Open access • Journal Article • DOI:10.1093/BIOMET/ASY018

Asymptotic post-selection inference for the Akaike information criterion

Ali Charkhi, Gerda Claeskens

Institutions: Katholieke Universiteit Leuven

Published on: 01 Sep 2018 - Biometrika (Oxford Academic)

Topics: Akaike information criterion, Asymptotic distribution, Model selection, Selection (genetic algorithm) and Confidence region

Asymptotic post-selection inference for Akaike's information criterion

Ali Charkhi and Gerda Claeskens
Operations Research & Business Statistics, University of Leuven, Naamsestraat 69, 3000 Leuven, Belgium.
ali.charkhi@kuleuven.be, gerda.claeskens@kuleuven.be

Abstract

Ignoring the model selection step in inference after selection is harmful. This paper studies the asymptotic distribution of estimators after model selection using the Akaike information criterion. First, we consider the classical setting in which a true model exists and is included in the candidate set of models. We exploit the overselection property of this criterion in the construction of a selection region, and obtain the asymptotic distribution of estimators and linear combinations thereof conditional on the selected model. The limiting distribution depends on the set of competitive models and on the smallest overparameterized model. Second, we relax the assumption about the existence of a true model, and obtain uniform asymptotic results. We use simulation to study the resulting post-selection distributions and to calculate confidence regions for the model parameters. We apply the method to data.
Key words: Akaike information criterion; confidence region; likelihood model; model selection; post-selection inference.

1 Introduction

Variable selection, model selection and estimation with a sparsity-enforcing penalty all induce uncertainty due to the process of selection, and they complicate subsequent inference.

We investigate post-selection inference for the Akaike information criterion (Akaike, 1973). The method is valid for variable selection in any likelihood-based model. We construct confidence intervals for regression parameters, or linear combinations thereof, conditional on the selected model, that have the correct coverage probabilities. The method involves rewriting the event of selection asymptotically as a number of inequalities that involve multivariate normal random variables. While the calculation of critical values might proceed exactly for one or two parameters, we develop a numerical approach that applies more generally. We focus explicitly on the classical low-dimensional setting, for which no such post-selection results are yet available.

The need to address the selection uncertainty has been pointed out several times (e.g., Kabaila, 1995, 1998; Hjort & Claeskens, 2003; Leeb & Pötscher, 2003, 2005, 2006; Danilov & Magnus, 2004; Kabaila & Leeb, 2006). Claeskens & Hjort (2008) approached the post-selection issue via model averaging, by simulation in a local misspecification framework. For model selection via sequential testing in nested models, Pötscher (1991) calculated the asymptotic distribution of the parameter estimator. Several advances have recently been made. The post-selection inference method of Berk et al. (2013) results for linear models in valid confidence intervals irrespective of the selection procedure, which can also be informal. Bachoc et al. (2015) generalized this method to prediction intervals. Since these methods are not specific to any selection procedure, the resulting confidence intervals might be quite conservative. Efron (2014) proposed to use a bagging, bootstrap aggregation, estimator and derived its variance, using normal quantiles to obtain confidence intervals. Ferrari & Yang (2014) assessed model uncertainty when performing F-tests in linear models via a so-called variable selection confidence set. Kabaila et al.

(2016) investigated the exact coverage and scaled expected length of certain model-averaged confidence intervals for a parameter of a linear regression model.

In selective inference one lets the data determine the selected model and the target of the parameter estimators. For the lasso, Lee et al. (2016) obtain exact post-selection inference by relating the selected set of active coefficients to a union of polyhedra. For forward selection and least angle regression in normal linear regression models, Taylor et al. (2016) study selective hypothesis tests and confidence intervals. Jansen (2014) studied the effect of the optimization on the expected values of the Akaike information criterion and Mallow's C_p in high-dimensional sparse models. Belloni et al. (2015) obtained uniformly valid confidence intervals in the presence of a sparse high-dimensional nuisance parameter.

We explain the methodology first in the traditional simple case of selection using the Akaike information criterion in a sequence of nested model, the so-called order selection problem. Next, this is extended to the practically more relevant selection from a general set of models, not necessarily nested and possibly all misspecified. When a true parametric model exists, only pointwise results can be obtained, while under misspecification and working with pseudo-true values that change per model, stronger, uniformly valid confidence intervals are constructed.

2 Post-AIC-selection in nested models

2.1 Selection properties of the AIC

Consider first a nested sequence of $K+1$ likelihood models $M_0 \subseteq \cdots \subseteq M_K$, for which the likelihood function L_n depends on a parameter vector $\theta^\top = (\theta_0^\top, \theta_1, \ldots, \theta_K) \in \Omega \subseteq \mathbb{R}^{a+K}$, where $\theta_0 \in \mathbb{R}^a$ denotes the parameter vector that is common to all models and hence is not subject to variable selection and n denotes the sample size. For ease of notation we assume that model M_i adds a single parameter to model M_{i-1}. Generalizations are straightforward.

We start by assuming that there is a single minimal true model M_{p_0} in the set of models $\mathcal{M}_{\text{nest}} = \{M_i : i = 0, \ldots, K\}$ in the sense that p_0 is the smallest model order for which all non-zero components of the true parameter vector ϑ are included. This assumption is relaxed in Section 4, where we do not require the existence of a true model, we allow for non-nested models and for model misspecification. In the current setting, models with indices $i < p_0$ are underparametrized, while models with $i > p_0$ are overparametrized. We denote by $\hat{\theta}'(i)$ the maximum likelihood estimator for the parameter vector $\theta^\top(i) = (\theta_0^\top, \ldots, \theta_i) \in \mathbb{R}^{a+i}$ in model M_i, $\hat{\theta}(i) = (\hat{\theta}'(i)^\top, 0_{K-i}^\top)^\top$, and by $\vartheta = \vartheta(p_0)$ the corresponding true value where $\vartheta_j = 0$ for $j > p_0$. Note that 0_l is a zero vector with length l.

The Akaike information criterion for model M_j in the model list $\mathcal{M}_{\text{nest}}$ is $\text{AIC}(M_j) = -2\ell_n\{\hat{\theta}(j)\} + 2(a+j)$ where $\ell_n = \log L_n$. The index of the selected model is $\hat{p}_0 = \min\{j : \text{AIC}(M_j) = \min_{0 \le i \le K} \text{AIC}(M_i)\}$. The idea behind the construction of post-selection inference is to rewrite the selection procedure in terms of a set of inequalities, which define a geometrical region in terms of random variables that can be easily simulated. For this purpose, we redefine $\hat{p}_0 = \min\{j \in \{0, \ldots, K\} : j = \arg\max_{j=0,\ldots,K} \text{AIC}^*(M_j)\}$ with $\text{AIC}^*(M_j) = 2[\ell_n\{\hat{\theta}(j)\} - \ell_n(\vartheta)] - 2j = 2\ell^*_{n,j} - 2j$.

Asymptotically, the probability of underselection is zero (Woodroofe, 1982, see Lemma A1 in the Appendix); see also Shibata (1976). Conditioning on $\hat{p}_0 = p$, we have that $\text{AIC}^*(M_p) - \text{AIC}^*(M_j) > 0$ for $j = p_0, \ldots, p-1$ and $\text{AIC}^*(M_p) - \text{AIC}^*(M_j) \ge 0$ for $j = p+1, \ldots, K$. For $n \to \infty$, there is joint convergence in distribution of $(\ell^*_{n,p_0}, \ldots, \ell^*_{n,K})$ to $(\sum_{i=1}^{a+p_0} Z_i^2, \ldots, \sum_{i=1}^{a+K} Z_i^2)/2$, with $Z_1, \ldots, Z_{a+K}$ independent and identically $N(0,1)$ variables (Woodroofe, 1982). By the continuous mapping theorem, asymptotically, when $\hat{p}_0 = p$, $(Z_1, \ldots, Z_{a+K}) \in \mathcal{A}_p(\mathcal{M}_{\text{nest}})$, which is called the selection region for

nested models and is defined by $\mathcal{A}_p(\mathcal{M}_{\text{nest}})$ equal to

$$\left\{ z \in \mathbb{R}^{a+K} : \bigcap_{j=p_0+1,\ldots,p} \left\{ \sum_{i=j}^{p} (z_{a+i}^2 - 2) > 0 \right\} \cap \bigcap_{j=p+1,\ldots,K} \left\{ \sum_{i=p+1}^{j} (z_{a+i}^2 - 2) \leq 0 \right\} \right\}.$$

Geometrically, the first set of $p - p_0 - 1$ strict inequalities specifies regions outside spheres, the last set of $K - p$ inequalities indicates regions inside certain other spheres, while the inequality $z_p^2 > 2$ determines the union of two half-spaces, namely $(-\infty, -2^{1/2}) \cup (2^{1/2}, +\infty)$.

The specific structure of the Akaike information criterion determines the form of the regions. Other selection methods define other regions, see Section 7 for examples. Lee et al. (2016, Lemma 5.1, Th. 5.2) characterize the lasso-selection procedure, for a given value of the ℓ_1-penalty, in terms of polyhedral sets; see also Taylor et al. (2016).

2.2 Distributional results

Inference post-selection deals with the distribution of the estimators in the selected model, conditional on the selection. In this paper we always mean selection of the model with the smallest Akaike information criterion value and by the post-selection estimator we mean the maximum likelihood estimator based on the selected model. We show that the limiting cumulative distribution function of $n^{1/2}\{\hat{\theta}(\hat{p}_0) - \vartheta\}$ conditional on the selected model can be described by a multivariate normal random variable Z that is for nested models conditioned on $Z \in \mathcal{A}_p(\mathcal{M}_{\text{nest}})$.

Due to the nature of the selection using Akaike's information criterion and the results of Pötscher (1991) and Leeb & Pötscher (2003) it can be shown that the selection of an overspecified model does not happen in a uniform way, but depends on the true parameter value ϑ. Hence, in sections 2 and 3, the results are pointwise. All proofs and assumptions are placed in the Appendix.

Define, for model M_i, the submatrix $J_{M_i}(\vartheta)$ of the Fisher information matrix $J(\vartheta)$ in the model with all parameters, see Assumption A4, and for a $(a + K)$ vector ν denote its subvector $\tilde{\nu}(i) = (\nu_1, \ldots, \nu_{a+i})^\top$. The indicator function $I(A) = 1$ if A is true, and $I(A) = 0$ otherwise.

Proposition 1. *Assume A1–A4. For a sequence of nested models $\mathcal{M}_{\text{nest}}$ with p_0 denoting the true model order, the asymptotic conditional cumulative distribution function of the post-selection estimator is*

$$\begin{aligned} F_p(t) &= \lim_{n\to\infty} P[n^{1/2}\{\hat{\theta}(p) - \vartheta\} \leq t \mid \hat{p}_0 = p, \mathcal{M}_{\text{nest}}] \\ &= P\{J_p^{-1/2}(\vartheta)\tilde{Z}(p) \leq \tilde{t}(p) \mid \tilde{Z}(p) \in \mathcal{A}_p^{(s)}(\mathcal{M}_{\text{nest}})\} I(t \in \mathcal{T}_p), \end{aligned} \tag{1}$$

with $p \geq p_0$ by Lemma 1, $Z = (Z_1, \ldots, Z_{a+K})^\top$, the region with simplified constraints $\mathcal{A}_p^{(s)}(\mathcal{M}_{\text{nest}}) = \{\tilde{z}(p) \in \mathbb{R}^{a+p} : \bigcap_{j=p_0+1,\ldots,p} \sum_{i=j}^{p} (z_{a+i}^2 - 2) > 0\}$ and $\mathcal{T}_p = \mathbb{R}^{a+p} \times (\mathbb{R}^+)^{K-p}$.

By the forms of $\mathcal{A}_p$ and $\mathcal{A}_p^{(s)}$, the limiting distribution of $n^{1/2}\{\hat{\theta}(p) - \vartheta\}$ conditional on selection in the set $\mathcal{M}_{\text{nest}}$ is symmetric and its density function is that of a truncated normal random variable. Let $\phi_p(\cdot \mid \mathcal{A}; V)$ denote the density of $V^{-1/2}\tilde{Z}(p)$, where $\tilde{Z}(p) \sim N_{a+p}(0, I_{a+p})$ is truncated such that $\tilde{Z}(p) \in \mathcal{A}$. In the case of selecting the true model, the conditioning event contains random variables that are independent of $\tilde{Z}(p_0)$ and hence may be omitted. Figure 3 depicts some of the limiting post-selection densities for an example of selecting the largest in a sequence of three nested models, while the smallest model is the true one. This example is continued in Section 3.1. For more details, see the Supplementary Material.

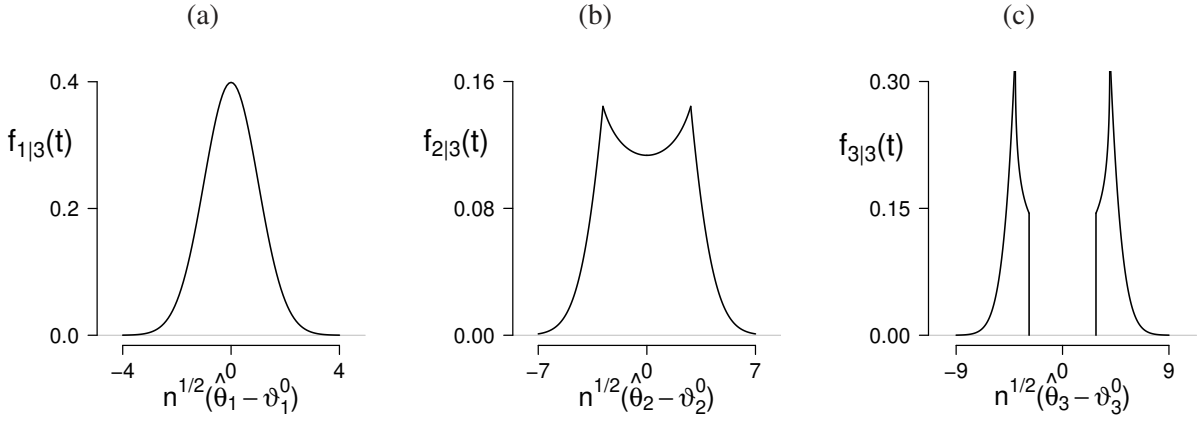

Figure 1: Marginal asymptotic densities $f_{j|3}$ $(j = 1, 2, 3)$ of $n^{1/2}(\hat{\theta}_j - \vartheta_j)$ conditional on $\hat{p}_0 = 3$ when $p_0 = 1$ and $J_3^{-1}(\vartheta)$ is a diagonal matrix with diagonal elements $(1, 4, 4)$.

Corollary 1. *Under the assumptions of Proposition 1, the limiting density of* $n^{1/2}\{\hat{\theta}(\hat{p}_0) - \vartheta\}$ *conditional on AIC-selection with* $\hat{p}_0 = p$ *from the set of nested models* $\mathcal{M}_{\text{nest}}$*, is* $f_p(t) = \phi_p\{\tilde{t}(p) \mid \mathcal{A}_p^{(s)}(\mathcal{M}_{\text{nest}}); J_p^{-1}(\vartheta)\}I(t \in \mathcal{T}_p)$*. When the true model is selected, i.e.,* $\hat{p}_0 = p_0$*, then* $f_{p_0}(t) = \phi_{p_0}\{\tilde{t}(p_0)\}I(t \in \mathcal{T}_p)$*.*

2.3 Confidence regions

A correct post-selection analysis incorporates the uncertainty associated with variable selection; we obtain confidence regions conditional on the selected model.

Corollary 2. *Under the assumptions of Proposition 1, an asymptotic* $100(1-\alpha)\%$ *Wald confidence ellipsoid conditionally on having selected a model with* $\hat{p}_0 = p$ *is*

$$\left\{\vartheta \in \mathbb{R}^{a+K} : n\{\hat{\theta}'(p) - \tilde{\vartheta}(p)\}^\top J_p(\vartheta)\{\hat{\theta}'(p) - \tilde{\vartheta}(p)\} \leq q_\alpha\right\},$$

where q_α *is defined such that* $1-\alpha$ *equals*

$$\int_{2(p-p_0)}^{q_\alpha}\int_{2(p-p_0)}^{w_1}\cdots\int_4^{w_{p-2}}\int_2^{w_{p-1}} \frac{f(w_p, \ldots, w_{p_0+1}, w_1)}{P\{\tilde{Z}(p) \in \mathcal{A}_p^{(s)}(\mathcal{M}_{\text{nest}})\}}\,\mathrm{d}w_p\,\mathrm{d}w_{p-1}\ldots\mathrm{d}w_{p_0+1}\,\mathrm{d}w_1; \tag{2}$$

$$f(w_p, \ldots, w_{p_0+1}, w_1) = \frac{\exp(-w_1/2)w_p^{-1/2}(w_1 - w_{p_0+1})^{-(a+p_0)/2-1}\prod_{i=1}^{p-p_0+1}(w_i - w_{i-1})^{-1/2}}{2^{\frac{a+p}{2}}\{\Gamma(1/2)\}^{p-p_0}\Gamma(\frac{a+p_0}{2})}.$$

In Section 2.4 we propose an accurate method to estimate q_α when exact computation is cumbersome. Clearly, the naive approach of using the quantile of a chi-square distribution is gives too low coverage. Confidence intervals for single components of ϑ require the calculation of marginal distributions.

Corollary 3. *Under the assumptions of Proposition 1, with* $\mathcal{R}_\alpha = \mathbb{R}^{j-1} \times [-q_{\alpha/2}, q_{\alpha/2}] \times \mathbb{R}^{a+p-j} \times (\mathbb{R}^+)^{K-p}$ *the asymptotic* $100(1-\alpha)\%$ *quantiles of the marginal distributions of* ϑ_j *with* $j = 1, 2, \ldots, a+p$ *satisfy* $\int_{\mathcal{R}_\alpha} f_p(t)\,\mathrm{d}t = 1-\alpha$*.*

2.4 Simulation based inference

Since the calculations are quite tedious, even in small dimensions, we present a method to simulate this conditional distribution, from which quantiles can then be obtained.

When $J(\vartheta)$ is unknown, we use a consistent estimator $\hat{J}\{\hat{\theta}(K)\}$. We use a Hamiltonian Monte Carlo method (Pakman & Paninski, 2014) to sample from a $(a+K)$-variate standard normal distribution subject to quadratic constraints that are also based on standard normal random variables. The resulting n' samples drawn from this density are placed in the $n' \times (a+K)$ matrix $\mathcal{Z}_{\mathcal{A}}$. Next, we multiply each row of $\tilde{\mathcal{Z}}_{\mathcal{A}}(p)$ by $\hat{J}_p^{-1/2}\{\hat{\theta}(K)\}$, which leads to n' samples from the limiting distribution of $n^{1/2}\{\hat{\theta}(\hat{p}_0) - \vartheta\}$; see Corollary 1.

The example in the Supplementary Material shows close agreement between the 95% quantiles q_α in (2) simulated via constrained χ^2 distributions and their exact values.

