Valid confidence intervals for post-model-selection predictors

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

Department of Mathematics, University Paul Sabatier, 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 Mathematics, University Paul Sabatier
**Department of Statistics, University of Vienna

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 confidence intervals for 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

In statistical practice, the model used for analysis is very often chosen after the data have been observed, either by ad-hoc methods or by more sophisticated model selection procedures. Inference following such a model selection step (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 for the true parameter can be dramatically smaller than the nominal one), and the construction of valid procedures is often non-trivial; see Leeb and Pötscher (2005, 2006, 2012), Kabaila and Leeb (2006), Pötscher (2009) and references therein for an introduction to the issues involved here. In these references, inference is focused on the true parameter of the data-generating model (or on components thereof). 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; see also Berk et al. (2013b) and Leeb et al. (2015). In this paper, we generalize the PoSI intervals to intervals for post-model-selection predictors.

Prediction following model selection is obviously also of great importance. In the case where the selected model is misspecified, parameter estimates are typically biased or at least difficult to interpret; cf. Remark 2.7. But even a misspecified model may perform well for prediction. In particular, Greenshtein and Ritov (2004) derive, under appropriate sparsity assumptions, feasible predictors that asymptotically perform as well as the (infeasible) best candidate predictor even if the available number of explanatory variables by far exceeds the sample size. These feasible predictors are also covered by the results in the present paper, among others. Like Greenshtein and Ritov (2004), our analysis does not rely on the assumption that the true data generating model is among the candidates for model selection. We develop confidence intervals for such predictors, that are easy to interpret and that are optimal in an appropriate sense; cf. Remarks 2.7(ii), 2.8, and 3.2, as well as Greenshtein and Ritov (2004). A further 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. (2015) as well as in Remarks 2.7 and 3.1 given further below.

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 $\hat{M}$ (more precisely, $\beta^{(n)}_{\hat{M}}$ represents the coefficients of the projection of the expected value vector of the dependent variable on the space spanned by the regressors included in $\hat{M}$); 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 2.8. 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 consider 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 a certain *'out-of-sample'* prediction error, namely the mean-squared prediction error, over all (infeasible) predictors of a future response y_0 that are of the form

2

$x_0'[\hat{M}]\gamma(\hat{M})$, when x_0 and the row-vectors of X are sampled from the same distribution; cf. Remark 3.2. In particular, this target does not suffer from the issues that plague the design-dependent coverage target, as discussed at the end of the preceding paragraph. Certain optimality properties of a feasible counterpart of $x_0'[\hat{M}]\beta^{(\star)}_{\hat{M}}$ are derived in Greenshtein and Ritov (2004), for a particular model selector $\hat{M}$ and under appropriate sparsity assumptions; a target closely related to $x_0'[\hat{M}]\beta^{(\star)}_{\hat{M}}$ is also studied in Leeb (2009). 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 are approximately valid for the target $x_0'[\hat{M}]\beta^{(\star)}_{\hat{M}}$, 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 'in-active' in the model $\hat{M}$. This may appear surprising at first sight but turns out to be inherent to the PoSI-approach (because of the need to take the maximum over all models M in (10)). In any case, 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. We also provide analytic results regarding the excess width of these intervals in an asymptotic setting where the number of regressors goes to infinity, see Section 2.3.

Inference post-model-selection is currently a very active area of research and we can only give a selection of work relevant for, or related to, this paper. Contemporary analyses of confidence sets for (components of) the true parameter of the underlying model include Andrews and Guggenberger (2009), Kabaila and Leeb (2006), Leeb and Pötscher (2005), Pötscher (2009), Pötscher and Schneider (2010), and Schneider (2015). These references also point to numerous earlier results. Also, the work of Lockhart et al. (2014), Wasserman and Roeder (2009), and Wasserman (2014) should be mentioned here. For the LASSO, in particular, a de-sparsifying method has recently been developed by Belloni et al. (2011, 2014), van de Geer et al. (2014), and Zhang and Zhang (2014). An alternative to the PoSI approach, that also focuses on $\beta^{(n)}_{\hat{M}}$ as the quantity of interest, is developed in Fithian et al. (2015), Lee et al. (2015), Lee and Taylor (2014), Tian and Taylor (2015), and Tibshirani et al. (2015). In these papers, confidence sets for $\beta^{(n)}_{\hat{M}}$ are considered that have a guaranteed coverage probability conditionally on the event that a particular model has been selected by the model selection procedures. In contrast to PoSI procedures, the confidence intervals obtained in these papers are specific to the model selection procedure used (the LASSO, in particular, being considered

in these references) and generally rely on certain geometric properties of the specific model selection procedure under consideration. In simulation experiments, we compare the confidence intervals proposed in these references with the intervals developed here and observe some interesting phenomena, see Section 4.3.

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 components of x_0 corresponding to the 'active' explanatory variables are available; moreover, we analyze properties of these intervals in an asymptotic framework where the model dimension increases; cf. Section 2.3. 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 when sample size increases. Finally, the results of a numerical study are reported in Section 4. Conclusions are drawn in Section 5. The proofs are given in Appendices A and B. In Appendix C we describe algorithms for computing the PoSI confidence intervals, that are comparable with those proposed by Berk et al. (2013a) in terms of computational complexity.

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

Consider the model

$$Y = \mu + U \tag{1}$$

where $\mu \in \mathbb{R}^n$ is unknown 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 $n \geq 1$. An important instance of this model arises when μ is known to reside in a lower dimensional linear subspace of $\mathbb{R}^n$, but we do not make such an assumption at this point. Apart from the data Y, we are given a (real) $n \times p$ matrix X, not necessarily of full column rank, the columns of which represent potential regressors. This setup allows for $p > n$ as well as for $1 \leq p \leq n$. The rank of X will be denoted by d. In this section the design matrix X is treated as fixed.

We consider fitting (potentially misspecified) linear models with design matrices that are obtained by deleting columns from X. Such a model will be represented by M, a subset of $\{1, ..., p\}$, where the elements of M index the columns of X that are retained. 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 others; 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 model M, we denote the corresponding least squares estimator by $\hat{\beta}_M$, i.e.,

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

4

where the inverse is to be interpreted as the Moore-Penrose inverse in case $X[M]$ does not have full column rank. For any given model M the corresponding least squares estimator $\hat{\beta}_M$ is obviously an unbiased estimator of

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

Note that $\hat{\beta}_M$ as well as $\beta_M^{(n)}$ reduce to 0 in case $M = \varnothing$.

As in Berk et al. (2013a) we further assume that, as an estimator for σ^2, we have available an (observable) random variable $\hat{\sigma}^2$ that is independent of $P_X Y$ and that is distributed as σ^2/r times a chi-square distributed random variable with r degrees of freedom $(1 \leq r < \infty)$, with P_X denoting orthogonal projection on the column space of X. This assumption is always satisfied in the important special case where one assumes that $d < n$ and $\mu \in \mathrm{span}(X)$ hold, upon choosing for $\hat{\sigma}^2$ the standard residual variance estimator obtained from regressing Y on X and upon setting $r = n - d$. However, otherwise it is not an innocuous assumption at all and this is further discussed in Section 2.2. Observe that our assumption 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 μ and σ as well as on sample size n and will be denoted by $P_{n,\mu,\sigma}$.

We are furthermore given a (non-empty) collection $\mathcal{M}$ of admissible models $M \subseteq \{1, ..., p\}$, the 'universe' of models considered by the researcher. Without loss of generality we will assume that any column of X appears as a regressor in at least one of the models M in $\mathcal{M}$, i.e., that $\bigcup \{M : M \in \mathcal{M}\} = \{1, ..., p\}$ holds (otherwise we can just redefine X by discarding all columns that do not appear in any of the models in $\mathcal{M}$); of course, we have excluded here the trivial and uninteresting case $\mathcal{M} = \{\varnothing\}$. For such a collection $\mathcal{M}$ it is easy to see that the assumed independence of $\hat{\sigma}^2$ and $P_X Y$ is in fact equivalent to independence of $\hat{\sigma}^2$ from the collection $\{\hat{\beta}_M : M \in \mathcal{M}\}$ of least squares estimators. While not really affecting the results, it proves useful to assume, throughout the following, that the empty model belongs to $\mathcal{M}$. We shall furthermore always assume that any non-empty $M \in \mathcal{M}$ is of full-rank in the sense that rank $X[M] = |M|$. We point out here that our assumptions on $\mathcal{M}$ imply that X can not have a zero column, and hence $d \geq 1$ must hold. An important instance of a collection $\mathcal{M}$ satisfying our assumptions is the collection of all full-rank submodels of $\{1, ..., p\}$ (enlarged by the empty model) provided that no column of X is zero; of course, there are many other examples, see, e.g., the list in Section 4.5 of Berk et al. (2013a).

A model selection procedure $\hat{M}$ is now a (measurable) rule that associates with every $(X, Y, \hat{\sigma}^2)$ a model $\hat{M}(X, Y, \hat{\sigma}^2) \in \mathcal{M}$. [Since $\mathcal{M}$ contains the empty model, the possibility that $\hat{M}(X, Y, \hat{\sigma}^2)$ is empty is allowed for.] In the following we shall, in abuse of notation, often write $\hat{M}$ for $\hat{M}(X, Y, \hat{\sigma}^2)$. Allowing explicitly dependence of $\hat{M}$ on $\hat{\sigma}^2$ is 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 $P_{n,\mu,\sigma}$ 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 non-standard quantity of interest studied in Berk et al. (2013a) is 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.

5

(2013a) do not allow for empty models $\hat{M}$.] 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 (real) $p \times 1$ vector x_0 and suppose we want to predict y_0 which is distributed as $N\left(\nu, \sigma^2\right)$, independently of Y. If one is forced to use a fixed model M for prediction, i.e., to use predictors of the form $x_0'[M]\gamma$, the predictor that would then typically be used is $x_0'[M]\hat{\beta}_M$, which can be viewed as an estimator of the infeasible predictor $x_0'[M]\beta_M^{(n)}$. Of course, for this predictor to be reasonable there must be some relation between the training data (X, Y) and (x_0, y_0). This is further discussed in Remark 2.8. 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 in turn 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)). A discussion of the merits of this target and its interpretation is postponed to Remarks 2.7 and 2.8 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 $x_0'[\hat{M}]\beta_{\hat{M}}^{(n)}$ 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 ($\hat{\sigma}$ of course representing the nonnegative square root of $\hat{\sigma}^2$), where

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

where $s_M = 0 \in \mathbb{R}^n$ for $M = \varnothing$ by our conventions, and where $K(x_0, M) = K(x_0, M, r) = K(x_0, M, r, X, \alpha, \mathcal{M})$ denotes a non-negative constant which may depend on x_0, M, r, X, α, and $\mathcal{M}$, 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 $[a - b, a + b]$ ($a \in \mathbb{R}$, $b \geq 0$). The motivation for the form of the confidence interval stems from the observation that for *fixed* M the interval $x_0'[M]\hat{\beta}_M \pm q_{r,1-\alpha/2}||s_M||\hat{\sigma}$ is a valid $1 - \alpha$ confidence interval for $x_0'[M]\beta_M^{(n)}$, where $q_{r,1-\alpha/2}$ is the $(1 - \alpha/2)$-quantile of Student's t-distribution with r degrees of freedom. Furthermore 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

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

Note that if one replaces $K(x_0, \hat{M})$ in (5) by $K_{naive} = q_{r,1-\alpha/2}$, 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 4, which is in line with the related results in Leeb et al. (2015).

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 $K_1(x_0) = K_1(x_0, r) = K_1(x_0, r, X, \alpha, \mathcal{M})$ (not depending on M) such that the resulting confidence interval (5) satisfies (7): Observe that

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

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

$$\left|\bar{s}_{\hat{M}}'\left(Y - \mu\right)\right|/\hat{\sigma} \leq \max_{M \in \mathcal{M}} \left|\bar{s}_M'\left(Y - \mu\right)\right|/\hat{\sigma}. \tag{9}$$

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

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

It is important to note that the probability on the left-hand side of the preceding display neither depends on μ nor on σ; it also depends on the estimator $\hat{\sigma}$ only through the 'degrees of freedom' parameter r: To see this note that $\bar{s}_M'\left(Y - \mu\right) = \bar{s}_M' P_X\left(Y - \mu\right)$, since $\bar{s}_M$ belongs to the column space of X. Consequently, the collection of all the quantities $\bar{s}_M'\left(Y - \mu\right)$ 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 $\left|\bar{s}_M'\left(Y - \mu\right)\right|/\hat{\sigma}$ does neither depend on μ 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, α, and $\mathcal{M}$. 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 in Appendix A and from the observation that, in view of our assumptions on $\mathcal{M}$, $\bar{s}_M' = 0$ for all $M \in \mathcal{M}$ 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. [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 $\bar{s}_M = 0$ holds for models M with $i \notin M$ if $x_0 = e_i$.] Finally, $K_{naive} \leq K_1\left(x_0\right)$ clearly holds provided $x_0 \neq 0$ (since $\bar{s}_M'\left(Y - \mu\right)/\hat{\sigma}$ 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 with values in $\mathcal{M}$, 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 with values in $\mathcal{M}$, and thus leads to 'universally valid post-selection inference' in case $\mathcal{M}$ is chosen to be the set of all full-rank submodels obtainable from X (enlarged by the empty set and provided X does not have a zero column); cf. Berk et al. (2013a), where similar guarantees are obtained for the components of $\beta_{\hat{M}}^{(n)}$.

