The Annals of Statistics
2019, Vol. 47, No. 3, 1475–1504
https://doi.org/10.1214/18-AOS1721

VALID CONFIDENCE INTERVALS FOR POST-MODEL-SELECTION PREDICTORS

BY FRANÇOIS BACHOC*, HANNES LEEB†,‡,[1] AND BENEDIKT M. PÖTSCHER†

*University Paul Sabatier**, *University of Vienna*† *and DataScience@Univie*‡

We consider inference post-model-selection in linear regression. In this setting, Berk et al. [*Ann. Statist.* **41** (2013a) 802–837] recently introduced a class of confidence sets, the so-called PoSI intervals, that cover a certain nonstandard 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.

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 nontrivial; see Leeb and Pötscher (2005, 2006, 2017), 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, Pötscher and Ewald (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

Received January 2016; revised May 2018.

[1]Supported in part by the Austrian Science Fund (FWF) projects P 28233-N32 and P 26354-N26.

MSC2010 subject classifications. Primary 62F25; secondary 62J05.

Key words and phrases. Inference post-model-selection, confidence intervals, optimal post-model-selection predictors, nonstandard targets, linear regression.

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 nonstandard targets; see the discussion in Remark 2.1 of Leeb, Pötscher and Ewald (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, that is, the quantity for which a confidence set is desired, is *not* the standard target, that is, the parameter in an overall model (or components thereof), but a *nonstandard* quantity of interest that depends on the selected model, and thus on the data. This nonstandard quantity of interest is denoted by $\beta_{\hat{M}}^{(n)}$ 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 nonstandard target $\beta_{\hat{M}}^{(n)}$ provides a certain vector of regression coefficients for those explanatory variables that are "active" in the model $\hat{M}$ (more precisely, $\beta_{\hat{M}}^{(n)}$ 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 equations (2.3) and (2.4) 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 $x_0'[\hat{M}]\beta_{\hat{M}}^{(n)}$. 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 $x_0'[\hat{M}]\beta_{\hat{M}}^{(n)}$ the design-dependent (nonstandard) coverage target, because different design matrices in the training data typically result in different values of $x_0'[\hat{M}]\beta_{\hat{M}}^{(n)}$ even if both training data sets lead to selection of the same model $\hat{M}$. We construct PoSI confidence intervals for $x_0'[\hat{M}]\beta_{\hat{M}}^{(n)}$ 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 $x_0'[\hat{M}]\beta_{\hat{M}}^{(n)}$ 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 $x_0'[\hat{M}]\beta^{(\star)}_{\hat{M}}$; see (3.1). We call $x_0'[\hat{M}]\beta^{(\star)}_{\hat{M}}$ the design-independent (nonstandard) 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 $x_0'[\hat{M}]\gamma(\hat{M})$, 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, that is, 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 (2.10)). In any case, this is problematic in situations when, after having selected a given model, only the "active" components of x_0 are observed, for example, 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 2.4.

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 (2016). 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 desparsifying method has recently been developed by Belloni, Chernozhukov and Hansen (2011, 2014), van de Geer et al. (2014) and Zhang and Zhang (2014). Another strand of literature that, like the PoSI approach, also focuses on $\beta_{\hat{M}}^{(n)}$ as the quantity of interest, is developed in Fithian, Sun and Taylor (2015), Lee et al. (2016), Lee and Taylor (2014), Tian and Taylor (2015) and Tibshirani et al. (2015, 2016): In these papers, confidence sets for $\beta_{\hat{M}}^{(n)}$ are considered that have a guaranteed coverage probability *conditionally* on the event that a particular model has been selected by the model selection procedure. 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 Appendix G. As prompted by a referee, we point out here that in the presence of a large number of regressors PoSI intervals (including intervals considered in the present paper) typically are computationally more burdensome than the confidence intervals proposed in Lee et al. (2016) for the LASSO with a fixed value for the tuning parameter; see, however, also the discussion towards the end of Appendix G.

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 2.4. 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. The results of a numerical study are reported in Section 4. Conclusions are drawn in Section 5. Appendix A, which like all the Appendices can be found in the Supplementary Material (Bachoc, Leeb and Pötscher (2018)), contains some comments on the assumptions made on the error variance. The proofs of the results in Sections 2 and 3 are given in Appendices B and C. Appendix D contains some comments on and extensions of the results in Section 3. In Appendix E 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. Appendix F contains details concerning the numerical calculations used for the results in Section 4. Finally, Appendix G reports simulation results regarding the comparison of the confidence intervals of Lee et al. (2016) with those proposed in the present paper. In the following, any reference to an Appendix points to the Supplementary Material (Bachoc, Leeb and Pötscher (2018)).

2. Confidence intervals for the design-dependent nonstandard target.

2.1. *The framework.* Consider the model

$$Y = \mu + U, \tag{2.1}$$

where $\mu \in \mathbb{R}^n$ is unknown and U follows an $N(0, \sigma^2 I_n)$-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. The design matrix X is treated as fixed throughout Section 2.

