The Annals of Statistics
2020, Vol. 48, No. 1, 440–463
https://doi.org/10.1214/19-AOS1815

UNIFORMLY VALID CONFIDENCE INTERVALS POST-MODEL-SELECTION

BY FRANÇOIS BACHOC[1], DAVID PREINERSTORFER[2] AND LUKAS STEINBERGER[3]

[1]*Institut de Mathématiques de Toulouse, Université Paul Sabatier, francois.bachoc@math.univ-toulouse.fr*
[2]*ECARES, Université libre de Bruxelles, david.preinerstorfer@ulb.ac.be*
[3]*Department of Mathematical Stochastics, University of Freiburg, lukas.steinberger@stochastik.uni-freiburg.de*

We suggest general methods to construct asymptotically uniformly valid confidence intervals post-model-selection. The constructions are based on principles recently proposed by Berk et al. (*Ann. Statist.* **41** (2013) 802–837). In particular, the candidate models used can be misspecified, the target of inference is model-specific, and coverage is guaranteed for *any* data-driven model selection procedure. After developing a general theory, we apply our methods to practically important situations where the candidate set of models, from which a working model is selected, consists of fixed design homoskedastic or heteroskedastic linear models, or of binary regression models with general link functions. In an extensive simulation study, we find that the proposed confidence intervals perform remarkably well, even when compared to existing methods that are tailored only for specific model selection procedures.

1. Introduction. Fitting a statistical model to data is often preceded by a model selection step, and practically always has to face the possibility that the candidate set of models from which a model is selected does not contain the true distribution. The construction of valid statistical procedures in such situations is quite challenging, even if the candidate set of models does contain the true distribution (cf. Leeb and Pötscher (2005, 2006, 2008), Kabaila and Leeb (2006) and Pötscher (2009), and the references given in that literature), and has recently attained a considerable amount of attention. In a Gaussian homoskedastic location model and fitting possibly misspecified linear candidate models to data, Berk et al. (2013) have shown how one can obtain valid confidence intervals post-model-selection for (nonstandard) model-dependent targets of inference in finite samples (cf. also the discussion in Leeb, Pötscher and Ewald (2015), and related results obtained for prediction post-model-selection in Bachoc, Leeb and Pötscher (2019)). In this setup, their approach leads to valid confidence intervals post-model-selection *regardless* of the specific model selection procedure applied. This aspect is of fundamental importance, because many model selection procedures used in practice are almost impossible to formalize: researchers typically use combinations of visual inspection and numerical algorithms, and sometimes they simply select models that let them reject many hypotheses, that is, they are hunting for significance. These often unreported and informal practices of model selection prior to conducting the actual analysis may also play a key role in the current crisis of reproducibility. Thus, to establish and popularize statistical methods that are in some sense robust to 'bad practice' is highly desirable.

The methods discussed in Berk et al. (2013) are based on the assumption that the true distribution is Gaussian and homoskedastic. Furthermore, they only consider situations where linear models are fit to data. It is of substantial interest to generalize this approach, and to obtain generic methods for constructing confidence intervals post-model-selection that are widely applicable beyond the Gaussian homoskedastic model considered in Berk et al. (2013). In the

Received September 2017; revised November 2018.
MSC2010 subject classifications. Primary 62F12, 62F25; secondary 62F35, 62J02.
Key words and phrases. Inference post-model-selection, uniform asymptotic inference, regression.

present article, we develop a general asymptotic theory for the construction of uniformly valid confidence sets post-model-selection. These results are applicable whenever the estimation error can be expanded as the sum of independent centered random vectors and a remainder term that is negligible relative to the variance of the leading term. Such a representation typically follows from standard first order linearization arguments, and can therefore be obtained in many situations.

Our confidence intervals can either be based on consistent estimators of the variance of the previously mentioned sum, or, more importantly, if such estimators are not available (which is usually the case when all working models are misspecified), can be based on variance estimators that consistently *overestimate* their targets. We also present results that allow one to obtain such estimators in general and demonstrate their construction in specific applications, where they often coincide with well-known sandwich-type estimators. This overcomes another limitation present in Berk et al. (2013), namely the assumption that there exists an unbiased (and chi-square distributed) or uniformly consistent estimator of the variance of the observations; cf. the discussion in Remark 2.1 of Leeb, Pötscher and Ewald (2015) and in Appendix A of Bachoc, Leeb and Pötscher (2019). Our usage of variance estimators that overestimate their targets, while leading to more conservative inference, renders the approach applicable to the fully misspecified setting. Moreover, the suggested conservative estimators usually have the property that their bias vanishes if the selected model is correct (cf. Remark 2.7 and Section A.2.1 in the Supplementary Material (Bachoc, Preinerstorfer and Steinberger (2019))).

Another important aspect of the results obtained is that they are valid uniformly over wide classes of potential underlying distributions, which is particularly important as this guarantees that the results provide a better description of finite sample properties than "pointwise" asymptotic results (cf. Leeb and Pötscher (2003, 2005) and Tibshirani et al. (2018) for a discussion of related issues in a model selection context).

We apply our general theory to three important modeling situations: First, we consider the case where linear homoskedastic models are fit to non-Gaussian homoskedastic data. This provides an extension of the results of Berk et al. (2013) to the non-Gaussian case, without requiring a consistent variance estimator. Next, we study the problem of fitting heteroskedastic linear models to non-Gaussian heteroskedastic data. This scenario necessitates a more careful choice of variance estimators and leads to an extension of the influential results of Eicker (1967) to the misspecified post-model-selection context. Our third application then considers the problem of fitting binary regression models to binary data. In this case, also the link function may be chosen in a data driven way. On a technical level, the third example is quite different from the previous ones, because here nontrivial existence and uniqueness questions concerning the targets of inference and the (quasi-) maximum likelihood estimators have to be addressed.

Our confidence intervals obtained in these specific situations are particularly convenient for practitioners, because they are structurally very similar to the confidence sets one would use in practice following the naive and invalid (see, e.g., Leeb, Pötscher and Ewald (2015), Bachoc, Leeb and Pötscher (2019)) approach that ignores that the model has been selected using the same data set. The main difference of our construction to the naive approach is the choice of a critical value: Quantiles from a standard normal or t-distribution are replaced by so-called POSI-constants (cf. Berk et al. (2013) and Section 2.5 below). Thus, the procedures are conceptually simple and easy to implement. Moreover, we provide mild and easily verifiable regularity conditions on observable quantities (e.g., the design or the link functions) under minimal restrictions on the unknown data generating process.

Finally, in a series of numerical examples, we illustrate that the proposed confidence intervals are valid also in small samples and in high dimension while their lengths appear to

be practically reasonable when compared to naive (and invalid) procedures. Furthermore, we compare our methods to those of Tibshirani et al. (2018) and Taylor and Tibshirani (2018), and find that our intervals are often shorter than their competitors, even when we study the exact same scenarios for which those competing methods were tailored for and even though our confidence intervals offer much stronger theoretical guarantees.

The structure of the present article is as follows: We first develop a general asymptotic theory for the construction of uniformly valid confidence sets post-model-selection in Section 2. In Section 3, we apply our theoretical results to the three previously mentioned modeling scenarios. Of course, the selection of examples in Section 3 is by no means exhaustive. But besides covering three very important modeling frameworks, Section 3 serves as an illustration of how the general theory developed in Section 2 can be applied. An outline of the numerical results is presented in Section 4. Additional results, remarks and discussion, details of the simulations, as well as all the proofs are collected in Sections A, B, C, D and E of the Supplementary Material (Bachoc, Preinerstorfer and Steinberger (2019)).

1.1. *Informal illustration of the proposed methodology.* To provide the reader with a concrete application of the general theory to be laid out in Section 2, we shall now *informally* preview one of our applications of Section 3, that is, valid confidence intervals post-model-selection in the case of fitting binary regression models to binary observations: Suppose we observe the realization of a random n-vector Y with independent components $Y_1, \dots, Y_n$ taking only the values 0 or 1, and let X denote a fixed (nonrandom) $n \times p$ design matrix. We hardly ever *know* whether the data follow any binary regression model. Oftentimes, however, one has reason to *believe* that only a small number of the p columns from X, together with an appropriate choice of the response function h from some finite set $\mathcal{H}$, could provide a "reasonable" description of the unknown mechanism generating the data. In such a situation, the statistician would use some "model selection procedure" to select a subset $\hat{M} \subseteq \{1, \dots, p\}$ of columns (regressors) from X and a response function $\hat{h}$ from $\mathcal{H}$. Write $\hat{\mathbb{M}} = (\hat{M}, \hat{h})$. Any such selection eventually leads to a data-dependent approximating model

$$\mathbb{P}(Y_i = 1) \approx \hat{h}\big(X_i[\hat{M}]\beta_{\hat{\mathbb{M}}}\big), \quad i = 1, \dots, n, \tag{1.1}$$

where $X_i[\hat{M}]$ is the ith row of the matrix $X[\hat{M}]$ resulting from X by keeping only those columns whose indices are in $\hat{M}$, and $\beta_{\hat{\mathbb{M}}}$ is that value in $\mathbb{R}^{|\hat{M}|}$ which realizes the "best approximation" of the true distribution among all binary regression models with link function $\hat{h}^{-1}$ and design matrix $X[\hat{M}]$. Note that $\beta_{\hat{\mathbb{M}}}$ depends on $\hat{\mathbb{M}}$ and is thus data-dependent. Our results in Section 3.3 allow one to make valid confidence statements about $\beta_{\hat{\mathbb{M}}}$. Given $\alpha \in (0, 1)$, our confidence intervals for the jth coordinate of $\beta_{\hat{\mathbb{M}}}$ are of the form

$$\mathrm{CI}^{(j)}_{\alpha,\hat{\mathbb{M}}} := \hat{\beta}^{(j)}_{\hat{\mathbb{M}}} \pm B_\alpha \cdot \sqrt{(\hat{S}_{\hat{\mathbb{M}}})_{jj}},$$

with $\hat{\beta}_{\hat{\mathbb{M}}}$ the MLE based on the (quasi-)log-likelihood function in the selected model (1.1), $\hat{S}_{\hat{\mathbb{M}}}$ a "sandwich-type" variance estimator based on a suggestion of Fahrmeir ((1990), page 491), and B_α a critical value (standard quantiles ignoring model selection would not be valid) that, besides α, only depends on sample size n and on the full collection of potential candidate models from which $\hat{\mathbb{M}}$ was selected (i.e., the range of $\hat{\mathbb{M}}$). The resulting intervals are asymptotically valid in the sense that their joint coverage probability $\mathbb{P}(\beta^{(j)}_{\hat{\mathbb{M}}} \in \mathrm{CI}^{(j)}_{\alpha,\hat{\mathbb{M}}}$, for all $j = 1, \dots, |\hat{M}|)$ is asymptotically bounded below by $1 - \alpha$, and the convergence is uniform over a large class of true unknown data generating distributions. Moreover, coverage is guaranteed for *any* model selection procedure, and the constant B_α is fast to compute, even for large values of n and p, and also for large collections of candidate models.

1.2. *Related work.* The present article is devised in the spirit of Berk et al. (2013), in the sense that we aim at inference post-model-selection that is valid irrespective of the employed model selection procedure. Very recently, Rinaldo et al. (2016) have investigated a classical sample spitting procedure that is also independent of the underlying selection method. However, they consider only the i.i.d. case, thereby excluding, for instance, fixed design regression. Several other authors have proposed inference procedures post-model-selection that are tailored towards specific selection methods and for specific modeling situations. In the context of fitting linear regression models to Gaussian data, methods that provide valid confidence sets post-model-selection, and that are constructed for specific model selection procedures (e.g., forward stepwise, least-angle-regression or the lasso) and for targets of inference similar to those considered in the present article, have been recently obtained by Lee and Taylor (2014), Fithian, Sun and Taylor (2015), Lee et al. (2016) and Tibshirani et al. (2016). Tibshirani et al. (2018) extended the approach of Tibshirani et al. (2016) to non-Gaussian data by obtaining uniform asymptotic results. Furthermore, valid inference post-model-selection on conventional regression parameters under sparsity conditions was considered, among others, by Belloni, Chernozhukov and Hansen (2011, 2014), van de Geer et al. (2014) and Zhang and Zhang (2014).

2. Inference post-model-selection: A general asymptotic theory.

2.1. *Framework, problem description and approach.* Consider a situation where we observe a data set $y \in \mathbb{R}^{n\times\ell}$ that is a realization of an unknown probability distribution $\mathbb{P}_n$ on the Borel sets of the sample space $\mathbb{R}^{n\times\ell}$. We denote the ith row of the data vector (matrix) y by $y_i \in \mathbb{R}^{1\times\ell}$, so that $y = (y_1', \ldots, y_n')'$, and write $\mathbb{P}_{i,n}$ for the marginal distribution corresponding to that row. Throughout, we assume that the data generating distribution is of product form, that is $\mathbb{P}_n = \bigotimes_{i=1}^n \mathbb{P}_{i,n}$. Suppose further that one wants to conduct inference on $\mathbb{P}_n$, and intends to use as a working model an element of M_n, a set consisting of d nonempty sets of distributions $\mathbb{M}_{1,n}, \ldots, \mathbb{M}_{d,n}$ on the Borel sets of $\mathbb{R}^{n\times\ell}$. Throughout d is fixed, that is, does not depend on n. We emphasize that it is *not* assumed that $\mathbb{P}_n$ is contained in any of the sets $\mathbb{M}_{j,n}$ for $j = 1, \ldots, d$. That is, the candidate set M_n might be *misspecified*.