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 in general, because $K_1(x_0)$ will depend on *all* components of x_0 (and not only on those appearing in $x_0[\hat{M}]$) due the maximum figuring in (10) and our assumptions on $\mathcal{M}$. 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 $x_0[M]$, and thus can be computed in case (ii). Of course, $K_2(x_0[M], M)$ also depends on r, X, α, and $\mathcal{M}$, 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 with values in $\mathcal{M}$, 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. The constant K_4 is obtained by applying a union bound to (10), whereas the constant K_3 is obtained by applying a more refined 'partial' union bound, see the proof of Proposition 2.3 for more detail. The constant K_5 is the Scheffé constant. Algorithms for computing these constants are discussed in Appendix C.

For $x_0 \in \mathbb{R}^p$ and $M \in \mathcal{M}$ define the distribution function F_{M,x_0}^* for $t \geq 0$ via

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

and via $F_{M,x_0}^* (t) = 0$ for $t < 0$. Here $c(M, \mathcal{M})$ denotes the number of models $M_* \in \mathcal{M}$ that satisfy $M_* \nsubseteq M$, V is a random vector that is uniformly distributed on the unit sphere in the column space of X, and $F_{Beta, 1/2, (d-1)/2}$ denotes the $Beta(1/2, (d-1)/2)$-distribution function, with the convention that in case $d = 1$ we use $F_{Beta, 1/2, 0}$ to denote the distribution function of pointmass at 1. In view of our assumptions on $\mathcal{M}$ it follows that $c(M, \mathcal{M}) \geq 1$ always holds, except in the case where $M = \{1, ..., p\}$ (and where this set belongs to $\mathcal{M}$). Next define the distribution function F_{M,x_0} via

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

8

where G denotes a nonnegative random variable such that G^2/d follows an F-distribution with (d, 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. The motivation for the definition of F^*_{M,x_0} and F_{M,x_0} originates from a partial union bound for 1 minus the probability appearing in (10), cf. the proof of the subsequent proposition.

Proposition 2.3. *Let $x_0 \in \mathbb{R}^p$ be arbitrary. For any $M \in \mathcal{M}$ 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 $M = \{1, ..., p\} \in \mathcal{M}$ 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 $M \in \mathcal{M}$. Furthermore,

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

and

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

hold whenever $M_1 \subseteq M_2$, $M_i \in \mathcal{M}$.

It is obvious that $K_3(x_0[M], M)$ depends, besides $x_0[M]$ and M, only on r, X, α, and $\mathcal{M}$, whereas K_4 only depends on r, d, α, and $\mathcal{M}$, and K_5 depends only on r, d, 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 $K_1(x_0) = K_3(x_0[M_{full}], M_{full}) = K_3(x_0, M_{full})$, provided $M_{full} := \{1, ..., p\}$ belongs to $\mathcal{M}$, and that $K_3(x_0[M], M) = K_4$ holds for any $M \in \mathcal{M}$ satisfying $|M| = 1$ and $\bar{s}_M \neq 0$. [Indeed, in this case, the probability appearing in (12) equals $1 - F_{Beta,1/2,(d-1)/2}(t^2)$ as can be seen from the proof of Proposition 2.3.] Similarly, $K_3(x_0[M], M) = K_4$ holds for any $M \in \mathcal{M}$ in case $d = 1$ as is not difficult to see. 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 – in the case $p = d$ – an upper-bound for their PoSI constant that does not depend on X, but only on d. [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 4.

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. We stress that the confidence intervals figuring in the subsequent corollary depend on x_0 only through $x_0[\hat{M}]$ and thus are feasible in case (ii).

Corollary 2.4. *Let $\hat{M}$ be an arbitrary model selection procedure with values in $\mathcal{M}$, 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 regarding extensions.

Remark 2.5. *(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) If 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 d degrees of freedom. 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 2.6. (i) All results carry over immediately to the case where μ can vary only in a subset $\mathfrak{M}$ of $\mathbb{R}^n$.

(ii) We have assumed that any non-empty $M \in \mathcal{M}$ is of full-rank. This assumption could easily be dropped, but this would lead to more unwieldy results.

2.1 On the merits of the non-standard targets

Remark 2.7. *(On the merits of $\beta_{\hat{M}}^{(n)}$ and $x_0'[\hat{M}]\beta_{\hat{M}}^{(n)}$ as targets for inference)* (i) As already noted, the (non-standard) coverage target in Berk et al. (2013a) is $\beta_{\hat{M}}^{(n)}$ (where these authors choose to represent it in what they call 'full model indexing'). While $\beta_{\hat{M}}^{(n)}$ has a clear technical meaning as the coefficient vector that provides the best approximation of μ 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 $\beta_{\hat{M}}^{(n)}$. Also note that, e.g., the meaning of the first component of the target $\beta_{\hat{M}}^{(n)}$ depends on the selected model $\hat{M}$. The target $x_0'[\hat{M}]\beta_{\hat{M}}^{(n)}$ considered in this paper, while again being random and sharing many of the properties of $\beta_{\hat{M}}^{(n)}$ just mentioned, seems to be somewhat more amenable to interpretation as a target for inference since it is simply the random convex combination $\sum_M x_0'[M]\beta_M^{(n)}\mathbf{1}(\hat{M} = M)$ 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) In the classical case, i.e., when $\mu = X\beta$ and $d = p \leq n$, 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 is $x_0'[\hat{M}]\beta_{\hat{M}}^{(n)}$ provided y_0 is independent of $\hat{M}$ (which will certainly 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 $x_0'[\hat{M}]\beta_{\hat{M}}^{(n)}$ even in the classical case. This is in contrast with the situation when, as in Berk et al. (2013a), one's interest focusses on parameters rather than predictors: Similar as before one can argue that in the classical case the true parameter β should be the target rather than $\beta_{\hat{M}}^{(n)}$ but there seems now to be little to justify the non-standard target $\beta_{\hat{M}}^{(n)}$ (as the preceding argument justifying the target $x_0'[\hat{M}]\beta_{\hat{M}}^{(n)}$ even in the classical case is obviously not applicable to the target $\beta_{\hat{M}}^{(n)}$).

(iii) In view of the preceding discussion it seems that the non-standard target $\beta_{\hat{M}}^{(n)}$ of Berk et al. (2013a) can only be justified in a non-classical setting where μ is not assumed to belong to the column space of X (implying $d < n$), or where $d < p$ holds (subsuming in particular the important case $p > n = d$), because in these cases β is no longer available as a target (being not defined or not uniquely defined). However, in a setting, where μ is not assumed to belong to the column space of X or where $p > n = d$ holds, the assumption on the variance estimator $\hat{\sigma}^2$ made in Berk et al. (2013a) (as well as in the present paper) becomes problematic and quite restrictive; see Remark 2.1(ii) in Leeb et al. (2015) as well as Section 2.2 below. Hence, there is some advantage in considering the targets $x_0'[\hat{M}]\beta_{\hat{M}}^{(n)}$ rather than $\beta_{\hat{M}}^{(n)}$ as the former has a justification in the classical as well as in the non-classical framework.

(iv) We note the obvious fact that if the target of inference is the standard target $x_0'\beta$ (assuming the classical case) 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 2.8. *(Optimality of the design-dependent target)* (i) The infeasible predictor $x_0'[M]\beta_M^{(n)}$ (for fixed M) is the best predictor for y_0 in the mean-squared error sense among all predictors of the form $x_0'[M]\gamma$ in case $y_0 | \nu, x_0 \sim N(\nu, \sigma^2)$ and (ν, x_0') is drawn from the empirical distribution of (μ_i, x_i') where x_i' denotes the i-th row of X ('in-sample prediction'). [More generally, this is so if (ν, x_0') is drawn from the empirical distribution of $(\mu_i + a_i, x_i')$ where a is a fixed vector orthogonal to the column space of X.] Otherwise, it does in general 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 $\hat{M}$.

2.2 On the assumptions on $\hat{\sigma}^2$

In line with Berk et al. (2013a) we have postulated the existence of an estimator $\hat{\sigma}^2$ that is independent of $P_X Y$ and is distributed as σ^2/r times a chi-square distributed random variable with r degrees of freedom ($1 \leq r < \infty$). As already noted earlier, if we assume that $d < n$ and $\mu = X\beta$ hold, such an estimator always exists and is given by the usual residual variance estimator obtained from the residuals $Y - P_X Y$. However, if $d = n$ holds (which typically is the case if $p > n$) or if μ is not known to belong to the column space of X, such an estimator is much harder to come by.

Consider first the case where $d = n$ holds. Then it is plain that such an estimator does not exist if it is to be only a function of Y (and X): Since here $P_X Y = Y$ holds, such an estimator would have to be independent of Y and thus constant with probability one, contradicting the requirement to be distributed as a positive multiple of a chi-square. In order to nevertheless be able to come up with an estimator $\hat{\sigma}^2$ with the desired properties, one is hence forced to assume that one has access to *additional* data beyond Y that are related to Y in an appropriate way. A prototypical situation where such a construction is possible is as follows: Assume that one has available additional data Y^* distributed as $N(\mu^*, \sigma^2 I_{n^*})$, independently of Y (for example, Y^* might have been obtained from splitting the original larger sample into Y and Y^*). Assume further that for Y^* one has available a (non-trivial) correct regression model (i.e., $\mu^* = X^*\delta$ with X^* of full column rank less than n^*). Obviously, then an estimator $\hat{\sigma}^2$ satisfying all the required properties can be constructed from this correct regression model for Y^*. However, this raises the question why one would be willing to assume a correct regression model for one part of the data, but would refuse to do so for the other part. [This might be defended by reference to a structural break in the mean, which then however would beg the question why the structural break would not also affect the variance σ^2.] Alternatively to the assumption $\mu^* = X^*\delta$, one could assume some 'smoothness' in μ^* and then use nonparametric estimators to produce $\hat{\sigma}^2$. Again the question arises why one would then not make a similar assumption for μ and use the nonparametric method also for the first (or the entire) sample. In the quite special situation where one has replicated observations in Y^* available, one can abandon the dependence on a correct model (or on smoothness assumptions) and nevertheless produce an estimator $\hat{\sigma}^2$ with the desired properties. All this granted, it seems that the desired assumptions on $\hat{\sigma}^2$ and the desire to treat the case $d = n$ are not completely at ease.

Second, in case $d < n$, but it is not assumed that $\mu = X\beta$ holds, it is not obvious how an estimator $\hat{\sigma}^2$ with the desired properties can be constructed without further assumptions (note that the residual variance estimator obtained from $Y - P_X Y$ while being independent from $P_X Y$ will in general *not* be guaranteed to follow the required distribution). One such assumption could be that we have available a correct model $\mu = Z\gamma$, where the column space of Z contains the column space of X with the rank of Z still less than n; we could then compute $\hat{\sigma}^2$ from this larger model (i.e., from $Y - P_Z Y$), the resulting estimator having the desired properties. While this assumption solves the existence problem for $\hat{\sigma}^2$, it raises the question why one would then still want to keep the model selection exercise restricted to submodels defined by the columns of X, when it is known that the correct, larger, model $\mu = Z\gamma$ holds

(and Z is available). Hence, we are led back essentially to the classical case with Z playing the role of X. Alternatively, the same constructions as in the preceding paragraph relying on an independent sample Y^* are available, but they again suffer from the limitations pointed out before.

The discussion in this section shows that outside of the framework $d < n$ and $\mu = X\beta$ (in which case $p = d$ can be assumed with little loss of generality) the assumptions on $\hat{\sigma}^2$ made in Berk et al. (2013a) as well as in the present paper are less than innocuous and will be satisfied only in quite special situations; cf. also the discussion in Remark 2.1(ii) in Leeb et al. (2015). [For this very reason, the first version of this paper was set in the classical framework.]

2.3 Behavior of the constants K_i as a function of p

In this section we provide some results on the size of the constants K_i that govern the length of the confidence intervals. In particular, these results help in answering the question how tight a bound for K_1 and K_2 is provided by K_3 or K_4.

2.3.1 Orthogonal designs

Berk et al. (2013a) show that in the case $p = d \leq n$ their PoSI constant becomes smallest for the case of orthogonal design (provided the model universe $\mathcal{M}$ is sufficiently rich, e.g., $\mathcal{M}$ contains all submodels) and has rate $\sqrt{\log p}$ as $p \to \infty$, at least in the known-variance case; cf. Proposition 5.5 in Berk et al. (2013a) (where the error term $o(d)$ given in this result should read $o(1)$). In the next proposition we study the order of magnitude of $K_1(x_0)$, the analogue of the PoSI constant and of the closely related constant $K_2(x_0[M], M)$ in the case of orthogonal design. Recall that $K_1(x_0)$ is only feasible if x_0 is observed in its entirety, while $K_2(x_0[M], M)$ is the ideal bound for $K_1(x_0)$ given only knowledge of $x_0[M]$. Note that in the following result some of the objects depend on p, but we do not always show this in the notation.

Proposition 2.9. *Consider the known-variance case (i.e., $r = \infty$ and $\hat{\sigma}^2 = \sigma^2$) and assume that for every $p \geq 1$ the model universe $\mathcal{M}$ used is the power set of $\{1, ..., p\}$. Let α, $0 < \alpha < 1$, be given, not depending on p.*

(a) For any $p \geq 1$ let $X = X(p)$ be an $n(p) \times p$ matrix with (non-zero) orthogonal columns. For any such sequence X one can find a corresponding sequence of (non-zero) $p \times 1$ vectors x_0 such that $K_1(x_0, \infty) = K_1(x_0, \infty, X, \alpha, \mathcal{M})$ satisfies

$$\liminf_{p \to \infty} K_1(x_0, \infty)/\sqrt{p} \geq \xi$$

where $\xi = \sup_{b>0} \phi(b)/\sqrt{1 - \Phi(b)} \approx 0.6363$. [$\phi$ and Φ denote the pdf and cdf of a standard normal variable, respectively.] Furthermore, for any sequence X as above one can find another sequence of (non-zero) $p \times 1$ vectors x_0 such that $K_1(x_0, \infty) = O(1)$.