3 Post-selection inference in general models

3.1 AIC selection in a set of non-nested models

Lemma 1 generalizes Lemma A1 in the Supplementary Material (Woodroofe, 1982) to an arbitrary set of models that contains at least one overparametrized model.

Lemma 1. *Under Assumptions A1–A4, the asymptotic probability that selection using the Akaike information criterion results in an underparametrized model from a set of models $\mathcal{M}$ that contains at least one overparametrized model is equal to zero.*

The distributional properties of the post-selection estimators depend on the candidate set of models $\mathcal{M}$. Indeed, another set $\mathcal{M}$ could have led to another selection. We define the selection matrix to indicate which variables appear in the set of models.

Definition 1. *The selection matrix $\zeta_{\mathcal{M}}$ is a $|\mathcal{M}| \times (a+K)$ matrix with $\{0,1\}$ elements, constructed as $\zeta_{\mathcal{M}} = (1^t_{a+K}\pi_1^t\pi_1, \ldots, 1^t_{a+K}\pi_M^t\pi_M)^\top$, where $|\mathcal{M}|$ is the number of models and π_m is a $|m| \times (a+K)$ projection matrix that selects those covariates that belong to model m.*

First consider $\mathcal{M} = \mathcal{M}_{\text{all}}$, the set of all possible submodels of a largest model. Denote by $\mathcal{M}_O \subseteq \mathcal{M}_{\text{all}}$ the set of all overparametrized models, including the true model, so the models in $\mathcal{M}_O$ are overlapping. In model M the estimator of ϑ is denoted by $\hat{\theta}(M)$, with zeros added for components not in M. For any vector ν, let $\tilde{\nu}(M)$ denote its subvector corresponding to the variables in model M. Under the orthogonality assumption A5, Proposition 2 is similar to the nested model case. Otherwise, we follow Vuong (1989) for testing in overlapping models. Define $\Sigma(\theta)$ as a partitioned matrix with i,jth block equal to $\Sigma_{M_i,M_j} = Q^{-1}_{M_i}(\theta)J_{ij}(\theta,\theta)Q^{-1}_{M_j}(\theta)$.

Proposition 2. *Assume A1–A4 and selection from $\mathcal{M}_{\text{all}}$. (i) If A5 holds, the selection region for model M is*

$$\mathcal{A}_M(\mathcal{M}_O) = \left\{ z \in \mathbb{R}^{a+K} : \{1_{(|\mathcal{M}_O|-1)} \otimes (1^t_K\pi_M^t\pi_M) - \zeta_{\mathcal{M}_O\setminus M}\}\{(z_1^2-2), \ldots, (z_{a+K}^2-2)\}^\top > 0 \right\}.$$

The conditional limiting cumulative distribution function of the post-selection estimator is

$$\begin{aligned} F_M(t) &= \lim_{n\to\infty} P[n^{1/2}\{\hat{\theta}(M) - \vartheta\} \le t \mid M_{\text{AIC}} = M, \mathcal{M}_{\text{all}}] \\ &= P\{J_M^{-1/2}(\vartheta)\tilde{Z}(M) \le \tilde{t}(M) \mid Z \in \mathcal{A}_M(\mathcal{M}_O)\} I(t \in \mathcal{T}_M) \end{aligned} \tag{3}$$

where $\mathcal{T}_M$ is $\mathbb{R}^{|M|} \times (\mathbb{R}^+)^{K-|M|}$ and $J_M(\vartheta)$, $\tilde{Z}(M)$ and $\tilde{t}(M)$ are submatrices of, respectively, $J(\vartheta)$, $Z = (Z_1, \ldots, Z_{a+K})$ and t, corresponding to the variables in model M.

(ii) If A5 does not hold, define $m = \sum_{M \in \mathcal{M}_O} |M|$ and let $W_{\mathrm{AIC},i}$ be a matrix partitioned in the same way as $\Sigma(\vartheta)$ with diagonal blocks corresponding to M_{AIC} and M_i equal to $Q_{M_{\mathrm{AIC}}}(\vartheta)$ and $-Q_{M_i}(\vartheta)$, and zero elsewhere. The selection region for model M_{AIC} is

$$\mathcal{A}_M(\mathcal{M}_O) = \{z \in \mathbb{R}^m : z^\top \Sigma^{1/2}(\vartheta) W_{\mathrm{AIC},i} \Sigma^{1/2}(\vartheta) z \geq 2(|M_{\mathrm{AIC}}| - |M_i|),\ M_i \in \mathcal{M}_O \backslash M_{\mathrm{AIC}}\}. \tag{4}$$

Let $\tilde{Z}(M)$ denote the subvector of $Z \sim N_m(0, I)$, $Z \in \mathcal{A}_M(\mathcal{M}_O)$ that contains only those components that correspond to components in the selected model M, then

$$F_M(t) \;=\; P\{J_M^{-1/2}(\vartheta)\tilde{Z}(M) \leq \tilde{t}(M) \mid Z \in \mathcal{A}_M(\mathcal{M}_O)\} I(t \in \mathcal{T}_M) \tag{5}$$

where $\mathcal{T}_M$ is $\mathbb{R}^{|M|} \times (\mathbb{R}^+)^{m-|M|}$.

The choice of $\mathcal{M}$ is important. Regarding (i), the constraint involves those Z_is corresponding to the parameters in the selected model M_{AIC} that are not in the smallest true model M_{pars}, hence no constraints are placed on the Z_i corresponding to parameters that occur in every model. Obviously, the selection affects the distribution of all parameters, even those common to all models. The effect of the set of models is illustrated by the following example. Let $K = 2$, $a = 1$ and M_0 be the smallest true model containing only θ_1. Assume that A5 holds and that the full model $M_{\mathrm{AIC}} = (\theta_1, \theta_2, \theta_3)$ is selected in both $\mathcal{M}_{\mathrm{nest}}$ and $\mathcal{M}_{\mathrm{all}}$. So, $\mathcal{A}_M(\mathcal{M}_{\mathrm{all}}) = \{z \in \mathbb{R}^3 : z_2^2 > 2, z_3^2 > 2, z_2^2 + z_3^2 > 4\}$ while $\mathcal{A}_M(\mathcal{M}_{\mathrm{nest}}) = \{z \in \mathbb{R}^3 : z_3^2 > 2, z_2^2 + z_3^2 > 4\}$. Figure 2 depicts these regions for both $\mathcal{M}_{\mathrm{nest}}$, shaded

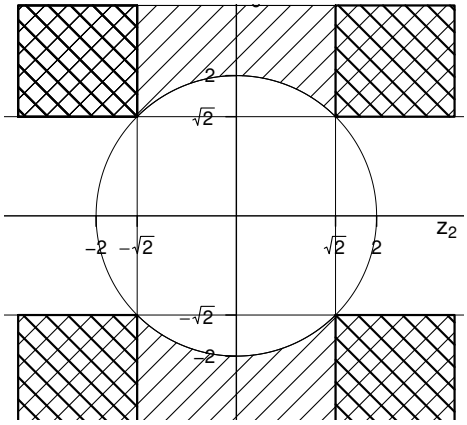

Figure 2: Allowable domain of Z_2 and Z_3 for nested model selection (shaded), and all subsets selection (double shaded) when AIC selects the full model.

area, and $\mathcal{M}_{\mathrm{all}}$, double shaded area. If one selects the full model in $\mathcal{M}_{\mathrm{nest}}$, then Z_2 is defined in $\mathbb{R}$ as long as $Z_2^2 + Z_3^2 > 4$, while selection in $\mathcal{M}_{\mathrm{all}}$ requires both Z_2 and Z_3 in $(-\infty, -2^{1/2}) \cup (2^{1/2}, \infty)$. The distribution of parameter estimators can be obtained by premultiplying $Z = (Z_1, Z_2, Z_3)$ by $J_{M_{\mathrm{AIC}}}^{1/2}(\vartheta)$. For the normal linear models $Y \sim N_n(X\vartheta, \sigma^2 I)$ and $M_{\mathrm{AIC}} \in \mathcal{M}_{\mathcal{O}}$, the distribution results are also exact for finite samples. In such models $J(\vartheta) = n^{-1}X^\top X/\sigma^2$, which does not depend on ϑ. For (ii) the main difference is that we need the joint distribution of the estimators in the different models and place constraints on the full vector.

3.2 Confidence regions

For any arbitrary set of models, $\mathcal{M}_{\text{arb}}$, with $\mathcal{M}_{\text{arb}} \cap \mathcal{M}_O \neq \emptyset$, due to Assumption A1, (3) still holds after replacing $\mathcal{A}_M(\mathcal{M}_O)$ with $\mathcal{A}_M(\mathcal{M}_{\text{arb}} \cap \mathcal{M}_O)$. With $M_{\text{AIC}} = M$ selected from $\mathcal{M}_{\text{arb}}$, the confidence region for ϑ is

$$C(q_\alpha) = \left\{ \theta \in \mathbb{R}^{a+K} : n\{\hat{\theta}'(M) - \tilde{\theta}(M)\}^\top J_M(\theta)\{\hat{\theta}'(M) - \tilde{\theta}(M)\} \leq q_\alpha \right\}, \tag{6}$$

with $\hat{\theta}'(M)$ the $|M|$-vector of non-zero values of $\hat{\theta}(M)$ and q_α determined by solving

$$\frac{\mathrm{P}\{(\sum_{i\in M} Z_i^2 \leq q_\alpha) \cap Z \in \mathcal{A}_M(\mathcal{M}_{\text{arb}} \cap \mathcal{M}_O)\}}{\mathrm{P}\{Z \in \mathcal{A}_M(\mathcal{M}_{\text{arb}} \cap \mathcal{M}_O)\}} = 1 - \alpha. \tag{7}$$

Let $f_M\{\tilde{t}(M)\} = \phi_M\{\tilde{t}(M) | \mathcal{A}_M(\mathcal{M}_{\text{arb}} \cap \mathcal{M}_O); J_M^{-1}(\vartheta)\}$ denote the density of $n^{1/2}\{\hat{\theta}'(M) - \tilde{\vartheta}(M)\}$, a truncated $|M|$-dimensional normal density. The quantile of its jth component is obtained via

$$\int_{\mathcal{R}_\alpha} f_M\{\tilde{t}(M)\}\, \mathrm{d}\tilde{t}(M) = 1 - \alpha,$$

where $\mathcal{R}_\alpha \subset \mathbb{R}^{|M|}$ restricts only the jth component to $[-q_{\alpha/2}, q_{\alpha/2}]$. The confidence interval for ϑ_j is $\hat{\theta}_j(M) \pm q_{\alpha/2} n^{-1/2}$.

While there is no uniform convergence of the distribution function in all settings (Leeb & Pötscher, 2003), for normal linear models using rectangular confidence regions and sequential testing, a uniform result regarding coverage has been obtained by Pötscher (1995). The following result holds for overspecified models. For models in the set $\mathcal{M}_O$ all parameter components that appear in the true model are nonzero, but there might be additional parameter components which might be zero or non-zero. However, the set $\mathcal{M}_O$ does not depend on the value of the true parameter ϑ. After conditioning on $M_{\text{AIC}} \in \mathcal{M}_O$, the set $C(q_\alpha)$ is random due to maximum likelihood estimation in the selected model.

Proposition 3. *Assume A1–A4, and that $Q_n(\theta)$ in (A2) is continuous over a compact set Θ that contains ϑ. The confidence region $C(q_\alpha)$ from (6) is such that* $\lim_{n\to\infty} \inf_{\vartheta\in\Theta} P_\vartheta\{\vartheta \in C(q_\alpha) \mid M_{\text{AIC}} \in \mathcal{M}_O\} = 1-\alpha$. *When $\mathcal{A}_M(\mathcal{M}_{\text{arb}})$ replaces $\mathcal{A}_M(\mathcal{M}_{\text{arb}} \cap \mathcal{M}_O)$ in (7) to obtain a value $\tilde{q}_\alpha$,* $\lim_{n\to\infty} \inf_{\vartheta\in\Theta} P\{\vartheta \in C(\tilde{q}_\alpha) \mid M_{\text{AIC}} \in \mathcal{M}_O\} \geq 1 - \alpha$.

One limitation of the Akaike information criterion is that the selection of an overspecified model does not happen in a uniform way (Leeb & Pötscher, 2003). Hence, this result cannot be strengthened. If the selected model is underparametrized, correct inference can be obtained for the pseudo-true values instead; see Section 4. For a predetermined number of steps in a forward selection, least angle regression and lasso in linear additive error models, Tibshirani et al. (2015) obtain asymptotic results which are uniformly valid for a specific class of non-normal errors. For a comparison between two models, Andrews & Guggenberger (2009) use a local neighborhood to deal with the overselection and to obtain uniform results for parameters that were not subject to selection. Chernozhukov et al. (2015) performed uniformly valid inference on a low-dimensional parameter when there is selection in a high-dimensional vector of nuisance parameters. See also Belloni et al. (2015) for using least absolute deviation in high dimensional regression.

Inference after selection depends on (i) the set of models $\mathcal{M}$ specified by the researcher and (ii) the smallest true model M_{pars}, in nested models p_0, via $\mathcal{A}_M(\mathcal{M} \cap \mathcal{M}_O)$. In $\mathcal{M}_{\text{nest}}$ and $\mathcal{M}_{\text{all}}$ one could take the smallest model for M_{pars}. If this model is true or overparametrized, Propositions 1 and 2 hold and the

asymptotic confidence intervals can be calculated exactly. If the smallest model is underparametrized, the structure of the additional constraints $\mathcal{A}_M(\mathcal{M})\backslash\mathcal{A}_M(\mathcal{M}\cap\mathcal{M}_O)$ is such that the resulting distribution of the parameters is longer-tailed. This leads to conservative confidence intervals, especially for the parameters which are truly non-zero. In practice we calculate the constraints based on the selected model and $\mathcal{A}_M(\mathcal{M}_{\rm arb})$.

For case (i), in $\mathcal{M}_{\rm all}$ the number of constraints equals $2^{K-|M_0|}-1$. Here, we show that $\mathcal{A}_M(\mathcal{M}_O)$ can be reduced to the set $\{z\in\mathbb{R}^{a+K}:\bigcap_{i\in M_{\rm AIC}\backslash M_{\rm pars}}(z_i^2>2)\cap\bigcap_{i\notin M_{\rm AIC}\backslash M_{\rm pars}}(z_i^2<2)\}$ without losing information. Let $\mathcal{I}_{M_{\rm AIC}}$ denote the set consisting of all subsets of the indices in $M_{\rm AIC}\backslash M_{\rm pars}$, referring to the redundant selected parameters, and denote by $\mathcal{I}^c_{M_{\rm AIC}}$ the set of all subsets of the indices in $\{1,\ldots,a+K\}\backslash M_{\rm AIC}$, referring to the variables that were not selected. Then

$$\mathcal{A}_{M_{\rm AIC}}(\mathcal{M}\cap\mathcal{M}_O)=\left\{\begin{array}{c} z\in\mathbb{R}^{a+K}:\bigcap_{i\in M_{\rm AIC}\backslash M_{\rm pars}}\{z_i^2>2\},\bigcap_{i\in\{1,\ldots,a+K\}\backslash M_{\rm AIC}}\{-z_i^2>-2\},\\ \bigcap_{I\in\mathcal{I}_{M_{\rm AIC}}}\bigcap_{J\in\mathcal{I}^c_{M_{\rm AIC}}}\{\sum_{i\in I}z_i^2-\sum_{j\in J}z_i^2>2(|I|-|J|)\}\end{array}\right\}.$$

The first two sets of constraints consist, respectively, of $|M_{\rm AIC}|-|M_{\rm pars}|$ and $K-|M_{\rm AIC}|$ elements. The third set only involves constraints that are summations of the constraints in the first two sets and does not add any new restrictions on z. The constraint set for any $\mathcal{M}_{\rm arb}$ can be simplified as long as some constraints can be implied by summing other constraints. Removing redundant constraints is not always possible, for example for $\mathcal{M}_{\rm nest}$.

3.3 Inference for linear combinations

For inference for linear combinations $x^t\vartheta$ after model selection, we rewrite (3) as

$$\begin{aligned}F(t) &= \lim_{n\to\infty}\mathrm{P}[n^{1/2}\,\tilde{x}^t(M)\{\hat{\theta}'(M)-\tilde{\vartheta}(M)\}\le t\mid M_{\rm AIC}=M,\mathcal{M}]\\ &= \mathrm{P}\{\tilde{x}^t(M)\,J_M^{-1/2}(\vartheta)\tilde{Z}(M)\le t\mid\mathcal{A}_M(\mathcal{M}\cap\mathcal{M}_O)\},\end{aligned}\tag{8}$$

where $\tilde{x}(M)$ are the covariates corresponding to M. The asymptotic distribution of the estimated linear combination $x^t\vartheta$ is simulated via (8).

When the sample size is small and the diagonal entries of $J(\hat{\theta})$ are large, it may happen that an underparametrized model is selected. In this case the coverage probability of confidence regions of a linear combination of the parameters, or a transformation thereof in generalized linear models, may be smaller than the nominal value. In case of suspected underselection, one can use

$$\lim_{n\to\infty}\mathrm{P}[n^{1/2}x^t\{\hat{\theta}(M_{\rm full})-\vartheta\}\le t\mid M_{\rm AIC}=M,\mathcal{M}]=\mathrm{P}\{x^tJ^{-1/2}(\vartheta)Z_{a+K}\le t\mid\mathcal{A}_M(\mathcal{M})\},\tag{9}$$

where $M_{\rm full}$ is the full model. This differs from (8) in using all parameters, not just the selected parameters. This procedure differs from assuming that the full model is selected, since, for example in $\mathcal{M}_{\rm all}$, $\mathcal{A}_M(\mathcal{M}_{\rm all})$ contains $z_i^2>2$ for the parameters which are selected and $z_i^2<2$ for those which are not selected, whereas $\mathcal{A}_{M_{\rm full}}(\mathcal{M}_{\rm all})$ contains $z_i^2>2$ for all parameters, which leads to a long-tailed distribution. The probability of underselection disappears asymptotically. The valid confidence intervals of Bachoc et al. (2015) target the true value for the selected model, not the true value $x^t\vartheta$. While in their case underparametrized selection is not an issue, there is no guarantee that their proposed confidence interval is valid for the true value.