For each model $\mathbb{M} \in \mathsf{M}_n$ one has to define a corresponding target of inference $\theta^*_{\mathbb{M},n} = \theta^*_{\mathbb{M},n}(\mathbb{P}_n)$, say, which we take as given throughout the present section. Furthermore we assume that for every $\mathbb{M}_{j,n} \in \mathsf{M}_n$ the target is an element of a Euclidean space of finite dimension $m(\mathbb{M}_{j,n})$ which does not depend on n. As an example in the case $\ell = 1$, consider the situation where $\mathbb{P}_n$ has mean vector $\mu_n \in \mathbb{R}^n$ and $\mathbb{M} \in \mathsf{M}_n$ is given by the collection of all n-dimensional normal distributions with covariance matrix proportional to identity and mean $X_{\mathbb{M}}\beta$, for different values of $\beta \in \mathbb{R}^{m(\mathbb{M})}$, and where $X_{\mathbb{M}}$ is an $n \times m(\mathbb{M})$ matrix obtained by selecting certain columns from a given fixed design matrix $X \in \mathbb{R}^{n\times p}$. In this setting, Berk et al. (2013) consider the target $\theta^*_{\mathbb{M},n}(\mathbb{P}_n) = (X_{\mathbb{M}}'X_{\mathbb{M}})^{-1}X_{\mathbb{M}}'\mu_n$ (cf. also Section 3 for more on this and further examples). In the general case, $\theta^*_{\mathbb{M},n}$ will typically be the value of the parameter that corresponds to the projection of $\mathbb{P}_n$ onto $\mathbb{M}$ w.r.t. some measure of closeness, for example, the Kullback–Leibler divergence, or the Hellinger-distance. Note that such a projection might not uniquely exist, or might not exist at all, and that in each application sufficient conditions—on $\mathbb{P}_n$ and/or the candidate set M_n of models—need to be imposed to obtain well defined targets. Note also that the target is model-specific, that is, it depends on $\mathbb{M}$. Lastly we emphasize that defining and working with such projections as targets of inference in potentially misspecified models has a long-standing tradition in statistics, dating back at least to Huber (1967), and we confer the reader to this strand of literature for further discussion.

Given data y the statistician now has two problems to solve: (i) model selection, that is, the statistician needs to choose an "appropriate" working model from the candidate set

M_n; and (ii) statistical inference post-model-selection, that is, given the selected model, the statistician typically wants to conduct inference on the target in this model. Note that this target is random, as it depends on the data via the model selection procedure used. We do not contribute anything new to how models can be selected from data. We take a model selection procedure as given, and denote the model selection procedure used by $\hat{\mathbb{M}}_n : \mathbb{R}^{n\times\ell} \to \mathsf{M}_n$ (measurable). That is, the quantity $\hat{\mathbb{M}}_n(y)$ denotes the selected model upon observing y. We also assume that for every model $\mathbb{M} \in \mathsf{M}_n$ an estimator $\hat{\theta}_{\mathbb{M},n} : \mathbb{R}^{n\times\ell} \to \mathbb{R}^{m(\mathbb{M})}$ (measurable) of the corresponding target $\theta^*_{\mathbb{M},n}$ is available.

Summarizing, the statistician selects the model using $\hat{\mathbb{M}}_n$, and estimates $\theta^*_{\hat{\mathbb{M}}_n,n}$ using $\hat{\theta}_{\hat{\mathbb{M}}_n,n}$. In this article, we address the question how valid confidence intervals can be constructed for the coordinates of the target $\theta^*_{\hat{\mathbb{M}}_n,n}$. Our approach is as follows:

1. Given $\alpha \in (0,1)$, we construct confidence intervals $\mathrm{CI}^{(j)}_{1-\alpha,\mathbb{M}}$ for the jth component $\theta^{*(j)}_{\mathbb{M},n}$ of $\theta^*_{\mathbb{M},n}$, for every $j = 1,\dots,m(\mathbb{M})$ and every $\mathbb{M} \in \mathsf{M}_n$ such that

$$\liminf_{n\to\infty} \mathbb{P}_n\big(\theta^{*(j)}_{\mathbb{M},n} \in \mathrm{CI}^{(j)}_{1-\alpha,\mathbb{M}} \text{ for all } j = 1,\dots,m(\mathbb{M}) \text{ and all } \mathbb{M} \in \mathsf{M}_n\big)$$

is not smaller than $1-\alpha$.

2. For a model selection procedure $\hat{\mathbb{M}}_n$, our suggested confidence intervals are then obtained via

$$\mathrm{CI}^{(j)}_{1-\alpha,\hat{\mathbb{M}}_n} \quad \text{for } j = 1,\dots,m(\hat{\mathbb{M}}_n).$$

From the coverage property in Part 1, we obtain

$$\liminf_{n\to\infty} \mathbb{P}_n\big(\theta^{*(j)}_{\hat{\mathbb{M}}_n,n} \in \mathrm{CI}^{(j)}_{1-\alpha,\hat{\mathbb{M}}_n} \text{ for all } j = 1,\dots,m(\hat{\mathbb{M}}_n)\big) \geq 1-\alpha.$$

As already discussed in the Introduction, the fact that our approach does not restrict the model selection procedure used is important. It is precisely this aspect that allows practitioners to obtain valid confidence intervals post-model-selection in situations where a wide variety of (formal or informal) mechanisms have been incorporated to select the model.

2.2. *Discussion.* The above framework is certainly somewhat abstract, but its generality is necessary to achieve the scope of the present paper, which is the development of results for the construction of confidence intervals post-model-selection that are widely applicable. In particular, apart from allowing for a misspecified candidate set of models, the framework allows the marginals $\mathbb{P}_{i,n}$ for $i = 1,\dots,n$ to be nonidentical. This property is not just a mere technical aspect, but is necessary if one wants to cover situations such as fixed-design regression models.

Most importantly, we work with a sequence $\mathbb{P}_n$ of data generating mechanisms, the marginals of which also depend on n. Again, this is not a technical nuisance. Rather, this aspect ensures that the results obtained can be used to construct uniformly valid confidence intervals post-model-selection: Suppose $\mathbb{P}_n$, the distribution that generated the data y, is known to be an element of a set $\mathbf{P}_n$. The set $\mathbf{P}_n$ describes the assumptions one is willing to impose on the unknown distribution in a particular modeling scenario, and will typically be large and potentially nonparametric. Suppose further that one wants to work with a candidate set of models M_n (possibly misspecified, that is, $\mathbf{P}_n \nsubseteq \bigcup_{\mathbb{M}\in\mathsf{M}_n} \mathbb{M}$) and corresponding model specific targets $\theta^*_{\mathbb{M},n}$ as above, and that the goal is to construct confidence sets post-model-selection. Under weak assumptions on $\mathbf{P}_n$, the general results developed in this paper allow one to construct confidence intervals such that

$$\liminf_{n\to\infty} \mathbb{P}_n\big(\theta^{*(j)}_{\hat{\mathbb{M}}_n,n} \in \mathrm{CI}^{(j)}_{1-\alpha,\hat{\mathbb{M}}_n} \text{ for all } j = 1,\dots,m(\hat{\mathbb{M}}_n)\big) \geq 1-\alpha$$

holds for any (measurable) model selection procedure $\hat{\mathbb{M}}_n$, and for every sequence of distributions $\mathbb{P}_n$ that satisfies $\mathbb{P}_n \in \mathbf{P}_n$ for every $n \in \mathbb{N}$. Certainly, this then implies

$$\liminf_{n\to\infty} \inf_{\mathbb{P}_n\in\mathbf{P}_n} \mathbb{P}_n\big(\theta^{*(j)}_{\hat{\mathbb{M}}_n,n} \in \mathrm{CI}^{(j)}_{1-\alpha,\hat{\mathbb{M}}_n} \text{ for all } j = 1,\dots,m(\hat{\mathbb{M}}_n)\big) \geq 1-\alpha,$$

that is, asymptotic validity of the constructed confidence sets *uniformly* over $\mathbf{P}_n$. That the development of results that hold uniformly over large classes of distributions is important, in particular so in the context of inference post-model-selection, is well understood (see Leeb and Pötscher (2003, 2005)). One recent article that studies uniform coverage properties post-model-selection is Tibshirani et al. (2018). Merits of uniform results in contrast to pointwise asymptotic results are discussed in their Section 1.1. Tibshirani et al. (2018) consider a setup similar to the example we consider in Section 3.1 and for specific model selectors, but compared to our results uniform validity is established only over substantially smaller sets of distributions, and they need to impose stronger conditions on the design matrices, which rule out some important cases our results allow for, for example, polynomial trends. See also Section 4.1 for numerical results and comparisons.

2.3. *Notation*. Before we proceed to our general theory and the corresponding basic assumption, we introduce some notation that is used throughout this article: A normal distribution with mean μ and (possibly singular) covariance matrix Σ is denoted by $N(\mu,\Sigma)$. For $\alpha \in (0,1)$ and a covariance matrix Γ we denote by $K_{1-\alpha}(\Gamma)$ the $1-\alpha$-quantile of the distribution of the supremum-norm $\|Z\|_\infty$ of $Z \sim N(0,\Gamma)$. The correlation matrix corresponding to a covariance matrix Σ is denoted by $\mathrm{corr}(\Sigma) = \mathrm{diag}(\Sigma)^{\dagger/2}\Sigma\,\mathrm{diag}(\Sigma)^{\dagger/2}$, where $\mathrm{diag}(\Sigma)$ denotes the diagonal matrix obtained from Σ by setting all off-diagonal elements equal to 0, $A^\dagger$ denotes the Moore–Penrose inverse of a matrix A, $A^{1/2}$ denotes the symmetric nonnegative definite square root of a symmetric nonnegative definite matrix A, and where we abbreviate $[A^\dagger]^{1/2}$ by $A^{\dagger/2}$. The smallest and largest eigenvalue of a real symmetric matrix A is denoted by $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$, respectively. For a vector v with coordinates $v^{(1)},\dots,v^{(l)}$ we also use the symbol $\mathrm{diag}(v)$ to denote the diagonal matrix with first diagonal entry $v^{(1)}$, second $v^{(2)}$, and so on. The operator norm of a matrix A (w.r.t. the Euclidean norm) is denoted by $\|A\|$, and the Euclidean norm of a vector v is denoted by $\|v\|$. Furthermore, A_{ii}, the ith diagonal element of a quadratic matrix A, is occasionally abbreviated as A_i. We also identify the indicator function $\mathbb{1}_B$ of a set B with the set B itself, whenever there is no risk of confusion. Weak convergence of a sequence of probability measures $\mathbb{Q}_n$ to $\mathbb{Q}$ is denoted by $\mathbb{Q}_n \Rightarrow \mathbb{Q}$. The image measure induced by a random variable (or vector) x defined on a probability space $(F,\mathcal{F},\mathbb{Q})$ is denoted by $\mathbb{Q}\circ x$. If not stated otherwise, limits are taken as $n\to\infty$. For a sequence $(a_n)_{n\in\mathbb{N}}$, we say that a property *holds eventually* if there exists a positive integer n_0 such that the property holds for every a_n with $n \geq n_0$. The expectation operator and the variance-covariance operator w.r.t. $\mathbb{P}_n$ is denoted by $\mathbb{E}_n$ and $\mathbb{V}_n$, respectively; and the expectation operator and the variance-covariance operator w.r.t. $\mathbb{P}_{i,n}$ is denoted by $\mathbb{E}_{i,n}$ and $\mathbb{V}_{i,n}$, respectively.