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 $\{1, \ldots, p\}$, where the elements of M index the columns of X that are retained. We use the following notation: For $M \subseteq \{1, \ldots, p\}$, we write M^c for the complement of M in $\{1, \ldots, p\}$. 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 $M = \{j_1, \ldots, j_m\}$ 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 dimension $l \times m$ 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 $(v'[M])'$, that is, $v[M] = (v_{j_1}, \ldots, v_{j_m})'$ 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 $\hat{\beta}_M$, that is,

$$\hat{\beta}_M = \big(X[M]'X[M]\big)^{-1} X[M]'Y, \tag{2.2}$$

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

$$\beta_M^{(n)} = \big(X[M]'X[M]\big)^{-1} X[M]'\mu. \tag{2.3}$$

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 $\mu \in \operatorname{span}(X)$ 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 Appendix A. 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 $P_{n,\mu,\sigma}$ (see also Appendix D.4).

We are furthermore given a (nonempty) collection $\mathcal{M}$ of admissible models $M \subseteq \{1, \dots, p\}$, 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 $\mathcal{M}$, that is, $\bigcup\{M : M \in \mathcal{M}\} = \{1, \dots, p\}$ 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 nonempty $M \in \mathcal{M}$ is of full-rank in the sense that $\operatorname{rank} X[M] = |M|$. We point out here that our assumptions on $\mathcal{M}$ imply that X cannot 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 $\{1, \dots, p\}$ (enlarged by the empty model) provided that no column of X is zero. Of course, there are many other examples; see, for example, 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 (possibly empty) model $\hat{M}(X, Y, \hat{\sigma}^2) \in \mathcal{M}$. 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 (2.2) with M replaced by $\hat{M}$.

The nonstandard quantity of interest studied in Berk et al. (2013a) is the random vector (with random dimension) $\beta^{(n)}_{\hat{M}}$ obtained by replacing M by $\hat{M}$ in (2.3). 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 $N(\nu, \sigma^2)$, independently of Y. If one is forced to use a fixed model M for prediction, that is, 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 $x_0'[M]\beta^{(n)}_M$. 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

$$x_0'[\hat{M}]\beta_{\hat{M}}^{(n)}. \tag{2.4}$$

The quantity in (2.4) will be our target for inference throughout Section 2 and will be called the *design-dependent (nonstandard) target* (to emphasize that it depends on the design matrix X apart from its dependence on $\hat{M}$, cf. (2.3)). 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. Throughout Section 2, we are interested in confidence intervals for the design-dependent target $x_0'[\hat{M}]\beta_{\hat{M}}^{(n)}$ that are of the form

$$\mathrm{CI}(x_0) = x_0'[\hat{M}]\hat{\beta}_{\hat{M}} \pm K(x_0, \hat{M})\|s_{\hat{M}}\|\hat{\sigma}, \tag{2.5}$$

where $\|\cdot\|$ denotes the Euclidean norm ($\hat{\sigma}$ of course representing the nonnegative square root of $\hat{\sigma}^2$), where

$$s_M' = x_0'[M]\big(X[M]'X[M]\big)^{-1}X[M]', \tag{2.6}$$

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 nonnegative 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. Finally, note that $\mathrm{CI}(x_0)$ constitutes a confidence interval for the predictor $x_0'[\hat{M}]\beta_{\hat{M}}^{(n)}$, and should not be mistaken for a prediction interval for a new response y_0.

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

$$\inf_{\mu \in \mathbb{R}^n, \sigma > 0} P_{n,\mu,\sigma}\big(x_0'[\hat{M}]\beta_{\hat{M}}^{(n)} \in \mathrm{CI}(x_0)\big) \geq 1-\alpha. \tag{2.7}$$

Note that if one replaces $K(x_0, \hat{M})$ in (2.5) by $K_{\mathrm{naive}} = q_{r,1-\alpha/2}$, then the confidence interval (2.5) 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 (2.7) as can be seen from the numerical results in Section 4, which is in line with the results in Leeb, Pötscher and Ewald (2015).

2.2. *The various confidence intervals.* 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 former case will arise if measuring all the components of x_0 is not too costly, whereas the latter case will 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, for example, out of cost considerations. For example, in a medical application one may want to avoid measuring prognostic variables that require invasive procedures or that incur high monetary costs; see, for example, Castera et al. (2015). Cost considerations in the context of model selection or prediction are also common in fields such as industrial process control or engineering (Jaupi (2014), Souders and Stenbakken (1991)).