4 Confidence regions when all models are misspecified

4.1 Limiting distribution of estimators

The results in this section do not require any assumption about the existence of a true model, are uniformly valid, and apply to general parametric likelihood models. In order to obtain uniformly valid results we consider the setting where there is no true parameter vector, either because the true density of the data does not belong to a parametric family or because all models are misspecified. We assume the observations to be represented by a triangular array $\{Y_{ni} : i = 1, \ldots, n, n \in \mathbb{N}\}$, where there is independence between the rows, i.e., different sample sizes n, and within the rows, i.e., for $i \neq j$, Y_{ni} and Y_{nj} are independent. Regression models are included, as observations may have different distributions. The true joint density of $(Y_{n1}, \ldots, Y_{nn})$ is g_n, with distribution function G_n. All probabilities are computed under the true distribution, so $\mathrm{P} = \mathrm{P}_{G_n}$. The data are modeled via models $M_{n,j} = \{\prod_{i=1}^n f_{j,i}(y_i;\theta_j) : \theta_j \in \Theta_j \subset \mathbb{R}^{m_j}\}$. Thus m_j is the number of parameters in model $M_{n,j}$. All models are collected in the set $\mathcal{M}_n = \{M_{n,1}, \ldots, M_{n,J}\}$. When there is no confusion, we omit the subscript n in the notation. We assume for each $n \in \mathbb{N}$ that $\int g_n(y) \log g_n(y) dy < \infty$. This defines the class of true distributions $\mathcal{G}_n$.

Regarding the models, assume that for each $i \in \mathbb{N}$ and each $j = 1, \ldots J$, $f_{j,i}(\cdot;\theta_j)$ is measurable for all $\theta_j \in \Theta_j$, a compact set, $f_{j,i}(y_i;\cdot)$ is continuous on Θ_j almost surely and continuously differentiable on Θ_j. Then for every model there exists (White, 1994, Th 2.12) an estimator $\hat{\theta}_{n,j}$, maximizing $\prod_{i=1}^n f_{j,i}(y_i;\theta_j)$ over Θ_j. If $E_{G_n}\{n^{-1}\sum_{i=1}^n \log f_{j,i}(y_i;\theta_j)\}$ has an identifiable unique maximizer over Θ_j, this maximizer is called the pseudo-true value $\vartheta_n^*(M_j)$. This value depends on the true joint density, the model densities, and on the sample size. We define two vectors of length $m' = \sum_{j=1}^J m_j$, $\vartheta_{n,\mathcal{M}}^* = \{\vartheta_n^{*t}(M_1), \ldots, \vartheta_n^{*t}(M_K)\}^\top$ and $\hat{\theta}_{n,\mathcal{M}} = \{\hat{\theta}_n^t(M_1), \ldots, \hat{\theta}_n^t(M_K)\}^\top$.

Lemma 2. *Let $\{Y_{ni} : i = 1, \ldots, n, n \in \mathbb{N}\backslash 0\}$ form a triangular array consisting of independent random variables. Assume that (i) for all components of the vector $\vartheta_{n,\mathcal{M}}^*$, here stated for the kth such component of θ_j corresponding to model M_j, for all $G_n \in \mathcal{G}_n$ with $A = \{y_i \in \mathbb{R} : |(\partial/\partial\theta_k)\log f_{j,i}\{y_i;\theta_n^*(M_j)\}| > \varepsilon n Q_{M_j,kk}\{\vartheta_n^*(M_j)\}\}$, and for all $\varepsilon > 0$,*

$$\lim_{n\to\infty}\sum_{i=1}^n \int_A \left[\frac{\partial}{\partial\theta_k}\log f_{j,i}\{y_i;\vartheta_n^*(M_j)\}\right]^2 / [nQ_{M_j,kk}\{\vartheta_n^*(M_j)\}] dG_{ni}(y_i) = 0.$$

and (ii) denoting $\Sigma_{M_j}\{\vartheta_n^*(M_j)\} = Q_{M_j}^{-1}\{\vartheta_n^*(M_j)\}J_{jj}\{\vartheta_n^*(M_j),\vartheta_n^*(M_j)\}Q_{M_j}^{-1}\{\vartheta_n^*(M_j)\}$,

$$\lim_{n\to\infty}\max_{i=1,\ldots,n} P_{G_n}\left((\Sigma_{M_j,kk})^{-1/2}n^{-1/2}[Q_{M_j}^{-1}\{\vartheta_n^*(M_j)\}]_{kk}\Big|\frac{\partial}{\partial\theta_k}\log f_{j,i}\{y_i;\vartheta_n^*(M_j)\}\Big| > \varepsilon\right) = 0$$

Define $\mathcal{W}_n \sim N_{m'}\{0, \Sigma(\vartheta_{n,\mathcal{M}}^*)\}$ where $\Sigma(\vartheta_{n,\mathcal{M}}^*)$ is a $m' \times m'$ matrix with ijth block, with dimensions $m_i \times m_j$, equal to $Q_{M_i}^{-1}\{\vartheta_n^*(M_i)\}J_{ij}\{\vartheta_n^*(M_i),\vartheta_n^*(M_j)\}Q_{M_j}^{-1}\{\vartheta_n^*(M_j)\}$, then

$$\lim_{n\to\infty}\sup_{t\in\mathbb{R}^{m'}}\sup_{G_n\in\mathcal{G}_n}\left|P\{n^{1/2}(\hat{\theta}_{n,\mathcal{M}} - \vartheta_{n,\mathcal{M}}^*) \leq t\} - P(\mathcal{W}_n \leq t)\right| = 0.$$

A pivot is needed in order to construct confidence regions. In general, the variance $\Sigma(\vartheta_{n,\mathcal{M}}^*)$ of $\mathcal{W}_n$ might depend on $\vartheta_{n,\mathcal{M}}^*$. When there is an estimator $\hat{\Sigma}$ of Σ such that

$$\lim_{n\to\infty}\sup_{G_n\in\mathcal{G}_n}\mathrm{P}_{G_n}(\|\hat{\Sigma}_n - \Sigma\| > \varepsilon) = 0,$$

with $\|A\|$ denoting the Euclidean matrix operator norm of A, then, with $\mathcal{Z}_{m'} \sim N_{m'}(0, I_{m'})$

$$\lim_{n\to\infty} \sup_{G_n \in \mathcal{G}_n} \sup_{t\in\mathbb{R}^{m'}} |\mathrm{P}\{\hat{\Sigma}_n^{-1/2} n^{-1/2}(\hat{\theta}_{n,\mathcal{M}} - \vartheta^*_{n,\mathcal{M}}) \le t\} - \mathrm{P}(\mathcal{Z}_{m'} \le t)| = 0.$$

The model determines whether or not the variance may be estimated well. White (1994, Sec 8.3) gives some general conditions for consistent estimation of the variance. One requirement is that

$$n^{-1} \sum_{i=1}^{n} E(s)E(s^\top) \to 0,$$

with s the vector of length m' consisting of subvectors $(\partial/\partial\theta_k) \log f_{ki}(Y_i; \vartheta^*_k)$, for $k = 1, \ldots, K$. This assumption holds, for example, when the models are correctly specified. Under misspecification, White (1994, Sec 8.3) showed that the empirical estimator for $\Sigma(\vartheta^*_{n,\mathcal{M}})$ might overestimate the covariance matrix, leading to conservative confidence intervals.

4.2 Selection region in a misspecified setting

When $\mathcal{M}$ consists of misspecified models, calculating the selection event requires additional care. Define $\ell_{n,M_j}(y, \theta_j) = \sum_{i=1}^n \log f_{j,i}(y_i, \theta_j)$. When model M_{AIC} is selected, then for all $M \in \mathcal{M}\backslash M_{\text{AIC}}$, $2[\ell_{n,M_{\text{AIC}}}\{y, \hat{\theta}_n(M_{\text{AIC}})\} - \ell_{n,M}\{y, \hat{\theta}_n(M)\}] \ge 2(|M_{\text{AIC}}| - |M|)$. When both models, M_{AIC} and M, are correctly specified, the difference of log-likelihoods can be characterized asymptotically by chi-squared random variables. However, when there is misspecification this difference can diverge to $+\infty$ or $-\infty$, depending on the assumptions about the models. For strictly non-nested models the difference always diverges (Vuong, 1989, Th. 5.1). When the selected model is always best, there is no restriction on parameter estimators. See also Cox & Hinkley (1974, Sec. 9.3) for the asymptotic behavior of likelihood ratio tests in non-nested settings. For overlapping models having some common parameters, the log-likelihood difference converges to some random variable if one of the models is correctly specified, and otherwise diverges. Under misspecification of all models, the only setting where the asymptotic distribution can be used to characterize the selection event is for nested models under similarity of the likelihoods (Vuong, 1989, Assumption A8). This means that $\ell_{n,M_k}\{y, \vartheta^*_n(M_k)\} = \ell_{n,M_l}\{y, \vartheta^*_n(M_l)\}$ for $k, l = 1, \ldots, K$. For an arbitrary set of models we impose the same similarity assumption and assume that $\mathcal{M}$ includes a model $M_{\text{s}} = M_{\text{small}}$ which is nested in all other models. If we were to perform a likelihood ratio test, under this assumption it would correspond to testing whether the smaller model can be considered equal versus worse than the larger model (Vuong, 1989, Lemma 7.1). We first compare each model with the smallest model and then we use the obtained regions from each comparison to compute the final selection region using pairwise comparisons. By imposing similarity, the calculated quantiles to be used in the confidence regions are larger than without similarity since, as explained earlier, the log-likelihood difference diverges otherwise and there is no restriction on the parameter estimators. For all $M \in \mathcal{M}\backslash M_{\text{s}}$,

$$\begin{aligned}
&2[\ell_{n,M}\{y, \hat{\theta}_n(M)\} - \ell_{n,M_s}\{y, \hat{\theta}_n(M_{\text{s}})\}] \\
&\quad = n\{\hat{\theta}_n(M) - \vartheta^*_n(M)\}^\top Q_M\{\vartheta^*_n(M)\}\{\hat{\theta}(M) - \vartheta^*_n(M)\} \\
&\qquad -n\{\hat{\theta}_n(M) - \vartheta^*_n(M_{\text{s}})\}^\top Q\{\vartheta^*_n(M_s)\}\{\hat{\theta}_n(M_{\text{s}}) - \vartheta^*_n(M_{\text{s}})\} + o_P(1) \\
&\quad = n(\hat{\theta}_{n,\mathcal{M}} - \vartheta^*_{n,\mathcal{M}})^\top W_{M,M_{\text{s}}}(\hat{\theta}_{n,\mathcal{M}} - \vartheta^*_{n,\mathcal{M}}) + o_P(1), \qquad (10)
\end{aligned}$$

where $W_{M,M_{\text{s}}}$ is a block-diagonal matrix partitioned in the same way as Σ, with the diagonal block referring to model M equal to $Q_M\{\vartheta^*_n(M)\}$ and that referring to model M_{s} equal to $-Q_{M_s}\{\vartheta^*_n(M_{\text{s}})\}$,

and zero elsewhere. If the models are already nested, there is no need to compare each model with the smallest model. The asymptotic counterpart of the selection event is

$$\mathcal{A}_{M_{\text{AIC}}}(\mathcal{M}) = \{z \in \mathbb{R}^{m'} : z^\top \Sigma^{1/2}(W_{M_{\text{AIC}},M_{\text{s}}} - W_{M,M_{\text{s}}})\Sigma^{1/2} z \geq 2(|M_{\text{AIC}}| - |M|), \\ M \in \mathcal{M}\backslash M_{\text{AIC}}\}. \tag{11}$$

Proposition 4. *Let the assumptions of Lemma 2 hold. For a set of models with $\mathcal{A}_{M_{\text{AIC}}}(\mathcal{M})$ from (11) it holds that*

$$\lim_{n\to\infty} \sup_{G_n \in \mathcal{G}_n} \sup_{t \in \mathbb{R}^{|M_{\text{AIC}}|}} \Big| P[n^{1/2}\{\hat{\theta}(M_{\text{AIC}}) - \vartheta^*(M_{\text{AIC}})\} \leq t \mid M_{\text{AIC}}] \\ - P\{\Sigma^{1/2} Z \leq t \mid \mathcal{A}_{M_{\text{AIC}}}(\mathcal{M})\} \Big| = 0 \tag{12}$$

As noted by Tibshirani et al. (2015), uniform convergence in distribution can be translated to uniformly valid confidence sets. The following proposition clarifies this statement. The proof is similar to the proof of Proposition 4, using the fact that a continuous mapping preserves uniform convergence.

Proposition 5. *Let the assumptions of Lemma 2 hold and let the set of models $\mathcal{M}$ contain a smallest model which is nested in all models. Define the set*

$$C^*(q_\alpha) = \{\theta \in \mathbb{R}^{|M_{\text{AIC}}|} : n\{\hat{\theta}(M_{\text{AIC}}) - \theta(M_{\text{AIC}})\}^\top \Sigma_{M_{\text{AIC}}}(\vartheta^*_{M_{\text{AIC}}})^{-1}\{\hat{\theta}(M_{\text{AIC}}) - \theta(M_{\text{AIC}})\} \leq q_\alpha\},$$

where q_a is determined by solving

$$P\Big\{[\tilde{Z}^\top \quad (M_{\text{AIC}})\Sigma_{M_{\text{AIC}}}(\vartheta^*_{M_{\text{AIC}}})^{-1}\tilde{Z}(M_{\text{AIC}}) \leq q_\alpha] \cap \{Z \in \mathcal{A}_{M_{\text{AIC}}}(\mathcal{M})\}\Big\} \\ = P\{Z \in \mathcal{A}_{M_{\text{AIC}}}(\mathcal{M})\}(1-\alpha).$$

Then $\lim_{n\to\infty} \sup_{G_n \in \mathcal{G}_n} \sup_{\alpha \in [0,1]} \big| P_{G_n}\{\vartheta^*(M_{\text{AIC}}) \in C^*(q_\alpha) \mid M_{\text{AIC}}\} - (1-\alpha)\big| = 0.$

5 Simulation study

5.1 Parameters in linear models

While the proposed method is applicable in general likelihood models, in order to compare it with existing methods, we present simulation results for linear models. Results for generalized linear models and for other settings are placed in the Supplementary Material.

The data were generated from a regression model $Y_i = \sum_{j=1}^{10} \vartheta_j x_{ji} + \varepsilon_i, \quad i = 1, \ldots, n$, with $\varepsilon_i \sim N(0,1)$. The true value for the parameters is $\vartheta^\top = (2{\cdot}25, -1{\cdot}1, 2{\cdot}43, -2{\cdot}24, 2{\cdot}5, 0_5^\top)$, with 0_5 a vector of all zeros with length 5. We set $x_{1i} = 1$ and $(x_{2i}, \ldots, x_{10,i})^\top \sim N(0_9, \Omega)$ where Ω is a positive definite matrix with diagonal elements equal to 1 and off-diagonal entries equal to $0{\cdot}25$. The sample size is either 30 or 100.

Three different model sets were considered. Let ζ^i_{all} be the selection matrix when the first i parameters are present in all models. We take ζ^3_{all} which is a $2^7 \times 10$ matrix and ζ^6_{all} which is $2^4 \times 10$ matrix, and ζ_{arb} which contains 14 rows, arbitrarily chosen from ζ^3_{all}.

We are interested in inference for the parameters in the selected model. In order to facilitate the comparison, the simulations were run until model M with parameters $(\vartheta_1, \ldots, \vartheta_6, \vartheta_8)$ had been selected 3000 times. For each of those simulation runs the Fisher information matrix is estimated in the full model by $\hat{J}(\hat{\theta})$, leading to the submatrix $\hat{J}_M(\hat{\theta})$. When A5 does not hold one should use (5) to calculate the

n	method	ϑ_j	ζ^3_{all}		ζ^6_{all}		ζ_{arb}	
30	PostAIC	ϑ_4	$[-2{\cdot}85, -1{\cdot}64]$	98	$[-2{\cdot}68, -1{\cdot}78]$	92	$[-2{\cdot}85, -1{\cdot}64]$	97
		ϑ_6	$[-0{\cdot}60, \ 0{\cdot}62]$	94	$[-0{\cdot}45, \ 0{\cdot}45]$	93	$[-0{\cdot}60, \ 0{\cdot}62]$	96
		ϑ_8	$[-0{\cdot}60, \ 0{\cdot}61]$	94	$[-0{\cdot}60, \ 0{\cdot}60]$	95	$[-0{\cdot}61, \ 0{\cdot}62]$	96
	PoSI	ϑ_4	$[-2{\cdot}98, -1{\cdot}51]$	99	$[-2{\cdot}89, -1{\cdot}57]$	99	$[-2{\cdot}97, -1{\cdot}52]$	99
		ϑ_6	$[-0{\cdot}73, \ 0{\cdot}75]$	99	$[-0{\cdot}66, \ 0{\cdot}66]$	99	$[-0{\cdot}71, \ 0{\cdot}73]$	99
		ϑ_8	$[-0{\cdot}73, \ 0{\cdot}74]$	98	$[-0{\cdot}66, \ 0{\cdot}67]$	97	$[-0{\cdot}72, \ 0{\cdot}73]$	99
	Naive	ϑ_4	$[-2{\cdot}67, -1{\cdot}82]$	89	$[-2{\cdot}68, -1{\cdot}79]$	91	$[-2{\cdot}66, -1{\cdot}83]$	89
		ϑ_6	$[-0{\cdot}42, \ 0{\cdot}43]$	69	$[-0{\cdot}44, \ 0{\cdot}44]$	92	$[-0{\cdot}41, \ 0{\cdot}42]$	71
		ϑ_8	$[-0{\cdot}42, \ 0{\cdot}43]$	70	$[-0{\cdot}44, \ 0{\cdot}45]$	75	$[-0{\cdot}41, \ 0{\cdot}43]$	71
100	PostAIC	ϑ_4	$[-2{\cdot}54, -1{\cdot}94]$	99	$[-2{\cdot}46, -2{\cdot}02]$	94	$[-2{\cdot}55, -1{\cdot}93]$	99
		ϑ_6	$[-0{\cdot}30, \ 0{\cdot}31]$	95	$[-0{\cdot}22, \ 0{\cdot}22]$	95	$[-0{\cdot}31, \ 0{\cdot}32]$	96
		ϑ_8	$[-0{\cdot}30, \ 0{\cdot}31]$	95	$[-0{\cdot}29, \ 0{\cdot}30]$	95	$[-0{\cdot}31, \ 0{\cdot}31]$	97
	PoSI	ϑ_4	$[-2{\cdot}58, -1{\cdot}90]$	100	$[-2{\cdot}54, -1{\cdot}94]$	99	$[-2{\cdot}57, -1{\cdot}90]$	99
		ϑ_6	$[-0{\cdot}33, \ 0{\cdot}34]$	98	$[-0{\cdot}30, \ 0{\cdot}30]$	99	$[-0{\cdot}33, \ 0{\cdot}34]$	98
		ϑ_8	$[-0{\cdot}34, \ 0{\cdot}34]$	98	$[-0{\cdot}29, \ 0{\cdot}31]$	95	$[-0{\cdot}33, \ 0{\cdot}34]$	98
	Naive	ϑ_4	$[-2{\cdot}46, -2{\cdot}02]$	93	$[-2{\cdot}46, -2{\cdot}02]$	93	$[-2{\cdot}46, -2{\cdot}02]$	92
		ϑ_6	$[-0{\cdot}22, \ 0{\cdot}22]$	66	$[-0{\cdot}22, \ 0{\cdot}22]$	94	$[-0{\cdot}21, \ 0{\cdot}22]$	67
		ϑ_8	$[-0{\cdot}22, \ 0{\cdot}22]$	66	$[-0{\cdot}22, \ 0{\cdot}23]$	69	$[-0{\cdot}22, \ 0{\cdot}22]$	65

Table 1: Simulation study with 3000 runs of AIC selection. Average confidence intervals and coverage percentages for $\vartheta_4, \vartheta_6, \vartheta_8$ using different selection matrices ζ corresponding to different model sets $\mathcal{M}$ and different sample sizes n for the proposed method, the method of Berk et al. (2013) and for a naive approach that treats the selected model as given and ignores selection.

confidence intervals. However, we used (3) instead, resulting in good approximations. Quantiles of the limiting asymptotic distribution for each setting were obtained via simulation. See the Supplementary Material for the code. In each simulation run we compute the lower and upper limit of the confidence interval and report the averaged confidence intervals along with the coverage percentages. Table 1 presents the results for ϑ_4, ϑ_6 and ϑ_8. Results for the other parameters are not presented to save space.

