{n,4}} P_{n,\beta,\sigma}\left(|e_1| \geq (1-\delta_{n,2}) \delta_{n,1}; \hat{M} = M \middle| X \right) \\
&\quad + \sup_{x_0 \in S(M), \| \beta[M^c] \| / \sigma < \delta_{n,4}} P_{n,\beta,\sigma}\left(|e_1| \geq (1-\delta_{n,2}) \delta_{n,1}; \hat{M} = M \middle| X \right) \\
&\leq \sup_{\| \beta[M^c] \| / \sigma \geq \delta_{n,4}} P_{n,\beta,\sigma}\left(\hat{M} = M \middle| X \right) \\
&\quad + \sup_{x_0 \in S(M), \| \beta[M^c] \| / \sigma < \delta_{n,4}} P_{n,\beta,\sigma}\left(|e_1| \geq (1-\delta_{n,2}) \delta_{n,1} | X \right) \\
&\leq o_p(1) + \sup_{x_0 \in S(M), \| \beta[M^c] \| / \sigma < \delta_{n,4}} P_{n,\beta,\sigma}\left(|e_1| \geq (1-\delta_{n,2}) \delta_{n,1} | X \right). \qquad (35)
\end{aligned}$$

Using the Cauchy-Schwartz inequality we obtain for $x_0 \in S(M)$

$$\begin{aligned} |e_1| &\leq \left\|x_0'[M]\right\| \|\beta[M^c]/\sigma\| \left\| n^{1/2} \left(\left(X'[M]X[M]\right)^{-1} X'[M]X[M^c] - \left(\Sigma[M,M]\right)^{-1} \Sigma[M,M^c] \right) \right\| \\ &\leq \|\beta[M^c]/\sigma\| \, B_n(X) \end{aligned}$$

where $B_n(X) \geq 0$ is $O_p(1)$, this following from Condition 3.4 and positive definiteness of $\Sigma[M,M]$. This shows that the second term on the far right-hand side of (35) is bounded by

$$\mathbf{1}\left(\delta_{n,4} B_n(X) \geq (1-\delta_{n,2})\,\delta_{n,1}\right).$$

If we set now, for example, $\delta_{n,1} = \delta_{n,4}^{1/2}$, we see that the above quantity converges to zero in probability as $n \to \infty$, implying that $Q_{n,2}$ converges to zero in probability as $n \to \infty$. This completes the proof. ■

Proof of Theorem 3.6: (a) Use Lemma B.1 with $W(x_0, X, M)$ equal to $K_1(x_0, r)\|s_M\|$ ($K_2(x_0[\hat{M}], \hat{M}, r)\|s_M\|$, $K_3(x_0[\hat{M}], \hat{M}, r)\|s_M\|$, $K_4(r)\|s_M\|$, or $K_5(r)\|s_M\|$, respectively) and $\tilde{\sigma}_1 = \tilde{\sigma}_2 = \hat{\sigma}$ and combine this with Proposition 2.1 (Corollaries 2.2, 2.4, respectively). Note that $r = r_n \to \infty$ because of Condition 3.5, and hence $\hat{\sigma}$ satisfies (29).

(b) Let $\tilde{\sigma}_2$ be a sequence of random variables such that, conditionally on X, $\tilde{\sigma}_2^2$ is independent of $\hat{\beta}$ and is distributed as σ^2/r^* times a chi-squared distributed random variable with r^* degrees of freedom with the convention that $\tilde{\sigma}_2 = \sigma$ in case $r^* = \infty$. [Such a sequence exists: Possibly after redefining the relevant random variables on a sufficiently rich probability space we may find a sequence $(Z_i)_{i\in\mathbb{N}}$ of i.i.d. standard Gaussian random variables that is independent of Y and X. Then define $\tilde{\sigma}_2^2 = \sigma^2 \sum_{i=1}^{r^*} Z_i^2/r^*$ if $r^* < \infty$ and set $\tilde{\sigma}_2^2 = \sigma^2$ otherwise.] In view of Remark 2.9 we have that Proposition 2.1 (Corollaries 2.2, 2.4, respectively) also hold if the confidence interval (5) for the target $x_0'[\hat{M}]\beta_{\hat{M}}^{(n)}$ uses $\tilde{\sigma}_2$ instead of $\hat{\sigma}$ and uses the constants $K_1(x_0, r^*)$ ($K_2(x_0[\hat{M}], \hat{M}, r^*)$, $K_3(x_0[\hat{M}], \hat{M}, r^*)$, $K_4(r^*)$, or $K_5(r^*)$, respectively). Now apply Lemma B.1 with $W(x_0, X, M)$ equal to $K_1(x_0, r^*)\|s_M\|$ ($K_2(x_0[\hat{M}], \hat{M}, r^*)\|s_M\|$, $K_3(x_0[\hat{M}], \hat{M}, r^*)\|s_M\|$, $K_4(r^*)\|s_M\|$, or $K_5(r^*)\|s_M\|$, respectively) and with $\tilde{\sigma}_1 = \tilde{\sigma}$. Note that $\tilde{\sigma}_1$ satisfies (29) by assumption, while $\tilde{\sigma}_2$ satisfies it because $r^* \to \infty$ has been assumed. ■