(b) Let $\gamma \in [0, 1)$ be given. Then $K_2(x_0[M], M, \infty) = K_2(x_0[M], M, \infty, X, \alpha, \mathcal{M})$ satisfies

$$\liminf_{p \to \infty} \inf_{x_0 \in \mathbb{R}^p} \inf_{X \in \mathsf{X}(p)} \inf_{M \in \mathcal{M}, |M| \leq \gamma p} K_2(x_0[M], M, \infty)/\sqrt{p} \geq \xi\sqrt{1 - \gamma},$$

where $\mathsf{X}(p) = \bigcup_{n \geq p} \{X : X \text{ is } n \times p \text{ with non-zero orthogonal columns}\}$.

13

The lower bounds given in the preceding proposition clearly also apply to $K_3(x_0[M], M, \infty)$ and $K_4(\infty)$ a fortiori. Part (a) of the above proposition shows that, even in the orthogonal case, the growth of $K_1(x_0, \infty)$ is – in the worst-case w.r.t. x_0 – of the order $\sqrt{p}$. This is in contrast to the above mentioned result of Berk et al. (2013a) for the PoSI constant. Part (a) also shows that there are other choices for x_0 such that $K_1(x_0, \infty)$ stays bounded. In this context also recall that $K_1(x_0, \infty)$ with x_0 equal to a $p \times 1$ standard basis vector coincides with a PoSI1 constant and thus equals the $(1-\alpha)$-quantile of the distribution of the absolute value of a standard normal variable in the orthogonal case. Part (b) goes on to show that regardless of x_0 and X the growth of the constants $K_2(x_0[M], M, \infty)$ is of the order $\sqrt{p}$ (except perhaps for very large submodels M).

2.3.2 Order of magnitude of K_3 and K_4

The next proposition, which exploits results in Zhang (2013), shows that $K_4(\infty)$ is a tight upper bound for $K_3(x_0[M], M, \infty)$ at least if p is large. It also provides the growth rates for $K_4(\infty)$ and $K_3(x_0[M], M, \infty)$. As before, the dependence of several objects on p (or n) will not always be shown in the notation. For the following recall the constants $c(M, \mathcal{M})$ defined after (12).

Proposition 2.10. *Consider the known-variance case (i.e., $r = \infty$ and $\hat{\sigma}^2 = \sigma^2$) and assume that for every $p \geq 1$ a (non-empty) model universe $\mathcal{M} = \mathcal{M}_p$ is given that satisfies (i) $\bigcup \{M : M \in \mathcal{M}\} = \{1, ..., p\}$, (ii) $\varnothing \in \mathcal{M}$, (iii) $c(M, \mathcal{M}) \geq \tau |M|$ for every $M \in \mathcal{M}$ with $M \neq \{1, \ldots, p\}$, where $\tau > 0$ is a given number (neither depending on M, $\mathcal{M}$, nor p), and (iv) $|\mathcal{M}| \to \infty$ as $p \to \infty$. For $n \in \mathbb{N}$, the set of positive integers, let $\mathsf{X}_{n,p}(\mathcal{M})$ denote the set of all $n \times p$ matrices of rank $\min(n, p)$ with the property that $X[M]$ has full column-rank for every $\varnothing \neq M \in \mathcal{M}$. Furthermore, let α, $0 < \alpha < 1$, be given (neither depending on p nor n). Let $n(p) \in \mathbb{N}$ be a sequence such that $n(p) \to \infty$ for $p \to \infty$ and such that $\mathsf{X}_{n(p),p}(\mathcal{M}) \neq \varnothing$ for every $p \geq 1$. Then we have*

$$\lim_{p \to \infty} \sup_{M \in \mathcal{M}, M \neq \{1,...,p\}} \sup_{x_0 \in \mathbb{R}^p} \sup_{X \in \mathsf{X}_{n(p),p}(\mathcal{M})} |1 - (K_3(x_0[M], M, \infty)/K_4(\infty))| = 0, \qquad (18)$$

where $K_3(x_0[M], M, \infty) = K_3(x_0[M], M, \infty, X, \alpha, \mathcal{M})$ and $K_4(\infty) = K_4(\infty, \min(n(p), p), \alpha, \mathcal{M})$. Furthermore,

$$K_4(\infty) \Big/ \sqrt{\min(n(p), p) \left(1 - |\mathcal{M}|^{-2/(\min(n(p),p)-1)}\right)} \to 1$$

as $p \to \infty$.

Remark 2.11. (i) $\mathsf{X}_{n(p),p}(\mathcal{M}) \neq \varnothing$ implies $\mathsf{X}_{n,p}(\mathcal{M}) \neq \varnothing$ for $n \geq n(p)$.

(ii) $\mathsf{X}_{n(p),p}(\mathcal{M})$ is certainly non-empty for $n(p) \geq p$, but – depending on $\mathcal{M}$ – this can already be true for $n(p)$ much smaller than p.

The assumptions (i)-(iv) on $\mathcal{M}$ in the preceding proposition are shown in the next corollary to be always satisfied in the important case where $\mathcal{M}$ is of the form $\{M \subseteq \{1, \ldots, p\} : |M| \leq m_p\}$. Furthermore, in the special case where $\mathcal{M}$ is the universe of all submodels, a simple formula for the growth rate of $K_4(\infty)$ is found.

Corollary 2.12. *Consider the known-variance case (i.e., $r = \infty$ and $\hat{\sigma}^2 = \sigma^2$) and let α, $0 < \alpha < 1$, be given (neither depending on p nor n). Let $m_p \in \mathbb{N}$ satisfy $1 \leq m_p \leq p$ for every $p \geq 1$ and define the set $\mathcal{M}(m_p) = \{M \subseteq \{1, \ldots, p\} : |M| \leq m_p\}$. Then $\mathcal{M}(m_p)$ satisfies (i)-(iv) in Proposition 2.10 with $\tau = 1/3$. Consequently, for $n(p)$ as in Proposition 2.10, (18) holds with $\mathcal{M}$ replaced by $\mathcal{M}(m_p)$ and*

$$K_4(\infty)/\sqrt{\min(n(p), p)\left(1 - \left(\sum_{k=0}^{m_p} \binom{p}{k}\right)^{-2/(\min(n(p), p) - 1)}\right)} \to 1$$

as $p \to \infty$. In particular, if $m_p = p$ for all $p \geq 1$, we necessarily have $n(p) \geq p$ and

$$K_4(\infty)/\sqrt{p} \to \sqrt{3}/2$$

as $p \to \infty$.

In the important case, where $p = d \leq n$ and $\mathcal{M}$ is the entire power set of $\{1, ..., p\}$, the preceding corollary shows that $K_4(\infty)$ (and hence a fortiori all the constants $K_1(x_0, \infty)$,..., $K_3(x_0[M], M, \infty)$) are 'bounded away' from the Scheffé constant K_5 which clearly satisfies $K_5/\sqrt{p} \to 1$ for $p \to \infty$. This is in line with a similar finding in Berk et al. (2013a), Section 6.3, for their PoSI constant.

Remark 2.13. In the proof of Proposition 2.3 union bounds were used to obtain 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 now collect 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 $n \times p$ design matrices X with $p = d = 2$ and vectors x_0 such that $K_4 = K_1(x_0)$ in case $\mathcal{M}$ is the universe of all submodels. Hence, in this case the union bounds used in the proof of Proposition 2.3 are all exact. Furthermore, in the known-variance case with $p = d \leq n$ and where $\mathcal{M}$ again is the universe of all submodels, the propositions given above entail that $K_4(\infty) \sim \sqrt{p}\sqrt{3}/2 \approx 0.866\sqrt{p}$ while $K_1(x_0, \infty) \succeq \xi\sqrt{p}$ with $\xi \approx 0.6363$ is possible; e.g., as the worst-case behavior in the orthogonal case, or with $x_0 = e_i$ and the design matrices constructed in the proof of Theorem 6.2 in Berk et al. (2013a) (recall that $K_1(e_i, \infty)$ coincides with a PoSI1 constant). This again shows that there is little room for improving K_3 and K_4. [Further evidence in that direction is provided by the observation that the proof of Theorem 6.3 in Berk et al. (2013a) implies that $K_1^*/\sqrt{p}$ tends to $\sqrt{3}/2$ in probability as $p \to \infty$, where K_1^* is an analogue of $K_1(x_0, \infty)$ that 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.]

Remark 2.14. The results for $p \to \infty$ in this subsection as well as the related results in Berk et al. (2013a) should be taken with a grain of salt as they obviously are highly non-uniform w.r.t. α: Note that – for fixed n and p – any one of the constants K_i will vary in the entire interval $(0, \infty)$ as α varies in $(0, 1)$ (except for degenerate cases), while the limits in the results in question do not depend on α at all.

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

In this section we again consider the model (1), but now assume that $\mu = X\beta$ for some unknown $\beta \in \mathbb{R}^p$ holds and that the $n \times p$ matrix X is random, with X independent of U, where U again follows an $N\left(0, \sigma^2 I_n\right)$-distribution with $0 < \sigma < \infty$. We also assume that X has full column rank almost surely (implying $p \leq n$) 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 preceding 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 $P_X Y$ (or, equivalently, of $\hat{\beta} = (X'X)^{-1}X'Y$) and is distributed as σ^2/r times a chi-squared distributed random variable with r degrees of freedom ($1 \leq r < \infty$). The collection $\mathcal{M}$ of admissible models will be assumed to be the power set of $\{1, \ldots, p\}$ in this section for convenience, but see Remark 3.11 for possible extensions. Observe that 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 $\tilde{\sigma}$ 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 the target studied in this section, 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

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

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. For purpose of comparison we point out that, under the assumption $\mu = X\beta$ maintained in the present section, $\beta_M^{(n)}$ defined in (3) can be rewritten as $\beta_M^{(n)} = \beta[M] + \left(X[M]'X[M]\right)^{-1} X[M]'X[M^c]\beta[M^c]$.

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 2.7(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 2.8 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 (in case that $\mathcal{M}$ is the power set of $\{1, \ldots, p\}$ as is the case in the present section), 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 $\delta > 0$,*

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

17

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 $2 \le i \le 5$ depend on x_0 only through $x_0[\hat{M}]$ (or not on x_0 at all).

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

$$\inf_{x_0 \in \mathbb{R}^p, \beta \in \mathbb{R}^p, \sigma > 0} P_{n,\beta,\sigma} \left(x_0'[\hat{M}]\beta^{(\star)}_{\hat{M}} \in CI(x_0) \Big| X \right) \ge (1 - \alpha) + o_p(1), \qquad (19)$$

where the $o_p(1)$ term above depends only on X and converges to zero in probability as $n \to \infty$. Relation (19) 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

$$\sup_{\beta \in \mathbb{R}^p, \sigma > 0} P_{n,\beta,\sigma}(|\tilde{\sigma}/\sigma - 1| \ge \delta \,|X\,) \xrightarrow{p} 0 \qquad (20)$$

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 $n \to \infty$. 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 relation (19) holds with $CI(x_0)$ replaced by $CI^*(x_0)$.

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 (20) 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

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

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

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

holds. See also Leeb (2009), where prediction intervals for y_0 are studied in a similar setting.

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 $\beta^{(\star)}_M$. 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 $\beta^{(\star)}_M$, 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 $\mathcal{M}$ (and $\mathcal{M}$ satisfies the basic assumptions made in Section 2).

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   Numerical study

In this section, we present a numerical study of the lengths and the minimal coverage probabilities of various confidence intervals. We begin, in Section 4.1, with an investigation of the length of the confidence intervals introduced in Section 2, including the 'naive' confidence interval that ignores the model selection step, as a function of the selected model. In Section 4.2 we then evaluate numerically the minimal coverage probabilities of these confidence intervals. As model selectors we consider here AIC, BIC, LASSO, SCAD (Fan and Li (2001)), and MCP (Zhang (2010)). Finally, in Section 4.3 we compare the confidence intervals introduced in Section 2 with the confidence interval proposed recently in Lee et al. (2015), which is specific to the LASSO model selector. Code for the computations in this section is available from the first author.

4.1   Lengths of confidence intervals

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 $K_{naive} = q_{r,1-\alpha/2}$, the $(1-\alpha/2)$-quantile of 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 $\alpha = 0.05$, $n = 29$, $d = p = 10$, $r = n - p$, $\sigma = 1$, and obtain X and x_0 from a data set of Rawlings et al. (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 et al. (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. (2015). 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}. It is easily checked that the so-obtained matrix X is indeed of full column rank (and $x_0 \neq 0$). Furthermore, the model universe $\mathcal{M}$ is chosen to be the power set of $\{1, ..., p\}$.