Confidence intervals from the method of Berk et al. (2013) are reported for sake of comparison. Their target for inference is the so-called non-standard target (Bachoc et al., 2015), namely the best coefficients within the selected model, not the standard target, the true values of the parameters (Berk et al., 2013, equation (3.2)). Simulation results in Leeb et al. (2015) showed that the coverage probability of such intervals for the standard target is lower than the nominal value for certain situations.

For ζ^3_{all} where ϑ_4 and ϑ_5 are truly non-zero, the conditional confidence intervals for the proposed method have simulated coverage probabilities higher than the nominal value 95%. This is because $\mathcal{A}_M(\mathcal{M}^3_{\text{all}})$, $Z_4^2 > 2$ and $Z_5^2 > 2$ in the constraint set, while Z_4 and Z_5 are truly unconstrained when taking $\mathcal{A}_M(\mathcal{M}^3_{\text{all}} \cap \mathcal{M}_O)$. For ϑ_6 and ϑ_8 which are truly zero, $Z_6^2 > 2$ and $Z_8^2 > 2$ are correct constraints. One may expect conservative confidence intervals for ϑ_6 and ϑ_8 because they are defined by multiplication of the corresponding rows in $\hat{J}_M^{1/2}(\hat{\theta})$ by $\tilde{Z}(M)$. The latter vector satisfies the constraints $\mathcal{A}_M(\mathcal{M}^3_{\text{all}})$ rather than $\mathcal{A}_M(\mathcal{M}^3_{\text{all}} \cap \mathcal{M}_O)$, so the distribution is longer-tailed than needed. For the current simulation, the settings considered lead to $\hat{J}_M^{1/2}(\hat{\theta})$ with small off-diagonal elements, so, the distribution of an estimator is mainly determined by its corresponding Z_i. For ζ^6_{all} the coverages almost equal the nominal values, especially for $n = 100$. Using ζ_{arb} leads to conservative confidence intervals for all parameters because of the additional constraints in $\mathcal{A}_M(\mathcal{M}_{\text{arb}})$, while theoretically the constraints should be $\mathcal{A}_M(\mathcal{M}_{\text{arb}} \cap \mathcal{M}_O)$.

The method of Berk et al. (2013) always yields conservative confidence intervals although there is

no guarantee that it always leads to valid confidence intervals for the true parameters. Naive confidence intervals for ϑ_4 have coverages almost equal to the nominal value while for ϑ_6 using ζ_{arb} and ζ^3_{all} and for ϑ_8 in all settings the coverage percentages are around 70%. This is the result of wrongly treating the selected model as given. For settings with small off-diagonal elements of $\hat{J}_M^{1/2}(\hat{\theta})$, the confidence intervals for the truly non-zero parameters are valid. Other simulation results are contained in the Supplementary Material. We find that the proposed method can be used even in underparametrized situations, where assumption A1 does not hold.

5.2 Linear combinations in linear models

The performance of the proposed method for linear combinations was investigated by simulations. Let $\vartheta^\top = (2.25, -1.1, 2.43, -1.24, 2.5, 0_8^\top)$ be the true values for the parameters in a linear model, with error standard deviation either 1 or 3. Four different selection matrices are considered, ζ^i_{all}, for $i \in \{3, 5, 8, 10\}$, indicating that the first i covariates are common to each model. The data generation processes are as in Section 5.1. For this simulation, we do not control the selected model because we are interested in a linear combination of the selected parameters. Table 2 shows the results. We compare the post-selection intervals with the smoothed bootstrap confidence intervals (Efron, 2014) and the intervals for post-selection predictions (Bachoc et al., 2015). The bootstrap samples consist of n draws with replacement from the main data set and we replicate this $B = 1000$ times. The non-ideal bootstrap when the number of replications is not equal to n^n biases the variance of the smoothed bootstrap estimator upward, so we use the bias-corrected version (Efron, 2014, remark J). The post-selection intervals for prediction have a target based on the selected model, so this might be different from the true prediction.

The choice of models with ζ^3_{all} as a selection matrix results in conservative confidence intervals due to conditioning on $\mathcal{A}_M(\mathcal{M}^3_{\text{all}})$, similar to before. For this selection matrix, the confidence intervals by the bootstrap method are shorter than by the proposed post-selection method. The bootstrap confidence intervals are not directly based on the selected model for the original data because a model is selected for each bootstrap sample.

The ideal situation is when the selection matrix is ζ^5_{all}, since all truly non-zero parameters are then forced to be in the model. The confidence intervals for the proposed method are always shorter than those for the competing methods and their coverages are almost equal to the nominal value. For ζ^8_{all} and ζ^{10}_{all} the situation is the same, though with wider intervals than with ζ^5_{all} for all methods, because more parameters are forced to be in the model, which increases the variability of the predictions. These confidence intervals are not wider than for ζ^3_{all}. Thus the variability of the prediction is more affected by the condition part than by forcing more variables into the model. The post-selection method for prediction (Bachoc et al., 2015) always leads to wider confidence intervals than the bootstrap method and the proposed method.

The coverages of the confidence intervals for the proposed method are always close to or higher than the nominal values, while the bootstrap method can have lower coverage probabilities than the nominal values. Moreover, the bootstrap method for all possible models is computationally intensive, because it needs B bootstrap samples and in each of them all candidate models are fit.

For the setting $\sigma = 3$ and $n = 30$ in ζ^3_{all}, we used the results in (9) instead of (8). In this setting the probability of selecting an underparametrized model is not zero due to a small sample size and large variance. The average length of the confidence interval was 9.9 and the coverage was around 90% when we used (8).

			ζ^3_{all}		ζ^5_{all}		ζ^8_{all}		ζ^{10}_{all}	
σ	n	method	length	cov.	length	cov.	length	cov.	length	cov.
1	30	PostAIC	3·11	97	2·61	95	2·90	94	3·08	94
		Boot	3·67	92	3·32	92	3·31	92	3·79	92
		PoSIp	4·38	100	4·39	100	5·36	100	6·00	100
	100	PostAIC	1·42	98	1·17	95	1·30	96	1·37	95
		Boot	1·25	94	1·25	94	1·30	94	1·33	93
		PoSIp	1·83	100	1·83	100	2·20	100	2·42	100
3	30	PostAIC	11·76	98	7·82	94	8·68	94	9·24	94
		Boot	11·46	92	9·95	92	9·94	92	11·37	92
		PoSIp	12·65	99	13·16	100	16·08	100	17·99	100
		PostAIC	4·25	98	3·50	95	3·90	96	4·12	95
		Boot	3·77	94	3·74	94	3·90	94	4·00	93
		PoSIp	5·47	100	5·48	100	6·60	100	7·26	100

Table 2: Simulation study with 3000 runs of selection with Akaike's information criterion. Average length of 95% confidence intervals and coverage percentages (cov.) for a linear combination of the parameters for different methods and model sets using the selection matrices ζ for different sample sizes.

6 Pima Indian diabetes data

We construct confidence intervals conditional on the selected model for a logistic regression model applied to the Pima Indian diabetes data set (Lichman, 2013). This data set consists of women at least 21 years old of Pima Indian heritage, living near Phoenix Arizona. We used 332 complete observations. The response is 0 if a test for diabetes is negative and is 1 for a positive test. We use seven covariates in the model, npreg: number of pregnancies, glu: plasma glucose concentration in an oral glucose tolerance test, bp: diastolic blood pressure, skin: triceps skin fold thickness in millimeter, bmi: body mass index, ped: diabetes pedigree function and age in years. See Smith et al. (1988) for more details about the data.

First, we consider bootstrap percentile and naive confidence intervals for the parameters in the full model when no selection is involved, see Table 3(b). We used 5000 bootstrap runs, each resampling the 332 women uniformly with replacement. Several intervals contain zero, which shows the possibility of using a smaller model.

Selection uses the set $\mathcal{M}_{\text{all}}$; an intercept is present in all models. This results in selecting four variables: npreg, glu, bmi and ped. Table 3(a) presents the unconditional confidence intervals for these parameters using the naive method with the post-selection confidence intervals that condition on the model selected using the Akaike information criterion. The naive method ignores the selection procedure which leads to the significance of the covariate ped, whereas the proposed method, which takes the selection uncertainty into account, concludes that this covariate is not individually significant at the 5% level. For logistic regression, to the best of our knowledge, there are no other post-selection methods to compare with.

7 Discussion and extensions

For one of the classic model selection methods, the Akaike information criterion (Akaike, 1973) we have provided an approach to deal with the selection uncertainty by performing inference conditional on the selected model. Our results have demonstrated that this inference depends not only on the selected model,

(a)

Method	npreg	glu	bmi	ped
Naive	[0·091, 0·269]	[0·028, 0·049]	[0·042, 0·129]	[0·305, 2·050]
PostAIC	[0·058, 0·299]	[0·022, 0·054]	[0·027, 0·142]	[−0·027, 2·358]

(b)

Method	npreg	glu	bp	skin	bmi	ped	age
Naive	[0·03, 0·26]	[0·03, 0·05]	[−0·03, 0·02]	[−0·03, 0·05]	[0·02, 0·14]	[0·24, 2·00]	[−0·02, 0·05]
Bootstrap	[−0·003, 0·30]	[0·03, 0·05]	[−0·03, 0·16]	[−0·03, 0·06]	[0·02, 0·15]	[0·005, 2·41]	[−0·02, 0·07]

Table 3: (a) Confidence intervals for the Pima Indian diabetes data with nominal level 95% ignoring (Naive) and including (PostAIC) model selection using Akaike's information criterion. (b) 95% Naive and bootstrap confidence intervals in the full model, without selection.

but also on the set of models from which the selection takes place, and on the smallest overparametrized model. The dependence on the set of models is not surprising, though has not received much attention so far.

The proposed method explicitly uses the overselection properties of Akaike's information criterion. For some selection properties under local misspecification, see Claeskens & Hjort (2004). For consistent selection criteria, e.g., the Bayesian information criterion, other approaches should be used, though effects of the selection remain present (Leeb & Pötscher, 2005). Other selection methods that are similar to Akaike's information criterion can be approached in the same way. Consider, for example, selection in an arbitrary set of models allowing for model misspecification, see Section 4, using Takeuchi's information criterion (Takeuchi, 1976) $\text{TIC}(M) = 2\ell_n\{\hat{\theta}(M)\} - 2\text{tr}\{Q_M(\vartheta^*)^{-1}J_M(\vartheta^*)\}$. For most practical settings the information matrices are estimated by their empirical counterparts $\hat{Q}_M(\hat{\theta}_M)$ and $\hat{J}_M(\hat{\theta}_M)$. We rewrite (10) for an arbitrary set of models containing M_{s} by replacing $|M|$ with $\text{tr}\{Q_M(\vartheta^*)^{-1}J_M(\vartheta^*)\}$ and proceed to calculate the asymptotic distribution of the parameters conditioned on the constraint set.

Another such example is the generalized information criterion introduced by Konishi & Kitagawa (1996). It considers functional estimators, such as M-estimators, and uses the influence function as part of the criterion, $\text{GIC}(M) = -2\ell_n\{\hat{\theta}(M)\} + (2/n)\sum_{i=1}^n \text{tr}\big\{\text{Infl}(Y_i)(\partial/\partial\theta_M^\top)\log f(Y_i;\hat{\theta}_M)\big\}$. Under some regularity conditions, the functional estimator has an asymptotic normal distribution, allowing to extend the results in Section 4.

Mallows' C_p (Mallows, 1973) for linear regression is $C_p(M) = \hat{\sigma}^{-2}\hat{\sigma}^2(M) + 2|M| - n$ where $\hat{\sigma}^2$ is the estimated variance in the full model while $\hat{\sigma}^2(M)$ uses model M. The model with the smallest C_p value is the best. In nested models one can easily show that when n tends to infinity, $C_p(M) - C_p(M^*) \sim \chi^2_q/q + 2q$ where $q = |M^*| - |M|$. In the same manner as for the Akaike information criterion, one can calculate the constraint set and hence the distribution of estimators for parameters in the selected model.

In forward stepwise selection, we start from a small model and embed it in a larger model containing one additional parameter. This procedure continues until adding a parameter does not decrease the Akaike information criterion. To be precise, in step t we embed model M_t in a number of bigger models, each adding one parameter. Define $\mathcal{M}_t$ to be this set of models. Model $M_{t+1} \in \mathcal{M}_t$ is selected when this model has a smaller criterion value than model M_t and it has the smallest criterion value amongst all models in $\mathcal{M}_t$. This means that $\text{AIC}(M_{t+1}) < \text{AIC}(M_t)$ and $\text{AIC}(M_{t+1}) < \text{AIC}(M)$ for all $M \in \mathcal{M}_t \setminus M_{t+1}$. These inequalities can be translated to constraints. The constraint set is the collection of all these constraints from all steps.

We explicitly dealt with low-dimensional parameters for which maximum likelihood estimators exist and Akaike's information crierion is well-defined. Other criteria are better suited for high-dimensional parameters.

Acknowledgement

The authors wish to thank the editor and reviewers. Support of the Research Foundation Flanders, University of Leuven and of the Interuniversity Attraction Poles Research Network of the Belgian Science Policy is acknowledged. The computational resources and services used in this work were provided by the Flemish Supercomputer Center, funded by the Hercules Foundation and the Flemish Government.

Appendix

Let $\mathcal{B}_K(\epsilon)$ denote a sphere in $a+K$ dimensions centered at ϑ with radius ϵ, and denote its complementary set by $\mathcal{B}_K^c(\epsilon)$.

A1 For each $\epsilon > 0$, as $n \to \infty$, $\sup_{\theta \in \mathcal{B}_K^c(\epsilon)}\{\ell_n(\theta) - \ell_n(\vartheta)\} \to -\infty$ in probability.

A2 There exist an $\epsilon_0 > 0$ such that $\ell_n(\theta)$ is twice continuously differentiable in $\mathcal{B}_K(\epsilon_0)$ for all n large enough. Define the score vector $U_n(\theta) = (\partial/\partial\theta)\ell_n(\theta)$ and the negative Hessian matrix $Q_n(\theta) = -(\partial^2/\partial\theta\partial\theta^\top)\ell_n(\theta)$.

A3 For some $0 < \epsilon_1 < \epsilon_0$ when $n \to \infty$, there exists a non-random positive definite continuous matrix $Q(\theta)$, for θ in $\mathcal{B}_K(\epsilon_1)$ such that $\sup_{\theta \in \mathcal{B}_K(\epsilon)} \operatorname{tr}\{Q_n(\theta)/n - Q(\theta)\} \to 0$ in probability.

A4 As $n \to \infty$, $n^{1/2}U_n(\vartheta)$ is asymptotically $N\{0, J(\vartheta)\}$.

A5 For $i \neq j$ and $M_i, M_j \in \mathcal{M}_O$, with the expectation with respect to the true distribution, $J_{ij}\{\theta(i), \theta(j)\} = E(\{\partial/\partial\theta(M_i)\}[\ell_n\{\theta(M_i)\}]\{\partial/\partial\theta(M_j)^\top\}[\ell_n\{\theta(M_j)\}]) = 0_{|M_i|\times|M_j|}$.

Assumptions A1–A4 are from Woodroofe (1982). Assumption A1 leads to the consistency of maximum likelihood estimators for θ in the model considered and its submodels. For the non-nested case A5 leads to a simplification (Vuong, 1989). In linear regression, assumption A5 is equivalent to having an orthogonal design matrix.

The next lemma is an extension of Lemma A in Vuong (1989) to more than two models.

Lemma 3. *Assume A1–A4. Fix any ordering of the models in $\mathcal{M}_O$ and denote $o = |\mathcal{M}_O|$. As $n \to \infty$, $n^{1/2}(\hat{\theta}_{\mathcal{M}_o} - \vartheta_{\mathcal{M}_o}) = n^{1/2}\{\hat{\theta}'(M_1)^\top - \vartheta(M_1)^\top, \ldots, \hat{\theta}'(M_o)^\top - \vartheta(M_o)^\top\}^\top \to N\{0, \Sigma(\vartheta)\}$ in distribution.*

Proof. Similar to Vuong (1989), a Taylor series expansions leads to

$$0 = n^{-1/2}U_{n,M_i}(\vartheta) + Q_{M_i}(\vartheta)n^{1/2}\{\hat{\theta}'(M_i) - \vartheta\} + o_P(1),\ M_i \in \mathcal{M}_O.$$

By the multivariate central limit theorem, there is convergence in distribution, for $n \to \infty$,

$$n^{-1/2}(U_{n,M_1}^\top, \ldots, U_{n,M_o}^\top)^\top \to N(0, \Sigma_u) \tag{13}$$

where Σ_u is a partitioned matrix with ijth block equal to $J_{ij}(\vartheta, \vartheta)$. The distribution of the estimators follows. □

When the models are correctly specified, $J_{ii}(\vartheta, \vartheta) = J_{M_i}(\vartheta) = Q_{M_i}(\vartheta)$. Lemma 3 is also valid for misspecified models and for models not in $\mathcal{M}_O$. In such case the true parameter is replaced by the pseudo-true parameter corresponding to the considered model.

Proof of Proposition 1. We show that (1) equals

$$\lim_{n\to\infty} \frac{\mathrm{P}([n^{1/2}\{\hat{\theta}'(p)-\tilde{\vartheta}(p)\}\le \tilde{t}(p)]\cap[2\ell^*_{n,p}-2p\ge 2\ell^*_{n,j}-2j,\ j\in\{p_0,p_0+1,\ldots,K\}])}{\mathrm{P}[2\ell^*_{n,p}-2p\ge 2\ell^*_{n,j}-2j,\ j\in\{p_0,p_0+1,\ldots,K\}]}.$$

From Lemma 3 there is joint convergence of the estimators in the different models. Next, since $\ell^*_{n,j}$ is a function of $\hat{\theta}'(j)$, namely

$$\ell^*_{n,j}=\frac{n}{2}\{\hat{\theta}'(p)-\vartheta(p)\}^\top J_p(\vartheta)\{\hat{\theta}'(p)-\vartheta(p)\}+o_P(1),$$

and since the probability of the event in the denominator is strictly positive, Slutsky's theorem and the continuous mapping theorem give joint convergence for both the numerator and denominator of the above expression to their asymptotic counterparts.