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 (2.5) satisfies (2.7): Observe that

$$x_0'[\hat{M}]\hat{\beta}_{\hat{M}} - x_0'[\hat{M}]\beta^{(n)}_{\hat{M}} = s'_{\hat{M}}(Y - \mu), \tag{2.8}$$

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

$$|\bar{s}'_{\hat{M}}(Y - \mu)|/\hat{\sigma} \leq \max_{M \in \mathcal{M}} |\bar{s}'_M(Y - \mu)|/\hat{\sigma}. \tag{2.9}$$

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

$$P_{n,\mu,\sigma}\Big(\max_{M \in \mathcal{M}} |\bar{s}'_M(Y - \mu)|/\hat{\sigma} \leq K_1(x_0)\Big) \geq 1 - \alpha. \tag{2.10}$$

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 $\bar{s}'_M(Y - \mu) = \bar{s}'_M P_X(Y - \mu)$, since $\bar{s}_M$ belongs to the column space of X. Consequently, the collection of all the quantities $\bar{s}'_M(Y - \mu)$ is jointly distributed as $N(0, \sigma^2 C)$, independently of $\hat{\sigma}^2 \sim (\sigma^2/r)\chi^2(r)$, where the covariance matrix C depends only on x_0 and X. Hence the joint distribution of the collection of ratios $|\bar{s}'_M(Y - \mu)|/\hat{\sigma}$ 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 (2.10), and $K_1(x_0)$ is the unique $(1 - \alpha)$-quantile of the distribution of the upper bound in (2.9). (This follows from Lemma B.1 in Appendix B 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 (2.10) 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, $K_{\mathrm{naive}} \le K_1(x_0)$ clearly holds provided $x_0 \neq 0$ (since $\bar{s}_M'(Y - \mu)/\hat{\sigma}$ follows Student's t-distribution with r degrees of freedom if $s_M \neq 0$).

As a consequence of (2.9) 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* (2.10). *Then the confidence interval* (2.5) *with $K(x_0, \hat{M})$ replaced by $K_1(x_0)$ satisfies the coverage property* (2.7).

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 $\beta_{\hat{M}}^{(n)}$. (In fact, the construction of $K_1(x_0)$ implies that the collection of intervals $x_0'[M]\hat{\beta}_M \pm K_1(x_0)\|s_M\|\hat{\sigma}$ with $M \in \mathcal{M}$ provides a simultaneous confidence band for $x_0'[M]\beta_M^{(n)}$.)

Consider next the 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 (2.10) and our assumptions on $\mathcal{M}$. A first solution is to define

$$K_2\bigl(x_0[M], M\bigr) = \sup\bigl\{K_1(x) : x[M] = x_0[M]\bigr\}, \tag{2.11}$$

and then to use the confidence interval (2.5) 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* (2.11). *Then the confidence interval* (2.5) *with $K(x_0, \hat{M})$ replaced by $K_2(x_0[\hat{M}], \hat{M})$ satisfies the coverage property* (2.7).

The computation of $K_2(x_0[\hat{M}], \hat{M})$ is more costly than that of $K_1(x_0)$. Indeed, it requires to embed the algorithm for computing $K_1(x_0)$ in an optimization procedure. Thus, for the cases where the resulting computational cost is prohibitive, we present in the subsequent proposition larger constants $K_3(x_0[\hat{M}], \hat{M})$, K_4 and K_5 that are simpler to compute. Algorithms for computing these constants are discussed in Appendix E. The constant K_4 is obtained by applying a union bound to (2.10), whereas K_3 is obtained by applying a more refined "partial" union bound. (More precisely, for $M \in \mathcal{M}$ the complement of the probability in (2.10) (with $K_1(x_0)$ replaced by a generic variable t) can be expressed as in (B.2) in Appendix B. For given $M \in \mathcal{M}$, and after conditioning on the variance estimator (represented by G there), we apply a union bound by decomposing the maximum over $\mathcal{M}$ into a maximum over the submodels of the given M and a maximum over the models not nested in M. A further union bound is applied to the latter group of models, giving rise to the bound (B.3) in Appendix B. Inspection of this bound shows that the probability appearing in (2.12) below springs from the submodels of M, whereas the models not nested in M give rise to the term in (2.12) involving the *Beta*-distribution function.)

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

$$F^*_{M,x_0}(t) = 1 - \min\Big[1, \Pr\Big(\max_{M_* \in \mathcal{M}, M_* \subseteq M} |\bar{s}'_{M_*} V| > t\Big) + c(M, \mathcal{M})\big(1 - F_{\mathrm{Beta},1/2,(d-1)/2}(t^2)\big)\Big] \tag{2.12}$$

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 $M_* \not\subseteq M$, V is a random vector that is uniformly distributed on the unit sphere in the column space of X, and $F_{\mathrm{Beta},1/2,(d-1)/2}$ denotes the Beta$(1/2, (d-1)/2)$-distribution function, with the convention that in case $d = 1$ we use $F_{\mathrm{Beta},1/2,0}$ 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 $M = \{1, \ldots, p\}$ (and when this set belongs to $\mathcal{M}$). Next, define the distribution function F_{M,x_0} via

$$F_{M,x_0}(t) = \mathbb{E}_G F^*_{M,x_0}(t/G), \tag{2.13}$$

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 $K_3(x_0[M], M)$ we define next: For any $x_0 \in \mathbb{R}^p$ and any $M \in \mathcal{M}$, define $K_3(x_0[M], M)$ to be the smallest constant K satisfying

$$F_{M,x_0}(K) \geq 1 - \alpha. \tag{2.14}$$

Furthermore, set $K_4 = K_3(x_0[\varnothing], \varnothing)$. Finally, K_5 is the Scheffé constant, that is, the $(1-\alpha)$-quantile of G (Scheffé (1959)); see the corresponding discussion in Section 4.8 of Berk et al. (2013a). Recall that $1-\alpha \in (0,1)$ has been assumed.