2.4. *Main assumption*. Our methods for constructing uniformly valid confidence intervals post-model-selection are developed under a high-level condition imposed on the stacked vector of estimators $\hat{\theta}_n = (\hat{\theta}'_{\mathbb{M}_1,n},\dots,\hat{\theta}'_{\mathbb{M}_d,n})'$ centered at the corresponding stacked vector of targets $\theta^*_n = (\theta^{*\prime}_{\mathbb{M}_1,n},\dots,\theta^{*\prime}_{\mathbb{M}_d,n})'$. In this section, we denote the dimension of $\hat{\theta}_n$ by $k := \sum_{j=1}^d m(\mathbb{M}_{j,n})$, which does not depend on n. The condition is as follows:

CONDITION 1. There exist Borel measurable functions $g_{i,n} : \mathbb{R}^{1\times\ell} \to \mathbb{R}^k$ for $i = 1, \ldots, n$, and $\Delta_n : \mathbb{R}^{n\times\ell} \to \mathbb{R}^k$, possibly depending on θ_n^*, such that

$$\hat{\theta}_n(y) - \theta_n^* = \sum_{i=1}^{n} g_{i,n}(y_i) + \Delta_n(y), \tag{2.1}$$

for $y \in \mathbb{R}^{n\times\ell}$, where, writing $r_n(y) := \sum_{i=1}^n g_{i,n}(y_i)$, it holds for every $i \in \{1, \ldots, n\}$ and every $j \in \{1, \ldots, k\}$ that

$$\mathbb{E}_{i,n}\big(g_{i,n}^{(j)}\big) = 0 \quad \text{and} \quad 0 < \mathbb{V}_n\big(r_n^{(j)}\big) < \infty. \tag{2.2}$$

Furthermore, for every coordinate $j \in \{1, \ldots, k\}$ we have

$$\mathbb{V}_n^{-1}\big(r_n^{(j)}\big) \sum_{i=1}^{n} \int_{\mathbb{R}^{1\times\ell}} \big[g_{i,n}^{(j)}\big]^2 \big\{\big|g_{i,n}^{(j)}\big| \geq \varepsilon \mathbb{V}_n^{\frac{1}{2}}\big(r_n^{(j)}\big)\big\}\, d\mathbb{P}_{i,n} \to 0 \tag{2.3}$$

for every $\varepsilon > 0$

and

$$\mathbb{P}_n\big(\big|\mathbb{V}_n^{-1/2}\big(r_n^{(j)}\big)\Delta_n^{(j)}\big| \geq \varepsilon\big) \to 0 \quad \text{for every } \varepsilon > 0.$$

Clearly, an expansion as in Equation (2.1) of Condition 1 is satisfied in many applications, and can typically be obtained by a standard linearization argument (see Section 3.3 for an example and Remark A.7 in the Supplementary Material for further discussion). We emphasize that the two last assumptions in Condition 1 are formulated in terms of rescaled summands, which, in applications, can be exploited to circumvent restrictive compactness assumptions on moments of the distribution generating the data or the design (e.g., in Sections 3.1 and 3.2 we do not need to restrict variance parameters to a compact set—as opposed to the conditions used by, for example, Eicker (1967) or Tibshirani et al. (2018); and in Section 3.3, we do not require the smallest singular value of the design matrix to diverge to infinity—as opposed to, for example, Lv and Liu (2014)).

REMARK 2.1. The careful reader will have noticed, that the functions $g_{i,n} : \mathbb{R}^{1\times\ell} \to \mathbb{R}^k$ in Condition 1 do not depend on all of the observation matrix $y \in \mathbb{R}^{n\times\ell}$, but only on its ith row $y_i \in \mathbb{R}^{1\times\ell}$. This is crucial. In the sequel, however, it will be convenient to also consider $g_{i,n}$ as a function on the full sample space $\mathbb{R}^{n\times\ell}$. Thus, we sometimes identify $g_{i,n}$ with the composition $g_{i,n} \circ \pi_{i,n} : \mathbb{R}^{n\times\ell} \to \mathbb{R}^k$, where $\pi_{i,n} : \mathbb{R}^{n\times\ell} \to \mathbb{R}^{1\times\ell}$ is the coordinate projection $\pi_{i,n}(y) = y_i$.

Before proceeding to the main results, we briefly highlight the most important consequence of Condition 1 for our method of constructing confidence sets post-model-selection. The first step of the approach outlined in Section 2.1 required the construction of confidence intervals for *each* coordinate of the stacked vector of targets θ_n^*. Naturally, such confidence intervals will be centered at the respective coordinates of $\hat{\theta}_n$. Our construction of such intervals is based on part one of the subsequent Lemma 2.2 which provides an asymptotic approximation to the distribution

$$\mathbb{P}_n \circ \big[\operatorname{diag}\big(\mathbb{V}_n(r_n)\big)^{\dagger/2}\big(\hat{\theta}_n - \theta_n^*\big)\big]. \tag{2.4}$$

One can not expect, in general, that the distribution in the previous display converges weakly to a limiting distribution as $n \to \infty$, simply because the correlations may not stabilize. However, under Condition 1 we can show that the distributions are "well approximated" by the

sequence of Gaussian distributions $N(0, \operatorname{corr}(\mathbb{V}_n(r_n)))$. Being "well approximated" is understood in the sense that

$$d_w\big(\mathbb{P}_n \circ \big[\operatorname{diag}\big(\mathbb{V}_n(r_n)\big)^{\dagger/2}\big(\hat{\theta}_n - \theta_n^*\big)\big], N\big(0, \operatorname{corr}\big(\mathbb{V}_n(r_n)\big)\big)\big) \to 0. \tag{2.5}$$

Here d_w denotes a distance metrizing weak convergence of probability measures on the Borel sets of the respective Euclidean space the dimension of which is not shown in the notation (cf. the discussion in Dudley ((2002), page 393) for specific examples). Note that in case $\operatorname{corr}(\mathbb{V}_n(r_n))$ is convergent this reduces to weak convergence. Furthermore, in the second part of Lemma 2.2, and defining under Condition 1 the matrix

$$S_n(y) := \sum_{i=1}^{n} g_{i,n}(y_i) g_{i,n}'(y_i), \tag{2.6}$$

we show that a suitable approximation statement continues to hold if $\mathbb{V}_n(r_n)$ is replaced by S_n: the d_w-distance between

$$\mathbb{P}_n \circ \big[\operatorname{diag}(S_n)^{\dagger/2}\big(\hat{\theta}_n - \theta_n^*\big)\big], \tag{2.7}$$

and the sequence of (random) Gaussian distributions $N(0, \operatorname{corr}(S_n))$ converges to 0 in $\mathbb{P}_n$-probability. This latter property is instrumental for our approach to constructing covariance estimators, as will be explained after the lemma.

LEMMA 2.2. *Under Condition 1 the convergence* (2.5) *holds and, for every* $\varepsilon > 0$, *we have*

$$\mathbb{P}_n\big(d_w\big(\mathbb{P}_n \circ \big[\operatorname{diag}(S_n)^{\dagger/2}\big(\hat{\theta}_n - \theta_n^*\big)\big], N\big(0, \operatorname{corr}(S_n)\big)\big) \geq \varepsilon\big) \to 0.$$

Remarkably, the previous result is obtained without requiring a *joint* Lindeberg-type condition on the random vectors $g_{i,n}$, but only requires "marginal" Lindeberg-type conditions. Further discussion on Condition 1 and on how it can be verified is provided in Section A.1.1 of the Supplementary Material (Bachoc, Preinerstorfer and Steinberger (2019)). Lemma 2.2 is proved in Section D.1 of the Supplementary Material using tightness arguments, a result in Pollak (1972), and Raikov's theorem (cf. the statement in Gnedenko and Kolmogorov (1954) on page 143, originally published in Raikov (1938)).

At first sight, one might be tempted to think that one can now immediately use S_n as a covariance estimator to construct confidence intervals as envisioned in Section 2.1. However, we emphasize that S_n is in general *not* an estimator of $\mathbb{V}_n(r_n)$. Typically $g_{i,n}$ depends on θ_n^*, which is unknown, and thus S_n is *infeasible*. Hence, while Lemma 2.2 presents a first step towards the construction of confidence sets post-model-selection, the construction of suitable covariance estimators is another step that we need to address. We nevertheless note that although Lemma 2.2 does not answer how such estimators can be obtained, it suggests that in applications one might use as an estimator for $\mathbb{V}_n(r_n)$ a "suitable" predictor for S_n, for example, by using "suitable" predictors for the unobserved components $g_{i,n}$. This is discussed in detail in the following section.

2.5. *Confidence intervals post-model-selection.* In this subsection, we shall now present our general asymptotic results for the construction of valid confidence intervals post-model-selection under Condition 1. We consider two different situations: (i) a situation where a consistent estimator of $\mathbb{V}_n(r_n)$ is available; (ii) a situation where a consistent estimator of $\mathbb{V}_n(r_n)$ is *not* available, but it is possible to construct estimators that "consistently overestimate" the diagonal entries of $\mathbb{V}_n(r_n)$. Concrete examples of such consistent or "consistently overestimating" estimators are also provided, based on approximating the summands $g_{i,n}$ appearing in Condition 1.

Given $\mathbb{M} = \mathbb{M}_{j,n} \in \mathsf{M}_n$ we abbreviate $\rho(\mathbb{M}) := \sum_{l=1}^{j-1} m(\mathbb{M}_{l,n})$, where sums over an empty index set are to be interpreted as 0.

2.5.1. *Confidence intervals based on consistent estimators of* $\mathbb{V}_n(r_n)$. Our first result considers the construction of confidence intervals post-model-selection under Condition 1, and under the additional assumption that it is possible to construct a consistent estimator $\hat{S}_n$ of $\mathbb{V}_n(r_n)$. The latter assumption is certainly very restrictive, due to possible misspecification of the model, and is relaxed substantially in Section 2.5.2.

THEOREM 2.3. *Let* $\alpha \in (0,1)$, *suppose Condition 1 holds, and let* $\hat{S}_n : \mathbb{R}^{n\times\ell} \to \mathbb{R}^{k\times k}$ *be a sequence of Borel-measurable functions such that*

$$\mathbb{P}_n\big(\|\operatorname{corr}(\hat{S}_n) - \operatorname{corr}(\mathbb{V}_n(r_n))\| + \|\operatorname{diag}(\mathbb{V}_n(r_n))^{-1}\operatorname{diag}(\hat{S}_n) - I_k\| \geq \varepsilon\big) \tag{2.8}$$

converges to 0 *for every* $\varepsilon > 0$, *or equivalently, that for every* $\varepsilon > 0$

$$\mathbb{P}_n\big(\|\operatorname{corr}(\hat{S}_n) - \operatorname{corr}(S_n)\| + \|\operatorname{diag}(S_n)^{\dagger}\operatorname{diag}(\hat{S}_n) - I_k\| \geq \varepsilon\big) \to 0. \tag{2.9}$$

Define for every $\mathbb{M} \in \mathsf{M}_n$ *and every* $j = 1,\dots,m(\mathbb{M})$ *the confidence interval*

$$\mathrm{CI}^{(j),\mathrm{est}}_{1-\alpha,\mathbb{M}} = \hat{\theta}^{(j)}_{\mathbb{M},n} \pm \sqrt{[\hat{S}_n]_{\rho(\mathbb{M})+j}}\,K_{1-\alpha}\big(\operatorname{corr}(\hat{S}_n)\big).$$

Then, $\mathbb{P}_n(\theta^{*(j)}_{\mathbb{M},n} \in \mathrm{CI}^{(j),\mathrm{est}}_{1-\alpha,\mathbb{M}}$ *for all* $\mathbb{M} \in \mathsf{M}_n$ *and all* $j = 1,\dots,m(\mathbb{M}))$ *converges to* $1-\alpha$ *as* $n \to \infty$. *In particular, for every* (*measurable*) *model selection procedure* $\hat{\mathbb{M}}_n$, *we have*

$$\liminf_{n\to\infty} \mathbb{P}_n\big(\theta^{*(j)}_{\hat{\mathbb{M}}_n,n} \in \mathrm{CI}^{(j),\mathrm{est}}_{1-\alpha,\hat{\mathbb{M}}_n} \text{ for all } j = 1,\dots,m(\hat{\mathbb{M}}_n)\big) \geq 1-\alpha. \tag{2.10}$$

Theorem 2.3 is based on the assumption that an estimator $\hat{S}_n$ is available that consistently estimates $\mathbb{V}_n(r_n)$. Coming back to the discussion at the end of Section 2.4, the vectors $g_{i,n}(y_i)$ appearing in the definition of S_n are typically *not* observable, because they will depend on the unknown target θ^*_n, that is, they are, more explicitly, of the form $g_{i,n}(y_i,\theta^*_n)$. In such cases S_n is not a feasible candidate for $\hat{S}_n$ in the previous theorem, and therefore one will, in most cases, naturally try to obtain predictors $\hat{g}_{i,n}(y)$ for $g_{i,n}(y_i)$ by replacing the unknown target by its estimator $\hat{\theta}_n$, that is, by setting $\hat{g}_{i,n}(y) = g_{i,n}(y_i,\hat{\theta}_n(y))$. The subsequent proposition now provides conditions on predictors $\hat{g}_{i,n}(y)$, which, if satisfied, immediately allow the construction of a consistent estimator $\hat{S}_n$ of $\mathbb{V}_n(r_n)$ by replacing each $g_{i,n}(y_i)$ in equation (2.6) by its predictor $\hat{g}_{i,n}(y)$. In the result the predictor $\hat{g}_{i,n}(y)$ may be of the form $g_{i,n}(y_i,\hat{\theta}_n(y))$ as discussed above, but the proposition is not restricted to that particular case. Again, the conditions are assumptions concerning the large sample behavior of the marginals only, which facilitates their verification in practice.

PROPOSITION 2.4. *Suppose Condition 1 is satisfied, and let* $\hat{g}_{i,n} : \mathbb{R}^{n\times\ell} \to \mathbb{R}^k$ *be Borel measurable for* $i = 1,\dots,n$ *and for every* n. *Suppose that for every* $j = 1,\dots,k$ *and for every* $\varepsilon > 0$ *it holds that*

$$\mathbb{P}_n\left(\sum_{i=1}^n \big(g^{(j)}_{i,n} - \hat{g}^{(j)}_{i,n}\big)^2 \Big/ \sum_{i=1}^n \big(g^{(j)}_{i,n}\big)^2 \geq \varepsilon\right) \to 0, \tag{2.11}$$

or equivalently that

$$\mathbb{P}_n\left(\sum_{i=1}^n \big(g^{(j)}_{i,n} - \hat{g}^{(j)}_{i,n}\big)^2 \Big/ \sum_{i=1}^n \mathbb{V}_n\big(g^{(j)}_{i,n}\big) \geq \varepsilon\right) \to 0. \tag{2.12}$$

Then the convergence in (2.9) *is satisfied for* $\hat{S}_n = \sum_{i=1}^n \hat{g}_{i,n}\hat{g}'_{i,n}$.