To obtain the selection set let $S_j=\{s\in\mathbb{R}^{a+K}: s_i=0, \text{ for } i=a+j,\ldots,a+K\}$ for $j=p_0,\ldots,K$. Woodroofe (1982) showed that $(\ell^*_{n,p_0},\ldots,\ell^*_{n,K})$ converges in distribution to $(\ell^*_{p_0},\ldots,\ell^*_K)$ as $n\to\infty$, where for $j=p_0,\ldots,K$, $\ell^*_j=\sup_{s\in S_j}\{s'Y-s'J(\vartheta)s/2\}$, where $Y\sim N\{0,J(\vartheta)\}$. Then $\ell^*_j=0.5\sum_{i=1}^{a+j}Z_i^2$, for $j=p_0,\ldots,K$, where $Z_1,\ldots,Z_{a+j}$ are independent and identically distributed standard normal random variables. Lemma 1 and Assumptions A1–A4 imply that $n^{1/2}J_j^{1/2}(\vartheta)\{\hat{\theta}'(j)-\tilde{\vartheta}(j)\}$ converges in distribution to $\tilde{Z}(j)$ as $n\to\infty$. Parameters not in the selected model are set to zero, which leads to the region $\mathcal{T}_p$. Since $\tilde{Z}(p)$ and $(Z_{p+1},\ldots,Z_K)$ are independent, and for $t\in\mathcal{T}_p$,

$$\begin{aligned}F_p(t) &= \mathrm{P}\{J_p^{-1/2}(\vartheta)\tilde{Z}(p)\le\tilde{t}(p)\mid Z\in\mathcal{A}_p(\mathcal{M}_{\text{nest}})\}\\ &= \mathrm{P}\big[J_p^{-1/2}(\vartheta)\tilde{Z}(p)\le\tilde{t}(p)\mid \bigcap_{j=p_0,\ldots,p-1}\{\sum_{i=j+1}^{p}Z_{a+i}^2>2(p-j)\}\big].\end{aligned}\tag{14}$$

□

Proof of Corollary 2. From Proposition 1, with $\hat{p}_0=p$, q_α is equivalently found via

$$\mathrm{P}\big[(\sum_{i=1}^{a+p}Z_i^2\le q_\alpha)\cap\bigcap_{j=p_0,\ldots,p-1}\{\sum_{i=j+1}^{p}Z_{a+i}^2>2(p-j)\}\big]/\mathrm{P}\{\tilde{Z}(p)\in\mathcal{A}_p^{(s)}(\mathcal{M}_{\text{nest}})\}=1-\alpha.$$

The denominator can be calculated by Lemma A1 in the Supplementary Material. To calculate the numerator, we first find the joint density of $(W_p,\ldots,W_{p_0+1},W_1)$ where $W_j=\sum_{i=a+j}^{a+p}Z_i^2$, $W_1=\sum_{i=1}^{p}Z_i^2$ and $Z_i^2\sim\chi_1^2$ for all $i=1,\ldots,a+p$. So, $Z_{a+i}^2=W_{i-1}-W_i$ for $i=p_0+1,\ldots,p-1$ and $Z_{a+p}^2=W_p$ with $\sum_{i=1}^{a+p_0}Z_i^2=W_1-W_{p_0+1}\sim\chi^2_{a+p_0}$. The joint distribution of $(W_p,\ldots,W_{p_0+1},W_1)$ is obtained via a transformation of the distribution of $(Z^2_{a+p},Z^2_{a+p-1},\ldots,Z^2_{a+p_0+1},\sum_{i=1}^{a+p_0}Z_i^2)$,

$$f(w_p,\ldots,w_{p_0+1},w_1)=\frac{\exp(-w_1/2)w_p^{-1/2}(w_1-w_{p_0+1})^{-(a+p_0)/2-1}\prod_{i=1}^{p-p_0+1}(w_i-w_{i-1})^{-1/2}}{2^{\frac{a+p}{2}}\{\Gamma(1/2)\}^{p-p_0}\Gamma(\frac{a+p_0}{2})}.$$

The region of integration follows from $\mathcal{A}_p^{(s)}(\mathcal{M}_{\text{nest}})$ and the fact that $W_i\le W_j$ for $i>j$. □

Proof of Lemma 1. Denote the smallest true model by M_{pars}. For all $M'\notin\mathcal{M}_O$, by assumption A1,

$$\begin{aligned}\mathrm{P}(M_{\text{AIC}}=M') &\le \mathrm{P}\left\{\text{AIC}^*(M')\ge\max_{M\in\mathcal{M}_O}\text{AIC}^*(M)\right\}\le\mathrm{P}\left\{\text{AIC}^*(M')\ge\text{AIC}^*(M_{\text{pars}})\right\}\\ &= \mathrm{P}\left[\ell_n\{\hat{\theta}(M')\}-|M'|\ge\ell_n\{\hat{\theta}(M_{\text{pars}})\}-|M_{\text{pars}}|\right]\\ &= \mathrm{P}\left[\ell_n\{\hat{\theta}(M')\}-\ell_n\{\vartheta(M_{\text{pars}})\}-|M'|\ge\ell_n\{\hat{\theta}(M_{\text{pars}})\}-\ell_n\{\vartheta(M_{\text{pars}})\}-|M_{\text{pars}}|\right]\\ &\to 0.\end{aligned}$$

□

Proof of Proposition 2. (i) Define $S_j = \{s \in \mathbb{R}^{a+K} : s_i = 0,\ i \notin M\}$ and $\ell^*_{n,M_i} = \ell_n\{\hat{\theta}(M_i)\} - \ell_n(\vartheta)$ where $M_i \in \mathcal{M}_O$. Similar to Proposition 1 we can show that for $M_i \in \mathcal{M}_O$, $\ell^*_{n,M_i} \to 0.5\sum_{j\in M_i} Z_j^2$ in distribution. Now, the condition part can be calculated by

$$\sum_{j\in M} Z_j^2 - 2|M| > \sum_{j\in M_i} Z_j^2 - 2|M_i|, \quad M_i \in \mathcal{M}_O \backslash M,$$

which is equivalent to $Z \in \mathcal{A}_M(\mathcal{M}_O)$.

(ii) By Lemma 3 there is joint convergence in distribution of the estimators in the different models. The constraint set can be calculated by pairwise comparisons of the AIC* values. To do so, write

$$\ell_n\{\hat{\theta}(M_{\rm i})\} = \ell_n(\vartheta) + \frac{n}{2}\{\hat{\theta}(M_i) - \vartheta\}^\top Q_{M_i}(\vartheta)\{\hat{\theta}(M_i) - \vartheta\} + o_P(1)$$

from which it follows that $\ell^*_{n,i} = \frac{n}{2}\{\hat{\theta}(M_i) - \vartheta\}^\top Q_{M_i}(\vartheta)\{\hat{\theta}(M_i) - \vartheta\} + o_P(1)$.

Then, since $\text{AIC}^*(M_{\rm AIC}) \geq \text{AIC}^*(M_i)$ is equivalent to $2(\ell^*_{n,\rm AIC} - \ell^*_{n,i}) \geq 2(|M_{\rm AIC}| - |M_{\rm i}|)$ it follows that

$$n(\hat{\theta}_{\mathcal{M}_O} - \vartheta_{\mathcal{M}_O})^\top W_{\rm AIC,i}(\hat{\theta}_{\mathcal{M}_O} - \vartheta_{\mathcal{M}_O}) + o_P(1) - 2(|M_{\rm AIC}| - |M_i|) \geq 0. \tag{15}$$

By using Lemma 3 and the continuous mapping theorem, the asymptotic counterpart of (15) can be written as $Z^\top\Sigma^{1/2}W_{\rm AIC,i}\Sigma^{1/2}Z \geq 2(|M_{\rm AIC}| - |M_i|), \quad M_i \in \mathcal{M}_O$, which results in the stated selection region and limiting distribution. □

Proof of Proposition 3. (i) Using Theorems 1 and 2 of Sweeting (1980),

$$n^{1/2}\{\hat{\theta}'(M) - \tilde{\vartheta}(M)\}^\top J_M^{1/2}(\vartheta) \longrightarrow \tilde{Z}(M),$$

uniformly in distribution over the compact set Θ. This leads to having $\lim_{n\to\infty}\inf_{\vartheta\in\Theta} \mathrm{P}_\vartheta\{\vartheta \in C_\alpha(\vartheta)\} = 1-\alpha$. (ii) When $\mathcal{M}_O$ is not known, we use $\mathcal{A}_M(\mathcal{M}_{\rm arb})$ in (7) instead of $\mathcal{A}_M(\mathcal{M}_{\rm arb} \cap \mathcal{M}_O)$, which defines the value $\tilde{q}_\alpha$. Since $\mathcal{A}_M(\mathcal{M}_{\rm arb}) \subset \mathcal{A}_M(\mathcal{M}_{\rm arb} \cap \mathcal{M}_O)$, $\tilde{q}_\alpha \geq q_\alpha$, which leads to a conservative confidence region. □

Proof of Lemma 2. For every $j = 1, \ldots, J$ and every component k of the vector $\hat{\theta}_{n,\mathcal{M}}(M_j)$, it holds that

$$n^{1/2}([\hat{\theta}_{n,\mathcal{M}}(M_j)]_k - [\vartheta^*_{n,\mathcal{M}}(M_j)]_k) = \sum_{i=1}^n Q^{-1}_{M_j}\{\vartheta^*_n(M_j)\}n^{-1/2}\frac{\partial}{\partial\theta_k}\log f_{j,i}\{Y_i, \vartheta^*_n(M_j)\} + o_P(1).$$

By assumption (i), which is a Lindeberg assumption for all $G_n \in \mathcal{G}_n$, we obtain a uniform limiting normality result for each of the components of $n^{1/2}(\hat{\theta}_{n,\mathcal{M}} - \vartheta^*_{n,\mathcal{M}})$. Under assumption (ii) the data are in a so-called null triangular array format, to which Corollary 2 of Pollak (1972) applies, resulting in a joint asymptotic normality for the vector combining all such components. □

Proof of Proposition 4. Define the events $B = [n^{1/2}\{\hat{\theta}(M_{\rm AIC}) - \vartheta^*(M_{\rm AIC})\} \leq t]$ and

$$C = \cap_{M\in\mathcal{M}}\{n(\hat{\theta}_{\mathcal{M}} - \vartheta^*_{\mathcal{M}})^\top(W_{M_{\rm AIC},M_{\rm s}} - W_{M_{\rm i},M_{\rm s}})(\hat{\theta}_{\mathcal{M}} - \vartheta^*_{\mathcal{M}}) \geq 2(|M_{\rm AIC}| - |M|)\} + o_P(1).$$

Using the results of Lemma 2 and the continuous mapping theorem, the difference between

$$\mathrm{P}[n^{1/2}\{\hat{\theta}(M_{\rm AIC}) - \vartheta^*(M_{\rm AIC})\} \leq t \mid \hat{M} = M_{\rm AIC}]\mathrm{P}(B\cap C)/\mathrm{P}(C)$$

and $\mathrm{P}[\{\Sigma_{M_{\rm AIC}}(\vartheta^*_{M_{\rm AIC}})^{1/2}\tilde{Z}(M_{\rm AIC}) \leq t\} \cap \{Z \in \mathcal{A}_{M_{\rm AIC}}(\mathcal{M})\}]/\mathrm{P}\{Z \in \mathcal{A}_{M_{\rm AIC}}(\mathcal{M})\}$ converges uniformly to 0. □

References

AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory, Tsahkadsor, Armenia, USSR, September 2-8, 1971*, B. Petrov & F. Cski, eds. Budapest: Akadémiai Kiadó, 267–281.

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

BACHOC, F., LEEB, H. & PÖTSCHER, B. (2015). Valid confidence intervals for post-model-selection predictors. *arXiv: 1412.4605* .

BELLONI, A., CHERNOZHUKOV, V. & KATO, K. (2015). Uniform post selection inference for least absolute deviation regression models and other z-estimation problems. *Biometrika* **102**, 77–94.

BERK, R., BROWN, L., BUJA, A., ZHANG, K. & ZHAO, L. (2013). Valid post-selection inference. *The Annals of Statistics* **41**, 802–837.

CHERNOZHUKOV, V., HANSEN, C. & SPINDLER, M. (2015). Valid post-selection and post-regularization inference: An elementary, general approach. *Annual Review of Economics* **7**, 649–688.

CLAESKENS, G. & HJORT, N. L. (2004). Goodness of fit via nonparametric likelihood ratios. *Scandinavian Journal of Statistics* **31**, 487–513.

CLAESKENS, G. & HJORT, N. L. (2008). *Model Selection and Model Averaging*. Cambridge: Cambridge University Press.

COX, D. R. & HINKLEY, D. V. (1974). *Theoretical Statistics*. London: Chapman and Hall.

DANILOV, D. & MAGNUS, J. R. (2004). On the harm that ignoring pretesting can cause. *Journal of Econometrics* **122**, 27–46.

EFRON, B. (2014). Estimation and accuracy after model selection. *Journal of the American Statistical Association* **109**, 991–1007.

FERRARI, D. & YANG, Y. (2014). Confidence sets for model selection by F-testing. *Statistica Sinica* **25**, 1637–1658.

HJORT, N. L. & CLAESKENS, G. (2003). Frequentist model average estimators. *Journal of the American Statistical Association* **98**, 879–899.

JANSEN, M. (2014). Information criteria for variable selection under sparsity. *Biometrika* **101**, 37–55.

KABAILA, P. (1995). The effect of model selection on confidence regions and prediction regions. *Econometric Theory* **11**, 537–549.

KABAILA, P. (1998). Valid confidence intervals in regression after variable selection. *Econometric Theory* **14**, 463–482.

KABAILA, P. & LEEB, H. (2006). On the large-sample minimal coverage probability of confidence intervals after model selection. *Journal of the American Statistical Association* **101**, 619–629.

KABAILA, P., WELSH, A. H. & ABEYSEKERA, W. (2016). Model-averaged confidence intervals. *Scandinavian Journal of Statistics* **43**, 35–48.

KONISHI, S. & KITAGAWA, G. (1996). Generalized information criteria in model selection. *Biometrika* **83**, 875–890.

LEE, J. D., SUN, D. L., SUN, Y. & TAYLOR, J. E. (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics* **44 (3)**, 907–927.

LEEB, H. & PÖTSCHER, B. (2005). Model selection and inference: Facts and fiction. *Econometric Theory* **21**, 22–59.

LEEB, H., PÖTSCHER, B. & EWALD, K. (2015). On various confidence intervals post-model-selection. *Statistical Science* **30(2)**, 216–227.

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

LEEB, H. & PÖTSCHER, B. M. (2006). Can one estimate the conditional distribution of post-model-selection estimators? *The Annals of Statistics* **34**, 2554–2591.

LICHMAN, M. (2013). UCI Machine Learning Repository, University of California, Irvine, School of Information and Computer Sciences.

MALLOWS, C. (1973). Some comments on Cp. *Technometrics* **15**, 661–675.

PAKMAN, A. & PANINSKI, L. (2014). Exact Hamiltonian Monte Carlo for truncated multivariate Gaussians. *Journal of Computational and Graphical Statistics* **23(2)**, 518–542.

POLLAK, M. (1972). A note on infinitely divisible random vectors. *Annals of Mathematical Statistics* **43**, 673–675.

PÖTSCHER, B. (1991). Effects of model selection on inference. *Econometric Theory* **7(2)**, 163–185.

PÖTSCHER, B. (1995). Comment on The effect of model selection on confidence regions and prediction regions, by P. Kabaila. *Econometric Theory* **11**, 550–559.

SHIBATA, R. (1976). Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika* **63**, 117–126.

SMITH, J. W., EVERHART, J. E., DICKSON, W. C., KNOWLER, W. C. & JOHANNES, R. S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. *Proc. Symp. on Computer Applications and Medical Care* , 261–265.

SWEETING, T. J. (1980). Uniform asymptotic normality of the maximum likelihood estimator. *The Annals of Statistics* **8**, 1375–1381.

TAKEUCHI, K. (1976). Distribution of information statistics and criteria for adequacy of models. *Biometrika* **153**, 12–18.

TAYLOR, J., LOCKHART, R., TIBSHIRANI, R. J. & TIBSHIRANI, R. (2016). Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association* **111**, 600–620.

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

VUONG, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* **57(2)**, 307–333.

WHITE, H. (1994). *Estimation, Inference and Specification Analysis*. New York: Cambridge University Press.

WOODROOFE, M. (1982). On model selection and the arc-sine laws. *The Annals of Statistics* **10**, 1182–1194.

Supplementary Material: Asymptotic post-selection inference for Akaike's information criterion

This supplement contains a rewriting of results of Woodroofe (1982), exact calculations for an example, the selection matrix for one of the simulation settings and additional simulation results.

A Additional lemma

The following Lemma is adapted from Woodroofe (1982). The zero probability of underestimation is a special case of our Lemma 1, while the exact expressions for overestimation are obtained by rewriting the generalized arc-sine probabilities of Woodroofe (1982).

Lemma 4. *Under assumptions (A1)–(A4), in the nested models case for the model order $\widehat{p}_0$ selected such that the Akaike information criterion is minimized for the corresponding model, it holds that*

$$\pi_p = \lim_{n\to\infty} P(\widehat{p}_0 = p) = \begin{cases} 0 & \text{if} \quad a \le p < p_0, \\ g_{p-p_0} q_{K-p} & \text{if} \quad p_0 \le p \le K, \end{cases}$$

where $g_0 = q_0 = 1$ and with $\mathfrak{R}_i = \{(r_1, \ldots, r_i) \in \mathbb{N}^i : r_1 + 2r_2 + \ldots + ir_i = i\}$, $a_j = P(\chi^2_j > 2j)$,

$$g_i = \sum_{\mathfrak{R}_i} \left\{ \prod_{j=1}^{i} \frac{1}{r_j!} \left(\frac{a_j}{j} \right)^{r_j} \right\} \text{ and } q_i = \sum_{\mathfrak{R}_i} \left\{ \prod_{j=1}^{i} \frac{1}{r_j!} \left(\frac{1-a_j}{j} \right)^{r_j} \right\}.$$

B A worked-out illustrative example

Let $\mathrm{erf}(x) = 2\pi^{-1/2} \int_0^x \exp(-w^2)\,\mathrm{d}w$ denote the 'error function' and let $\mathrm{erfc}(x) = 1 - \mathrm{erf}(x)$. Assume $\emptyset = M_0 \subset \ldots \subset M_3$ with the true value $\vartheta = (\vartheta_1, 0, 0)^\top$ and three situations for the 3×3 matrix

$J^{-1/2}(\vartheta)$,

$$(a)\begin{pmatrix}1&0&0\\0&2&0\\0&0&2\end{pmatrix},\quad (b)\begin{pmatrix}1&0&0\\0&2&0.5\\0&0.5&2\end{pmatrix},\quad (c)\begin{pmatrix}1&0.9&0.9\\0.9&2&0.5\\0.9&0.5&2\end{pmatrix}.$$

The Akaike information criterion is used to select a model from the set of three nested models $\mathcal{M}_{\text{nest}} = \{M_1, M_2, M_3\}$. Consider the situation that the smallest value of the Akaike information criterion is attained for the full model, thus $\widehat{p}_0 = 3$. In this case $\mathcal{A}_3 = \{(z_1, z_2, z_3) : z_3^2 > 2, z_3^2 + z_2^2 > 4\}$. Using Lemma A.4 with $p_0 = 1$, $K = 3$ and $\widehat{p}_0 = 3$ results in $P(Z \in \mathcal{A}_3) = 0.08$.