PROPOSITION 2.3. *Let $x_0 \in \mathbb{R}^p$ be arbitrary. Then we have the following*:

(a) *$K_3(x_0[M], M)$ exists and is well defined. If $M = \{1, \ldots, p\} \in \mathcal{M}$ and $x_0 = 0$, then $K_3(x_0[M], M) = 0$ (and F_{M,x_0} is the c.d.f. of pointmass at zero). If $M \neq \{1, \ldots, p\}$ or $x_0 \neq 0$, then* (i) *$0 < K_3(x_0[M], M) < \infty$ holds, and* (ii) *equality holds in (2.14) if and only if $K = K_3(x_0[M], M)$.*

(b) *For every $M \in \mathcal{M}$, we have*

$$K_2\big(x_0[M], M\big) \le K_3\big(x_0[M], M\big) \le K_4 \le K_5. \tag{2.15}$$

Furthermore,

$$K_2\big(x_0[M_2], M_2\big) \le K_2\big(x_0[M_1], M_1\big), \tag{2.16}$$

$$K_3\big(x_0[M_2], M_2\big) \le K_3\big(x_0[M_1], M_1\big) \tag{2.17}$$

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 $K_1(x_0) = K_3(x_0[M_{\text{full}}], M_{\text{full}}) = K_3(x_0, M_{\text{full}})$, provided $M_{\text{full}} := \{1, \ldots, p\}$ 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 (2.12) equals $1 - F_{\text{Beta},1/2,(d-1)/2}(t^2)$ 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. 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 (2.16) and (2.17) 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 $\|s_{M_1}\| \le \|s_{M_2}\|$ for $M_1 \subseteq M_2$; cf. Figure 1 in Section 4.

Proposition 2.3 implies that (2.15) holds with $\hat{M}$ replacing M, which 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) discussed at the beginning of Section 2.2.

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* (2.5) *with $K(x_0, \hat{M})$ replaced by $K_3(x_0[\hat{M}], \hat{M})$ (K_4, or K_5, resp.) satisfies the coverage property* (2.7).

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 derived in Section 2 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 results in Section 2 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 nonempty $M \in \mathcal{M}$ is of full-rank. This assumption could be dropped, but this would lead to more unwieldy results.

(iii) Since the development in Section 2 is based on the bound (2.9), it is obvious that all results in Section 2 also hold if $\hat{M} = \hat{M}(X, Y, \bar{\sigma}^2)$ for some arbitrary estimator $\bar{\sigma}^2$, that may differ from the estimator $\hat{\sigma}^2$ that governs the length of the confidence intervals considered.

2.3. *On the merits of the nonstandard targets.*

REMARK 2.7. (i) As already noted, the (nonstandard) coverage target in Berk et al. (2013a) is $\beta^{(n)}_{\hat{M}}$ (where these authors choose to represent it in what they call "full model indexing"). While $\beta^{(n)}_{\hat{M}}$ has a clear technical meaning as the coefficient vector that provides the best approximation of μ by elements of the form $X[\hat{M}]\gamma$ w.r.t. the Euclidean distance, adopting this quantity as the target for inference confronts one 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, for example, 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, in our opinion, to be somewhat more amenable to interpretation 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 and which all have one and the same dimension, not depending on the data).

(ii) In the classical case, that is, 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 $y_0 \sim N(x_0'\beta, \sigma^2)$ 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 $\hat{M}$ is only a function of X and Y). (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: Similarly 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 less to justify the nonstandard 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 nonstandard target $\beta_{\hat{M}}^{(n)}$ of Berk et al. (2013a) mainly has a justification in a nonclassical 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, Pötscher and Ewald (2015) as well as Appendix A. 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 nonclassical 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 (On the 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 $y_0|v, x_0 \sim N(v, \sigma^2)$ and (v, x_0') is drawn from the empirical distribution of (μ_i, x_i') where x_i' denotes the ith row of X ("in-sample prediction"). (More generally, this is so if (v, 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 $\hat{M}$.

2.4. *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.

2.4.1. *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, for example, $\mathcal{M}$ contains all submodels) and then 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. Furthermore, ϕ and Φ denote the p.d.f. and c.d.f. of a standard normal variable, respectively, and $\|x\|_0$ denotes the l_0-norm.