2.5.2. *Confidence intervals based on estimators that consistently overestimate the diagonal entries of* $\mathbb{V}_n(r_n)$. Due to an asymptotically nonnegligible bias term arising from misspecification of the model, it is typically difficult to obtain an estimator $\hat{S}_n$ satisfying the condition in Theorem 2.3 (see Remark 2.7 and Section A.2.1 in the Supplementary Material (Bachoc, Preinerstorfer and Steinberger (2019)) for details). Nevertheless, it is often still possible to construct estimators of the diagonal entries of the matrix $\mathbb{V}_n(r_n)$ that, while possibly inconsistent, asymptotically *overestimate* their targets; for a corresponding constructive result see Proposition 2.6 below. Similarly, it is in general not difficult to find an estimator of $K_{1-\alpha}(\mathrm{corr}(S_n))$ that consistently overestimates that quantity, see the discussion and the result following Proposition 2.6 below concerning upper bounds on the function $K_{1-\alpha}(\cdot)$ over the set of all correlation matrices (using this upper bound, although leading to wider confidence intervals, also leads to substantial computational advantages). Based on such estimators it is then possible to construct asymptotically valid confidence intervals post-model-selection, even though the candidate set of models might be (severely) misspecified. This is the content of the subsequent result, which, together with Proposition 2.6 below, is the main theoretical result in this section.

THEOREM 2.5. *Let* $\alpha \in (0,1)$, *and suppose Condition 1 is satisfied. For every* n *and every* $j = 1, \ldots, k$ *let* $\hat{v}_{j,n}^2 \geq 0$ *be an estimator of* $\mathbb{V}_n(r_n^{(j)})$, *and let* $\hat{K}_n \geq 0$ *be an estimator of* $K_{1-\alpha}(\mathrm{corr}(\mathbb{V}_n(r_n)))$, *such that the sequence*

$$\kappa_n = \frac{K_{1-\alpha}(\mathrm{corr}(\mathbb{V}_n(r_n)))}{\hat{K}_n} \max_{j=1,\ldots,k} \sqrt{\frac{[\mathbb{V}_n(r_n)]_j}{\hat{v}_{j,n}^2}},$$

satisfies

$$\mathbb{P}_n(\kappa_n \geq 1+\varepsilon) \to 0 \quad \textit{for every } \varepsilon > 0 \tag{2.13}$$

(*implicitly including that* $\mathbb{P}_n(\kappa_n$ *is well defined*$) \to 1$) *or, equivalently, that the condition in* (2.13) *holds with* κ_n *replaced by*

$$\frac{K_{1-\alpha}(\mathrm{corr}(S_n))}{\hat{K}_n} \max_{j=1,\ldots,k} \sqrt{\frac{[S_n]_j}{\hat{v}_{j,n}^2}}.$$

For every $\mathbb{M} \in \mathsf{M}_n$ *and every* $j = 1, \ldots, m(\mathbb{M})$, *define the confidence interval*

$$\mathrm{CI}_{1-\alpha,\mathbb{M}}^{(j),\mathrm{oest}} = \hat{\theta}_{\mathbb{M},n}^{(j)} \pm \sqrt{\hat{v}_{\rho(\mathbb{M})+j,n}^2}\,\hat{K}_n.$$

Then, for every (*measurable*) *model selection procedure* $\hat{\mathbb{M}}_n$, *we have*

$$\liminf_{n\to\infty} \mathbb{P}_n\big(\theta_{\hat{\mathbb{M}}_n,n}^{*(j)} \in \mathrm{CI}_{1-\alpha,\hat{\mathbb{M}}_n}^{(j),\mathrm{oest}} \textit{ for all } j = 1, \ldots, m(\hat{\mathbb{M}}_n)\big) \geq 1-\alpha.$$

In the important special case where $\hat{K}_n \geq K_{1-\alpha}(\mathrm{corr}(\mathbb{V}_n(r_n)))$ *holds eventually, the condition in equation* (2.13) *is implied by the condition that for every* $j = 1, \ldots, k$ *it holds that*

$$\mathbb{P}_n\Big(\sqrt{[S_n]_j/\hat{v}_{j,n}^2} \geq 1+\varepsilon\Big) \to 0 \quad \textit{for every } \varepsilon > 0, \tag{2.14}$$

or equivalently, that for every $j = 1, \ldots, k$ *it holds that*

$$\mathbb{P}_n\Big(\sqrt{[\mathbb{V}_n(r_n)]_j/\hat{v}_{j,n}^2} \geq 1+\varepsilon\Big) \to 0 \quad \textit{for every } \varepsilon > 0. \tag{2.15}$$

The preceding theorem operates under the assumption that estimators are available that consistently overestimate the diagonal entries of $\mathbb{V}_n(r_n)$ and $K_{1-\alpha}(\mathrm{corr}(\mathbb{V}_n(r_n)))$. The following result now shows how such estimators for the diagonal entries of $\mathbb{V}_n(r_n)$ can be obtained. To construct an estimator $\hat{K}_n$ that eventually satisfies $\hat{K}_n \geq K_{1-\alpha}(\mathrm{corr}(\mathbb{V}_n(r_n)))$ (as required for the special case of Theorem 2.5) one can numerically compute the upper bound in Lemma 2.8 below. The subsequent result considers the case where the vectors $g_{i,n}$ from Condition 1 are well approximated in the sense of the condition appearing in Proposition 2.4, but where the approximating quantities are now *unobservable* due to nonstochastic additive error terms. These additive error terms typically are bias terms due to misspecification of the model. This is further discussed after the proposition.

PROPOSITION 2.6. *Suppose Condition 1 is satisfied, and let* $\tilde{g}_{i,n} : \mathbb{R}^{n\times\ell} \to \mathbb{R}^k$ *and* $\hat{g}_{i,n} : \mathbb{R}^{n\times\ell} \to \mathbb{R}^k$ *be Borel measurable for* $i = 1, \ldots, n$ *and for every* n. *Suppose that for every* $j = 1, \ldots, k$ *and for every* $\varepsilon > 0$ *the condition* (2.11), *or equivalently* (2.12), *is satisfied. Suppose further that there exist real numbers* $a_{i,n}^{(j)}$ *such that for* $y \in \mathbb{R}^{n\times\ell}$

$$\tilde{g}_{i,n}^{(j)}(y) = \hat{g}_{i,n}^{(j)}(y) + a_{i,n}^{(j)}$$

holds for every $n \in \mathbb{N}$, $i \in \{1, \ldots, n\}$ *and* $j \in \{1, \ldots, k\}$. *Then the statement in* (2.15) *is satisfied for* $\hat{v}_{j,n}^2 = \sum_{i=1}^n [\tilde{g}_{i,n}^{(j)}]^2$ $(j = 1, \ldots, k)$.

The proposition is developed for situations where random variables $\tilde{g}_{i,n}^{(j)}$ are observed, that can be decomposed as the sum of unobserved random variables $\hat{g}_{i,n}^{(j)}$, which satisfy (2.11), and unobserved real numbers $a_{i,n}^{(j)}$. In contrast to the situation in Proposition 2.4, now the (unobservable) random variables $\hat{g}_{i,n}^{(j)}$ can not be used for the construction of estimators. Nevertheless, the proposition shows how suitable variance estimators can then still be constructed based on the observed quantities $\tilde{g}_{i,n}^{(j)}$. Confidence intervals post-model-selection can then be obtained via Theorem 2.5. Besides being suitable for situations where random variables satisfying (2.11) are not observed (otherwise one could use Proposition 2.4 to obtain consistent estimators), Proposition 2.6 is particularly geared towards the case where the nonstochastic additive components $a_{i,n}$ are nonnegligible in the sense that

$$\frac{\sum_{i=1}^n [a_{i,n}^{(j)}]^2}{\mathbb{V}_n(r_n^{(j)})} \not\to 0 \quad \text{holds for some } j \in \{1, \ldots, k\}.$$

For if the nonstochastic additive components are negligible in this sense, a consistent estimator of $\mathbb{V}_n(r_n)$ in the sense of (2.8) can be constructed.

REMARK 2.7. Using the bound $(g_{i,n}^{(j)} - \tilde{g}_{i,n}^{(j)})^2 \leq 2(g_{i,n}^{(j)} - \hat{g}_{i,n}^{(j)})^2 + 2[a_{i,n}^{(j)}]^2$, it is easy to verify that if the nonstochastic additive components $a_{i,n}$ are negligible in the previously defined sense, then $\tilde{g}_{i,n}$ satisfies the assumptions of $\hat{g}_{i,n}$ appearing in Proposition 2.4. As a consequence, the estimator $\tilde{S}_n = \sum_{i=1}^n \tilde{g}_{i,n}\tilde{g}_{i,n}'$ satisfies (2.8), and one can construct confidence intervals based on this estimator as discussed in Theorem 2.3. Note that $\hat{v}_{j,n}^2 = [\tilde{S}_n]_j$.

Let us finally consider an upper bound on $K_{1-\alpha}(\Gamma)$ as required in the special case of Theorem 2.5 above. The bound we shall discuss is based on the quantity $B_\alpha(q, N)$, for $q, N \in \mathbb{N}$, defined as the smallest $t > 0$ such that

$$\mathbb{E}_G\big(\min\big(1, \big[1 - F_{\mathrm{Beta},1/2,(q-1)/2}\big(t^2/G^2\big)\big] \cdot N\big)\big) \leq \alpha, \tag{2.16}$$

where $F_{\text{Beta},1/2,(q-1)/2}$ is the cumulative distribution function of the Beta$(1/2,(q-1)/2)$ distribution, and where G^2 follows a chi-squared distribution with q degrees of freedom. The quantity $B_\alpha(q,N)$ corresponds to the quantity K_4 of Bachoc, Leeb and Pötscher (2019) in the known variance case (for a discussion of numerical algorithms for obtaining $B_\alpha(q,N)$ in practice we confer the reader to that reference). We have (Berk et al. (2013), Bachoc, Leeb and Pötscher (2019)) that $B_\alpha(q,N)$ is larger than all the $1-\alpha$ quantiles of random variables of the form $\max_{i=1,\ldots,N}|v_i'\epsilon|$, where $v_1,\ldots,v_N$ are column vectors of $\mathbb{R}^q$ with $\|v_i\|\leq 1$ and where $\epsilon\sim N(0,I_q)$; furthermore, for fixed α and N the function $q\mapsto B_\alpha(q,N)$ is monotonically increasing.

Asymptotic approximations of $B_\alpha(q,N)$ for large q and N are provided in Berk et al. (2013), Zhang (2017) and Bachoc, Leeb and Pötscher (2019). In particular, as $q,N\to\infty$, $B_\alpha(q,N)[q(1-N^{-2/(q-1)})]^{-1/2}\to 1$, from Proposition 2.10 in Bachoc, Leeb and Pötscher (2019), itself building on results from Berk et al. (2013) and Zhang (2017).

An often useful upper bound on $K_{1-\alpha}(\Gamma)$ with Γ a $k\times k$-dimensional correlation matrix is provided in the following lemma:

LEMMA 2.8. *For every $\alpha\in(0,1)$ and every $k\times k$ correlation matrix Γ we have $K_{1-\alpha}(\Gamma)\leq B_\alpha(\text{rank}(\Gamma),k)$.*

In a particular application, it might of course be possible to obtain better upper bounds by exploiting structural properties of the specific correlation matrix Γ at hand; cf. Section 3.1. Using the upper bound of Lemma 2.8 can also be very useful in situations where the complexity of computing $K_{1-\alpha}(\Gamma)$ is prohibitive (see Section B.1.3 in the Supplementary Material (Bachoc, Preinerstorfer and Steinberger (2019)) for an example).

3. Applications. In this section, we now apply the general results obtained in Section 2 to some important special cases that are frequently encountered in practice. We consider situations of the following type:

1. The underlying distribution $\mathbb{P}_n$ is assumed to be an element of a set of distributions $\mathbf{P}_n$.
2. A model $\hat{\mathbb{M}}_n$ is selected in a data-driven way from a candidate set M_n, which is potentially misspecified, that is, $\mathbf{P}_n\not\subseteq\bigcup_{\mathbb{M}\in\mathsf{M}_n}\mathbb{M}$.
3. One aims at constructing confidence intervals for all coordinates of the model-specific target parameter $\theta^*_{\hat{\mathbb{M}}_n,n}$.