Let $f_{j|3}$ denote the limiting density of $n^{1/2}(\widehat{\theta}_j - \vartheta_j)$ conditional on $\widehat{p}_0 = 3$, then for case (a) $f^{(a)}_{1|3}(t) = \phi(t)$ where ϕ is the standard normal density function,

$$\{0.16(2\pi)^{1/2}\}f^{(a)}_{2|3}(t) = \begin{cases}\exp(-t^2/8)\text{erfc}\left\{(2-t^2/8)^{1/2}\right\}, & t \in (-2^{3/2}, 2^{3/2})\\ \exp(-t^2/8)\text{erfc}(1), & t \in \mathbb{R}\backslash[-2^{3/2}, 2^{3/2}]\\ 0 & \text{otherwise},\end{cases}$$

$$\{0.16(2\pi)^{1/2}\}f^{(a)}_{3|3}(t) = \begin{cases}\exp(-t^2/8)\text{erfc}\left\{(2-t^2/8)^{1/2}\right\} & t \in (-4, -2^{3/2}) \cup (2^{3/2}, 4)\\ \exp(-t^2/8) & t \in (-\infty, -4] \cup [4, \infty)\\ 0 & \text{otherwise},\end{cases}$$

For case (b) where there is correlation between the second and third component of the estimator, we only calculate $f_{3|3}$, with similar results for $f_{2|3}$. The limiting distribution of $n^{1/2}(\widehat{\theta}_3 - \vartheta_3)$ conditional on $\widehat{p}_0 = 3$ is the distribution of the third row in $J^{-1/2}(\vartheta)Z_3$ which is $T = 0{\cdot}5Z_2 + 2Z_3$. We define

$$\begin{aligned} g_1(t) &= \text{erf}\{17^{1/2} - t4(2/17)^{1/2}\},\\ g_2(t) &= \text{erf}\{17^{1/2} + t4(2/17)^{1/2}\},\\ g_3(t) &= \text{erf}\{(2 - 2t^2/17)^{1/2}\}.\end{aligned}$$

By tedious calculations, we find the distribution of $n^{1/2}(\widehat{\theta}_3 - \vartheta_3)$ conditional on $\widehat{p}_0 = 3$ as follows,

$$\{0.08(34\pi)^{1/2}\}f^{(b)}_{3|3}(t) \tag{16}$$
$$= \begin{cases}\exp(-2t^2/17)\,(2 - g_1(t) - g_2(t)) & t \in (-2^{-1/2}3, 2^{-1/2}3)\\ \exp(-2t^2/17)\,(2 - g_2(t) - g_3(t)) & t \in [2^{-1/2}3, 2^{-1/2}5)\\ \exp(-2t^2/17)\,(2 - g_1(t) - g_3(t)) & t \in (-2^{-1/2}5, -2^{-1/2}3]\\ \exp(-2t^2/17)\,(2 - g_1(t) - g_2(t) - 2g_3(t)) & t \in (-17^{-1/2}, -2^{-1/2}5] \cup [2^{-1/2}5, 17^{1/2})\\ \exp(-2t^2/17)\,(2 - g_1(t) - g_2(t)) & t \in (-\infty, -17^{-1/2}] \cup [17^{-1/2}, \infty)\\ 0 & \text{otherwise}.\end{cases}$$

For case (c) we calculate for $f_{3|3}$ the distribution of $W = T + 0{\cdot}9Z_1$ where T has a density function as in (16). Hence for case (c),

$$f^{(c)}_{3|3}(w) = \int_{\mathcal{D}(T)} f^{(b)}_{3|3}(t)\phi_{0\cdot9}(w - t)\,\mathrm{d}t, \tag{17}$$

where $\mathcal{D}(T)$ is the domain of random variable T and $\phi_{0\cdot9}$ is the density of a normal random variable with standard deviation $0{\cdot}9$.

In the naive approach, often out of convenience, one wrongly assumes that $\widehat{p}_0$ is deterministic, not random, and hence one constructs the confidence interval for the parameters using an assumed asymptotic normal distribution of the maximum likelihood estimators. For instance, with $J_p^{-1/2}(\vartheta)$ as in (a),

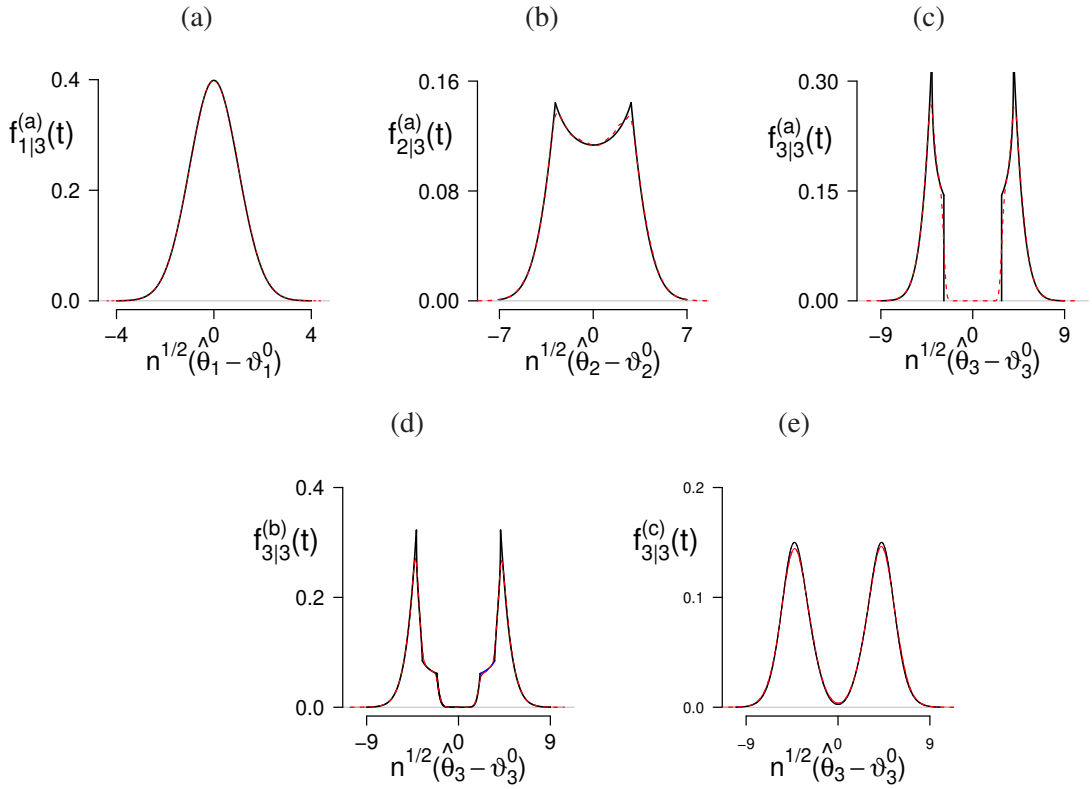

Figure 3: Marginal asymptotic density of $n^{1/2}(\hat{\theta}_j - \vartheta_j)$ conditional on $\hat{p}_0 = 3$ when $p_0 = 1$ for $j = 1, 2, 3$ and for case (a) in panels (a)–(c), for the third component of case (b) in panel (d) and for that of case (c) in panel (e). Dashed line: kernel density estimate using the simulated values; solid line: exact asymptotic density.

the naive 95% confidence interval for ϑ_3 is $\widehat{\theta}_3 \pm 1{\cdot}96(2n^{-1/2})$. Clearly, this confidence interval does not consider the uncertainty of model selection. Rather, we should use the quantile of the symmetric conditional distribution. The exact $0{\cdot}975\%$ quantile is $5{\cdot}75$ while the simulated one is $5{\cdot}77$. Hence, the conditional confidence interval is $\widehat{\theta}_3 \pm 5{\cdot}75n^{-1/2}$, clearly showing the overoptimism in the meaning of having a too narrow interval, when neglecting the model selection uncertainty. It should be noted that for case (a) the limiting probability of $n^{1/2}(\widehat{\theta}_3 - \vartheta_3)$ in $[-2^{3/2}, 2^{3/2}]$ is zero. The density function is not only bimodal, but also has quite some curvature. The simulation method captures these properties almost perfectly. For this information matrix, the limiting probability of $n^{1/2}(\widehat{\theta}_3 - \vartheta_3)$ to be in $[0, 2^{-1/2}3]$ is equal to $0{\cdot}0141$ while based on our sampling method, we find $0{\cdot}0139$. Again, the naive confidence interval $\widehat{\theta}_3 \pm 1{\cdot}96(4{\cdot}25/n)^{1/2}$ is too narrow as compared to the conditional confidence interval $\widehat{\theta}_3 \pm 5{\cdot}93n^{-1/2}$ where $5{\cdot}93$ is the exact $0{\cdot}975\%$ quantile ($5{\cdot}94$ based on the simulated distribution).

The diversity of the shape of the density functions after model selection is illustrated with case (c) for $n^{1/2}(\widehat{\theta}_3 - \vartheta_3)$. The plot of the exact limiting density is based on numerical integration from (17). The $97{\cdot}5\%$ quantile is equal to $6{\cdot}48$, again larger than the unconditional value $1{\cdot}96(5{\cdot}06^{1/2}) = 4{\cdot}41$.

For simultaneous confidence regions we compute q_α in equation (3) in the paper via constrained χ^2 distributions. An exact calculation is possible when the difference between the number of selected and true parameters is less than three. Table 4 presents the simulated 95% quantiles for some values of p_0 and p. Using these values in equation (3) in the paper gives coverages, showing close agreement with

(p_0, p)	simulated quantile	coverage	χ^2 quantile
(2,3)	10·78	95·0	7·81
(3,4)	12·29	94·8	9·49
(1,3)	12·17	94·9	7·81
(2,4)	13·76	94·8	9·49
(3,5)	15·36	94·9	11·07
(10,12)	25·28	95·0	21·03
(28,30)	47·97	94·9	43·77

Table 4: Simulated quantiles and their exact coverage percentages along with unconditional quantiles of χ^2 distributions.

the nominal value. The unconditional quantiles from χ^2 distributions are obviously too small, resulting in too optimistic inference, that is, too low coverage probabilities.

C Selection matrix for $\mathcal{M}_{\text{arb}}$ in Section 4

$$\zeta_{\text{arb}} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}_{14\times 10}.$$

D Effect of p_0 on confidence intervals in nested models

This simulation study illustrates that in nested models considering the smallest model as the true model leads to confidence intervals with higher coverage probabilities than the nominal value.

Take $\theta = (2{\cdot}25, -1{\cdot}1, 2{\cdot}43, -1{\cdot}24, 2{\cdot}5, 0_3)^\top$ as the true parameters in a linear regression model, thus $p_0 = 5$, $a = 1$ and $K = 7$. $\mathcal{M}_{\text{nest}}$ consists of 8 nested models; the smallest model contains only an intercept, the biggest model is the model with all covariates. The sample size varies in $\{30, 100, 300\}$. All other settings are as in Section 4·1.

For each sample size we generate data until each of the model orders 5, 6, 7 and 8 has been selected 3000 times. While in this simulation we know the true order is 5, we ignore this information by considering all possible values for the true order which are smaller or equal to $\widehat{p}_0$. A confidence interval for

each parameter of the selected model is calculated. For example, when $\widehat{p}_0 = 6$, this means six intervals. Confidence intervals for the post-selection method of Berk et al. (2013) are reported for comparison. The confidence intervals and their coverage have been calculated as in Section 4.1.

Tables 5–**??** present the confidence intervals for $\theta_1, \ldots, \theta_8$ in different settings under different assumptions for p_0. For moderate and relatively large sample sizes, 100 and 300, when we assume $p_0 = 5$ the simulation shows the validity of the proposed method for each of the selected orders $\hat{p}$. Smaller values of the assumed p_0 lead to wider intervals. For $n = 30$ the coverage probabilities decrease by increasing $\hat{p}$ which is due to a too small sample size for an accurate estimation of the full 8×8 information matrix. When $\widehat{p} = p_0$, the confidence intervals correspond to the naive confidence intervals, which have coverage probabilities close to the nominal value for $\theta_1, \ldots, \theta_5$ while for the other parameters they fail to produce the correct intervals by ignoring the constraints in the selection procedure.

		$\hat{p}$							
n	p_0	5		6		7		8	
30	1	[1·85, 2·66]	95	[1·85, 2·64]	94	[1·85, 2·64]	93	[1·86, 2·65]	93
	2	[1·85, 2·65]	95	[1·85, 2·64]	94	[1·86, 2·64]	93	[1·87, 2·64]	92
	3	[1·85, 2·65]	95	[1·86, 2·64]	93	[1·86, 2·64]	93	[1·87, 2·64]	92
	4	[1·85, 2·65]	95	[1·86, 2·63]	93	[1·86, 2·64]	92	[1·87, 2·64]	92
	5	[1·86, 2·64]	95	[1·86, 2·63]	93	[1·86, 2·63]	92	[1·87, 2·64]	91
	6	-	-	[1·87, 2·63]	93	[1·87, 2·63]	92	[1·87, 2·63]	91
	7	-	-	-	-	[1·87, 2·63]	91	[1·88, 2·63]	91
	8	-	-	-	-	-	-	[1·88, 2·63]	90
	PoSI	[1·63, 2·87]	100	[1·65, 2·85]	99	[1·65, 2·85]	99	[1·67, 2·84]	99
100	1	[2·05, 2·45]	96	[2·05, 2·45]	95	[2·05, 2·45]	95	[2·05, 2·45]	94
	2	[2·05, 2·45]	95	[2·05, 2·45]	95	[2·05, 2·45]	95	[2·05, 2·45]	94
	3	[2·05, 2·45]	95	[2·05, 2·45]	95	[2·05, 2·45]	95	[2·05, 2·45]	94
	4	[2·05, 2·45]	95	[2·05, 2·45]	95	[2·05, 2·45]	95	[2·05, 2·45]	94
	5	[2·05, 2·45]	95	[2·05, 2·45]	95	[2·05, 2·44]	95	[2·05, 2·45]	94
	6	-	-	[2·05, 2·45]	95	[2·05, 2·44]	95	[2·05, 2·45]	94
	7	-	-			[2·05, 2·44]	95	[2·05, 2·45]	94
	8	-	-			-	-	[2·05, 2·45]	94
	PoSI	[1·93, 2·54]	100	[1·96, 2·54]	99	[1·96, 2·53]	99	[1·96, 2·54]	99
300	1	[2·14, 2·37]	96	[2·14, 2·36]	96	[2·14, 2·36]	96	[2·14, 2·36]	95
	2	[2·14, 2·37]	96	[2·14, 2·36]	96	[2·14, 2·36]	96	[2·14, 2·36]	95
	3	[2·14, 2·37]	96	[2·14, 2·36]	96	[2·14, 2·36]	96	[2·14, 2·36]	95
	4	[2·14, 2·37]	96	[2·14, 2·36]	96	[2·14, 2·36]	96	[2·14, 2·36]	95
	5	[2·14, 2·37]	96	[2·14, 2·36]	96	[2·14, 2·36]	96	[2·14, 2·36]	95
	6	-	-	[2·14, 2·36]	96	[2·14, 2·36]	96	[2·14, 2·36]	95
	7	-	-	-	-	[2·14, 2·36]	96	[2·14, 2·36]	95
	8	-	-	-	-	-	-	[2·14, 2·36]	95
	PoSI	[2·09, 2·42]	100	[2·09, 2·41]	99	[2·09, 2·41]	100	[2·09, 2·41]	99

Table 5: Average simulated post-selection confidence intervals when Akaike's information criterion is used for selection, for θ_1, together with the average coverage percentage for different scenarios and different assumptions regarding p_0, and the post-selection interval by Berk et al. (2013).

		$\hat{p}$							
n	p_0	5		6		7		8	
30	1	$[-1.59, -0.61]$	97	$[-1.58, -0.63]$	95	$[-1.57, -0.62]$	95	$[-1.58, -0.63]$	93
	2	$[-1.55, -0.65]$	95	$[-1.55, -0.66]$	94	$[-1.54, -0.65]$	92	$[-1.55, -0.66]$	91
	3	$[-1.55, -0.66]$	95	$[-1.55, -0.67]$	93	$[-1.54, -0.66]$	92	$[-1.54, -0.66]$	91
	4	$[-1.54, -0.66]$	95	$[-1.54, -0.67]$	93	$[-1.54, -0.66]$	92	$[-1.54, -0.67]$	91
	5	$[-1.53, -0.67]$	94	$[-1.54, -0.68]$	93	$[-1.53, -0.67]$	92	$[-1.54, -0.67]$	90
	6	–	–	$[-1.53, -0.69]$	92	$[-1.53, -0.67]$	91	$[-1.53, -0.67]$	90
	7	–	–	–	–	$[-1.52, -0.68]$	91	$[-1.53, -0.68]$	90
	8	–	–	–	–	–	–	$[-1.52, -0.69]$	89
	PoSI	$[-1.78, -0.42]$	100	$[-1.77, -0.44]$	99	$[-1.76, -0.43]$	99	$[-1.77, -0.44]$	99
100	1	$[-1.34, -0.86]$	97	$[-1.34, -0.86]$	96	$[-1.34, -0.86]$	96	$[-1.34, -0.86]$	96
	2	$[-1.32, -0.87]$	96	$[-1.32, -0.88]$	95	$[-1.32, -0.88]$	94	$[-1.32, -0.87]$	95
	3	$[-1.32, -0.88]$	96	$[-1.32, -0.88]$	95	$[-1.32, -0.88]$	94	$[-1.32, -0.87]$	95
	4	$[-1.32, -0.88]$	95	$[-1.32, -0.88]$	94	$[-1.32, -0.88]$	94	$[-1.32, -0.87]$	95
	5	$[-1.31, -0.88]$	95	$[-1.32, -0.88]$	94	$[-1.32, -0.88]$	94	$[-1.32, -0.88]$	95
	6	–	–	$[-1.32, -0.88]$	94	$[-1.32, -0.88]$	93	$[-1.32, -0.88]$	95
	7	–	–	–	–	$[-1.32, -0.88]$	93	$[-1.32, -0.88]$	94
	8	–	–	–	–	–	–	$[-1.32, -0.88]$	94
	PoSI	$[-1.41, -0.78]$	99	$[-1.42, -0.78]$	99	$[-1.42, -0.78]$	99	$[-1.42, -0.78]$	99
300	1	$[-1.24, -0.96]$	98	$[-1.24, -0.96]$	97	$[-1.24, -0.97]$	97	$[-1.24, -0.96]$	96
	2	$[-1.23, -0.98]$	97	$[-1.23, -0.97]$	96	$[-1.23, -0.97]$	95	$[-1.23, -0.97]$	95
	3	$[-1.23, -0.98]$	97	$[-1.22, -0.97]$	96	$[-1.23, -0.98]$	95	$[-1.23, -0.97]$	95
	4	$[-1.22, -0.98]$	97	$[-1.22, -0.97]$	96	$[-1.23, -0.98]$	95	$[-1.23, -0.97]$	95
	5	$[-1.22, -0.98]$	96	$[-1.22, -0.98]$	96	$[-1.23, -0.98]$	95	$[-1.23, -0.97]$	95
	6	–	–	$[-1.22, -0.98]$	95	$[-1.23, -0.98]$	95	$[-1.23, -0.97]$	95
	7	–	–	–	–	$[-1.23, -0.98]$	95	$[-1.22, -0.97]$	94
	8	–	–	–	–	–	–	$[-1.22, -0.98]$	94
	PoSI	$[-1.28, -0.93]$	100	$[-1.28, -0.92]$	99	$[-1.28, -0.92]$	99	$[-1.28, -0.92]$	99

Table 6: Average simulated post-selection confidence intervals for θ_2, together with the average coverage percentage for different scenarios and different assumptions regarding p_0. Also given are the results of the post-selection interval by Berk et al. (2013).