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 $\{1, \ldots, p\}$. Let α, $0 < \alpha < 1$, be given, not depending on p. Set $\xi = \sup_{b>0} \phi(b)/\sqrt{1 - \Phi(b)} \approx 0.6363$.*

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

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

Furthermore, for any sequence X *as above one can find another sequence of* (*nonzero*) $p \times 1$ *vectors* x_0 *such that* $K_1(x_0, \infty) = O(1)$ (*e.g., any sequence of* (*nonzero*) $p \times 1$ *vectors* x_0 *satisfying* $\sup_p \|x_0\|_0 < \infty$ *will do*).

(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*

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

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

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).

2.4.2. *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 (2.12).

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* (*nonempty*) *model universe* $\mathcal{M} = \mathcal{M}_p$ *is given that satisfies* (i) $\bigcup\{M : M \in \mathcal{M}\} = \{1, \ldots, p\}$, (ii) $\varnothing \in \mathcal{M}$, (iii) $c(M, \mathcal{M}) \geq \tau|\mathcal{M}|$ *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

$$\lim_{p\to\infty} \sup_{M\in\mathcal{M}, M\neq\{1,\dots,p\}} \sup_{x_0\in\mathbb{R}^p} \sup_{X\in\mathsf{X}_{n(p),p}(\mathcal{M})} \big|1 - \big(K_3\big(x_0[M], M, \infty\big) / K_4(\infty)\big)\big| = 0, \tag{2.18}$$

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,

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

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 nonempty 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 Corollary B.7 in Appendix B to be always satisfied in the important case where $\mathcal{M}$ is of the form $\{M \subseteq \{1, \dots, p\} : |M| \leq m_p\}$. 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 in that corollary.

In the important case, where $p = d \leq n$ and $\mathcal{M}$ is the entire power set of $\{1, \dots, p\}$, Corollary B.7 shows that $K_4(\infty)$ (and hence a fortiori all the constants $K_1(x_0, \infty), \dots, K_3(x_0[M], M, \infty)$) 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 2.12. 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 B.4 in Appendix B 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 and Corollary B.7 entail that $K_4(\infty) \sim \sqrt{p}\sqrt{3}/2 \approx 0.866\sqrt{p}$ while $K_1(x_0, \infty) \geq \xi\sqrt{p}$ with $\xi \approx 0.6363$ is possible; for example, 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 (2.10) (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 2.13. The results for $p \to \infty$ in this subsection (and Corollary B.7 in Appendix B) as well as the related results in Berk et al. (2013a) should be taken with a grain of salt as they obviously are highly nonuniform 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 nonstandard target. In this section, we again consider the model (2.1), 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 $N(0, \sigma^2 I_n)$-distribution with $0 < \sigma < \infty$. We also assume that X has full column rank almost surely (implying $p \le n$) and that each row of X is distributed according to a common p-dimensional distribution $\mathcal{L}$ (not depending on n) 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 ($1 \le r < \infty$). 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 3.8 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 $P_{n,\beta,\sigma}$ (see also Appendix D.4).

In this section, we shall consider asymptotic results for $n \to \infty$ but where p is held constant (for an extension to the case where p is allowed to diverge with n see Appendix D.3). 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.) However, recall that $\mathcal{L}$, and hence Σ, are not allowed to depend on n.

If M_1 and M_2 are subsets of $\{1,\dots,p\}$ 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 $u_0 \sim N(0, \sigma^2)$. For a given model $M \subseteq \{1,\dots,p\}$, we consider the (infeasible) predictor $x_0'[M]\beta_M^{(\star)}$ where

$$\beta_M^{(\star)} = \beta[M] + \big(\Sigma[M, M]\big)^{-1}\Sigma\big[M, M^c\big]\beta\big[M^c\big], \tag{3.1}$$

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 $M = \{1,\dots,p\}$. 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 (2.3) can be rewritten as $\beta_M^{(n)} = \beta[M] + (X[M]'X[M])^{-1}X[M]'X[M^c]\beta[M^c]$. Given a model selection procedure $\hat{M} = \hat{M}(X, Y, \hat{\sigma}^2)$, 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. 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 $\hat{M}$ is only a function of X and Y). 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 (On the 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 D.2 in Appendix D.2 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 $\delta > 0$,

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

in probability as $n \to \infty$.

Condition 3.3 is very mild and typically holds for model selection procedures such as AIC- and BIC-based procedures as well as Lasso-type procedures. (This can be established along the lines of the proof of Corollary 5.4(a) in Leeb and Pötscher (2003).) 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 $\hat{\sigma}^2_{\mathrm{OLS}}$ 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* $\mathrm{CI}(x_0)$ *be the confidence interval* (2.5) *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* $\mathrm{CI}(x_0)$ *satisfies*

$$\inf_{x_0 \in \mathbb{R}^p, \beta \in \mathbb{R}^p, \sigma > 0} P_{n,\beta,\sigma}\big(x_0'[\hat{M}]\beta^{(\star)}_{\hat{M}} \in \mathrm{CI}(x_0) | X\big) \ge (1 - \alpha) + o_p(1), \tag{3.2}$$

where the $o_p(1)$ *term above depends only on* X *and converges to zero in probability as* $n \to \infty$. *Relation* (3.2) *a fortiori holds if the confidence interval* $\mathrm{CI}(x_0)$ *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*

$$\sup_{\beta \in \mathbb{R}^p, \sigma > 0} P_{n,\beta,\sigma}\big(|\tilde{\sigma}/\sigma - 1| \ge \delta | X\big) \xrightarrow{p} 0 \tag{3.3}$$