Lemma B.2. *Suppose that the maintained model assumptions of Section 3 are satisfied and that $X'X/n \to \Sigma$ in probability for $n \to \infty$. Assume further that Condition 3.3 holds and define $\hat{\sigma}_{\hat{M}}^2 = ||Y - X[\hat{M}]\hat{\beta}_{\hat{M}}||^2/(n - |\hat{M}|)$ for $n > p$. Then $\hat{\sigma}_{\hat{M}}^2$ satisfies condition (19).*

Proof: Clearly

$$Y - X[\hat{M}]\hat{\beta}_{\hat{M}} = Y - P_{X[\hat{M}]}Y = P_{X[\hat{M}]^\perp}U + P_{X[\hat{M}]^\perp}X[\hat{M}^c]\beta[\hat{M}^c] = A + B,$$

where $P_{X[\hat{M}]^\perp}$ denotes orthogonal projection on the orthogonal complement of the column space of $X[\hat{M}]$. By the triangle inequality we hence have

$$\begin{aligned} \left| \left(\hat{\sigma}_{\hat{M}}^2/\sigma \right) - 1 \right| &\leq \left| \left(n - |\hat{M}|\right)^{-1/2} \|A/\sigma\| - 1 \right| + \left(n - |\hat{M}|\right)^{-1/2} \|B/\sigma\| \\ &\leq \left| \left(n - |\hat{M}|\right)^{-1/2} \|A/\sigma\| - 1 \right| + \left(n - |\hat{M}|\right)^{-1/2} \left\| X[\hat{M}^c]\beta[\hat{M}^c]/\sigma \right\|. \end{aligned}$$

We now bound the probability in (19) by the sum of the probabilities that the first and second term on the r.h.s. of the preceding display, respectively, exceed $\delta/2$. Because p is fixed there is a fixed finite number of possible models $\hat{M}$ and thus for $\delta > 0$ we have the bound for the first term

$$\begin{aligned}
& \sup_{\beta\in\mathbb{R}^p,\sigma>0} P_{n,\beta,\sigma}\left(\left|\left(n-|\hat{M}|\right)^{-1/2}\|A/\sigma\|-1\right|\geq\delta/2\middle|X\right)\\
= & \sup_{\beta\in\mathbb{R}^p,\sigma>0}\sum_M P_{n,\beta,\sigma}\left(\left|(n-|M|)^{-1/2}\left\|P_{X[M]^\perp}U/\sigma\right\|-1\right|\geq\delta/2,\hat{M}=M\middle|X\right)\\
\leq & \sum_M \sup_{\beta\in\mathbb{R}^p,\sigma>0} P_{n,\beta,\sigma}\left(\left|(n-|M|)^{-1/2}\left\|P_{X[M]^\perp}U/\sigma\right\|-1\right|\geq\delta/2\middle|X\right).
\end{aligned}$$

Note that the probabilities in the upper bound on the far r.h.s. of the preceding display do actually neither depend on β nor σ and are each of the form $\Pr(|W/w-1|\geq\delta)$ where W is distributed as the square root of a chi-squared random variable with w^2 degrees of freedom. Since $w^2 = n-|M|$ goes to infinity for $n\to\infty$ and any fixed M, and since the sum has a fixed finite number of terms, we can conclude that the upper bound converges to zero in probability as $n\to\infty$.