E PostAIC confidence intervals for linear combinations in nested models

Four different scenarios for the true parameters are considered,

$$\begin{aligned}
\text{Scenario 1} &: \theta = (2{\cdot}25, -1{\cdot}1, 2{\cdot}43, -1{\cdot}24, 2{\cdot}5)^\top,\\
\text{Scenario 2} &: \theta = (2{\cdot}25, -1{\cdot}1, 2{\cdot}43, -1{\cdot}24, 2{\cdot}5, 0_3)^\top,\\
\text{Scenario 3} &: \theta = (2{\cdot}25, -1{\cdot}1, 2{\cdot}43, -1{\cdot}24, 2{\cdot}5, 0_{12})^\top,\\
\text{Scenario 4} &: \theta = (2{\cdot}25, 0_7)^\top.
\end{aligned}$$

In Scenario 1 the largest model is the true model. Scenarios 2 and 3 are dealing with the true model somewhere in between but with different numbers of candidate models and redundant variables. In Scenario 4 the smallest model is the true model. The error standard deviation varies in the set $\{0{\cdot}5, 1, 3\}$ and other settings for data generation process are the same as in the previous section. We use $\mathcal{M}_{\text{nest}}$ to select a model which is used to make a confidence interval for $x^\top\theta$, with x an out-of-sample new observation. We run the simulation 3000 times for all settings.

Table **??** presents the average length of the intervals over 3000 runs with their coverage percentages

		$\hat{p}$							
n	p_0	5		6		7		8	
30	1	[1·90, 2·96]	99	[1·92, 2·94]	96	[1·92, 2·93]	95	[1·94, 2·93]	95
	2	[1·94, 2·92]	98	[1·96, 2·91]	95	[1·95, 2·90]	94	[1·97, 2·91]	93
	3	[1·98, 2·87]	97	[1·99, 2·87]	93	[1·98, 2·87]	92	[2·00, 2·87]	91
	4	[1·99, 2·87]	96	[2·00, 2·87]	93	[1·99, 2·86]	92	[2·00, 2·87]	91
	5	[2·00, 2·86]	96	[2·00, 2·86]	92	[1·99, 2·86]	91	[2·00, 2·87]	91
	6	–	–	[2·01, 2·85]	92	[1·99, 2·85]	91	[2·01, 2·86]	91
	7	–	–	–	–	[2·00, 2·85]	90	[2·01, 2·86]	90
	8	–	–	–	–	–	–	[2·02, 2·85]	90
	PoSI	[1·75, 3.11]	100	[1·77, 3.10]	99	[1·76, 3.09]	99	[1·78, 3.10]	99
100	1	[2·17, 2·69]	99	[2·17, 2·69]	98	[2·17, 2·68]	97	[2·17, 2·68]	97
	2	[2·19, 2·67]	98	[2·19, 2·67]	97	[2·19, 2·67]	96	[2·19, 2·67]	96
	3	[2·21, 2·65]	96	[2·21, 2·65]	95	[2·21, 2·65]	95	[2·20, 2·65]	94
	4	[2·21, 2·65]	96	[2·21, 2·65]	95	[2·21, 2·65]	94	[2·20, 2·65]	94
	5	[2·21, 2·65]	95	[2·21, 2·65]	95	[2·21, 2·65]	94	[2·20, 2·65]	94
	6	–	–	[2·21, 2·64]	94	[2·21, 2·65]	94	[2·21, 2·65]	94
	7	–	–	–	–	[2·21, 2·65]	94	[2·21, 2·65]	94
	8	–	–	–	–	–	–	[2·21, 2·65]	94
	PoSI	[2·12, 2·75]	100	[2·11, 2·74]	99	[2·11, 2·75]	99	[2·11, 2·75]	99
300	1	[2·28, 2·58]	98	[2·28, 2·58]	98	[2·29, 2·58]	97	[2·29, 2·57]	98
	2	[2·29, 2·57]	97	[2·29, 2·57]	97	[2·30, 2·57]	96	[2·29, 2·57]	97
	3	[2·30, 2·55]	96	[2·30, 2·56]	95	[2·31, 2·56]	95	[2·30, 2·56]	95
	4	[2·30, 2·55]	96	[2·30, 2·55]	95	[2·31, 2·56]	95	[2·30, 2·56]	95
	5	[2·31, 2·55]	95	[2·31, 2·55]	95	[2·31, 2·56]	95	[2·30, 2·56]	95
	6	–	–	[2·31, 2·55]	95	[2·31, 2·56]	95	[2·30, 2·56]	95
	7	–	–	–	–	[2·31, 2·56]	95	[2·31, 2·56]	95
	8	–	–	–	–	–	–	[2·31, 2·56]	95
	PoSI	[2·25, 2·60]	99	[2·25, 2·61]	99	[2·25, 2·61]	99	[2·25, 2·61]	99

Table 7: Average simulated post-selection confidence intervals for θ_3, together with the average coverage percentage for different scenarios and different assumptions regarding p_0. Also given are the results of the post-selection interval by Berk et al. (2013).

for the proposed method, the post-selection prediction method and the smoothed bootstrap method for different settings. In scenario 4 where the true model is the smallest model, the asymptotic method gives accurate results as expected. For $n = 30$, the bootstrap method underestimates the confidence intervals (low coverages) except in scenario 4 in which it gives acceptable coverage probabilities but the length of these intervals is about 1·5 times larger than those of the proposed method. In scenarios 1 and 2 for $n = 100$ the bootstrap coverages are a bit lower than 95% but still acceptable and the lengths are smaller than from the proposed method, which is conservative. For scenarios 3 and 4 the proposed method performs better than the other methods, especially for high values of σ. The post-selection prediction intervals are always wider and their coverage probabilities are always close to one. The reason for this is that this method does not specify the specific selection procedure and in this simulation study, we used the corresponding code that assumes that all subsets of a largest model are used in the selection.

		$\hat{p}$							
n	p_0	5		6		7		8	
30	1	$[-1{\cdot}81, -0{\cdot}66]$	99	$[-1{\cdot}79, -0{\cdot}70]$	97	$[-1{\cdot}78, -0{\cdot}71]$	96	$[-1{\cdot}76, -0{\cdot}72]$	96
	2	$[-1{\cdot}78, -0{\cdot}69]$	98	$[-1{\cdot}76, -0{\cdot}72]$	97	$[-1{\cdot}75, -0{\cdot}73]$	96	$[-1{\cdot}74, -0{\cdot}74]$	95
	3	$[-1{\cdot}74, -0{\cdot}74]$	97	$[-1{\cdot}72, -0{\cdot}76]$	95	$[-1{\cdot}72, -0{\cdot}76]$	95	$[-1{\cdot}71, -0{\cdot}77]$	93
	4	$[-1{\cdot}68, -0{\cdot}80]$	95	$[-1{\cdot}68, -0{\cdot}81]$	93	$[-1{\cdot}68, -0{\cdot}80]$	92	$[-1{\cdot}67, -0{\cdot}80]$	90
	5	$[-1{\cdot}67, -0{\cdot}81]$	95	$[-1{\cdot}67, -0{\cdot}81]$	93	$[-1{\cdot}68, -0{\cdot}81]$	92	$[-1{\cdot}67, -0{\cdot}81]$	90
	6	–	–	$[-1{\cdot}66, -0{\cdot}82]$	92	$[-1{\cdot}67, -0{\cdot}81]$	92	$[-1{\cdot}67, -0{\cdot}81]$	90
	7	–	–	–	–	$[-1{\cdot}66, -0{\cdot}82]$	91	$[-1{\cdot}66, -0{\cdot}81]$	89
	8	–	–	–	–	–	–	$[-1{\cdot}66, -0{\cdot}82]$	89
	PoSI	$[-1{\cdot}92, -0{\cdot}56]$	100	$[-1{\cdot}91, -0{\cdot}57]$	99	$[-1{\cdot}91, -0{\cdot}57]$	99	$[-1{\cdot}90, -0{\cdot}58]$	99
100	1	$[-1{\cdot}53, -0{\cdot}95]$	99	$[-1{\cdot}51, -0{\cdot}96]$	99	$[-1{\cdot}51, -0{\cdot}97]$	98	$[-1{\cdot}51, -0{\cdot}97]$	97
	2	$[-1{\cdot}51, -0{\cdot}97]$	98	$[-1{\cdot}50, -0{\cdot}98]$	98	$[-1{\cdot}50, -0{\cdot}98]$	98	$[-1{\cdot}49, -0{\cdot}98]$	97
	3	$[-1{\cdot}49, -0{\cdot}99]$	98	$[-1{\cdot}48, -0{\cdot}99]$	97	$[-1{\cdot}48, -1{\cdot}00]$	97	$[-1{\cdot}48, -1{\cdot}00]$	96
	4	$[-1{\cdot}46, -1{\cdot}02]$	95	$[-1{\cdot}46, -1{\cdot}02]$	96	$[-1{\cdot}46, -1{\cdot}02]$	95	$[-1{\cdot}46, -1{\cdot}02]$	94
	5	$[-1{\cdot}45, -1{\cdot}02]$	95	$[-1{\cdot}46, -1{\cdot}02]$	95	$[-1{\cdot}46, -1{\cdot}02]$	95	$[-1{\cdot}46, -1{\cdot}02]$	94
	6	–	–	$[-1{\cdot}45, -1{\cdot}02]$	95	$[-1{\cdot}46, -1{\cdot}02]$	95	$[-1{\cdot}46, -1{\cdot}02]$	94
	7	–	–	–	–	$[-1{\cdot}45, -1{\cdot}02]$	95	$[-1{\cdot}46, -1{\cdot}02]$	94
	8	–	–	–	–	–	–	$[-1{\cdot}46, -1{\cdot}02]$	94
	PoSI	$[-1{\cdot}55, -0{\cdot}92]$	100	$[-1{\cdot}55, -0{\cdot}92]$	99	$[-1{\cdot}55, -0{\cdot}92]$	99	$[-1{\cdot}56, -0{\cdot}92]$	99
300	1	$[-1{\cdot}40, -1{\cdot}08]$	99	$[-1{\cdot}40, -1{\cdot}08]$	99	$[-1{\cdot}40, -1{\cdot}09]$	99	$[-1{\cdot}40, -1{\cdot}09]$	98
	2	$[-1{\cdot}39, -1{\cdot}09]$	98	$[-1{\cdot}39, -1{\cdot}09]$	98	$[-1{\cdot}39, -1{\cdot}09]$	98	$[-1{\cdot}39, -1{\cdot}10]$	98
	3	$[-1{\cdot}38, -1{\cdot}10]$	98	$[-1{\cdot}38, -1{\cdot}10]$	97	$[-1{\cdot}38, -1{\cdot}10]$	97	$[-1{\cdot}38, -1{\cdot}11]$	97
	4	$[-1{\cdot}36, -1{\cdot}11]$	96	$[-1{\cdot}37, -1{\cdot}12]$	96	$[-1{\cdot}37, -1{\cdot}11]$	95	$[-1{\cdot}37, -1{\cdot}12]$	95
	5	$[-1{\cdot}36, -1{\cdot}12]$	96	$[-1{\cdot}37, -1{\cdot}12]$	96	$[-1{\cdot}37, -1{\cdot}12]$	95	$[-1{\cdot}37, -1{\cdot}12]$	95
	6	–	–	$[-1{\cdot}36, -1{\cdot}12]$	95	$[-1{\cdot}37, -1{\cdot}12]$	95	$[-1{\cdot}37, -1{\cdot}12]$	95
	7	–	–	–	–	$[-1{\cdot}36, -1{\cdot}12]$	95	$[-1{\cdot}37, -1{\cdot}12]$	95
	8	–	–	–	–	–	–	$[-1{\cdot}37, -1{\cdot}12]$	95
	PoSI	$[-1{\cdot}41, -1{\cdot}06]$	99	$[-1{\cdot}42, -1{\cdot}06]$	99	$[-1{\cdot}42, -1{\cdot}06]$	99	$[-1{\cdot}42, -1{\cdot}06]$	99

Table 8: Average simulated post-selection confidence intervals for θ_4, together with the average coverage percentage for different scenarios and different assumptions regarding p_0. Also given are the results of the post-selection interval by Berk et al. (2013).

F Poisson regression

To investigate the performance of the proposed method in generalized linear models, we consider Poisson regression where the response values are generated from

$$Y_i = \text{Pois}\big\{\exp(\sum_{j=1}^{10}\theta_j x_{ji})\big\}, \quad i = 1, \cdots, n,$$

$x_{1i} = 1$ and for $j = 2, \ldots, 10$, x_{ji} are generated independently from Uniform$[-1, 1]$. The sample size varies as before and $\theta = (1{\cdot}25, -1{\cdot}1, 1{\cdot}43, -1{\cdot}24, 1{\cdot}5, 0_5)^\top$. Three different selection matrices are considered, ζ_i, $i \in \{1, 3, 5\}$ which force the first i covariates in the model. There were no under-parametrized models selected, also for the small sample size. The simulation runs until for each setting the model $(\theta_1, \ldots, \theta_5, \theta_7, \theta_9)^\top$ had been selected 3000 times. The confidence intervals for the superfluous parameters are presented in Table **??**.

The results for the proposed method are similar as in the previous examples. For ζ_5 the simulated coverage probabilities show the validity of the proposed method. Because this selection matrix considers

		$\hat{p}$							
n	p_0	5		6		7		8	
30	1	[1·81, 3.20]	100	[1·91, 3.09]	98	[1·93, 3.06]	97	[1·95, 3.05]	96
	2	[1·84, 3.17]	100	[1·94, 3.06]	98	[1·95, 3.04]	96	[1·97, 3.03]	95
	3	[1·87, 3.14]	99	[1·97, 3.03]	97	[1·98, 3.01]	095	[2·00, 3.00]	94
	4	[1·92, 3.09]	99	[2·01, 2·99]	96	[2·02, 2·98]	93	[2·03, 2·97]	93
	5	[2·07, 2·94]	95	[2·07, 2·93]	93	[2·06, 2·93]	91	[2·07, 2·93]	91
	6	–	–	[2·08, 2·92]	92	[2·07, 2·93]	90	[2·07, 2·93]	90
	7	–	–	–	–	[2·07, 2·92]	89	[2·07, 2·92]	90
	8	–	–	–	–	–	–	[2·08, 2·92]	90
	PoSI	[1·82, 3.19]	100	[1·83, 3.17]	99	[1·83, 3.17]	99	[1·84, 3.16]	99
100	1	[2·15, 2·85]	100	[2·20, 2·80]	99	[2·21, 2·79]	99	[2·22, 2·78]	98
	2	[2·17, 2·83]	100	[2·22, 2·79]	99	[2·22, 2·78]	98	[2·23, 2·77]	98
	3	[2·18, 2·82]	100	[2·23, 2·77]	98	[2·24, 2·76]	98	[2·24, 2·75]	97
	4	[2·21, 2·79]	99	[2·25, 2·75]	97	[2·26, 2·74]	96	[2·26, 2·74]	96
	5	[2·28, 2·72]	95	[2·28, 2·72]	95	[2·28, 2·72]	94	[2·28, 2·72]	94
	6	–	–	[2·29, 2·72]	94	[2·28, 2·72]	94	[2·28, 2·72]	93
	7	–	–	–	–	[2·28, 2·72]	94	[2·28, 2·72]	93
	8	–	–	–	–	–	–	[2·28, 2·71]	93
	PoSI	[2·18, 2·82]	100	[2·19, 2·82]	99	[2·18, 2·82]	99	[2·18, 2·81]	99
300	1	[2·30, 2·70]	100	[2·33, 2·67]	99	[2·34, 2·67]	99	[2·34, 2·66]	99
	2	[2·31, 2·69]	100	[2·34, 2·66]	99	[2·34, 2·66]	99	[2·35, 2·66]	98
	3	[2·32, 2·68]	100	[2·35, 2·65]	99	[2·35, 2·65]	98	[2·35, 2·65]	98
	4	[2·34, 2·67]	99	[2·36, 2·64]	98	[2·36, 2·64]	97	[2·36, 2·64]	97
	5	[2·38, 2·62]	96	[2·37, 2·62]	95	[2·37, 2·63]	95	[2·38, 2·63]	95
	6	–		[2·38, 2·62]	95	[2·38, 2·63]	95	[2·38, 2·63]	95
	7	–		–	–	[2·38, 2·62]	95	[2·38, 2·63]	95
	8	–		–	–	–	–	[2·38, 2·63]	95
	PoSI	[2·33, 2·68]	100	[2·32, 2·68]	99	[2·32, 2·68]	99	[2·32, 2·68]	99

Table 9: Average simulated post-selection confidence intervals for θ_5, together with the average coverage percentage for different scenarios and different assumptions regarding p_0. Also given are the results of the post-selection interval by Berk et al. (2013).

all the non-zero parameters in the model and all truly zero parameters are under selection, the coverage probabilities are close to 95%. Other selection matrices lead to more conservative confidence intervals for the parameters due to conditioning on the selected model. The naive unconditional confidence intervals are always tighter than those of the proposed method and their coverage probabilities are much lower than the nominal value.