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* $\mathrm{CI}^*(x_0)$ *denote the modified confidence interval which is obtained 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*) *in* (2.5) (*while keeping* $\hat{M}$ *unchanged*). *Then relation* (3.2) *holds with* $\mathrm{CI}(x_0)$ *replaced by* $\mathrm{CI}^*(x_0)$.

Theorem 3.6(a) shows that for any $x_0 \in \mathbb{R}^p$ the interval $\mathrm{CI}(x_0)$ 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 $K_5(r)$, resp.) 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 $K_5(\infty)$, resp.). Measurability issues regarding Theorem 3.6 are discussed in Appendix D.1.

Condition (3.3) is a uniform consistency property. It is clearly satisfied by $\hat{\sigma}^2_{\mathrm{OLS}}$ (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 C.2 in Appendix C 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. Furthermore, a remark similar to Remark 2.6(iii) also applies here.

REMARK 3.8 (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).

4. Numerical study. We next 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 Appendix G we compare the intervals introduced in Section 2 with those proposed recently in Lee et al. (2016), which are specific to the LASSO model selector. Code for the computations is available from the first author.

4.1. *Lengths of confidence intervals.* We consider the lengths of the confidence intervals obtained from (2.5) standardized by $\hat{\sigma}$, that is, 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 $K_{\text{naive}} = q_{r,1-\alpha/2}$, 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 $K_{\text{naive}} \le K_1(x_0) \le \cdots \le K_5$ (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, Pantula and Dickey (1998) concerning the peak flow rate of watersheds. This data set contains a 30×10 design matrix X_{Raw} corresponding to ten explanatory variables. For a description of these variables, see Appendix F. This data set is also studied in Kabaila and Leeb (2006) and Leeb, Pötscher and Ewald (2015). We refer to it as

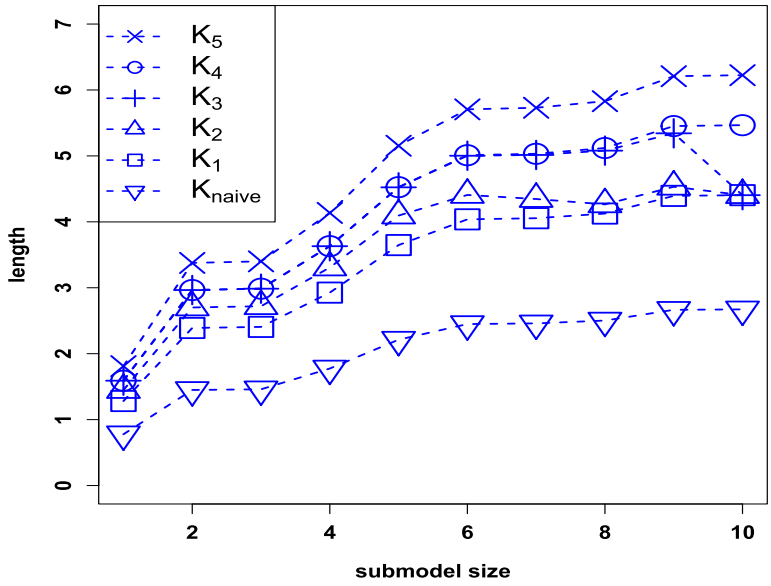

FIG. 1. *Standardized lengths of various confidence intervals as function of model size. Dashed lines are added to improve readability.*

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 $\{1, \ldots, p\}$.

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, we compute the standardized lengths of the confidence intervals only for M belonging to the family $\{\{1\}, \ldots, \{1, \ldots, 10\}\}$ 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 $\{1, \ldots, p\}$.) 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 E. However, computing $K_2(x_0[M], M)$ for $M \neq \{1, \ldots, 10\}$ necessitates to compute $\sup\{K_1(x) : x[M] = x_0[M]\}$. We approximate this supremum by using a three-step Monte Carlo procedure described in Appendix F.

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 (e.g., 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 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. (2.18). 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 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 $d = p < n$, that is, 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 $x_0'[\hat{M}]\beta^{(n)}_{\hat{M}}$ on the one hand as well as for the target $x_0'[\hat{M}]\beta^{(\star)}_{\hat{M}}$ on the other hand. The constants K_1, K_3 and K_4

are computed based on $\mathcal{M}$ equal to the power set of $\{1, \ldots, p\}$. 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)). Tuning parameters of the latter three procedures are chosen by cross-validation. For all 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, that is, we do not use a restricted universe of models but use $\mathcal{M}$ equal to the power set of $\{1, \ldots, p\}$. (Additional simulations with no intercept term and no protected explanatory variable lead to results very similar to the ones given in Table 1 below.) Computational details regarding these procedures can be found in Appendix F.