The scenarios we discuss in this section are all concerned with the case $\ell=1$,[1] that is, one observes a realization of a random n-vector $Y_n=(Y_{1,n},\ldots,Y_{n,n})'$ defined on some probability space $(\Omega,\mathcal{A},\mathbb{P})$, whose distribution under $\mathbb{P}$ coincides with $\mathbb{P}_n\in\mathbf{P}_n$ (we write $\mathbb{E}$ and $\mathbb{V}$ to denote the expectation and variance-covariance operator with respect to $\mathbb{P}$). In Section 3.1, we consider the case where the candidate set M_n consists of fixed design homoskedastic linear models. In this framework, the model selection problem is equivalent to a subset-selection problem of regressors. Here, the model-specific target we consider is the coefficient vector of the projection of the mean vector $\mu_n=\mathbb{E}(Y_n)\in\mathbb{R}^n$ onto the model-specific fixed regressor matrix. In such a setup, confidence intervals post-model-selection have also been suggested in Tibshirani et al. (2018), but for specific model selection methods. Our approach can also be used to obtain confidence intervals in their setup, and requires less assumptions on the

[1]The case $\ell>1$ is of interest, for example, in a regression problem with random design where one observes a data matrix $(y_i,x_{i1},\ldots,x_{ip})_{i=1}^n$ which is a realization of a probability distribution $\mathbb{P}_n$ on the sample space $\mathbb{R}^{n\times(p+1)}$.

set of distributions over which uniformity is achieved and on the design matrices allowed. In Section 3.2, we then discuss the case where M_n consists of fixed design heteroskedastic linear models. While the model-specific target is the same as in the homoskedastic case, the construction of confidence sets is more complicated as the heteroskedasticity needs to be taken into account. The results of this section can be viewed as an extension of the influential results of Eicker (1967) (see also White (1980)) to the potentially misspecified, post-model-selection context. Comparable results do not exist to the best of our knowledge. Finally, in Section 3.3, we consider the situation where M_n consists of binary regression models. We allow for situations where both the regressors and the link function are chosen in a data-driven way. In each candidate model, the model-specific target vector is here obtained as a minimizer of the Kullback–Leibler divergence. For numerical results concerning the methods discussed in Sections 3.1 and 3.3 see Section 4 as well as Section B of the Supplementary Material.

3.1. *Inference post-model-selection when fitting fixed design linear models to homoskedastic data.* One important application of our general theory is the case where homoskedastic linear regression models are fit to data. The feasible set for the true underlying distribution $\mathbb{P}_n$ we can allow for in this setup is denoted as $\mathbf{P}_n^{(\mathrm{lm})}(\delta,\tau)$, where $\delta>0$ and $\tau\geq 1$, and is defined as follows: the distribution $\mathbb{P}_n$ of the random n-vector $Y_n=(Y_{1,n},\ldots,Y_{n,n})'$ is an element of $\mathbf{P}_n^{(\mathrm{lm})}(\delta,\tau)$ if and only if the n coordinates of Y_n are independent, homoskedastic (i.e., the variances of the coordinates are equal to some $\mathbb{V}(Y_{i,n})=\sigma_n^2\in(0,\infty)$, for all $i=1,\ldots,n$), and

$$\max_{i=1,\ldots,n}\big[\mathbb{E}\big(\big|Y_{i,n}-\mathbb{E}(Y_{i,n})\big|^{2+\delta}\big)\big]^{\frac{2}{2+\delta}}\leq\tau\sigma_n^2.$$

Note that $\mathbf{P}_n^{(\mathrm{lm})}(\delta,\tau)$ is empty for $\delta>0$ and $\tau<1$, because then the inequality in the previous display can never be satisfied. Furthermore, observe that $\mathbf{P}_n^{(\mathrm{lm})}(\delta,\tau)$ contains the set of n-variate spherical normal distributions with unrestricted mean vector if $\mathbf{\Gamma}((3+\delta)/2)\leq(\tau/2)^{1+\delta/2}\sqrt{\pi}$, where $\mathbf{\Gamma}(\cdot)$ denotes the Gamma-function. For such a pair (δ,τ) the set $\mathbf{P}_n^{(\mathrm{lm})}(\delta,\tau)$ thus contains the Gaussian model considered in Berk et al. (2013). Finally, note that there is no restriction on the mean vector $\mu_n=\mathbb{E}(Y_n)\in\mathbb{R}^n$ of elements of $\mathbf{P}_n^{(\mathrm{lm})}(\delta,\tau)$.

We are interested in a situation where one works with candidate sets consisting of homoskedastic linear models. That is, a situation where one wants to conduct inference on the mean vector μ_n of the underlying distribution $\mathbb{P}_n$, and it is *assumed* by the practitioner that μ_n is an element of $\mathrm{span}(X_n)$, the column span of a design matrix $X_n\in\mathbb{R}^{n\times p}$, with p not depending on n, or it is *assumed* that μ_n is at least "well-approximated" by an element of that linear space; and that the practitioner *knows* (and takes into account in the construction of the confidence sets) that the observations have identical variances (for a situation where the observations are heteroskedastic see Section 3.2). In such a situation one then often tries to decide in a data-driven way *which* regressors to use, that is, one needs to solve a subset-selection problem. We assume that we are given a nonempty set $\mathcal{I}=\{M_1,\ldots,M_d\}$ of nonempty subsets of $\{1,2,\ldots,p\}$, that does not depend on n. Given $M\in\mathcal{I}$ we shall denote by $X_n[M]$ the matrix obtained from X_n by striking all columns whose index is not an element of M. We then consider for each $j\in\{1,\ldots,d\}$ a linear, homoskedastic candidate model $\mathbb{M}_{j,n}$ with fixed design $X_n[M_j]$, that is, the distribution of a random vector $z=(z_1,\ldots,z_n)'$ is an element of $\mathbb{M}_{j,n}$ if and only if there exists a $\beta\in\mathbb{R}^{|M_j|}$ such that the random (residual) vector $z-X_n[M_j]\beta$ has independent, homoskedastic coordinates with mean zero. Our candidate set of models is then given by $\mathsf{M}_n=\{\mathbb{M}_{j,n}:j=1,\ldots,d\}$.

We assume that X_n satisfies the following condition, where we denote the ith row of X_n by $X_{i,n}$:

CONDITION X1. Eventually $\operatorname{rank}(X_n) = p$, and for every $M \in \mathcal{I}$,

$$\max_{i=1,\dots,n} X_{i,n}[M]\big(X_n[M]'X_n[M]\big)^{-1}X_{i,n}[M]' \to 0. \tag{3.1}$$

REMARK 3.1. Condition X1 particularly holds if $\operatorname{rank}(X_n) = p$, eventually, and $\max_{i=1,\dots,n} X_{i,n}(X_n'X_n)^{-1}X_{i,n}' \to 0$. Moreover, it also holds in case $\|X_{i,n}\|$ is bounded and $\lambda_{\min}(\frac{1}{n}X_n'X_n)$ is bounded away from 0, which is typically the case in sufficiently balanced factorial designs, but Condition X1 is obviously much more general. For example, it also covers the important cases of polynomial regressors, trigonometric regressors, or mixed polynomial and trigonometric regressors (cf. the discussion in Eicker (1967), page 64). Finally, we point out that the condition in equation (3.1) is classical, and is necessary for asymptotic normality of the ordinary least-squares estimator in the fixed model $\mathbb{M}$ (see Huber (1973), Arnold (1980)).

The model-specific target of inference is then (eventually) defined as follows: Given $\mathbb{M} \in \mathsf{M}_n$ with a corresponding index set M, we let

$$\beta^*_{\mathbb{M},n} = \beta^*_{\mathbb{M},n}(\mathbb{P}_n) = \big(X_n[M]'X_n[M]\big)^{-1}X_n[M]'\mu_n, \tag{3.2}$$

that is, $\beta^*_{\mathbb{M},n}$ is the coefficient vector corresponding to the orthogonal projection of μ_n onto $\operatorname{span}(X_n[M])$.

We shall now describe how asymptotically uniformly valid confidence sets can be constructed post-model-selection for the target defined in equation (3.2) above: Given $\mathbb{M} \in \mathsf{M}_n$ with index set M, we estimate the corresponding target by the model-specific ordinary least-squares estimator, that is, by

$$\hat\beta_{\mathbb{M},n}(y) = \big(X_n[M]'X_n[M]\big)^{-1}X_n[M]'y; \tag{3.3}$$

let

$$\hat\sigma^2_{\mathbb{M},n}(y) = \frac{1}{n - m(\mathbb{M})}\sum_{i=1}^{n}\big(y_i - X_{i,n}[M]\hat\beta_{\mathbb{M},n}(y)\big)^2,$$

where $m(\mathbb{M})$ here coincides with $|M|$, the cardinality of M, and define for $\alpha \in (0,1)$ and $j = 1,\dots,m(\mathbb{M})$

$$\mathrm{CI}^{(j),\mathrm{lm}}_{1-\alpha,\mathbb{M}} = \hat\beta^{(j)}_{\mathbb{M},n} \pm \sqrt{\hat\sigma^2_{\mathbb{M},n}\big[\big(X_n[M]'X_n[M]\big)^{-1}\big]_j}\,K_{1-\alpha}\big(\operatorname{corr}(\Gamma_n)\big), \tag{3.4}$$

where the block-matrix Γ_n is defined via its (s,t)th block of dimension $|M_s| \times |M_t|$ given by

$$\begin{aligned}&\mathbb{E}_n\big[(\hat\beta_{\mathbb{M}_s,n} - \beta^*_{\mathbb{M}_s,n})(\hat\beta_{\mathbb{M}_t,n} - \beta^*_{\mathbb{M}_t,n})'\big]\\ &\quad= \sigma_n^2\big(X_n[M_s]'X_n[M_s]\big)^{-1}X_n[M_s]'X_n[M_t]\big(X_n[M_t]'X_n[M_t]\big)^{-1},\end{aligned}$$

for $s,t \in \{1,\dots,d\}$. Note that while Γ_n depends on σ_n^2, $\operatorname{corr}(\Gamma_n)$ is observed. Essentially, the construction in (3.4) coincides with the confidence intervals of Berk et al. (2013). However, there are two major differences. First of all, we here do not assume that the data are Gaussian, which is why we resort to asymptotic results. This is also the reason why our constant $K_{1-\alpha}$, the so called POSI constant, is the quantile of a maximum of Gaussian rather than t-distributed random variables, as is the case in Berk et al. (2013). Furthermore, we simply use the usual variance estimator $\hat\sigma^2_{\mathbb{M},n}$ which, in general, is not unbiased or uniformly consistent (due to potential misspecification) as required in Berk et al. (2013), but we still obtain uniformly valid inference asymptotically. This shows that the restrictive assumption of Berk et al. (2013), that there exists an unbiased or a uniformly consistent estimator for σ_n^2 (cf.

Proposition A.3 in the Supplementary Material, as well as the discussion in Remark 2.1 of Leeb, Pötscher and Ewald (2015) and in Appendix A of Bachoc, Leeb and Pötscher (2019)), is not needed for uniform asymptotic validity. If the estimator $\hat{\sigma}^2_{\mathbb{M},n}$ is used in the construction of Berk et al. (2013), then their confidence intervals asymptotically coincide with our procedure. We also point out that the classical variance estimator used here adapts to misspecification in the sense that it is consistent for σ_n^2 if a first order correct model is selected and it otherwise overestimates the target in the sense of Section 2.5.2 (cf. Remark 2.7, and Section A.2.1 of the Supplementary Material, where it is also shown that uniformly consistent estimators for σ_n^2 do *not* exist).

It is also worth noting that up to the choice of the last multiplicative factor $K_{1-\alpha}(\mathrm{corr}(\Gamma_n))$ in the definition of the confidence intervals above, that is, the POSI constant, this is just the usual confidence interval for the jth coordinate of the coefficient vector one would typically use in practice working with homoskedastic linear models, and by following the naive way of ignoring the data-driven model selection step. The crucial difference, however, is that the naive approach is invalid (see, e.g., Leeb, Pötscher and Ewald (2015), Bachoc, Leeb and Pötscher (2019)).

We now present the main result of this subsection, where we emphasize once more that the (measurable) model selection procedure $\hat{\mathbb{M}}_n$ is data-driven and unrestricted, and that some, or all of the candidate models in M_n may be misspecified,that is, $\mathbf{P}_n^{(\mathrm{lm})}(\delta,\tau) \not\subseteq \bigcup_{\mathbb{M}\in\mathsf{M}_n}\mathbb{M}$. Nevertheless it is possible to construct an asymptotically uniformly valid confidence set for the model-specific target vector $\beta^*_{\hat{\mathbb{M}}_n,n}$.

THEOREM 3.2. *Let $\alpha \in (0,1)$, $\delta > 0$ and $\tau \geq 1$, suppose Condition X1 holds, and let $\hat{\mathbb{M}}_n$ be a (measurable) model selection procedure, that is, a measurable map from the sample space $\mathbb{R}^n$ to M_n. Then*

$$\liminf_{n\to\infty} \inf_{\mathbb{P}_n\in\mathbf{P}_n^{(\mathrm{lm})}(\delta,\tau)} \mathbb{P}_n\big(\beta^{*(j)}_{\hat{\mathbb{M}}_n,n} \in \mathrm{CI}^{(j),\mathrm{lm}}_{1-\alpha,\hat{\mathbb{M}}_n} \text{ for all } j=1,\ldots,m(\hat{\mathbb{M}}_n)\big) \geq 1-\alpha.$$

The statement in Theorem 3.2 concerns simultaneous coverage of all coefficients of the model-dependent target parameter. In some applications, it may be of interest to construct confidence intervals only for single coefficients, that is, coefficients corresponding to a certain regressor. Of course, as simultaneous coverage implies individual coverage, the confidence intervals in the previous section achieve this goal a fortiori. However, shorter confidence intervals can be constructed if one only wants to achieve individual coverage. This is discussed in detail in Section A.2.2 of the Supplementary Material (Bachoc, Preinerstorfer and Steinberger (2019)).