G Under-parametrized model selection

As discussed before, for small sample sizes it might happen that a model with less parameters than the true model is selected. If this happens, the proposed method can still be used, although assumption A1 does not hold. Consider the true value for parameters in linear regression $\theta = (0{\cdot}25, -0{\cdot}1, 0{\cdot}43, -0{\cdot}24, 0{\cdot}5, 0_5)^\top$, sample size 30, error standard deviation equal to 2 and all other settings are as before. With the same notation as in the previous example, the selection matrix is ζ_1. We focus on three models which are

		$\hat{p}$					
n	p_0	6		7		8	
30	1	$[-0{\cdot}70, 0{\cdot}71]$	99	$[-0{\cdot}61, 0{\cdot}61]$	98	$[-0{\cdot}57, 0{\cdot}59]$	97
	2	$[-0{\cdot}68, 0{\cdot}69]$	99	$[-0{\cdot}59, 0{\cdot}59]$	97	$[-0{\cdot}55, 0{\cdot}57]$	96
	3	$[-0{\cdot}65, 0{\cdot}66]$	98	$[-0{\cdot}56, 0{\cdot}56]$	97	$[-0{\cdot}53, 0{\cdot}54]$	95
	4	$[-0{\cdot}62, 0{\cdot}63]$	96	$[-0{\cdot}53, 0{\cdot}53]$	95	$[-0{\cdot}50, 0{\cdot}52]$	93
	5	$[-0{\cdot}57, 0{\cdot}58]$	94	$[-0{\cdot}49, 0{\cdot}49]$	93	$[-0{\cdot}47, 0{\cdot}48]$	91
	6	$[-0{\cdot}42, 0{\cdot}43]$	73	$[-0{\cdot}43, 0{\cdot}43]$	88	$[-0{\cdot}42, 0{\cdot}44]$	88
	7	–	–	$[-0{\cdot}42, 0{\cdot}42]$	87	$[-0{\cdot}42, 0{\cdot}43]$	87
	8	–	–	–	–	$[-0{\cdot}41, 0{\cdot}42]$	86
	PoSI	$[-0{\cdot}66, 0{\cdot}68]$	98	$[-0{\cdot}67, 0{\cdot}67]$	99	$[-0{\cdot}65, 0{\cdot}67]$	98
100	1	$[-0{\cdot}36, 0{\cdot}36]$	99	$[-0{\cdot}31, 0{\cdot}31]$	99	$[-0{\cdot}30, 0{\cdot}30]$	98
	2	$[-0{\cdot}35, 0{\cdot}35]$	99	$[-0{\cdot}30, 0{\cdot}30]$	98	$[-0{\cdot}29, 0{\cdot}29]$	97
	3	$[-0{\cdot}34, 0{\cdot}33]$	98	$[-0{\cdot}29, 0{\cdot}29]$	98	$[-0{\cdot}28, 0{\cdot}28]$	96
	4	$[-0{\cdot}32, 0{\cdot}32]$	97	$[-0{\cdot}27, 0{\cdot}27]$	97	$[-0{\cdot}26, 0{\cdot}26]$	95
	5	$[-0{\cdot}29, 0{\cdot}29]$	95	$[-0{\cdot}25, 0{\cdot}25]$	95	$[-0{\cdot}24, 0{\cdot}24]$	93
	6	$[-0{\cdot}22, 0{\cdot}21]$	72	$[-0{\cdot}22, 0{\cdot}22]$	90	$[-0{\cdot}22, 0{\cdot}22]$	90
	7	–	–	$[-0{\cdot}22, 0{\cdot}22]$	89	$[-0{\cdot}22, 0{\cdot}22]$	90
	8	–	–	–	–	$[-0{\cdot}22, 0{\cdot}22]$	90
	PoSI	$[-0{\cdot}32, 0{\cdot}31]$	96	$[-0{\cdot}32, 0{\cdot}32]$	99	$[-0{\cdot}32, 0{\cdot}32]$	98
300	1	$[-0{\cdot}21, 0{\cdot}20]$	100	$[-0{\cdot}18, 0{\cdot}18]$	99	$[-0{\cdot}17, 0{\cdot}17]$	99
	2	$[-0{\cdot}20, 0{\cdot}20]$	99	$[-0{\cdot}17, 0{\cdot}17]$	99	$[-0{\cdot}17, 0{\cdot}16]$	99
	3	$[-0{\cdot}19, 0{\cdot}19]$	99	$[-0{\cdot}16, 0{\cdot}16]$	99	$[-0{\cdot}16, 0{\cdot}16]$	98
	4	$[-0{\cdot}18, 0{\cdot}18]$	98	$[-0{\cdot}15, 0{\cdot}15]$	98	$[-0{\cdot}15, 0{\cdot}15]$	97
	5	$[-0{\cdot}17, 0{\cdot}17]$	96	$[-0{\cdot}14, 0{\cdot}14]$	96	$[-0{\cdot}14, 0{\cdot}14]$	95
	6	$[-0{\cdot}12, 0{\cdot}12]$	74	$[-0{\cdot}12, 0{\cdot}13]$	92	$[-0{\cdot}13, 0{\cdot}13]$	92
	7	–	–	$[-0{\cdot}12, 0{\cdot}12]$	91	$[-0{\cdot}13, 0{\cdot}12]$	92
	8	–	–	–	–	$[-0{\cdot}13, 0{\cdot}12]$	92
	PoSI	$[-0{\cdot}18, 0{\cdot}18]$	97	$[-0{\cdot}18, 0{\cdot}18]$	99	$[-0{\cdot}18, 0{\cdot}18]$	99

Table 10: Average simulated post-selection confidence intervals for θ_6, together with the average coverage percentage for different scenarios and different assumptions regarding p_0. Also given are the results of the post-selection interval by Berk et al. (2013).

represented in the selection matrix and contain the following parameters,

$$\begin{aligned} \text{model 1} &: (\theta_1, \theta_3) \\ \text{model 2} &: (\theta_1, \theta_5) \\ \text{model 3} &: (\theta_1, \theta_2, \theta_5). \end{aligned}$$

The simulations were run until each of these models had been selected 3000 times. Table **??** illustrates that the proposed method is able to provide conditional confidence intervals even in possibly underparametrized models. The naive method's simulated coverage percentages are shown between parentheses. For model 3, the naive method performs poorly in terms of coverage.

H Naive method fails for the truly non-zero parameters

Inference for the truly non-zero parameters can fail because in the limit the estimators are defined as a multiplication of the corresponding row in $\widehat{J}_M^{1/2}(\widehat{\theta})$ to $\widetilde{Z}(M)$ with $\widetilde{Z}(M) \in \mathcal{A}_M(\mathcal{M})$. So, if one of

		$\hat{p}$			
n	p_0	7		8	
30	1	$[-0{\cdot}71, 0{\cdot}72]$	98	$[-0{\cdot}61, 0{\cdot}63]$	97
	2	$[-0{\cdot}69, 0{\cdot}70]$	98	$[-0{\cdot}60, 0{\cdot}61]$	96
	3	$[-0{\cdot}67, 0{\cdot}67]$	97	$[-0{\cdot}57, 0{\cdot}59]$	96
	4	$[-0{\cdot}65, 0{\cdot}65]$	96	$[-0{\cdot}55, 0{\cdot}56]$	95
	5	$[-0{\cdot}61, 0{\cdot}62]$	94	$[-0{\cdot}52, 0{\cdot}53]$	93
	6	$[-0{\cdot}57, 0{\cdot}57]$	90	$[-0{\cdot}48, 0{\cdot}49]$	89
	7	$[-0{\cdot}42, 0{\cdot}42]$	58	$[-0{\cdot}42, 0{\cdot}43]$	82
	8	–	–	$[-0{\cdot}41, 0{\cdot}42]$	81
	PoSI	$[-0{\cdot}66, 0{\cdot}66]$	96	$[-0{\cdot}66, 0{\cdot}67]$	98
100	1	$[-0{\cdot}37, 0{\cdot}37]$	99	$[-0{\cdot}32, 0{\cdot}33]$	98
	2	$[-0{\cdot}36, 0{\cdot}36]$	99	$[-0{\cdot}31, 0{\cdot}32]$	98
	3	$[-0{\cdot}35, 0{\cdot}35]$	98	$[-0{\cdot}30, 0{\cdot}31]$	97
	4	$[-0{\cdot}33, 0{\cdot}34]$	97	$[-0{\cdot}29, 0{\cdot}29]$	96
	5	$[-0{\cdot}32, 0{\cdot}32]$	96	$[-0{\cdot}27, 0{\cdot}28]$	94
	6	$[-0{\cdot}29, 0{\cdot}30]$	92	$[-0{\cdot}25, 0{\cdot}25]$	91
	7	$[-0{\cdot}21, 0{\cdot}22]$	48	$[-0{\cdot}22, 0{\cdot}22]$	84
	8	–	–	$[-0{\cdot}21, 0{\cdot}22]$	83
	PoSI	$[-0{\cdot}31, 0{\cdot}32]$	95	$[-0{\cdot}32, 0{\cdot}32]$	98
300	1	$[-0{\cdot}21, 0{\cdot}21]$	99	$[-0{\cdot}18, 0{\cdot}18]$	99
	2	$[-0{\cdot}21, 0{\cdot}21]$	99	$[-0{\cdot}18, 0{\cdot}18]$	99
	3	$[-0{\cdot}20, 0{\cdot}20]$	98	$[-0{\cdot}17, 0{\cdot}17]$	98
	4	$[-0{\cdot}19, 0{\cdot}19]$	97	$[-0{\cdot}16, 0{\cdot}16]$	97
	5	$[-0{\cdot}18, 0{\cdot}18]$	95	$[-0{\cdot}15, 0{\cdot}15]$	95
	6	$[-0{\cdot}17, 0{\cdot}17]$	92	$[-0{\cdot}14, 0{\cdot}14]$	93
	7	$[-0{\cdot}12, 0{\cdot}12]$	49	$[-0{\cdot}13, 0{\cdot}13]$	85
	8	–	–	$[-0{\cdot}12, 0{\cdot}12]$	85
	PoSI	$[-0{\cdot}18, 0{\cdot}18]$	93	$[-0{\cdot}18, 0{\cdot}18]$	98

Table 11: Average simulated post-selection confidence intervals for θ_7, together with the average coverage percentage for different scenarios and different assumptions regarding p_0. Also given are the results of the post-selection interval by Berk et al. (2013).

		$\hat{p}$	
n	p_0	8	
30	1	$[-0{\cdot}75, 0{\cdot}71]$	97
	2	$[-0{\cdot}74, 0{\cdot}69]$	97
	3	$[-0{\cdot}72, 0{\cdot}67]$	96
	4	$[-0{\cdot}70, 0{\cdot}65]$	95
	5	$[-0{\cdot}67, 0{\cdot}62]$	94
	6	$[-0{\cdot}64, 0{\cdot}59]$	92
	7	$[-0{\cdot}59, 0{\cdot}54]$	87
	8	$[-0{\cdot}44, 0{\cdot}39]$	46
	PoSI	$[-0{\cdot}68, 0{\cdot}64]$	94
100	1	$[-0{\cdot}38, 0{\cdot}38]$	99
	2	$[-0{\cdot}37, 0{\cdot}37]$	98
	3	$[-0{\cdot}36, 0{\cdot}36]$	98
	4	$[-0{\cdot}35, 0{\cdot}35]$	97
	5	$[-0{\cdot}34, 0{\cdot}34]$	96
	6	$[-0{\cdot}32, 0{\cdot}32]$	93
	7	$[-0{\cdot}30, 0{\cdot}30]$	87
	8	$[-0{\cdot}22, 0{\cdot}22]$	39
	PoSI	$[-0{\cdot}32, 0{\cdot}32]$	92
300	1	$[-0{\cdot}22, 0{\cdot}22]$	99
	2	$[-0{\cdot}21, 0{\cdot}21]$	99
	3	$[-0{\cdot}21, 0{\cdot}21]$	99
	4	$[-0{\cdot}20, 0{\cdot}20]$	98
	5	$[-0{\cdot}19, 0{\cdot}19]$	96
	6	$[-0{\cdot}18, 0{\cdot}18]$	93
	7	$[-0{\cdot}17, 0{\cdot}17]$	87
	8	$[-0{\cdot}12, 0{\cdot}13]$	38
	PoSI	$[-0{\cdot}18, 0{\cdot}18]$	89

Table 12: Average simulated post-selection confidence intervals for θ_8, together with the average coverage percentage for different scenarios and different assumptions regarding p_0. Also given are the results of the post-selection interval by Berk et al. (2013).

the Z_is is constrained and the corresponding element for this Z_i in $\widehat{J}_M^{1/2}(\widehat{\theta})$ is relatively big for one parameter, then the distribution of that parameter is highly effected by that Z_i.

Consider the settings in Section 4.1 but here Ω is defined as

$$\Omega_{ij} = \begin{cases} 0{\cdot}95 & i = 3, j = 4, \ldots, 9 \\ 0{\cdot}95 & j = 3, i = 4, \ldots, 9 \\ 1 & i = j \\ 0{\cdot}25 & \text{otherwise} \end{cases}$$

We use the function `nearPD` in R to find the nearest positive definite matrix for this Ω and use that matrix to generate the covariates. For $n = 100$, and ζ^3_{all} the naive confidence interval's coverage for θ_4 is only $0{\cdot}60$ while for the proposed method it is $0{\cdot}96$.

			Scenario 1		Scenario 2		Scenario 3		Scenario 4	
σ	n	method	length	coverage	length	coverage	length	coverage	length	coverage
0·5	30	PostAIC	1·17	98	1·26	99	1·72	97	0·57	94
		Bootstrap	0·78	90	0·98	91	6.51	89	0·87	93
		PoSIp	1·31	99	1·60	100	2·91	100	0·84	100
	100	PostAIC	0·59	99	0·63	99	0·66	98	0·28	95
		Bootstrap	0·42	93	0·50	94	0·68	95	0·42	96
		PoSIp	0·63	100	0·74	100	0·96	100	0·39	99
	300	PostAIC	0·34	99	0·35	99	0·37	98	0·15	95
		Bootstrap	0·24	95	0·29	95	0·38	96	0·24	97
		PoSIp	0·35	99	0·41	100	0·52	100	0·22	99
1	30	PostAIC	2·39	98	2·53	99	3.45	97	1·13	94
		Bootstrap	1·57	90	1·96	91	13.01	89	1·74	93
		PoSIp	2·62	99	3.19	100	5.82	100	1·69	100
	100	PostAIC	1·19	99	1·25	99	1·33	98	0·55	95
		Bootstrap	0·84	93	1·00	94	1·37	95	0·85	96
		PoSIp	1·26	100	1·47	100	1·92	100	0·78	99
	300	PostAIC	0·67	99	0·71	99	0·74	98	0·31	95
		Bootstrap	0·49	95	0·58	95	0·75	96	0·49	97
		PoSIp	0·71	99	0·82	100	1·04	100	0·44	99
3	30	PostAIC	6.98	98	7.56	98	10·32	97	3.40	94
		Bootstrap	4.79	90	5.91	91	39.04	89	5.23	94
		PoSIp	7.82	99	9.53	99	17.43	100	5.06	100
	100	PostAIC	3.57	99	3.76	99	4.00	98	1·65	95
		Bootstrap	2·51	95	2·98	94	4.10	95	2·54	96
		PoSIp	3.79	100	4.14	100	5.74	100	2·35	99
	300	PostAIC	2·02	99	2·12	99	2·22	99	0·93	95
		Bootstrap	1·46	95	1·73	95	2·25	96	1·46	97
		PoSIp	2·12	99	2·46	100	3.11	100	1·31	99

Table 13: Simulated average length of 95% confidence intervals and the coverage percentages for a linear combination of the parameters for different methods in nested models.

n	method	θ_j	ζ_1		ζ_3		ζ_5	
30	PostAIC	θ_7	$[-0{\cdot}46, 0{\cdot}48]$	98	$[-0{\cdot}44, 0{\cdot}45]$	96	$[-0{\cdot}42, 0{\cdot}43]$	94
		θ_9	$[-0{\cdot}47, 0{\cdot}48]$	97	$[-0{\cdot}45, 0{\cdot}46]$	96	$[-0{\cdot}42, 0{\cdot}43]$	95
	Naive	θ_7	$[-0{\cdot}29, 0{\cdot}31]$	56	$[-0{\cdot}30, 0{\cdot}31]$	56	$[-0{\cdot}30, 0{\cdot}31]$	56
		θ_9	$[-0{\cdot}30, 0{\cdot}31]$	55	$[-0{\cdot}30, 0{\cdot}31]$	55	$[-0{\cdot}30, 0{\cdot}31]$	55
100	PostAIC	θ_7	$[-0{\cdot}18, 0{\cdot}18]$	97	$[-0{\cdot}17, 0{\cdot}17]$	96	$[-0{\cdot}17, 0{\cdot}17]$	95
		θ_9	$[-0{\cdot}18, 0{\cdot}17]$	96	$[-0{\cdot}17, 0{\cdot}17]$	96	$[-0{\cdot}17, 0{\cdot}16]$	95
	Naive	θ_7	$[-0{\cdot}12, 0{\cdot}12]$	64	$[-0{\cdot}12, 0{\cdot}12]$	64	$[-0{\cdot}12, 0{\cdot}12]$	64
		θ_9	$[-0{\cdot}12, 0{\cdot}12]$	60	$[-0{\cdot}12, 0{\cdot}12]$	60	$[-0{\cdot}12, 0{\cdot}12]$	60
300	PostAIC	θ_7	$[-0{\cdot}09, 0{\cdot}09]$	97	$[-0{\cdot}09, 0{\cdot}09]$	97	$[-0{\cdot}09, 0{\cdot}09]$	95
		θ_9	$[-0{\cdot}09, 0{\cdot}09]$	97	$[-0{\cdot}09, 0{\cdot}09]$	97	$[-0{\cdot}09, 0{\cdot}08]$	96
	Naive	θ_7	$[-0{\cdot}06, 0{\cdot}06]$	67	$[-0{\cdot}06, 0{\cdot}06]$	66	$[-0{\cdot}06, 0{\cdot}06]$	67
		θ_9	$[-0{\cdot}07, 0{\cdot}06]$	67	$[-0{\cdot}07, 0{\cdot}06]$	66	$[-0{\cdot}07, 0{\cdot}06]$	67

Table 14: Averaged simulated confidence intervals and the simulated coverage percentages for parameters in Poisson regression.

	model 1		model 2		model 3	
θ_1	$[-0{\cdot}54, 1{\cdot}05]$	$96(0{\cdot}96)$	$[-0{\cdot}53, 1{\cdot}04]$	$96(95)$	$[-0{\cdot}52, 1{\cdot}01]$	$94(93)$
θ_2	$-$	$-$	$-$	$-$	$[-1{\cdot}52, 0{\cdot}71]$	$96(74)$
θ_3	$[-0{\cdot}26, 1{\cdot}86]$	$98(92)$	$-$	$-$	$-$	$-$
θ_5	$-$	$-$	$[-0{\cdot}22, 1{\cdot}89]$	$98(93)$	$[-0{\cdot}20, 1{\cdot}98]$	$98(88)$

Table 15: Average simulated PostAIC confidence intervals and their coverage percentage using a possibly under-parametrized selected model (coverage percentage of the naive intervals).