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 set $\tilde{\Sigma} = I_{\tilde{p}} + (2a + \tilde{p}a^2)E_{\tilde{p}}$ with $\tilde{p} = 9$, $a = 10$, and with $E_{\tilde{p}}$ the $\tilde{p} \times \tilde{p}$ matrix which has all entries equal to 1. For the third case, $\tilde{\Sigma}$ coincides with the identity matrix $I_{\tilde{p}}$, except that the zero elements in the last row and column of $I_{\tilde{p}}$ are replaced by the constant $c = \sqrt{0.8/(\tilde{p} - 1)}$ where $\tilde{p} = 9$. Similarly as in Berk et al. (2013a) and Leeb, Pötscher and Ewald (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); see Appendix F for more details. 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, Pötscher and Ewald (2015), which is described in detail in Appendix F. 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 C.1 in Appendix C, which entails 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 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.

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 $p = 10$

			Target							
			Design-dependent $x_0[\hat{M}]'\beta^{(n)}_{\hat{M}}$				**Design-independent** $x_0[\hat{M}]'\beta^{(\star)}_{\hat{M}}$			
Data set	n	**Model selector**	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

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 sig-

nificantly 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 where tuning parameters are chosen 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 (nonexhaustive) simulations we have repeated the above analysis for other values of X and x_0 and have found similar results.

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, that is, $x_0'[\hat{M}]\beta^{(n)}_{\hat{M}}$ and $x_0'[\hat{M}]\beta^{(\star)}_{\hat{M}}$, minimize a certain in-sample prediction error and, under additional assumptions relating the training period to the prediction period, a certain out-of-sample prediction error, respectively. For in-sample prediction, that is, for the target $x_0'[\hat{M}]\beta^{(n)}_{\hat{M}}$, our intervals are valid, in finite samples, irrespective of the model selection procedure that is being used. For out-of-sample prediction, that is, for the target $x_0'[\hat{M}]\beta^{(\star)}_{\hat{M}}$, the same is true asymptotically under very mild assumptions on the underlying model selector.

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 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, that is, 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 in Appendix E. 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 E.5 in Appendix E). Our algorithms are of similar computational complexity as those proposed in Berk et al. (2013a).

We furthermore have studied the behavior of the constants K_i 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, Pötscher and Ewald (2015), where inter alia "naive" confidence intervals for components of $\beta^{(n)}_{\hat{M}}$ were studied). Furthermore, studying in Appendix G the confidence intervals developed for model selection with the LASSO by Lee et al. (2016), 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.

Acknowledgements. We thank the referees and an Associate Editor for thoughtful feedback and constructive comments. The first author acknowledges constructive discussions with Lukas Steinberger and Nina Senitschnig.

SUPPLEMENTARY MATERIAL

Appendix: Proofs, algorithms, comments, details and extensions (DOI: 10.1214/18-AOS1721SUPP; .pdf). The Appendix contains the following material:

comments on the assumptions made on the error variance; proofs of the results given in Sections 2 and 3; additional material for Sections 2 and 3; descriptions of the algorithms for computing the PoSI confidence intervals; details concerning the numerical calculations for Section 4; additional simulation results.

REFERENCES

ANDREWS, D. W. K. and GUGGENBERGER, P. (2009). Hybrid and size-corrected subsampling methods. *Econometrica* **77** 721–762. MR2531360

BACHOC, F., LEEB, H. and PÖTSCHER, B. M. (2019). Supplement to "Valid confidence intervals for post-model-selection predictors." DOI:10.1214/18-AOS1721SUPP.

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

BELLONI, A., CHERNOZHUKOV, V. and HANSEN, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *Rev. Econ. Stud.* **81** 608–650. MR3207983

BERK, R., BROWN, L., BUJA, A., ZHANG, K. and ZHAO, L. (2013a). Valid post-selection inference. *Ann. Statist.* **41** 802–837. MR3099122

BERK, R., BROWN, L., BUJA, A., ZHANG, K. and ZHAO, L. (2013b). Valid post-selection inference. Unpublished version. Available at http://www-stat.wharton.upenn.edu/~lzhao/papers/MyPublication/24PoSI-submit.pdf.

CASTERA, L., CHAN, H. L. Y., ARRESE, M., AFDHAL, N., BEDOSSA, P., FRIEDRICH-RUST, M., HAN, K.-H. and PINZANI, M. (2015). EASL-ALEH clinical practice guidelines: Non-invasive tests for evaluation of liver disease severity and prognosis. *J. Hepatol.* **63** 237–264.

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

FITHIAN, W., SUN, D. and TAYLOR, J. (2015). Optimal inference after model selection. Available at arXiv:1410.2597.