3.2. *Inference post-model-selection when fitting fixed design linear models to heteroskedastic data.* The feasible sets for $\mathbb{P}_n$ we consider here again depend on two parameters $\delta > 0$ and $\tau \geq 1$ but, compared to the set $\mathbf{P}_n^{(\mathrm{lm})}(\delta,\tau)$ defined above, we now drop the requirement of homoskedasticity: the distribution of a random n-vector $Y_n = (Y_{1,n},\ldots,Y_{n,n})'$ is an element of $\mathbf{P}_n^{(\mathrm{het})}(\delta,\tau)$ if and only if the n coordinates of Y_n are independent, the variance $\sigma^2_{i,n} = \mathbb{V}(Y_{i,n}) \in (0,\infty)$ exists for every $i=1,\ldots,n$, and

$$\max_{i=1,\ldots,n}\big[\mathbb{E}\big(|Y_{i,n}-\mathbb{E}(Y_{i,n})|^{2+\delta}\big)\big]^{\frac{2}{2+\delta}} \leq \tau \min_{i=1,\ldots,n}\sigma^2_{i,n}.$$

Here, we consider a situation where one works with candidate sets consisting of heteroskedastic linear models, that is, where similar as in Section 3.1 one is interested in conducting inference on $\mu_n = \mathbb{E}(Y_n) \in \mathbb{R}^n$, and it is *assumed* that μ_n is (well approximated by) an element

of span(X_n), the column span of a design matrix $X_n \in \mathbb{R}^{n\times p}$ with p fixed; but where it is now taken into account that the observations may have different variances. We start with a set $\mathcal{I} = \{M_1, \dots, M_d\}$ as in Section 3.1, and we then define for each $j \in \{1, \dots, d\}$ the linear, heteroskedastic model $\mathbb{M}_{j,n}$ as follows: the distribution of a random vector $z = (z_1, \dots, z_n)'$ is an element of $\mathbb{M}_{j,n}$ if and only if there exists a $\beta \in \mathbb{R}^{|M_j|}$ such that the random (residual) vector $z - X[M_j]\beta$ has independent coordinates with positive finite variances and mean zero. The corresponding candidate set of models is then given by $\mathsf{M}_n = \{\mathbb{M}_{j,n} : j = 1, \dots, d\}$.

As in Section 3.1 we assume that X_n satisfies Condition X1, and define our model-specific target of inference as in equation (3.2). Again, we estimate the corresponding target by the model-specific ordinary least-squares estimator in (3.3). For variance estimation, we do no longer use the estimator as defined in Section 3.1, but now take into consideration, that the observations may be heteroskedastic. Therefore, we consider an approach based on estimators suggested by Eicker (1967). As in Section 3.1, the variance estimators used here are not uniformly consistent due to potential model misspecification, but overestimate their targets in the sense of Section 2.5.2. Furthermore, in contrast to the construction of Section 3.1, the construction of the confidence sets now needs to incorporate an upper bound for the POSI constant $K_{1-\alpha}(\mathrm{corr}(\Gamma_n))$, because here $\Gamma_n = \mathbb{V}_n[(\hat\beta'_{\mathbb{M}_1,n}, \dots, \hat\beta'_{\mathbb{M}_d,n})']$, and also $\mathrm{corr}(\Gamma_n)$, is unobserved and can not be estimated consistently due to potential misspecification. Define for every $\mathbb{M} \in \mathsf{M}_n$ with corresponding index set M the Eicker-estimator $\tilde S_{\mathbb{M},n}$ as

$$\big(X_n[M]'X_n[M]\big)^{-1} X_n[M]' \,\mathrm{diag}\big(\hat u^2_{1,\mathbb{M}}, \dots, \hat u^2_{n,\mathbb{M}}\big) X_n[M] \big(X_n[M]'X_n[M]\big)^{-1},$$

where, for $y \in \mathbb{R}^n$, we let $\hat u_{\mathbb{M}}(y) = (\hat u_{1,\mathbb{M}}(y), \dots, \hat u_{n,\mathbb{M}}(y))' = y - X_n[M]\hat\beta_{\mathbb{M},n}(y)$, and denote the jth diagonal entry ($j = 1, \dots, m(\mathbb{M})$) of $\tilde S_{\mathbb{M},n}$ by $\hat\sigma^2_{j,\mathbb{M},n}$. Finally, given $\alpha \in (0,1)$, we define for each $\mathbb{M} \in \mathsf{M}_n$ with corresponding index set M and for every $j = 1, \dots, m(\mathbb{M})$ the confidence sets

$$\mathrm{CI}^{(j),\mathrm{hlm}}_{1-\alpha,\mathbb{M}} = \hat\beta^{(j)}_{\mathbb{M},n} \pm \sqrt{\hat\sigma^2_{j,\mathbb{M},n}}\, B_\alpha\big(\min(k,p),k\big),$$

with $k = \sum_{\mathbb{M}\in\mathsf{M}_n} m(\mathbb{M})$, and where B_α is defined at the end of Section 2.5.2.

Note, similarly as in Section 3.1 above, that up to the choice of the last multiplicative factor $B_\alpha(\min(k,p),k)$, an upper bound for the corresponding POSI-constant, this is just the usual confidence interval for the jth coordinate of the coefficient vector one would typically use in practice working with heteroskedastic linear models by following the naive way of ignoring the data-driven model selection step. Our construction delivers an adjustment to that approach, which turns it, regardless of the (measurable) model selection procedure applied, into an asymptotically valid statistical procedure. The main result of this subsection is as follows.

THEOREM 3.3. *Let $\alpha \in (0,1)$, $\delta > 0$ and $\tau \ge 1$, suppose Condition X1 holds, and let $\hat{\mathbb{M}}_n$ be a (measurable) model selection procedure, that is, a measurable map from the sample space $\mathbb{R}^n$ to M_n. Then*

$$\liminf_{n\to\infty} \inf_{\mathbb{P}_n \in \mathbf{P}^{(\mathrm{het})}_n(\delta,\tau)} \mathbb{P}_n\big(\beta^{*(j)}_{\hat{\mathbb{M}}_n,n} \in \mathrm{CI}^{(j),\mathrm{hlm}}_{1-\alpha,\hat{\mathbb{M}}_n} \text{ for all } j = 1, \dots, m(\hat{\mathbb{M}}_n)\big) \ge 1-\alpha.$$

3.3. *Inference post-model-selection when fitting binary regression models to binary data.* The feasible sets $\mathbf{P}^{(\mathrm{bin})}_n(\tau)$ for $\mathbb{P}_n$ we consider here depend on a parameter $\tau \in (0, 1/4)$ and are defined as follows: the distribution of a random vector $Y_n = (Y_{1,n}, \dots, Y_{n,n})'$ is an element of $\mathbf{P}^{(\mathrm{bin})}_n(\tau)$ if and only if the n coordinates of Y_n are independent, each coordinate $Y_{i,n}$ takes on either 0 or 1, and $\mathbb{V}(Y_{i,n}) \ge \tau$. We consider a situation where binary regression

models are fit to binary data generated under one of the elements $\mathbb{P}_n \in \mathbf{P}_n^{(\mathrm{bin})}(\tau)$. It is important to point out, however, that unlike other work on misspecified binary regression (e.g., Ruud (1983), Kubkowski and Mielniczuk (2017)), we here do not assume that the true data generating process $\mathbb{P}_n$ is itself a binary regression model, but we consider the nonparametric case where every observation $Y_{i,n}$ may have its own success rate $p_{i,n} = \mathbb{P}(Y_{i,n} = 1)$, with the only restriction that $\mathbb{V}(Y_{i,n}) = p_{i,n}(1 - p_{i,n}) \geq \tau$. In binary regression, the maintained *modeling assumption* is that the probability of a success on the ith observation $(Y_{i,n} = 1)$, or equivalently its expectation, is given by $h(X_{i,n}\beta)$, for some $\beta \in \mathbb{R}^p$, some response function $h : \mathbb{R} \to (0, 1)$ and where $X_{i,n}$ is the ith row of a design matrix $X_n \in \mathbb{R}^{n\times p}$. Usually, when h is invertible, h^{-1} is called the link function. Thus, unlike the previous two examples, here we also have to make a choice for the response function h, in addition to selecting variables from X_n. Classical choices are the logit and the probit functions, but we allow also for other choices of response functions h, as long as they belong to a finite set $\mathcal{H} = \{h_1, \ldots, h_{d_1}\}$ of potential candidates, that does not depend on n. Together with the collection $\mathcal{I} = \{M_1, \ldots, M_{d_2}\} \subseteq 2^{\{1,\ldots,p\}} \setminus \varnothing$ of candidate regressor subsets, we can define for every $j_1 \in \{1, \ldots, d_1\}$ and $j_2 \in \{1, \ldots, d_2\}$ a candidate binary regression model $\mathbb{M}_{(j_1,j_2),n}$ as follows: the distribution of a random vector $z = (z_1, \ldots, z_n)'$ is an element of $\mathbb{M}_{(j_1,j_2),n}$ if and only if the n coordinates of z are independent, each coordinate z_i takes on either 0 or 1, and there exists a $\beta \in \mathbb{R}^{|M_{j_2}|}$ such that the mean of z_i equals $h_{j_1}(X_{i,n}[M_{j_2}]\beta)$ for $i = 1, \ldots, n$. Thus, our candidate set of size $d = d_1 \cdot d_2$ is given by

$$\mathsf{M}_n = \big\{\mathbb{M}_{(j_1,j_2),n} : j_1 \in \{1, \ldots, d_1\}, j_2 \in \{1, \ldots, d_2\}\big\}.$$

We need to impose some regularity conditions on the possible response functions $h \in \mathcal{H}$ and the design X_n. The conditions are formally stated as Conditions X2 and H, respectively, in Section A.3 of the Supplementary Material, where additional discussion can be found.

Note that since the design matrix $X_n \in \mathbb{R}^{n\times p}$ is fixed, a candidate model $\mathbb{M} \in \mathsf{M}_n$ can be identified with a pair $\mathbb{M} \triangleq (h, M) \in \mathcal{H} \times \mathcal{I}$. Estimating the parameter $\beta \in \mathbb{R}^{|M|}$ of a candidate model $\mathbb{M} \in \mathsf{M}_n$ is usually done by numerically maximizing the likelihood. The (quasi-)log-likelihood function for model $\mathbb{M} \triangleq (h, M)$ can be expressed as

$$\ell_{\mathbb{M},n}(y, \beta) = \sum_{i=1}^{n} \big[y_i \phi_1\big(X_{i,n}[M]\beta\big) + (1 - y_i)\phi_2\big(X_{i,n}[M]\beta\big)\big],$$

where $\phi_1(\gamma) = \log h(\gamma)$ and $\phi_2(\gamma) = \log(1 - h(\gamma))$, and $y = (y_1, \ldots, y_n)' \in \{0, 1\}^n$, $\beta \in \mathbb{R}^{|M|}$. Whenever Condition H(iii) holds, we denote the matrix of negative second derivatives of $\ell_{\mathbb{M},n}$ by

$$H_{\mathbb{M},n}(y, \beta) = -\frac{\partial^2 \ell_{\mathbb{M},n}(y, \beta)}{\partial\beta\partial\beta'} = X_n[M]' D_{\mathbb{M},n}(y, \beta) X_n[M],$$

where $D_{\mathbb{M},n}(y, \beta)$ is a diagonal matrix with ith diagonal entry equal to

$$-y_i \ddot{\phi}_1\big(X_{i,n}[M]\beta\big) - (1 - y_i)\ddot{\phi}_2\big(X_{i,n}[M]\beta\big).$$

Note that under Conditions X2(i) and H(iii), $H_{\mathbb{M},n}(y, \beta)$ is positive definite.

As our target of inference we take the model dependent vector $\beta^*_{\mathbb{M},n} \in \mathbb{R}^{|M|}$ that maximizes the expected log-likelihood $\beta \mapsto \mathbb{E}_n[\ell_{\mathbb{M},n}(\cdot, \beta)]$ under the true data generating distribution $\mathbb{P}_n \in \mathbf{P}_n^{(\mathrm{bin})}(\tau)$. If $\beta^*_{\mathbb{M},n}$ exists, then it is easy to see that it also minimizes the Kullback–Leibler divergence between the true data generating distribution $\mathbb{P}_n$ and the class of distributions specified by the working model $\mathbb{M} \in \mathsf{M}_n$. Focusing on the Kullback–Leibler minimizer has a longstanding tradition in the misspecification literature dating back at least to Huber (1967) (see also White (1982) and the references given therein). For references more specific

to generalized linear models see Fahrmeir (1990) and Lv and Liu (2014). That this target uniquely exists in the present context of binary regression is the subject of the following lemma.[2]

LEMMA 3.4. *Suppose that* $\operatorname{rank}(X_n) = p$ *and* H(i), (ii) *hold. Then, for every* $\mathbb{M} \in \mathsf{M}_n$ *and for every* $\mathbb{P}_n \in \bigcup_{\delta>0} \mathbf{P}_n^{(\mathrm{bin})}(\delta)$, *there exists a unique vector* $\beta^*_{\mathbb{M},n} = \beta^*_{\mathbb{M},n}(\mathbb{P}_n) \in \mathbb{R}^{m(\mathbb{M})}$, *such that*

$$\int_{\mathbb{R}^n} \ell_{\mathbb{M},n}\big(y, \beta^*_{\mathbb{M},n}(\mathbb{P}_n)\big)\, d\mathbb{P}_n(y) = \sup_{\beta \in \mathbb{R}^{m(\mathbb{M})}} \int_{\mathbb{R}^n} \ell_{\mathbb{M},n}(y, \beta)\, d\mathbb{P}_n(y).$$

Furthermore, it is well known that for some points in the sample space $\{0,1\}^n$ the MLE in the binary regression model does not exist (see, e.g., Wedderburn (1976)). But those samples have vanishing asymptotic probability. The following lemma establishes this asymptotic existence of the (quasi-) MLE $\hat\beta_{\mathbb{M},n}$ in the present setting, along with uniform consistency. Its proof is deferred to Section E.5 of the Supplementary Material (Bachoc, Preinerstorfer and Steinberger (2019)).