For the so chosen values of α, n, p, r, σ, X, x_0, and $\mathcal{M}$, 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; recall that we use $\mathcal{M}$ equal to the power set of $\{1, ..., p\}$.] 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 Appendix C. 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 C.1 in Appendix C, 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 C.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 C.1.
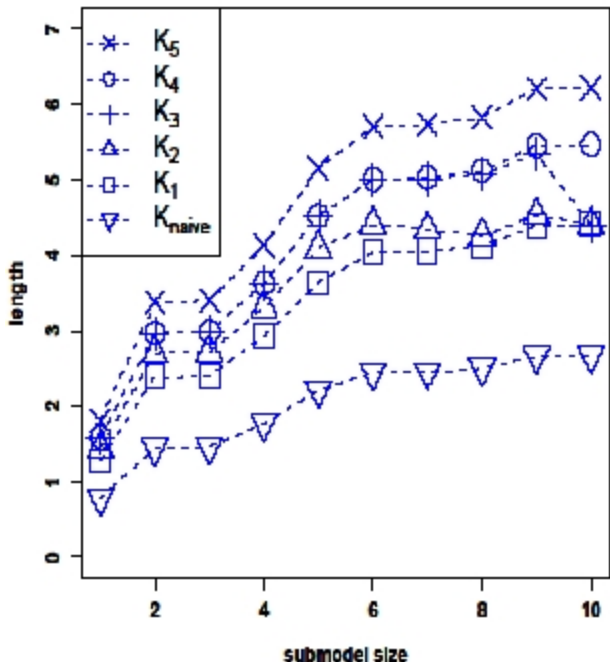
Figure 1: Standardized lengths of various confidence intervals as function of model size. Dashed lines are added to improve readability.

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 in Figure 1). However, as discussed after Proposition 2.3, the values of K_2 and K_3 decrease with increasing 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 typically 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

22

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 9; cf. (18). 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 increases 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.]

4.2   Minimal coverage probabilities

In this section we consider the case where $\mu = X\beta$ and $d = p < n$, i.e., the case where the given matrix X has full rank less than n and provides a correct linear model for the data Y. We then investigate the minimal coverage probabilities (the minimum being w.r.t. $\beta \in \mathbb{R}^p$ and $\sigma \in (0, \infty)$) of the intervals obtained from the constants K_{naive}, K_1, K_3, and K_4 when used as confidence intervals for the target $x_0'[\hat{M}]\beta_{\hat{M}}^{(n)}$ on the one hand as well as for the target $x_0'[\hat{M}]\beta_{\hat{M}}^{(\star)}$ on the other hand. The constants K_1, K_3, and K_4 are computed based on $\mathcal{M}$ equal to the power set of $\{1, ..., p\}$. We do not report results for 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 those 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 AIC-, BIC-procedures, the LASSO, SCAD (Fan and Li (2001)), and MCP (Zhang (2010)). For these five procedures, we always protect the first explanatory variable (which corresponds to an intercept term) from selection. However, note that 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 but use $\mathcal{M}$ equal to the power set of $\{1, ..., p\}$. [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.

For SCAD and MCP, we use the `ncvreg` package in `R`. With the function `cv.ncvreg()` (with parameters `SCAD` or `MCP`) the penalty is selected by 10-fold cross-validation and the corresponding estimated regression coefficients are computed. Like for the LASSO, the function `cv.ncvreg()` is applied for a regression of $\tilde{Y}$ on $\tilde{X}$, and the selected model is comprised of those regressors in $\tilde{X}$ for which the SCAD (or MCP) coefficients are non-zero, plus the first column of X.

The design matrix X and the vector x_0 are generated in the following manner: 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 $30 - 1 = 29$) of the variables in the 30×10 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 $\tilde{\Sigma} = (\mathbf{X}^{(\tilde{p})}(a))'(\mathbf{X}^{(\tilde{p})}(a))$. Hence, in this case, $\tilde{\Sigma} = I_{\tilde{p}} + (2a + \tilde{p}a^2)E_{\tilde{p}}$, with $E_{\tilde{p}}$ the $\tilde{p} \times \tilde{p}$ matrix with all components equal to 1. 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 $\tilde{\Sigma} = (\mathbf{X}^{(\tilde{p})}(c))'(\mathbf{X}^{(\tilde{p})}(c))$. Hence, in this case, $\tilde{\Sigma} = (\bar{e}_1(c), ..., \bar{e}_{\tilde{p}-1}(c), V(c))$, with $\bar{e}_i(c) = ((e_i^{(\tilde{p}-1)})', c)'$ with $e_i^{(\tilde{p}-1)}$ the i-th standard basis vector of $\mathbb{R}^{\tilde{p}-1}$ and with $V(c) = (c, ..., c, 1)'$. Similar as in Berk et al. (2013a) and Leeb et al. (2015), we refer to the data set obtained in the second case as the exchangeable data set (as the covariance matrix $\tilde{\Sigma}$ is permutation-invariant), and to the one obtained in the third case as the equicorrelated data set (as $\tilde{\Sigma}$ is the correlation matrix of a random vector, the last component of which has the same correlation with all the other components). 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. [It is easy to see that the mechanism just described generates matrices of full column rank almost surely. The matrices X actually generated were additionally checked to be of full column rank.]

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) 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. (2015), 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 full model (i.e., from a $N(X\beta, \sigma^2 I_n)$-distribution) using β and $\sigma = 1$ as the true parameters. [For invariance reasons it suffices to consider only the case where $\sigma = 1$.] 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 depending on the m_1 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 $\beta(1)$, $\beta(2)$, $\beta(3)$, 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, SCAD, or MCP model selectors are 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, when n increases. For $n = 100$, the results are thus almost identical for the two targets: For the five 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

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
	20	SCAD	0.90	0.99	1.00	1.00	0.45	0.77	0.84	0.84
	20	MCP	0.89	0.99	1.00	1.00	0.47	0.78	0.85	0.85
	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
	100	SCAD	0.88	0.99	1.00	1.00	0.88	0.99	1.00	1.00
	100	MCP	0.88	0.99	1.00	1.00	0.88	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
	20	SCAD	0.91	1.00	1.00	1.00	0.55	0.90	0.94	0.94
	20	MCP	0.91	1.00	1.00	1.00	0.54	0.89	0.94	0.94
	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
	100	SCAD	0.90	0.99	1.00	1.00	0.90	0.99	1.00	1.00
	100	MCP	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
	20	SCAD	0.88	0.99	1.00	1.00	0.67	0.92	0.95	0.96
	20	MCP	0.86	0.99	1.00	1.00	0.66	0.93	0.96	0.96
	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
	100	SCAD	0.88	0.99	1.00	1.00	0.89	1.00	1.00	1.00
	100	MCP	0.88	0.99	1.00	1.00	0.89	0.99	1.00	1.00

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

the design-independent target. [Note that, for the design-dependent target, for $n = 20$ and $n = 100$, the coverage probabilities of the 'naive' confidence interval are generally smaller for the equicorrelated data set than for the exchangeable data set. This can possibly be explained by the fact that Theorems 6.1 and 6.2 in Berk et al. (2013a) suggest that K_1 should be larger for the equicorrelated data set than for the exchangeable data set. Hence, for the equicorrelated data set, larger confidence intervals seem to be needed to have the required minimal coverage probability for all model selection procedures.] Furthermore, again for $n = 20$ and when AIC or BIC is used, 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 confidence intervals have 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 design-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 supremum in the display in Condition 3.3 is not small for the LASSO procedure, so that the asymptotics in Theorem 3.6 does not provide a good approximation for the finite-sample situation. Finally, for $n = 20$ and for the design-independent target, the results for the SCAD and MCP model selectors lie somewhere in between those of the AIC and BIC and those of the LASSO model selectors. Indeed, for SCAD and MCP, the confidence intervals often fail to have the required minimal coverage probabilities, but less severely than for the LASSO. We stress that the preceding conclusions hold for the LASSO, SCAD, and MCP procedures as implemented here that select the penalty by cross-validation. Other implementations of these procedures 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 depend on the values of X and x_0 used. In additional (non-exhaustive) simulations we have repeated the above analysis for other values of X and x_0 and have found similar results.

4.3   Comparison with the confidence interval of Lee et al. (2015)

In this section we now compare the confidence intervals of Section 2 with a confidence interval recently introduced in Lee et al. (2015). Again, we consider the case where $\mu = X\beta$ and $d = p < n$, and we focus on the design-dependent target $x_0'[\hat{M}]\beta_{\hat{M}}^{(n)}$. As in Lee et al. (2015)

we consider the known-variance case and set $\sigma = 1$ in this section. The confidence interval of Lee et al. (2015) is dedicated to the LASSO model selector and is given in the R package accompanying that paper for the case where x_0 is a standard basis vector. We hence assume in the following that x_0 is equal to the first standard basis vector e_1. The proposed interval is then conditionally valid for the design-dependent target in the following sense: Consider the model selector $\hat{M}$ obtained by selecting those explanatory variables for which the LASSO estimator has non-zero coefficients, with the penalty parameter λ in (4.1) of Lee et al. (2015) being fixed, independently of Y. Then the interval proposed by Lee et al. (2015), which we denote by $\bar{C}I$, satisfies, for any fixed X, for $x_0 = e_1$, and for any fixed $M \subseteq \{1, ..., p\}$ with $1 \in M$,

$$\inf_{\beta \in \mathbb{R}^p} P_{n,\beta,1} \left(x_0'[\hat{M}]\beta_{\hat{M}}^{(n)} \in \bar{C}I \Big| \hat{M} = M \right) = 1 - \alpha, \tag{21}$$

with the convention that the probability in the above display is 1 if $P_{n,\beta,1}(\hat{M} = M) = 0$. The computation of $\bar{C}I$ for a given value of $\hat{M}$ can be carried out without observing $x_0[\hat{M}^c]$, which is also the case for the confidence intervals obtained from $K_2, ..., K_5$, but not for that obtained from K_1. Furthermore, the computation of $\bar{C}I$ (when the conditioning additionally is also on the signs, see Lee et al. (2015) for details) entails a cost that grows linearly with p. Thus, $\bar{C}I$ can be implemented for significantly larger values of p than the confidence intervals based on $K_1, ..., K_4$ currently can be. We note here for later use that $\bar{C}I$ as given in Lee et al. (2015) is not defined on the event that a model $\hat{M}$ is selected that does not contain 1. Hence, we can not speak about unconditional coverage without amending the definition in Lee et al. (2015). [A possible amendment, consistent with our conventions and maximizing unconditional coverage among all possible amendments, is to set $x_0'[\hat{M}]\beta_{\hat{M}}^{(n)} = 0$ if $1 \notin \hat{M}$ and to set $\bar{C}I = \{0\}$ on this event. With such an amendment, $\bar{C}I$ then a fortiori has minimal unconditional coverage probability not less than $1 - \alpha$.]

Despite being specific to the LASSO model selector with fixed λ, we nevertheless find below that the confidence interval of Lee et al. (2015) is *not* shorter than those based on K_1, K_3, and K_4 (presumably due to the fact that (21) imposes a stricter requirement than requiring only correct unconditional coverage). In addition, we point out that this confidence interval can be very sensitive to deviations from the specific model selector it is designed for: In particular, we show that its coverage can break down, when the LASSO model selector is used but with a *data-dependent* penalty parameter λ selected by cross-validation. This is certainly in stark contrast to the confidence intervals obtained from K_1, K_3, and K_4, that are valid for *any* model selection procedure whatsoever.

We first present the results on confidence interval lengths. We conduct a Monte-Carlo study in the case $p = 10$, $n = 100$, and $\alpha = 0.05$. We use the function `fixedLassoInf` of the R package `selectiveInference` to construct the confidence interval of Lee et al. (2015). In line with the presentation in that paper, this function is designed for the cases where x_0 is one of the standard basis vectors of $\mathbb{R}^p$ and thus we set $x_0' = (1, 0, ..., 0)$ as already mentioned above. We consider two different settings for X and β. In the 'independent' setting, we sample independently 1000 values of X, β, and Y in the following way: We first sample the (transposes of the) rows of X and the vector β as $n + 1$ independent draws from the standard Gaussian distribution on $\mathbb{R}^p$. Given X and β, we then sample Y from the $N(X\beta, I_n)$-distribution. For

Setting	Lengths	Confidence interval			
		K_1	K_3	K_4	$\bar{C}I$
'Independent'	Median	0.46	0.78	0.78	0.43
	90%-quantile	0.51	0.85	0.85	1.06
'Correlated'	Median	0.56	0.81	0.81	1.42
	90%-quantile	0.90	1.30	1.30	14.3

Table 2: Medians and empirical quantiles of the lengths of the confidence intervals $\bar{C}I$ of Lee et al. (2015) and of those obtained from K_1, K_3, and K_4. The nominal coverage probability is $1 - \alpha = 0.95$, $n = 100$, and $p = 10$.

each of the 1000 values of X, β, and Y so obtained, we run the LASSO model selector, with λ fixed as a function of X as described at the beginning of Section 7 of Lee et al. (2015). We use the function `glmnet` of the `R` package `glmnet` to compute the LASSO model selector. Then, if the first variable is included in the selected model, we record the lengths of the confidence interval $\bar{C}I$ and of the confidence intervals obtained from K_1, K_3, and K_4 (where these three constants are computed with $r = \infty$ and $\mathcal{M}$ equal to the power set of $\{1, \ldots, p\}$). If not, we discard the realization of X, β, and Y (this is in line with the fact that the $\bar{C}I$ is only defined on the event where $1 \in \hat{M}$ as discussed above). In the 'correlated' setting we proceed as just described, with the only difference that the rows of X are sampled according to the Gaussian distribution with mean vector $0 \in \mathbb{R}^p$ and covariance matrix $[\exp(-|i - j|/10)]_{i,j=1}^{10}$ and that then β is sampled from a random vector b, so that Xb follows the standard Gaussian distribution within the column space of X. [We note that this mechanism almost surely generates matrices X that have full column rank.]

The medians and empirical 90%-quantiles of the confidence interval lengths distributions obtained that way are reported in Table 2. The conclusion is that there is no unilateral hierarchy of the two methods for confidence interval construction (that of Lee et al. (2015) and that using the intervals based on K_1, K_3, and K_4) in terms of median length. Depending on the situation, any of the two methods can provide the smallest median length. The 90%-quantiles, on the other hand, are always larger for the confidence interval $\bar{C}I$ of Lee et al. (2015) than for those obtained from K_1, K_3, and K_4. The feature that $\bar{C}I$ can be very long (with small but non-negligible probability) has also been noted in Lee et al. (2015), and is not shared by the confidence intervals obtained from K_1, K_3, and K_4. Note finally that we have obtained the same conclusions in other length simulations, which we do not report for the sake of brevity.