GREENSHTEIN, E. and RITOV, Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli* **10** 971–988. MR2108039

JAUPI, L. (2014). Variable selection methods for multivariate process monitoring. In *Proceedings of the World Congress of Engineering* 2014, *Vol. II* (S. I. Ao, L. Gelman, D. Hukins, A. Hunter and A. M. Korsunsky, eds.) 1116–1120.

KABAILA, P. and LEEB, H. (2006). On the large-sample minimal coverage probability of confidence intervals after model selection. *J. Amer. Statist. Assoc.* **101** 619–629. MR2256178

LEE, J. D. and TAYLOR, J. (2014). Exact post model selection inference for marginal screening. In *Advances in Neural Information Processing Systems* 27 (Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence and K. Q. Weinberger, eds.) 136–144. Curran Associates, Red Hook, NY.

LEE, J. D., SUN, D. L., SUN, Y. and TAYLOR, J. E. (2016). Exact post-selection inference, with application to the lasso. *Ann. Statist.* **44** 907–927. MR3485948

LEEB, H. (2009). Conditional predictive inference post model selection. *Ann. Statist.* **37** 2838–2876. MR2541449

LEEB, H. and PÖTSCHER, B. M. (2003). The finite-sample distribution of post-model-selection estimators and uniform versus nonuniform approximations. *Econometric Theory* **19** 100–142. MR1965844

LEEB, H. and PÖTSCHER, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory* **21** 21–59. MR2153856

LEEB, H. and PÖTSCHER, B. M. (2006). Can one estimate the conditional distribution of post-model-selection estimators? *Ann. Statist.* **34** 2554–2591. MR2291510

LEEB, H. and PÖTSCHER, B. M. (2017). Testing in the presence of nuisance parameters: Some comments on tests post-model-selection and random critical values. In *Big and Complex Data Analysis* (S. E. Ahmed, ed.) 69–82. Springer, Cham. MR3644121

LEEB, H., PÖTSCHER, B. M. and EWALD, K. (2015). On various confidence intervals post-model-selection. *Statist. Sci.* **30** 216–227. MR3353104

LOCKHART, R., TAYLOR, J., TIBSHIRANI, R. J. and TIBSHIRANI, R. (2014). A significance test for the lasso. *Ann. Statist.* **42** 413–468. MR3210970

PÖTSCHER, B. M. (2009). Confidence sets based on sparse estimators are necessarily large. *Sankhyā* **71** 1–18. MR2579644

PÖTSCHER, B. M. and SCHNEIDER, U. (2010). Confidence sets based on penalized maximum likelihood estimators in Gaussian regression. *Electron. J. Stat.* **4** 334–360. MR2645488

RAWLINGS, J. O., PANTULA, S. G. and DICKEY, D. A. (1998). *Applied Regression Analysis*: *A Research Tool*, 2nd ed. Springer, New York. MR1631919

SCHEFFÉ, H. (1959). *The Analysis of Variance*. Wiley, New York. MR0116429

SCHNEIDER, U. (2016). Confidence sets based on thresholding estimators in high-dimensional Gaussian regression models. *Econometric Rev.* **35** 1412–1455. MR3511026

SOUDERS, T. M. and STENBAKKEN, G. N. (1991). Cutting the high cost of testing. *IEEE Spectrum* **28** 48–51.

TIAN, X. and TAYLOR, J. (2015). Asymptotics of selective inferene. Available at arXiv:1501.03588.

TIBSHIRANI, R. J., RINALDO, A., TIBSHIRANI, R. and WASSERMAN, L. (2015). Uniform asymptotic inference and the bootstrap after model selection. Available at arXiv:1506.06266.

TIBSHIRANI, R. J., TAYLOR, J., LOCKHART, R. and TIBSHIRANI, R. (2016). Exact post-selection inference for sequential regression procedures. *J. Amer. Statist. Assoc.* **111** 600–620. MR3538689

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. MR3224285

WASSERMAN, L. (2014). Discussion: "A significance test for the lasso" [MR3210970]. *Ann. Statist.* **42** 501–508. MR3210975

WASSERMAN, L. and ROEDER, K. (2009). High-dimensional variable selection. *Ann. Statist.* **37** 2178–2201. MR2543689

ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. MR2604701

ZHANG, K. (2013). Rank-extreme association of Gaussian vectors and low-rank detection. Available at arXiv:1306.0623.

ZHANG, C.-H. and ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 217–242. MR3153940

F. BACHOC
DEPARTMENT OF MATHEMATICS
UNIVERSITY PAUL SABATIER
TOULOUSE 31062
FRANCE
E-MAIL: Francois.Bachoc@math.univ-toulouse.fr

H. LEEB
B. M. PÖTSCHER
DEPARTMENT OF STATISTICS
UNIVERSITY OF VIENNA
VIENNA A-1090
AUSTRIA
E-MAIL: Hannes.Leeb@univie.ac.at
Benedikt.Poetscher@univie.ac.at