LEMMA 3.5. *Suppose that Conditions* X2(i), (ii) *and* H(i), (ii), (iii) *hold and fix* $\tau \in (0, 1/4)$. *Then, for every* $n \in \mathbb{N}$, *every* $\mathbb{M} \in \mathsf{M}_n$ *and every* $\mathbb{P}_n \in \mathbf{P}_n^{(\mathrm{bin})}(\tau)$, *there exists a function* $\hat\beta_{\mathbb{M},n} : \{0,1\}^n \to \mathbb{R}^{m(\mathbb{M})}$ (*depending only on* n *and* $\mathbb{M}$) *and a set* $E_{\mathbb{M},\mathbb{P}_n,n} \subseteq \{0,1\}^n$, *such that*

$$\ell_{\mathbb{M},n}\big(y, \hat\beta_{\mathbb{M},n}(y) + \beta\big) < \ell_{\mathbb{M},n}\big(y, \hat\beta_{\mathbb{M},n}(y)\big) \quad \forall y \in E_{\mathbb{M},\mathbb{P}_n,n}, \forall \beta \neq 0$$

and

$$\inf_{\mathbb{M}\in\mathsf{M}_n} \inf_{\mathbb{P}_n \in \mathbf{P}_n^{(\mathrm{bin})}(\tau)} \mathbb{P}_n(E_{\mathbb{M},\mathbb{P}_n,n}) \xrightarrow[n\to\infty]{} 1.$$

Moreover, for the pseudo parameter $\beta^*_{\mathbb{M},n} \in \mathbb{R}^{m(\mathbb{M})}$ *of Lemma* 3.4, *we have*

$$\limsup_{n\to\infty} \sup_{\substack{\mathbb{M}\in\mathsf{M}_n \\ \mathbb{P}_n \in \mathbf{P}_n^{(\mathrm{bin})}(\tau)}} \mathbb{P}_n\Big(\big\|\big(X_n[M]' X_n[M]\big)^{1/2}\big(\hat\beta_{\mathbb{M},n} - \beta^*_{\mathbb{M},n}(\mathbb{P}_n)\big)\big\| > \delta\Big) \to 0,$$

as $\delta \to \infty$.

To construct asymptotically valid confidence intervals for the components of $\beta^*_{\mathbb{M},n}$, we need an estimate of the asymptotic covariance matrix of $\hat\beta_{\mathbb{M},n}$. In the misspecified setting, it is usually not possible to obtain a consistent estimator. We here follow the suggestion of Fahrmeir ((1990), page 491) who proposed a sandwich-type estimator for misspecified generalized linear models. This estimator fits with the general idea of Section 2.5.2. For $\mathbb{M} \in \mathsf{M}_n$, $\mathbb{M} \triangleq (h, M)$, define

$$\tilde S_{\mathbb{M},n} = \hat H_{\mathbb{M},n}^{-1} X_n[M]' \operatorname{diag}\big(\hat u_{1,\mathbb{M}}^2, \ldots, \hat u_{n,\mathbb{M}}^2\big) X_n[M] \hat H_{\mathbb{M},n}^{-1}, \tag{3.5}$$

where $\hat H_{\mathbb{M},n}(y) = H_{\mathbb{M},n}(y, \hat\beta_{\mathbb{M},n}(y))$,

$$\hat u_{i,\mathbb{M}}(y) = \frac{\dot h(\hat\gamma_{i,n,M}(y))}{h(\hat\gamma_{i,n,M}(y))(1 - h(\hat\gamma_{i,n,M}(y)))}\big(y_i - h\big(\hat\gamma_{i,n,M}(y)\big)\big)$$

[2]A similar claim is made in Theorem 5 of Lv and Liu (2014) and its proof is deferred to Version 1 of the arXiv preprint Lv and Liu (2010), where it appears to be the case that the *existence* issue has been ignored. For a complete proof of our Lemma 3.4 see Section E.3 of the Supplementary Material.

and $\hat{\gamma}_{i,n,M}(y) = X_{i,n}[M]\hat{\beta}_{\mathbb{M},n}(y)$, and denote the jth diagonal entry ($j = 1, \dots, m(\mathbb{M})$) of $\tilde{S}_{\mathbb{M},n}$ by

$$\hat{\sigma}^2_{j,\mathbb{M},n}. \tag{3.6}$$

Finally, given $\alpha \in (0,1)$, we define for each $\mathbb{M} \in \mathsf{M}_n$ and for every $j = 1, \dots, m(\mathbb{M})$ the confidence sets

$$\mathrm{CI}^{(j),\mathrm{bin}}_{1-\alpha,\mathbb{M}} = \hat{\beta}^{(j)}_{\mathbb{M},n} \pm \sqrt{\hat{\sigma}^2_{j,\mathbb{M},n}}\, B_\alpha\big(\min(k,n),k\big),$$

with $k = \sum_{\mathbb{M}\in\mathsf{M}_n} m(\mathbb{M})$, and where B_α is defined as in (2.16).

These confidence intervals have the same basic structure as in Section 3.2, in the sense that they use estimators $\hat{\sigma}^2_{j,\mathbb{M},n}$ for the asymptotic variances that consistently overestimate their respective target quantities and replace the usual Gaussian quantile by the correction constant $B_\alpha(\min(k,n),k)$ that adjusts for the effect of model selection. This leads to asymptotically valid inference post-model-selection, as stated in the following theorem.

THEOREM 3.6. *Let $\alpha \in (0,1)$ and $\tau \in (0, 1/4)$, suppose Conditions* X2 *and* H *hold, and let $\hat{\mathbb{M}}_n$ be a model selection procedure, that is, a map from the sample space $\{0,1\}^n$ to M_n. Then*

$$\liminf_{n\to\infty} \inf_{\mathbb{P}_n \in \mathbf{P}^{(\mathrm{bin})}_n(\tau)} \mathbb{P}_n\big(\beta^{*(j)}_{\hat{\mathbb{M}}_n,n} \in \mathrm{CI}^{(j),\mathrm{bin}}_{1-\alpha,\hat{\mathbb{M}}_n}\ \forall j = 1,\dots,m(\hat{\mathbb{M}}_n)\big) \geq 1-\alpha.$$

REMARK 3.7. It is important to note that if one decides a priori to use only the canonical link function, which, in the present case of binary regression, corresponds to the logistic response function $h^{(c)}(\gamma) := e^\gamma/(1+e^\gamma)$, then Theorem 3.6 holds with the POSI-constant $B_\alpha(\min(k,n),k)$ decreased to $B_\alpha(\min(k,p),k)$. See Corollary E.3 in Section E.7 of the Supplementary Material (Bachoc, Preinerstorfer and Steinberger (2019)).

4. Simulation study. In this section, we present the main findings of an extensive simulation study; see Section B of the Supplementary Material for details.

4.1. *Comparison with Tibshirani et al. (2018) and with the naive intervals.* We study the setting of Section 3.1, where linear models are fit to homoskedastic data. Furthermore, we address the well-specified case, where the true data generating process corresponds to one of the candidate models. We consider observations of $Y_n = X_n\beta + \sigma u$, where X_n is an $n \times p$ matrix (which will be randomly generated in the simulations), β is a $p \times 1$ vector, σ is positive and u is an $n \times 1$ vector with independent and identically distributed components which is also independent of X_n. We consider the least angle regression (LAR) model selector (Efron et al. (2004)) and compare the "POSI" confidence intervals of Section 3.1 with the "TG" (truncated Gaussian) intervals developed in Tibshirani et al. (2018), and with "naive" intervals that ignore the data driven model selection step. The TG intervals are specifically tailored for the LAR model selector. The model $\hat{\mathbb{M}}^{(k)}_n$ consists of those variables that were selected by running the LAR algorithm for k steps. This model selection setting has become a benchmark for post-model-selection inference simulation studies (Tibshirani et al. (2016, 2018)). As in Tibshirani et al. (2018), we seek inference for the variable that is selected in the final (kth) step of the LAR algorithm. More precisely, for $k = 1, 2, 3$, let $\hat{M}^{(k)}_n$ be the set of the k selected variables. Let $\beta^*_{\hat{\mathbb{M}}^{(k)}_n,n} = (X_n[\hat{M}^{(k)}_n]' X_n[\hat{M}^{(k)}_n])^{-1} X_n[\hat{M}^{(k)}_n]' X_n\beta$ (cf. (3.2) with $\mu_n = X_n\beta$). Then, for the model selector k, the *target of inference* is the $\hat{j}_k$th component of $\beta^*_{\hat{\mathbb{M}}^{(k)}_n,n}$, corresponding to the variable added at step k of the LAR algorithm

TABLE 1

Coverage proportion (cov.), median length (med.) and 90% quantile length (qua.) for the "POSI", "TG" and "naive" confidence intervals at nominal level $1-\alpha=0.9$. *The design matrix is generated with independent (upper half of the table) or correlated (lower half) columns. The errors u have normal (N), Laplace (L), uniform (U) or skewed normal (SN) distributions. For each setting, the "POSI" (resp. "TG", resp. "naive") intervals correspond to the first (resp. second, resp. third) row. The coverage proportions, median and quantile lengths are given for each of the three targets for the three first steps of the LAR algorithm. The last column provides the simultaneous coverage proportion of the three targets after step* 3

	Step 1			Step 2			Step 3			Simult.
u	Cov.	Med.	Qua.	Cov.	Med.	Qua.	Cov.	Med.	Qua.	Cov.
N	0.99	6.64	7.40	1.00	6.12	7.09	0.98	6.10	7.27	0.97
	0.88	5.50	20.54	0.91	11.09	55.08	0.90	24.82	130.31	0.76
	0.90	3.63	4.06	0.88	3.36	3.88	0.60	3.36	3.95	0.46
L	0.99	6.52	7.73	1.00	6.00	7.39	0.97	6.05	7.65	0.96
	0.90	5.26	19.57	0.91	10.39	50.47	0.89	24.97	123.77	0.74
	0.88	3.58	4.27	0.88	3.29	4.06	0.58	3.31	4.18	0.43
U	0.99	6.59	7.23	1.00	6.06	6.88	0.98	6.08	7.21	0.96
	0.88	5.26	17.35	0.90	11.72	70.71	0.88	24.92	154.58	0.74
	0.89	3.62	3.96	0.89	3.33	3.75	0.58	3.34	3.94	0.45
SN	1.00	6.53	7.52	0.98	5.98	7.11	0.98	5.99	7.23	0.96
	0.90	5.24	25.65	0.88	11.35	57.75	0.89	23.41	153.79	0.73
	0.90	3.59	4.13	0.87	3.30	3.90	0.57	3.31	3.94	0.44
N	0.99	6.50	7.39	1.00	8.31	13.27	1.00	11.32	16.87	0.99
	0.90	8.38	36.72	0.88	65.69	401.53	0.88	107.03	683.69	0.73
	0.71	3.31	3.78	0.82	4.24	6.81	0.83	5.77	8.55	0.49
L	0.98	6.36	7.65	1.00	8.24	13.04	1.00	11.12	16.50	0.98
	0.89	7.42	41.62	0.90	52.19	344.65	0.91	99.07	625.88	0.76
	0.69	3.25	3.90	0.84	4.18	6.73	0.82	5.66	8.42	0.49
U	0.98	6.48	7.04	1.00	8.22	12.75	1.00	11.23	16.73	0.98
	0.89	7.52	38.75	0.89	52.25	309.02	0.91	95.25	652.63	0.77
	0.71	3.30	3.60	0.83	4.19	6.46	0.82	5.73	8.53	0.49
SN	1.00	6.36	7.51	1.00	8.50	14.11	1.00	10.98	16.58	0.99
	0.91	7.90	35.24	0.92	57.76	338.34	0.90	97.81	639.70	0.78
	0.71	3.25	3.83	0.82	4.33	7.23	0.82	5.61	8.48	0.47

($\hat{j}_k \in \hat{M}_n^{(k)} \setminus \hat{M}_n^{(k-1)}$ with $\hat{M}_n^{(0)} = \varnothing$). This target is discussed in Berk et al. (2013), Bachoc, Leeb and Pötscher (2019) (in the general post-model-selection context) and in Tibshirani et al. (2016). In particular, if the target is zero, then adding the regressor obtained from the step k of the LAR procedure does not improve the approximation of the unknown mean $X_n\beta$ (compared to the approximation obtained from the regressors from the $k-1$ first steps of the LAR procedure).