We now demonstrate that the confidence interval $\bar{C}I$ of Lee et al. (2015) can have conditional coverage probability considerably smaller than the nominal one when λ is selected by cross-validation. Rather than evaluating the minimal conditional coverage probabilities conditional on $\hat{M} = M$ separately for every M satisfying $1 \in M$, which would be quite costly, we evaluate the minimal conditional coverage probability where conditioning is on the event that $1 \in \hat{M}$. We denote this quantity by $P_{cond,\min}$. A simple calculation shows that if we find that this latter minimal conditional coverage probability is smaller than $1 - \alpha$, then it follows

that we must have

$$\inf_{\beta \in \mathbb{R}^p} P_{n,\beta,1} \left(x_0'[\hat{M}]\beta_{\hat{M}}^{(n)} \in \bar{C}I \, \middle| \, \hat{M} = M \right) < 1 - \alpha$$

for at least some M satisfying $1 \in M$, showing that property (21) is violated. [To see this, note that $P_{n,\beta,1}(\cdot \,|\, 1 \in \hat{M})$ is a convex combination (over all M with $1 \in M$) of the probabilities $P_{n,\beta,1}(\cdot \,|\, \hat{M} = M)$, with the (nonnegative) weights summing to 1.]

In order to numerically evaluate the minimal conditional coverage probability $P_{cond,\min}$ we proceed as follows: We consider eight configurations given by all the possible combinations of $n = 20, 100$, $p = 2, 10$ and $\alpha = 0.05, 0.2$. Recall that $x_0' = (1, 0, ..., 0)$. For each of these eight configurations, the (transpose of the) rows of X are sampled once from the $N(0, \Sigma)$-distribution and then remain fixed throughout the minimal coverage probability evaluation. For $p = 2$, we take Σ to have 1 as the diagonal and 0.8 as the off-diagonal elements. For $p = 10$, we define $\mathbf{X}^{(p)}(c)$ as in Section 6.2 of Berk et al. (2013a), with $c = \sqrt{0.8/(p-1)}$ and we set $\Sigma = (\mathbf{X}^{(p)}(c))'(\mathbf{X}^{(p)}(c))$. For each of these eight configurations, we carry out the three-step minimal coverage probability evaluation of Section 4.2, with the same values of m_1, m_2 and I_1, I_2, I_3 (with the only difference that in the third step the coverage probability is reevaluated only for one value $\beta(1)$, say, where $\beta(1)$ corresponds to that value of β that gives the smallest estimate for the coverage probability in the second step). When we evaluate a conditional coverage probability for a given β in this process, we proceed as follows: We sample I_1 (or I_2 or I_3) values of Y from the $N(X\beta, I_n)$ distribution. For each value of Y we run the model selection procedure $\hat{M}$, where $\lambda = \hat{\lambda}$ is first selected by cross-validation with the `cv.glmnet` function of the `R` package `glmnet`, and where the `glmnet` function is then used to compute the selected model with the LASSO with penalty parameter $\hat{\lambda}$. Then, if the selected model does not contain the first explanatory variable, we discard the value of Y, and else, we record whether the design-dependent target belongs to $\bar{C}I$ or not. The conditional coverage probability is then obtained by taking the average number of times this is the case, over all the recorded events.

The evaluated minimal conditional coverage probabilities $\hat{P}_{cond,\min}$, say, are presented in Table 3. For comparison, we also provide similar evaluations of minimal conditional coverage probabilities, with the same above-described procedure, but with λ now fixed as a function of X as in the beginning of Section 7 of Lee et al. (2015). When λ is fixed, these minimal conditional coverage probabilities are approximately equal to the nominal level $1 - \alpha$, in agreement with the results of Lee et al. (2015). However, when λ is selected by cross-validation, the evaluated minimal conditional coverage probabilities can be way below the nominal level. In particular, these probabilities can be equal to 0.31 for a nominal level of 0.80 and to 0.86 for a nominal level of 0.95.

In addition, for λ selected by cross-validation and in all the configurations of n, p, and α, for the vector β leading to the minimal conditional coverage probability $\hat{P}_{cond,\min}$, we can also estimate the unconditional coverage probability $P_{n,\beta,1}(x_0'[\hat{M}]\beta_{\hat{M}}^{(n)} \in \bar{C}I)$ by $\hat{P}(1 \in \hat{M})\hat{P}_{cond,\min} + \hat{P}(1 \notin \hat{M})$. [Here we make use of the aforementioned amendment to $\bar{C}I$ in order to allow for a well-defined unconditional coverage probability.] In this estimate, $\hat{P}(1 \in \hat{M})$ is the proportion of times the first regressor belongs to $\hat{M}$, over the I_3 Monte-Carlo samples.

p	n	$1 - \alpha$	LASSO	
			Fixed λ	CV-selected λ
2	20	0.80	0.80	0.43
		0.95	0.95	0.93
	100	0.80	0.80	0.31
		0.95	0.95	0.86
10	20	0.80	0.79	0.79
		0.95	0.94	0.93
	100	0.80	0.79	0.70
		0.95	0.95	0.92

Table 3: Monte-Carlo estimates of the minimal conditional coverage probabilities (w.r.t. β and σ) of the confidence intervals of Lee et al. (2015), with the LASSO model selector where λ is either fixed or selected by cross validation.

The so evaluated unconditional coverage probabilities are 0.60 for $p = 2$, $n = 100$, $1 - \alpha = 0.80$ and 0.91 for $p = 2$, $n = 100$, $1 - \alpha = 0.95$, which implies that the confidence intervals of Lee et al. (2015) also have minimal *unconditional* coverage probabilities below the nominal level when λ is estimated by cross-validation.

The conclusion of this comparison, and particularly of the evaluations of minimal coverage probabilities, is that, although the confidence intervals of Lee et al. (2015) are conditionally valid and convenient to compute, their current applicability appears to be restricted to the case where the tuning parameter λ is fixed. [An extension of the strategy of Lee et al. (2015) to cross-validated versions of LASSO has recently been studied in Loftus and Taylor (2015) and Loftus (2015). This extension, however, comes with significantly higher computational cost.] This also highlights the benefit of the confidence intervals introduced in Section 2, which are intrinsically designed to be valid for any model selection procedure whatsoever.

5 Conclusion

We have extended the PoSI confidence intervals of Berk et al. (2013a) to PoSI intervals for predictors. The coverage targets of our intervals, i.e., $x_0'[\hat{M}]\beta_{\hat{M}}^{(n)}$ and $x_0'[\hat{M}]\beta_{\hat{M}}^{(\star)}$, respectively, minimize a certain in-sample prediction error and, under additional assumptions relating the training period to the prediction period, also a certain out-of-sample prediction error. For in-sample prediction, i.e., for the target $x_0'[\hat{M}]\beta_{\hat{M}}^{(n)}$, our intervals are valid, in finite samples, irrespective of the model selection procedure that is being used. For out-of-sample prediction, i.e., for the target $x_0'[\hat{M}]\beta_{\hat{M}}^{(\star)}$, the same is true asymptotically under very mild assumptions on the underlying model selector. See also Greenshtein and Ritov (2004) for optimality results related to the latter target and for its feasible counterpart, under appropriate sparsity conditions.

Two types of confidence intervals were studied here: The first one (corresponding to the constant $K_1(x_0, \hat{M})$) depends on all components of the vector x_0 (even if only a subset of these

31

components is 'active' in the selected model $\hat{M}$) and thus is feasible only if x_0 is observed completely. The intervals of the second type (corresponding to the constants $K_2(x_0[\hat{M}], \hat{M})$, $K_3(x_0[\hat{M}], \hat{M})$, and K_4) depend only on the active components in the selected model, i.e., on $x_0[\hat{M}]$. The constants K_2, K_3, and K_4 correspond to successively larger confidence intervals.

Computing the constant K_2 was found to be quite expensive in practice. For computing the remaining constants, simple algorithms were presented. The computational complexity of our algorithms for computing K_1 and K_3 is governed by the number of candidate models under consideration, limiting computations to a few million candidate models in practice. Computation of K_4 is easy and not limited by complexity constraints (see, however, the warning about numerical stability in Remark C.5 in Appendix C). Our algorithms are of similar computational complexity as those proposed in Berk et al. (2013a).

We furthermore have studied the behavior of these constants and of the corresponding confidence intervals through analytic results in a setting where model dimension is allowed to grow with sample size, and also through simulations. These results provide evidence that K_4, which is relatively cheap to compute, is a reasonably tight bound for the computationally more expensive constants K_1 to K_3. Furthermore, these results show that all the constants K_1 to K_4 are 'bounded away' from the Scheffé constant.

We have also provided simulation results regarding the coverage probabilities of the various intervals introduced in the paper. We find that the asymptotic results in Section 3 regarding the design-independent target already 'kick-in' at moderate sample sizes, and these results demonstrate that the PoSI confidence intervals for the predictors maintain the desired minimal coverage probability. The simulation study also shows that 'naive' confidence intervals, which ignore the data-driven model selection step and which use standard confidence procedures as if the selected model were correct and given a priori, are invalid also in the setting considered here (which is in line with earlier findings in Leeb et al. (2015), where inter alia 'naive' confidence intervals for components of $\beta_{\hat{M}}^{(n)}$ were studied). Furthermore, studying the confidence intervals developed for model selection with the LASSO by Lee et al. (2015), and others, we find that these intervals are invalid if the LASSO penalty is chosen by cross-validation. This contrasts the established fact that these intervals are valid (conditionally on the event that a given model is selected), if the penalty is fixed in advance.

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 $\left|a'_{l_1} w_1\right| = t_1$ 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 $w \in B$

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

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

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

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 $F(t) = \mathbb{E}_S \mathbb{E}_{S^*} \mathbf{1}(S^* \leq t/S) = \Pr(S S^* \leq t)$. Because $S^* \neq 0$ holds almost

surely by the assumption on F^*, we have $\Pr\left(SS^* = t\right) = \mathbb{E}_{S^*}\mathbb{E}_S \mathbf{1}\left(S = t/S^*\right)$. Since S has a continuous distribution function, we have $\mathbb{E}_S \mathbf{1}\left(S = t/S^*\right) = 0$ almost surely, implying that $\Pr\left(SS^* = t\right) = 0$. ∎

Proof of Proposition 2.3: To prove the first claim observe that in case $M = \{1, \ldots, p\}$ we have $c(M, \mathcal{M}) = 0$ and that $x_0 = 0$ implies $\bar{s}_{M_*} = 0$ for every M_*; thus 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 $M = \{1, \ldots, p\}$ but $x_0 \neq 0$, 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 in this case $F_{M,x_0}(t)$ reduces to $P_{n,\mu,\sigma}\left(\max_{M_* \in \mathcal{M}} \left|\bar{s}'_{M_*}(Y - \mu)\right|/\hat{\sigma} \leq t\right)$ (see (22) and (23) below) and since not all $\bar{s}_{M_*}$ can be zero (in view of our assumptions on $\mathcal{M}$).

In case $M \in \mathcal{M}$ 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 now $c(M, \mathcal{M}) \geq 1$ holds and since $F_{Beta,1/2,(d-1)/2}(0) = 0$. 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 $F_{Beta,1/2,(d-1)/2}(1) = 1$) 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 $M \in \mathcal{M}$ is a proper subset of $\{1, \ldots, p\}$ and $p = 1$ holds, then M is empty and $d = 1$ must hold, and hence F^*_{M,x_0} reduces to the indicator function of $[1, \infty)$. But then $F_{M,x_0}(t) = \Pr\left(G \leq t\right)$ 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 $M_* \in \mathcal{M}$ and hence we have

$$
P_{n,\mu,\sigma}\left(\max_{M_* \in \mathcal{M}} \left|\bar{s}'_{M_*}(Y - \mu)\right|/\hat{\sigma} > t\right)
$$

$$
= P_{n,\mu,\sigma}\left(\max_{M_* \in \mathcal{M}} \left|\bar{s}'_{M_*} P_X(Y - \mu)/\|P_X(Y - \mu)\|\right| > \left(\hat{\sigma}/\|P_X(Y - \mu)\|\right)t\right), \qquad (22)
$$

where $P_X(Y - \mu)/\|P_X(Y - \mu)\|$ and $\|P_X(Y - \mu)\|/\hat{\sigma}$ are independent since the random variables $P_X(Y - \mu)/\|P_X(Y - \mu)\|$, $\|P_X(Y - \mu)\|$, and $\hat{\sigma}$ are mutually independent. [Observe that $P_X(Y - \mu)$ is nonzero with probability 1 since $d \geq 1$ holds.] Consequently, the probability given above can be represented as

$$
\Pr\left(\max_{M_* \in \mathcal{M}} \left|\bar{s}'_{M_*}V\right| > t/G\right) \qquad (23)
$$