Turning to the second term we have, letting $\lambda_{\max}$ denote the largest eigenvalue of a symmetric matrix,

$$\begin{aligned}
& \sup_{\beta\in\mathbb{R}^p,\sigma>0} P_{n,\beta,\sigma}\left(\left(n-|\hat{M}|\right)^{-1/2}\left\|X[\hat{M}^c]\beta[\hat{M}^c]/\sigma\right\|\geq\delta/2\middle|X\right)\\
\leq & \sup_{\beta\in\mathbb{R}^p,\sigma>0} P_{n,\beta,\sigma}\left(\left(n-|\hat{M}|\right)^{-1/2}\lambda_{\max}^{1/2}\left(X'[\hat{M}^c]X[\hat{M}^c]\right)\left\|\beta[\hat{M}^c]/\sigma\right\|\geq\delta/2\middle|X\right)\\
\leq & \sup_{\beta\in\mathbb{R}^p,\sigma>0} P_{n,\beta,\sigma}\left(\lambda_{\max}^{1/2}\left(X'X/(n-p)\right)\left\|\beta[\hat{M}^c]/\sigma\right\|\geq\delta/2\middle|X\right)\\
\leq & \sum_{M\neq\{1,\ldots,p\}}\sup_{\beta\in\mathbb{R}^p,\sigma>0} P_{n,\beta,\sigma}\left(\hat{M}=M,\,\lambda_{\max}^{1/2}\left(X'X/(n-p)\right)\|\beta[M^c]/\sigma\|\geq\delta/2\middle|X\right).
\end{aligned}$$

Now, since $X'X/(n-p)$ converges to the positive definite matrix Σ in probability, we can find an event D_n, which has probability converging to 1 for $n\to\infty$, such that on this event $\lambda_{\max}(X'X/(n-p))$ is not larger than $4\lambda_{\max}(\Sigma)$. Hence, on D_n we can bound each supremum on the far r.h.s. of the preceding display by

$$\begin{aligned}
& \sup_{\beta\in\mathbb{R}^p,\sigma>0} P_{n,\beta,\sigma}\left(\hat{M}=M,\,\|\beta[M^c]/\sigma\|\geq\lambda_{\max}^{-1/2}(\Sigma)\,\delta/4\middle|X\right)\\
= & \sup\left\{P_{n,\beta,\sigma}(\hat{M}=M|X):\beta\in\mathbb{R}^p,\sigma>0,\|\beta[M^c]\|/\sigma\geq\lambda_{\max}^{-1/2}(\Sigma)\,\delta/4\right\},
\end{aligned}$$

which goes to zero in probability as $n\to\infty$ by Condition 3.3. ■

References

Berk, R., Brown, L., Buja, A., Zhang, K. and Zhao, L. (2013a). Valid post-selection inference. *Ann. Statist.*, **41** 802–837.

Berk, R., Brown, L., Buja, A., Zhang, K. and Zhao, L. (2013b). Valid post-selection inference. Unpublished version, URL `http://www-stat.wharton.upenn.edu/~lzhao/papers/MyPublication/24PoSI-submit.pdf`.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Ann. Statist.*, **32** 407–499.

Ewald, K. (2012). *On the influence of model selection on confidence regions for marginal associations in the linear model*. Master's thesis, University of Vienna.

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.

Leeb, H. and Pötscher, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory*, **21** 21–59.

Leeb, H. and Pötscher, B. M. (2006). Can one estimate the conditional distribution of post-model-selection estimators? *Ann. Statist.*, **34** 2554–2591.

Leeb, H. and Pötscher, B. M. (2012). Testing in the presence of nuisance parameters: Some comments on tests post-model-selection and random critical values. Working paper.

Leeb, H., Pötscher, B. M. and Ewald, K. (2013). On various confidence intervals post-model-selection. *Statist. Sci.* forthcoming.

Pötscher, B. M. (2009). Confidence sets based on sparse estimators are necessarily large. *Sankhya*, **71** 1–18.

Rawlings, J. (1998). *Applied Regression Analysis: A Research Tool*. Springer Verlag, New York, NY.

Scheffé, H. (1959). *The Analysis of Variance*. Wiley, New York.