We set $n=50$, $p=10$, $1-\alpha=0.9$ and repeat $N=500$ independent data generations, model selections and confidence interval computations. The setup is the same as in Tibshirani et al. (2018) (see Section B in the Supplementary Material for details). In Table 1, we report the coverage proportions, the median lengths and the 90% quantiles of the lengths for each of the nine procedures ("POSI", "TG" and "naive" for $k=1,2,3$), in different settings. We also report the proportions of times where the three targets corresponding to the regressors selected after step 3 of the LAR algorithm are simultaneously contained by the three respective confidence intervals.

The "POSI" confidence intervals always have target-specific and simultaneous coverage above the nominal level. The coverage proportions are large, which is so because these confi-

dence intervals offer strong guarantees: they are valid for any model selection procedure, and simultaneously over all the variables in the selected model. Turning to the "TG" confidence intervals, we observe that these intervals have coverage probabilities approximately equal to the nominal level when the three targets are considered separately, but their median lengths are often larger and never much smaller than the lengths of the "POSI" intervals. Furthermore, the 90% quantiles are always larger for the "TG" intervals, for which they can be very (and sometimes extremely) large. A theoretical explanation of this phenomenon was recently given by Kivaranovic and Leeb (2018). In contrast, the confidence intervals suggested in this paper are more robust, in the sense that their 90% quantile lengths are always less than twice as large as their median lengths. Finally, the "naive" intervals always have the smallest length, but can yield coverage proportions significantly below the nominal level. Hence, they are not valid.

The numerical results of Table 1 seem to favor the "POSI" confidence intervals suggested in this paper over the "TG" procedure. Indeed, we have seen that, even though the LAR model selector is used, the "POSI" confidence intervals have larger coverage proportions, remain valid when considered simultaneously, typically have smaller median lengths, and never exhibit very large quantile lengths. On top of this, the "POSI" confidence intervals are much more broadly applicable, as they have theoretical guarantees for any model selection procedure.

One needs to mention here that Tibshirani et al. (2018) also discuss a bootstrap version of their "TG" intervals. These bootstrap confidence intervals have similar coverage properties as the "TG" intervals, but much smaller median width. Their width seems to be comparable to the width of our "POSI" confidence intervals (cf. Tables 1 and 2 in Tibshirani et al. (2018)). The discussion of advantages of the "POSI" method over the "TG" intervals of the preceding paragraph (besides the comments concerning their smaller width) also applies to the "POSI" method and the bootstrapped "TG" intervals. Furthermore, this suggests that our "POSI" intervals could potentially be improved by using suitable bootstrap methods as well. However, answering this question goes beyond the scope of the present article.

4.2. *The case of 'significance hunting'.* In the same setting as above, we investigate a different model selection procedure which we call "significance hunting" and which is closely related to the SPAR procedure in Berk et al. (2013). We first sort all the possible candidate models $\mathbb{M} \in \mathsf{M}_n$ according to their penalized log-likelihood, and then select the model $\hat{\mathbb{M}}$ and index $\hat{j}$ that maximize the test statistics

$$|\hat{\beta}^{(j)}_{\mathbb{M},n}| / \sqrt{\hat{\sigma}^2_{\mathbb{M},n}\big[(X_n[M]'X_n[M])^{-1}\big]_j},$$

among the n_{best} models with largest penalized log-likelihood. The target of inference is now the $\hat{j}$th coordinate of $\beta^*_{\hat{\mathbb{M}},n}$. We set $n = 100$, $p = 5$, $1 - \alpha = 0.9$ and consider two settings for β. In the "zero" setting, we set $\beta = (0, \ldots, 0)'$. In the "nonzero" setting, we set $\beta = (2, -1, 0, 0, 1)'$. We consider the values $n_{\text{best}} = 5$ and $n_{\text{best}} = 20$. The errors are normally distributed. We consider the "POSI" and "naive" intervals (the "TG" ones are not designed for this setting). In Table 2, the coverage proportions are significantly lower than in Table 1, and closer to the nominal level for the "POSI" intervals. Hence, the confidence intervals suggested in this paper may have conservative coverage proportions for some model selection procedures (such as LAR) but this is somehow necessary, since there exist other model selection procedures (such as "significance hunting") for which the coverage proportions are close to the nominal level. Also, in Table 2 the coverage proportions of the "naive" intervals can become very small, even more so than in Table 1.

TABLE 2

Coverage proportion (cov.), median lengths (med.) and 90% quantile lengths (qua.) of the "POSI" (P) and "naive" (N) confidence intervals at level $1-\alpha=0.9$ for the "significance hunting" model selection procedure

		Cov.		Med.		Qua.	
n_{best}	β	P	N	P	N	P	N
20	zero	0.88	0.50	4.99	3.34	5.83	3.85
	nonzero	0.93	0.76	5.00	3.35	5.73	3.78
5	zero	0.92	0.59	4.87	3.25	5.35	3.55
	nonzero	0.94	0.80	4.94	3.30	5.41	3.58

4.3. *Further results*. In Section B.1 of the Supplementary Material, we provide all the details of the previous simulations as well as further discussions of the results. In Section B.13, we also provide the results of an additional simulation study (for the LAR model selector in linear regression) in the high-dimensional setting of Tibshirani et al. (2018) ($n=50$ and $p=1000$). We find that the "POSI" and "TG" intervals remain valid, that the length increase of the "POSI" intervals due to the high dimension is moderate, and that the "POSI" intervals are shorter than the "TG" ones. The "naive" intervals fail dramatically in this high-dimensional study.

In Section B.2 of the Supplementary Material, we also present simulations for the binary regression problem of Section 3.3, comparing our methods to a procedure suggested by Taylor and Tibshirani (2018) and to naive intervals. Furthermore, we investigate the effect of misspecification and we also consider the "significance hunting" procedure in the binary regression case. The overall picture is similar to the results for the linear model, with the additional aspect that the "POSI" intervals remain valid also under misspecification, whereas the coverage probabilities of the methods of, for example, Taylor and Tibshirani (2018) can be substantially below the nominal level.

5. Conclusion. We have presented a general theory for the construction of asymptotically valid confidence sets post-model-selection. Open questions that go beyond the scope of this article, but are currently under investigation, include the extension of the approach discussed here to dependent data; the applicability and performance of bootstrap procedures; and the theoretical study of procedures in the spirit of Berk et al. (2013) in the challenging situation when the number of models fit can grow with sample size.

Acknowledgements. We thank Hannes Leeb, Benedikt M. Pötscher and Ulrike Schneider for helpful comments. We are also grateful for the suggestions of two anonymous referees and an Associate Editor who helped to produce a considerably improved version of the paper.

The second author was supported by the Austrian Science Fund (FWF): P27398, the Danish National Research Foundation (grant DNRF 78, CREATES) and the Program of Concerted Research Actions (ARC) of the Université libre de Bruxelles.

The third author was supported by the Austrian Science Fund (FWF): P28233 and the German Research Foundation (DFG): RO 3766/401.

SUPPLEMENTARY MATERIAL

Supplement to "Uniformly valid confidence intervals post-model-selection" (DOI: 10.1214/19-AOS1815SUPP; .pdf). The supplement contains additional results, remarks and discussion, details of the simulations and all the proofs.

REFERENCES

ARNOLD, S. F. (1980). Asymptotic validity of F tests for the ordinary linear model and the multiple correlation model. *J. Amer. Statist. Assoc.* **75** 890–894. MR0600972

BACHOC, F., LEEB, H. and PÖTSCHER, B. M. (2019). Valid confidence intervals for post-model-selection predictors. *Ann. Statist.* **47** 1475–1504. MR3911119 https://doi.org/10.1214/18-AOS1721

BACHOC, F., PREINERSTORFER, D. and STEINBERGER, L. (2019). Supplement to "Uniformly valid confidence intervals post-model-selection." https://doi.org/10.1214/19-AOS1815SUPP.

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 https://doi.org/10.1093/restud/rdt044

BERK, R., BROWN, L., BUJA, A., ZHANG, K. and ZHAO, L. (2013). Valid post-selection inference. *Ann. Statist.* **41** 802–837. MR3099122 https://doi.org/10.1214/12-AOS1077

DUDLEY, R. M. (2002). *Real Analysis and Probability*. *Cambridge Studies in Advanced Mathematics* **74**. Cambridge Univ. Press, Cambridge. MR1932358 https://doi.org/10.1017/CBO9780511755347

EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–499. MR2060166 https://doi.org/10.1214/009053604000000067

EICKER, F. (1967). Limit theorems for regressions with unequal and dependent errors. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability* (*Berkeley, Calif.*, 1965/66) 59–82. Univ. California Press, Berkeley, CA. MR0214223

FAHRMEIR, L. (1990). Maximum likelihood estimation in misspecified generalized linear models. *Statistics* **21** 487–502. MR1087280 https://doi.org/10.1080/02331889008802259

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

GNEDENKO, B. V. and KOLMOGOROV, A. N. (1954). *Limit Distributions for Sums of Independent Random Variables*. Addison-Wesley, Cambridge, MA. MR0062975

HUBER, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability* (*Berkeley, Calif.*, 1965/66), *Vol. I*: *Statistics* 221–233. Univ. California Press, Berkeley, CA. MR0216620

HUBER, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *Ann. Statist.* **1** 799–821. MR0356373

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 https://doi.org/10.1198/016214505000001140

KIVARANOVIC, D. and LEEB, H. (2018). Expected length of post-model-selection confidence intervals conditional on polyhedral constraints. Preprint. Available at arXiv:1803.01665.

KUBKOWSKI, M. and MIELNICZUK, J. (2017). Active sets of predictors for misspecified logistic regression. *Statistics* **51** 1023–1045. MR3698499 https://doi.org/10.1080/02331888.2017.1290096

LEE, J. D. and TAYLOR, J. E. (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 https://doi.org/10.1214/15-AOS1371

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 https://doi.org/10.1017/S0266466603191050

LEEB, H. and PÖTSCHER, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory* **21** 21–59. MR2153856 https://doi.org/10.1017/S0266466605050036

LEEB, H. and PÖTSCHER, B. M. (2006). Performance limits for estimators of the risk or distribution of shrinkage-type estimators, and some general lower risk-bound results. *Econometric Theory* **22** 69–97. MR2212693 https://doi.org/10.1017/S0266466606060038

LEEB, H. and PÖTSCHER, B. M. (2008). Model selection. In *Handbook of Financial Time Series* (T. G. Andersen, R. A. Davis, J.-P. Kreiß and T. Mikosch, eds.) 785–821. Springer, New York, NY.

LEEB, H., PÖTSCHER, B. M. and EWALD, K. (2015). On various confidence intervals post-model-selection. *Statist. Sci.* **30** 216–227. MR3353104 https://doi.org/10.1214/14-STS507

LV, J. and LIU, J. S. (2010). Model selection principles in misspecified models. Preprint. Available at arXiv:1005.5483v1.

LV, J. and LIU, J. S. (2014). Model selection principles in misspecified models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 141–167. MR3153937 https://doi.org/10.1111/rssb.12023

POLLAK, M. (1972). A note on infinitely divisible random vectors. *Ann. Math. Stat.* **43** 673–675.

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

RAIKOV, D. (1938). On a connection between the central limit-law of the theory of probability and the law of great numbers. *Izv. Ross. Akad. Nauk Ser. Mat.* **2** 323–338.

RINALDO, A., WASSERMAN, L., G'SELL, M., LEI, J. and TIBSHIRANI, R. (2016). Bootstrapping and sample splitting for high-dimensional, assumption-free inference. Preprint. Available at arXiv:1611.05401.

RUUD, P. A. (1983). Sufficient conditions for the consistency of maximum likelihood estimation despite misspecification of distribution in multinomial discrete choice models. *Econometrica* **51** 225–228. MR0694460 https://doi.org/10.2307/1912257

TAYLOR, J. and TIBSHIRANI, R. (2018). Post-selection inference for ℓ_1-penalized likelihood models. *Canad. J. Statist.* **46** 41–61. MR3767165 https://doi.org/10.1002/cjs.11313

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 https://doi.org/10.1080/01621459.2015.1108848

TIBSHIRANI, R. J., RINALDO, A., TIBSHIRANI, R. and WASSERMAN, L. (2018). Uniform asymptotic inference and the bootstrap after model selection. *Ann. Statist.* **46** 1255–1287. MR3798003 https://doi.org/10.1214/17-AOS1584

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 https://doi.org/10.1214/14-AOS1221

WEDDERBURN, R. W. M. (1976). On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. *Biometrika* **63** 27–32. MR0408092 https://doi.org/10.1093/biomet/63.1.27

WHITE, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* **48** 817–838. MR0575027 https://doi.org/10.2307/1912934

WHITE, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50** 1–25. MR0640163 https://doi.org/10.2307/1912526

ZHANG, K. (2017). Spherical cap packing asymptotics and rank-extreme detection. *IEEE Trans. Inform. Theory* **63** 4572–4584. MR3666977 https://doi.org/10.1109/TIT.2017.2700202

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 https://doi.org/10.1111/rssb.12026