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 $M \in \mathcal{M}$

and $t \geq 0$

$$\Pr\left(\max_{M_* \in \mathcal{M}} \left|\vec{s}'_{M_*} V\right| > t/G\right) = \int \Pr\left(\max_{M_* \in \mathcal{M}} \left|\vec{s}'_{M_*} V\right| > t/g\right) dF_G(g)$$

$$\leq \int \min\left[1, \Pr\left(\max_{M_* \in \mathcal{M}, M_* \subseteq M} \left|\vec{s}'_{M_*} V\right| > t/g\right) + \Pr\left(\max_{M_* \in \mathcal{M}, M_* \not\subseteq M} \left|\vec{s}'_{M_*} V\right| > t/g\right)\right] dF_G(g)$$

$$\leq \int \min\left[1, \Pr\left(\max_{M_* \in \mathcal{M}, M_* \subseteq M} \left|\vec{s}'_{M_*} V\right| > t/g\right) + \sum_{M_* \in \mathcal{M}, M_* \not\subseteq M} \Pr\left(\left|\vec{s}'_{M_*} V\right| > t/g\right)\right] dF_G(g)$$

$$= \int \min\left[1, \Pr\left(\max_{M_* \in \mathcal{M}, M_* \subseteq M} \left|\vec{s}'_{M_*} V\right| > t/g\right) + \sum_{M_* \in \mathcal{M}, M_* \not\subseteq M} \Pr\left(\left(\vec{s}'_{M_*} V\right)^2 > t^2/g^2\right)\right] dF_G(g)$$

$$\leq \int \left(1 - F^*_{M,x_0}\left(t/g\right)\right) dF_G(g) = \mathbb{E}_G\left(1 - F^*_{M,x_0}\left(t/G\right)\right) = 1 - F_{M,x_0}\left(t\right), \tag{24}$$

where F_G here denotes the c.d.f. of G. The last inequality follows from the fact that $\Pr\left(\left(\vec{s}'_{M_*} V\right)^2 > t^2/g^2\right)$ is either equal to zero (if $\bar{s}_{M_*} = 0$) or is equal to $1 - F_{Beta,1/2,(d-1)/2}\left(t^2/g^2\right)$ (if $\bar{s}_{M_*} \neq 0$) as is easy to see; for the case where M is the empty set also observe that $\Pr\left(\max_{M_* \in \mathcal{M}, M_* \subseteq M} \left|\vec{s}'_{M_*} V\right| > t/g\right) = 0$ for $t \geq 0$ because $\bar{s}_\emptyset = 0$. In view of (10) the chain of inequalities in (22)-(24) 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 $\Pr\left(\max_{M_* \in \mathcal{M}, M_* \subseteq M_1} \left|\vec{s}'_{M_*} V\right| > t\right)$ for $t \geq 0$ are equal to zero if M_1 is empty)

$$\Pr\left(\max_{M_* \in \mathcal{M}, M_* \subseteq M_2} \left|\vec{s}'_{M_*} V\right| > t\right) + c(M_2, \mathcal{M})\left(1 - F_{Beta,1/2,(d-1)/2}\left(t^2\right)\right)$$

$$\leq \Pr\left(\max_{M_* \in \mathcal{M}, M_* \subseteq M_1} \left|\vec{s}'_{M_*} V\right| > t\right) + \Pr\left(\max_{M_* \in \mathcal{M}, M_* \subseteq M_2, M_* \not\subseteq M_1} \left|\vec{s}'_{M_*} V\right| > t\right)$$

$$+ c(M_2, \mathcal{M})\left(1 - F_{Beta,1/2,(d-1)/2}\left(t^2\right)\right)$$

$$\leq \Pr\left(\max_{M_* \in \mathcal{M}, M_* \subseteq M_1} \left|\vec{s}'_{M_*} V\right| > t\right) + \sum_{M_* \in \mathcal{M}, M_* \subseteq M_2, M_* \not\subseteq M_1} \Pr\left(\left|\vec{s}'_{M_*} V\right| > t\right)$$

$$+ c(M_2, \mathcal{M})\left(1 - F_{Beta,1/2,(d-1)/2}\left(t^2\right)\right)$$

$$\leq \Pr\left(\max_{M_* \in \mathcal{M}, M_* \subseteq M_1} \left|\vec{s}'_{M_*} V\right| > t\right) + c(M_1, \mathcal{M})\left(1 - F_{Beta,1/2,(d-1)/2}\left(t^2\right)\right).$$

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

$$1 - F_{\emptyset,x_0}(t) = \mathbb{E}_G\left(1 - F^*_{\emptyset,x_0}(t/G)\right) = \mathbb{E}_G\left(\left(1 - F^*_{\emptyset,x_0}(t/G)\right)\mathbf{1}\left(t \leq G\right)\right)$$

$$\leq \mathbb{E}_G\mathbf{1}\left(t \leq G\right) = 1 - \Pr\left(G \leq t\right),$$

which proves $K_4 \leq K_5$. ∎

35

Lemma A.4. *Assume $p = 2$ and $n \geq 2$. Then there exists a design matrix X with full column rank and a vector x_0 such that $K_4 = K_1(x_0)$ for $\mathcal{M}$ the power set of $\{1, 2\}$.*

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 (24) for $M = \varnothing$ and all $t \geq 0$. Inspection of (24) 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

$$X = X^{(2)} = \left[\begin{array}{cc} 1 & \cos(2\pi/3) \\ 0 & \sin(2\pi/3) \end{array}\right]$$

and $x'_0 = x_0^{(2)'} = (\cos(4\pi/3), \sin(4\pi/3)) X^{(2)}$: 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 $\left\{\max_{\varnothing \neq M_* \subseteq \{1,2\}} \left|\bar{s}'_{M_*} V\right| > c\right\}$ is either the entire space or is the disjoint union of the events $\{|\bar{s}'_{\{1\}} V| > c\}$, $\{|\bar{s}'_{\{2\}} V| > c\}$ and $\{|\bar{s}'_{\{1,2\}} V| > c\}$. In the case $n > 2$ simply set

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

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.

Proof of Proposition 2.9: (a) The distribution of $\omega = \max_{M \in \mathcal{M}} \left|\bar{s}'_M (Y - \mu)\right| / \sigma$ does clearly not change if X is replaced by AX, where A is an orthogonal $n \times n$ matrix. Furthermore, scaling the columns of X and the corresponding columns of x'_0 by the same (column-specific) positive constants does not alter $\max_{M \in \mathcal{M}} \left|\bar{s}'_M (Y - \mu)\right| / \sigma$. Hence, we may assume w.l.o.g. that X consist of the first p standard basis vectors of $\mathbb{R}^n$. Then choose x'_0 as the $1 \times p$ vector $(1, \ldots, 1)$. It follows that ω can be written as

$$\max_{M \in \mathcal{M}} \left|\sum_{i \in M} Z_i\right| / \sqrt{|M|}$$

where Z_i are i.i.d. standard normal and where we use the convention that the expression in the display is zero if $|M| = 0$. For a positive real b define the random set $\tilde{M} = \{i \in \{1, ..., p\} : Z_i \geq b\}$. Since for any realization of the random variables Z_i we have that $\tilde{M} \in \mathcal{M}$, we must have

$$\omega / \sqrt{p} \geq p^{-1/2} \left|\sum_{i \in \tilde{M}} Z_i\right| / \sqrt{\left|\tilde{M}\right|} = \left|p^{-1} \sum_{i=1}^{p} Z_i \mathbf{1}(Z_i \geq b)\right| / \sqrt{p^{-1} \sum_{i=1}^{p} \mathbf{1}(Z_i \geq b)}.$$

By the law of large numbers we obtain that the r.h.s. converges to $\phi(b)/\sqrt{1 - \Phi(b)}$ almost surely. Because $K_1(x_0, \infty)/\sqrt{p}$ is the $(1 - \alpha)$-quantile of $\omega/\sqrt{p}$ with α independent of p and since $b > 0$ was arbitrary, the first claim follows. The second claim follows immediately by choosing x_0 equal to a $p \times 1$ standard basis vector and by noting that then ω is distributed as the absolute value of a standard normal variable.

(b) For the same reasons as given at the beginning of the proof of part (a) we have for every p that

$$
\begin{aligned}
&\inf_{x_0 \in \mathbb{R}^p} \inf_{X \in \mathsf{X}(p)} \inf_{M \in \mathcal{M}, |M| \leq \gamma p} K_2(x_0[M], M, \infty, X, \alpha, \mathcal{M})/\sqrt{p} \\
&= \inf_{x_0 \in \mathbb{R}^p} \inf_{n \geq p} \inf_{M \in \mathcal{M}, |M| \leq \gamma p} K_2(x_0[M], M, \infty, (I_p, 0_{p \times (n-p)})', \alpha, \mathcal{M})/\sqrt{p} \qquad (25) \\
&= \inf_{x_0 \in \mathbb{R}^p} \inf_{M \in \mathcal{M}, |M| \leq \gamma p} K_2(x_0[M], M, \infty, I_p, \alpha, \mathcal{M})/\sqrt{p}
\end{aligned}
$$

where I_p is the identity matrix of dimension p. By the monotonicity property (16) the r.h.s. of (25) equals

$$
\inf_{x_0 \in \mathbb{R}^p} \inf_{M \in \mathcal{M}, |M| = \lfloor \gamma p \rfloor} K_2(x_0[M], M, \infty, I_p, \alpha, \mathcal{M})/\sqrt{p}. \qquad (26)
$$

Now fix an arbitrary $x_0 \in \mathbb{R}^p$ and $M \in \mathcal{M}$ with $|M| = \lfloor \gamma p \rfloor$. Define x_0^* via $x_{0i}^* = x_{0i}$ for $i \in M$ and set $x_{0i}^* = 1$ else. Then

$$
K_2(x_0[M], M, \infty, I_p, \alpha, \mathcal{M}) \geq K_1(x_0^*, \infty, I_p, \alpha, \mathcal{M}) \qquad (27)
$$

holds and the latter quantity is the $(1 - \alpha)$-quantile of

$$
\omega^* = \max_{M^* \in \mathcal{M}} \left| \sum_{i \in M^*} x_{0i}^* Z_i \right| / \sqrt{\sum_{i \in M^*} (x_{0i}^*)^2}.
$$

For a positive real b define now the random set $\check{M} = \{i \notin M : Z_i \geq b\}$. Similar as above we then conclude that

$$
\omega^*/\sqrt{p} \geq \left| p^{-1} \sum_{i \notin M} Z_i \mathbf{1}(Z_i \geq b) \right| / \sqrt{p^{-1} \sum_{i \notin M} \mathbf{1}(Z_i \geq b)}.
$$

While the r.h.s. of the above display depends on M, its distribution does not as it coincides with the distribution of

$$
A_p = \left| p^{-1} \sum_{i=\lfloor \gamma p \rfloor + 1}^{p} Z_i \mathbf{1}(Z_i \geq b) \right| / \sqrt{p^{-1} \sum_{i=\lfloor \gamma p \rfloor + 1}^{p} \mathbf{1}(Z_i \geq b)}.
$$

Consequently, $K_1(x_0^*, \infty, I_p, \alpha, \mathcal{M})/\sqrt{p}$ as the $(1 - \alpha)$-quantile of $\omega^*/\sqrt{p}$ is not smaller than the corresponding quantile of A_p, and this is true independently of the choice of x_0 and of M with $|M| = \lfloor \gamma p \rfloor$. Since A_p converges to $\sqrt{1 - \gamma}\phi(b)/\sqrt{1 - \Phi(b)}$ almost surely and α does not depend on p, we can conclude that

$$
\liminf_{p \to \infty} \inf_{x_0 \in \mathbb{R}^p} \inf_{M \in \mathcal{M}, |M| = \lfloor \gamma p \rfloor} K_1(x_0^*, \infty, I_p, \alpha, \mathcal{M})/\sqrt{p} \geq \sqrt{1 - \gamma}\phi(b)/\sqrt{1 - \Phi(b)}.
$$

Since $b > 0$ was arbitrary, the proof is then complete in view of (25), (26), and (27). $\blacksquare$

Lemma A.6. *Let α, $0 < \alpha < 1$, be a fixed number, let $N \in \mathbb{N}$, and $d \in \mathbb{N}$. Let $K(N, d, \alpha)$ denote the $(1 - \alpha)$-quantile of the distribution function given by*

$$1 - \mathbb{E}_G \left(\min \left[1, N \left(1 - F_{Beta, 1/2, (d-1)/2} \left(t^2 / G^2 \right) \right) \right] \right)$$

for $t \geq 0$ and by 0 for $t < 0$. Here G is a nonnegative random variable such that G^2 follows a chi-square distribution with d degrees of freedom. Then as $\min(N, d) \to \infty$

$$K(N, d, \alpha) / \sqrt{d \left(1 - N^{-2/(d-1)} \right)} \to 1.$$

Proof: The cdf in the lemma is the cdf of GW, where W is independent of G, is nonnegative, and has distribution function given by

$$1 - \min \left[1, N \left(1 - F_{Beta, 1/2, (d-1)/2} \left(t^2 \right) \right) \right]$$

for $t \geq 0$. Observe that $G/\sqrt{d}$ converges to 1 in probability as $d \to \infty$. To complete the proof it thus suffices to show that $W / \sqrt{\left(1 - N^{-2/(d-1)} \right)}$ converges to 1 in probability as $\min(N, d) \to \infty$: For $t > 1$ we have that

$$\Pr \left(W / \sqrt{\left(1 - N^{-2/(d-1)} \right)} > t \right) \leq N \left(1 - F_{Beta, 1/2, (d-1)/2} \left(t^2 \left(1 - N^{-2/(d-1)} \right) \right) \right).$$

But the r.h.s. of the preceding display has been shown in Zhang (2013) to converge to zero as $\min(N, d) \to \infty$, cf. (A.5) and (A.6) in that paper. For $t < 1$ we have

$$\Pr \left(W / \sqrt{\left(1 - N^{-2/(d-1)} \right)} \leq t \right) = 1 - \min \left[1, N \left(1 - F_{Beta, 1/2, (d-1)/2,} \left(t^2 \left(1 - N^{-2/(d-1)} \right) \right) \right) \right],$$

and hence it suffices to show that

$$N \left(1 - F_{Beta, 1/2, (d-1)/2,} \left(t^2 \left(1 - N^{-2/(d-1)} \right) \right) \right) \to \infty$$

as $\min(N, d) \to \infty$. But this has again been established in Zhang (2013), see (C.4) and (C.5) in that paper. ∎

Proof of Proposition 2.10: Observe that $K_4(\infty)$ is always positive and hence the ratio in (18) is well-defined. In view of (12) and (13) and the assumptions on $\mathcal{M}$ we have for $M \in \mathcal{M}$ with $M \neq \{1, \ldots, p\}$, for $X \in \mathsf{X}_{n(p), p}(\mathcal{M})$, and for $x_0 \in \mathbb{R}^p$ that $K_3(x_0[M], M, \infty)$ is not less than $K(\lfloor \tau |\mathcal{M}| \rfloor, \min(n(p), p), \alpha)$ in the notation of Lemma A.6, where we note that $\lfloor \tau |\mathcal{M}| \rfloor \geq 1$ holds at least from a certain p onwards. Since $K_3(x_0[M], M, \infty) \leq K_4(\infty)$ always holds, and since $K_4(\infty) = K(|\mathcal{M}| - 1, \min(n(p), p), \alpha)$ in the notation of Lemma A.6, it suffices to show that

$$K(\lfloor \tau |\mathcal{M}| \rfloor, \min(n(p), p), \alpha) / K(|\mathcal{M}| - 1, \min(n(p), p), \alpha) \to 1$$

as $p \to \infty$. Note that $d = \min(n(p), p) \to \infty$ as $p \to \infty$ by the assumption on $n(p)$, and that $|\mathcal{M}| \to \infty$ as $p \to \infty$. By Lemma A.6 we thus need to show that

$$A_d(|\mathcal{M}|) := \left(1 - (|\mathcal{M}| - 1)^{-2/(d-1)} \right) / \left(1 - \lfloor \tau |\mathcal{M}| \rfloor^{-2/(d-1)} \right) \to 1$$

38

as $p \to \infty$. Observe that $\tau < 1$ must hold, and thus $|\mathcal{M}| - 1 \geq \lfloor \tau |\mathcal{M}| \rfloor > 1$ holds for large p. This, in particular, implies $A_d(|\mathcal{M}|) \geq 1$ for large p. But then for large p

$$1 \leq A_d(|\mathcal{M}|) \leq \left(1 - |\mathcal{M}|^{-2/(d-1)}\right) / \left(1 - ((\tau/2)|\mathcal{M}|)^{-2/(d-1)}\right) =: B_d(|\mathcal{M}|)$$

holds since also $\lfloor \tau |\mathcal{M}| \rfloor \geq (\tau/2)|\mathcal{M}| > 1$ is satisfied for large p. It thus suffices to show that $B_d(|\mathcal{M}|) \to 1$ for $p \to \infty$. Let $f > 2/\tau$ be a real number. Then $|\mathcal{M}| \geq f$ holds for large p. Because $B_d(x)$ is monotone decreasing in x for $x > 2/\tau > 1$ and for every d as is easily checked by inspection of the derivative, we have that $B_d(f) \geq B_d(|\mathcal{M}|)$ holds for large p. But now it is easily checked (Hôpital's rule) that $B_d(f)$ converges to $\log f / (\log f + \log(\tau/2))$ as $p \to \infty$ (and thus $d \to \infty$). Making f arbitrarily large, $\log f / (\log f + \log(\tau/2))$ approaches 1. This completes the proof of (18). The second claim follows immediately from $K_4 = K(|\mathcal{M}| - 1, \min(n(p), p), \alpha)$, the preceding lemma, the observation that $B_d(|\mathcal{M}|) \to 1$, and that

$$1/B_d(|\mathcal{M}|) \leq \left(1 - (|\mathcal{M}| - 1)^{-2/(d-1)}\right) / \left(1 - |\mathcal{M}|^{-2/(d-1)}\right) \leq 1$$

holds. ∎

Proof of Corollary 2.12: Properties (i), (ii), and (iv) are obvious. In case $m_p = p$, we have that $\mathcal{M}(m_p)$ is the power set and hence $|\mathcal{M}(m_p)| = 2^{m_p} = 2^p$. But then we have $|M| < m_p = p$ for $M \neq \{1, \dots, p\}$. Consequently,

$$c(M, \mathcal{M}(m_p)) = |\mathcal{M}(m_p)| - 2^{|M|} \geq 2^{m_p} - 2^{m_p - 1} = (1/2)|\mathcal{M}(m_p)| \geq (1/3)|\mathcal{M}(m_p)|.$$

Next consider the case where $m_p < p$. Then certainly $|\mathcal{M}(m_p)| \geq 2^{m_p + 1} - 1$ holds. Now, for $M \in \mathcal{M}(m_p)$ we have $|M| \leq m_p$ (and $M \neq \{1, \dots, p\}$) and thus

$$\begin{aligned} c(M, \mathcal{M}(m_p)) &= |\mathcal{M}(m_p)| - 2^{|M|} \geq |\mathcal{M}(m_p)| - 2^{m_p} = |\mathcal{M}(m_p)|(1 - 2^{m_p}/|\mathcal{M}(m_p)|) \\ &\geq |\mathcal{M}(m_p)|\left(1 - 2^{m_p}/(2^{m_p + 1} - 1)\right) \geq (1/3)|\mathcal{M}(m_p)|. \end{aligned}$$

Thus (iii) with $\tau = 1/3$ holds. The next claim now follows from Proposition 2.10 since $|\mathcal{M}(m_p)| = \sum_{k=0}^{m_p} \binom{p}{k}$. The final claim is then a trivial consequence, since $|\mathcal{M}(p)| = 2^p$ in this case. Note that $\mathsf{X}_{n(p),p}(\mathcal{M}(p)) \neq \varnothing$ by the assumptions on $n(p)$, implying $n(p) \geq p$. ∎

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 $\tilde{\sigma}_2 = \tilde{\sigma}_{2,n}$ (which may be functions of σ) satisfying*

$$\sup_{\beta \in \mathbb{R}^p, \sigma > 0} P_{n,\beta,\sigma}\left(|(\tilde{\sigma}_i/\sigma) - 1| > \delta \,|\, X\right) \to 0 \tag{28}$$

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

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

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

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

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

$$Q_n = \sup_{x_0 \in S(M), \beta \in \mathbb{R}^p, \sigma > 0, W \in \mathcal{W}} \left| P_{n,\beta,\sigma} \left(|\sigma e_1 + \sigma e_2| \leq W_M \tilde{\sigma}_1; \hat{M} = M \,\middle|\, X \right) \right.$$
$$\left. - P_{n,\beta,\sigma} \left(|\sigma e_2| \leq W_M \tilde{\sigma}_2; \hat{M} = M \,\middle|\, X \right) \right| \tag{29}$$

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

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

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 $|\Pr(A \cap C) - \Pr(B \cap C)| \leq \Pr(A^c \cap B \cap C) + \Pr(A \cap B^c \cap C)$ we can bound the absolute value inside the supremum in (29) by

$$P_{n,\beta,\sigma} \left(|\sigma e_1 + \sigma e_2| > W_M \tilde{\sigma}_1; |\sigma e_2| \leq W_M \tilde{\sigma}_2; \hat{M} = M \,\middle|\, X \right)$$
$$+ P_{n,\beta,\sigma} \left(|\sigma e_1 + \sigma e_2| \leq W_M \tilde{\sigma}_1; |\sigma e_2| > W_M \tilde{\sigma}_2; \hat{M} = M \,\middle|\, X \right). \tag{30}$$

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

$$P_{n,\beta,\sigma} \left(\tilde{\sigma}_1 \left(W_M - \delta_{n,1} \right)_+ \leq |\sigma e_2| \leq \tilde{\sigma}_2 W_M \,\middle|\, X \right)$$
$$+ P_{n,\beta,\sigma} \left(\tilde{\sigma}_2 W_M \leq |\sigma e_2| \leq \tilde{\sigma}_1 \left(W_M + \delta_{n,1} \right) \,\middle|\, X \right)$$
$$+ 2 P_{n,\beta,\sigma} \left(|\sigma e_1| \geq \tilde{\sigma}_1 \delta_{n,1}; \hat{M} = M \,\middle|\, X \right). \tag{31}$$

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

$$\sup_{\beta,\sigma} P_{n,\beta,\sigma}\left(\max_{i=1,2}|(\tilde{\sigma}_i/\sigma)-1| > \delta_{n,2}\,\middle|\, X\right) \to 0 \tag{32}$$

in probability as $n \to \infty$. This can easily be seen from a diagonal sequence argument. Now, using (29), (30), (31), and (32), we have

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

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 $c_n\,(x_0, X) = x_0'[M]\left(n^{-1}X[M]'X[M]\right)^{-1}x_0[M]$. By Condition 3.4 the variance $c_n\,(x_0, X)$ converges to $c(x_0) = x_0'[M]\,(\Sigma[M, M])^{-1}\,x_0[M] > 0$ in probability, and in fact even uniformly in $x_0 \in S(M)$. Since $\Sigma[M, M]$ is obviously positive definite, $0 < c_* \le c(x_0) \le c^* < \infty$ must hold for all $x_0 \in S(M)$. Consequently,

$$\sup_{x_0\in S(M)}\left|\left(c_n^{1/2}\,(x_0, X)\,/c^{1/2}\,(x_0)\right) - 1\right| \tag{33}$$

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 (33) is less than $\delta_{n,3}$ has probability converging to 1. On this event $\inf_{x_0\in S(M)} c_n\,(x_0, X)$ is then positive for sufficiently large n and we have

41

on D_n and for sufficiently large n

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

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

$$
\begin{aligned}
Q_{n,2} \;\leq\; & \sup_{x_0 \in S(M), \| \beta[M^c] \| / \sigma \geq \delta_{n,4}} P_{n,\beta,\sigma} \left(|e_1| \geq (1 - \delta_{n,2}) \delta_{n,1}; \hat{M} = M \, \middle| \, X \right) \\
& + \sup_{x_0 \in S(M), \| \beta[M^c] \| / \sigma < \delta_{n,4}} P_{n,\beta,\sigma} \left(|e_1| \geq (1 - \delta_{n,2}) \delta_{n,1}; \hat{M} = M \, \middle| \, X \right) \\
\leq \; & \sup_{\| \beta[M^c] \| / \sigma \geq \delta_{n,4}} P_{n,\beta,\sigma} \left(\hat{M} = M \, \middle| \, X \right) \\
& + \sup_{x_0 \in S(M), \| \beta[M^c] \| / \sigma < \delta_{n,4}} P_{n,\beta,\sigma} \left(|e_1| \geq (1 - \delta_{n,2}) \delta_{n,1} \, \middle| \, X \right) \\
\leq \; & o_p(1) + \sup_{x_0 \in S(M), \| \beta[M^c] \| / \sigma < \delta_{n,4}} P_{n,\beta,\sigma} \left(|e_1| \geq (1 - \delta_{n,2}) \delta_{n,1} \, \middle| \, X \right). \qquad (34)
\end{aligned}
$$

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

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

where $B_n\left(X\right) \geq 0$ is $O_p\left(1\right)$, 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 (34) is bounded by

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

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 $K_1\left(x_0, r\right)\|s_M\|$ $(K_2(x_0[\hat{M}], \hat{M}, r)\|s_M\|$, $K_3(x_0[\hat{M}], \hat{M}, r)\|s_M\|$, $K_4\left(r\right)\|s_M\|$, or $K_5\left(r\right)\|s_M\|$, 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 (28).

(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 2.5 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 $K_1\left(x_0, r^*\right)$ $(K_2(x_0[\hat{M}], \hat{M}, r^*)$, $K_3(x_0[\hat{M}], \hat{M}, r^*)$, $K_4\left(r^*\right)$, or $K_5\left(r^*\right)$, respectively). Now apply Lemma B.1 with $W(x_0, X, M)$ equal to $K_1\left(x_0, r^*\right)\|s_M\|$ $(K_2(x_0[\hat{M}], \hat{M}, r^*)\|s_M\|$, $K_3(x_0[\hat{M}], \hat{M}, r^*)\|s_M\|$, $K_4\left(r^*\right)\|s_M\|$, or $K_5\left(r^*\right)\|s_M\|$, respectively) and with $\tilde{\sigma}_1 = \tilde{\sigma}$. Note that $\tilde{\sigma}_1$ satisfies (28) 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 (20).*

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

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

We now bound the probability in (20) 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

$$\sup_{\beta \in \mathbb{R}^p, \sigma > 0} P_{n,\beta,\sigma} \left(\left| \left(n - |\hat{M}| \right)^{-1/2} \|A/\sigma\| - 1 \right| \geq \delta/2 \,\middle|\, X \right)$$

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

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

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 $\Pr\left(|W/w - 1| \geq \delta \right)$ 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,

$$\sup_{\beta \in \mathbb{R}^p, \sigma > 0} P_{n,\beta,\sigma} \left(\left(n - |\hat{M}| \right)^{-1/2} \left\| X[\hat{M}^c]\beta[\hat{M}^c]/\sigma \right\| \geq \delta/2 \,\middle|\, X \right)$$

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

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

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

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 $\lambda_{\max}\left(X'X/(n-p) \right)$ 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

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

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

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

C   Appendix: Algorithms for computing the confidence intervals

In this appendix we consider the setting of Section 2. In particular, recall that X is a fixed $n \times p$ matrix of rank $d \geq 1$. Let Q be a $n \times d$ 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, cf. Section 5.1 in Berk et al. (2013a). We then have $\tilde{Y} = \tilde{\mu} + \tilde{U}$ with $\tilde{\mu} = Q'\mu$ and $\tilde{U} = Q'U \sim N\left(0, \sigma^2 I_d\right)$. It is now easy to see that $P_{\tilde{X}}\tilde{Y} = Q'P_X Y$ and $QP_{\tilde{X}}\tilde{Y} = P_X Y$ hold. In particular, the independence of the *given* $\hat{\sigma}^2$ from the projection of the data vector on the space spanned by the regressor holds whether we work with the original data or with the data in canonical coordinates. Furthermore, setting $\tilde{s}'_M = x'_0[M](\tilde{X}[M]'\tilde{X}[M])^{-1}\tilde{X}[M]'$ for $\varnothing \neq M \in \mathcal{M}$ with $\mathcal{M}$ as in Section 2 and $\tilde{s}'_M = 0 \in \mathbb{R}^d$ for $M = \varnothing$, it follows that $\|s'_M\| = \|\tilde{s}'_M\|$ and $s'_M(Y - \mu) = \tilde{s}'_M(\tilde{Y} - \tilde{\mu})$. 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 $\tilde{X}$ (but using the originally given $\hat{\sigma}^2$ in both cases). 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. In the important case $d = p(\leq n)$ the matrices Q and $\tilde{X}$ can be obtained, for example, from a SVD or a QR decomposition of X, cf. Section 5.1 in Berk et al. (2013a). [In case $p \geq n = d$, one can always set $Q = I_n$.]

The following algorithm for computing $K_1(x_0)$ 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 from the arguments used to prove (23) in Appendix A we see that, in case $x_0 \neq 0$, $K_1(x_0)$ is the solution to

$$\mathbb{E}_G \Pr\left(\max_{M \in \mathcal{M}} \left|\bar{\tilde{s}}'_M V\right| \leq t/G \,\middle|\, G\right) = 1 - \alpha,$$

where V here is uniformly distributed on the unit sphere of $\mathbb{R}^d$, independently of G (and G is as in Section 2). The algorithm now replaces the (conditional) 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 $F^\sharp_{d,r}$ the c.d.f. of G; note that then $F^\sharp_{d,r}(t) = F_{d,r}\left(t^2/d\right)$, where $F_{d,r}$ denotes the c.d.f. of an F-distribution with (d,r)-degrees of freedom.

Algorithm C.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 $\mathbb{R}^d$. Calculate the quantities $c_i = \max_{M \in \mathcal{M}} \left|\bar{\tilde{s}}'_M V_i\right|$ 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*

$$\frac{1}{I} \sum_{i=1}^{I} F^\sharp_{d,r}\left(\frac{K}{c_i}\right) = 1 - \alpha. \tag{35}$$

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

Note that for $x_0 \neq 0$ at least one of the vectors $\bar{\bar{s}}_M$, $M \in \mathcal{M}$, is non-zero, implying that the quantities c_i are all non-zero with probability 1; hence the terms $F_{d,r}^\sharp \left(\frac{K}{c_i}\right)$ 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 (35) exists, is unique and positive. The costly factor in Algorithm C.1 is the maximization involved in the computation of the quantities c_i, while searching for the value of K that solves (35), for example by bisection searches, incurs only negligible cost. In our simulations, computing $K_1(x_0)$ for $p = d = 10$ (with $\mathcal{M}$ the power set of $\{1, ..., p\}$) and $I = 10,000$ takes around one second on a personal computer; and around 10 minutes for $p = d = 20$ and $I = 1,000$. In case $\mathcal{M}$ is the power set of $\{1, ..., p\}$, the complexity of Algorithm C.1 will be exponential in p and thus will be feasible only for moderately large values of p. In relation to this we mention that Berk et al. (2013b) found their algorithm (which is similar to Algorithm C.1 as noted above) to be 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 longer running time is due to the fact that Berk et al. (2013b) have to search over $p2^{p-1}$ unit vectors, while we have to search only over 2^p unit vectors.]

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 C.1, and in case M is empty we have $K_3(x_0[M], M) = K_4$, which can be computed by Algorithm C.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 $\bar{\bar{s}}_M$ 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. $F_{d,r}^\sharp$, 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. $F_{d,r}^\sharp$. Recall that $c(M, \mathcal{M})$ has been defined subsequent to (12).

Algorithm C.2. *Suppose that $M \in \mathcal{M}$ 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 $\mathbb{R}^d$, and calculate the quantities $c_i = \max_{M_* \in \mathcal{M}, M_* \subseteq M} \left| \bar{\bar{s}}_{M_*}' V_i \right|$ with $\bar{\bar{s}}_{M_*}$ as defined above. In case $d > 1$, find m_* as the smallest value such that*

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

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

$$\left(1 - F_{Beta,1/2,(d-1)/2}(m_*^2)\right) \frac{j}{J} = \left(1 - F_{Beta,1/2,(d-1)/2}(m_j^2)\right) \tag{36}$$

46

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

$$1 - \alpha = F_{d,r}^\sharp \left(\frac{K}{m_J} \right) - \frac{1}{I} \sum_{i:c_i > m_J} \left(F_{d,r}^\sharp \left(\frac{K}{m_J} \right) - F_{d,r}^\sharp \left(\frac{K}{c_i} \right) \right)$$

$$+ c\left(M, \mathcal{M} \right) \left(1 - F_{Beta,1/2,(d-1)/2} \left(m_J^2 \right) \right) \frac{1}{J} \sum_{j=1}^{J-1} \left(F_{d,r}^\sharp \left(\frac{K}{m_j} \right) - F_{d,r}^\sharp \left(\frac{K}{m_J} \right) \right). \quad (37)$$

In case $d = 1$, $K_3(x_0[M], M)$ is the (uniquely determined and positive) constant K that solves

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

Note that m_* exists, 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 (36). But in this case the r.h.s. of (37) reduces to $F_{d,r}^\sharp (K)$ anyway and hence there is no need for solving equation (36).] Furthermore, note that the r.h.s. of (37) can be written as

$$F_{d,r}^\sharp \left(\frac{K}{m_J} \right) \left[1 - \frac{1}{I} \sum_{i=1}^{I} \mathbf{1} \left(c_i > m_J \right) - c\left(M, \mathcal{M} \right) \left(1 - F_{Beta,1/2,(d-1)/2} \left(m_J^2 \right) \right) \frac{J-1}{J} \right]$$

$$+ \frac{1}{I} \sum_{i:c_i > m_J} F_{d,r}^\sharp \left(\frac{K}{c_i} \right) + c\left(M, \mathcal{M} \right) \left(1 - F_{Beta,1/2,(d-1)/2} \left(m_J^2 \right) \right) \frac{1}{J} \sum_{j=1}^{J-1} F_{d,r}^\sharp \left(\frac{K}{m_j} \right).$$

Observing that the expression in brackets is nonnegative (in fact, positive) because of the definition of m_J, we see that the r.h.s. of (37) is strictly increasing in K. Furthermore, inspection of the r.h.s. of (37) shows that it is zero for $K = 0$ and converges to one for $K \to \infty$. Consequently, equation (37) has a unique solution for K, which necessarily is positive. We note that in Algorithm C.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 $1 - F_{Beta,1/2,(d-1)/2} \left(t^2 \right)$ 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 (37) 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 (37) 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.

47

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 C.1 and C.2.

Algorithm C.3. *In case $d > 1$, choose $J \in \mathbb{N}$, $J > 1$, and find the values $m_1, ..., m_J$ so that, for $j = 1, ..., J$,*

$$c\left(\varnothing, \mathcal{M}\right)\left(1 - F_{Beta,1/2,(d-1)/2}\left(m_j^2\right)\right) = \frac{j}{J}. \tag{38}$$

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

$$\frac{1}{J} \sum_{m_j > 0} F_{d,r}^{\sharp}\left(\frac{K}{m_j}\right) = 1 - \alpha. \tag{39}$$

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

$$F_{1,r}^{\sharp}\left(K\right) = 1 - \alpha. \tag{40}$$

Note that in case $d > 1$ the constants m_j always exist and are unique; they are all positive in case $\mathcal{M} \neq \{\varnothing, \{1, \ldots, p\}\}$ (as then $c\left(\varnothing, \mathcal{M}\right) > 1$ must hold in view of our assumptions on $\mathcal{M}$), and they are positive for $j = 1, \ldots, J - 1$ in case $\mathcal{M} = \{\varnothing, \{1, \ldots, p\}\}$. Consequently, the solution K of (39) exists, is unique and positive. In case $d = 1$ the solution of (40) 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 (39) is replaced by

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

with the convention that $m_0 = 1$.

Remark C.4. For the computation of the constants $K_1(x_0, \infty)$, $K_3(x_0[M], M, \infty)$, and $K_4\left(\infty\right)$ (cf. Remark 2.5) one can use the above algorithms with the only modification that the distribution function $F_{d,r}^{\sharp}$ is replaced by the distribution function of the square root of a chi-squared-distributed random variable with d degrees of freedom.

Remark C.5. When the collection $\mathcal{M}$ becomes large (e.g., if $\mathcal{M}$ is the power set of $\{1, ..., p\}$ in case $p = d \leq n$ and d is larger than 20), Algorithms C.1 or C.2 may not be tractable, but Algorithm C.3 can still be as it does not require the costly step of searching over the model universe $\mathcal{M}$. However, it is reported in Berk et al. (2013b) that, for about $d \geq 40$, it can be problematic to compute the extreme quantiles in (38) with standard routines. In this case, one can of course always use the Scheffé constant K_5. In practice, one may also consider in such cases (since p is large) to use rule-of-thumb constants smaller than K_5 that are based on asymptotic considerations such as Corollary 2.12: For example, if $p = d \leq n$, $\mathcal{M}$ is the power set of $\{1, ..., p\}$, but p is very large, this corollary could be read as suggesting to use the constant $K_6 = 0.866 K_5$ in (5). A similar advice is given in the framework of Berk et al. (2013b). [In case $p > n$ and $\mathcal{M}$ is as in Corollary 2.12, this corollary can be used to

provide appropriate substitutes for K_6.] However, we would like to issue a warning here: The asymptotic results for $p \to \infty$ like Corollary 2.12 and the related results in Berk et al. (2013b) and Berk et al. (2013a) are highly non-uniform w.r.t. α (cf. Remark 2.14), showing that rule-of-thumb approximations such as K_6 have to be taken with a grain of salt; see also the warning expressed at the end of Section 5.2 of Berk et al. (2013b).

References

ANDREWS, D. W. K. and GUGGENBERGER, P. (2009). Hybrid and size-corrected subsampling methods. *Econometrica,* **77** 721–762.

BELLONI, A., CHERNOZHUKOV, V. and HANSEN, C. (2011). Inference for high-dimensional sparse econometric models. *Advances in Economics and Econometrics. 10th World Congress of the Econometric Society, Volume III,* 245–295.

BELLONI, A., CHERNOZHUKOV, V. and HANSEN, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *Rev. Econom. Stud.,* **81** 608–650.

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.

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.,* **96** 1348–1360.

FITHIAN, W., SUN, D. and TAYLOR, J. (2015). Optimal inference after model selection. Preprint.

GREENSHTEIN, E. and RITOV, Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli,* **10** 971–988.

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.

LEE, J. D., SUN, D. L., SUN, Y., and TAYLOR, J. E. (2015). Exact post-selection inference, with application to the lasso. *To appear in Ann. Statist.*

LEE, J. D. and TAYLOR, J. (2014). Exact post model selection inference for marginal screening. ArXiv:1402.5596v2.

LEEB, H. (2009). Conditional predictive inference post model selection. *Ann. Statist.*, **37** 2838–2876.

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. (2015). On various confidence intervals post-model-selection. *Statist. Sci.*, **30** 216–227.

LOCKHART, R., TAYLOR, J., TIBSHIRANI, R. J. and TIBSHIRANI, R. (2014). A significance test for the LASSO. *Ann. Statist.*, **42** 413–468.

LOFTUS, J. R. (2015). Selective inference after cross-validation. ArXiv:1511.08866.

LOFTUS, J. R. and TAYLOR, J. (2015). Selective inference in regression models with groups of variables. ArXiv:1511.01478.

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

PÖTSCHER, B. M. and SCHNEIDER, U. (2010). Confidence sets based on penalized maximum likelihood estimators in Gaussian regression. *Electron. J. Statist.*, **4** 334–360.

RAWLINGS, J. O., PANTULA, S. G. and DICKEY, D. A. (1998). *Applied Regression Analysis: A Research Tool.* 2nd ed. Springer Verlag, New York, NY.

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

SCHNEIDER, U. (2015). Confidence sets based on thresholding estimators in high-dimensional Gaussian regression models. *Econometric Reviews.* Forthcoming.

TIAN, X. and TAYLOR, J. (2015). Asymptotics of selective inferene. Preprint.

TIBSHIRANI, R. J., RINALDO, A., TIBSHIRANI, R. and WASSERMAN, L. (2015). Uniform asymptotic inference and the bootstrap after model selection. Preprint.

VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. and DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.*, **42** 1166–1202.

WASSERMAN, L. (2014). Discussion: "A significance test for the LASSO". *Ann. Statist.*, **42** 501–508.

WASSERMAN, L. and ROEDER, K. (2009). High-dimensional variable selection. *Ann. Statist.*, **37** 2178–2201.

ZHANG, C. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, **38** 894–942.

ZHANG, C.-H. and ZHANG, S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. Roy. Statist. Soc. Ser. B*, **76** 217–242.

ZHANG, K. (2013). Rank-extreme association of gaussian vectors and low-rank detection. Working paper, available at `http://arxiv.org/abs/1306.0623`.