Uniformly valid confidence intervals post-model-selection

François Bachoc[†], **David Preinerstorfer**[‡] **and Lukas Steinberger**[*]

[†] *Institut de Mathématiques de Toulouse*
Université Paul Sabatier

[‡] *CREATES*
Aarhus University

[*] *Department of Statistics and Operations Research*
University of Vienna

Abstract: We suggest general methods to construct asymptotically uniformly valid confidence intervals post-model-selection. The constructions are based on principles recently proposed by Berk et al. (2013). In particular the candidate models used can be misspecified, the target of inference is model-specific, and coverage is guaranteed for *any* data-driven model selection procedure. After developing a general theory we apply our methods to practically important situations where the candidate set of models, from which a working model is selected, consists of fixed design homoskedastic or heteroskedastic linear models, or of binary regression models with general link functions.

1. Introduction

Fitting a statistical model to data is often preceded by a model selection step, and practically always has to face the possibility that the candidate set of models from which a model is selected does not contain the true distribution. The construction of valid statistical procedures in such situations is quite challenging, even if the candidate set of models contains the true distribution (cf. Leeb and Pötscher (2005, 2006, 2008), Kabaila and Leeb (2006) and Pötscher (2009), and the references given in that literature), and has recently attained a considerable amount of attention. In a Gaussian homoskedastic location model and fitting possibly misspecified linear candidate models to data, Berk et al. (2013) have shown how one can obtain valid confidence intervals post-model-selection for (non-standard) model-dependent targets of inference in finite samples (cf. also the discussion in Leeb et al. (2015), and related results obtained for prediction post-model-selection in Bachoc et al. (2014)). In this setup, their approach leads to valid confidence intervals post-model-selection *regardless* of the specific model selection procedure applied. This aspect is of fundamental importance, because many model selection procedures used in practice are almost impossible to formalize: researchers typically use combinations of graphical procedures and numerical algorithms, and sometimes they simply select models that let them reject many hypotheses, i.e., they are hunting for significance. These often unreported and informal practices of model selection prior to conducting the actual analysis may also play a key role in the current crisis of reproducibility. Thus, to establish and popularize statistical methods that are in some sense robust to 'bad practice' is highly desirable. Of course, protecting against the negative effects of any possible form of model selection has the drawback that the obtained confidence sets might be conservative for specific model selection procedures. In the context of fitting linear regression models to Gaussian data, methods that provide valid confidence sets post-model-selection, and are constructed for specific model selection procedures and for targets of inference similar to those considered in the present article, have been recently obtained by Tibshirani et al. (2014), Lee and Taylor (2014), Fithian et al. (2015), Lee et al. (2016). Tibshirani et al. (2015) extended the approach of Tibshirani et al. (2014) to non-Gaussian data by obtaining uniform asymptotic results. Furthermore, valid inference post-model-selection on conventional regression parameters under sparsity conditions was considered, among others, by Belloni et al. (2011, 2014); van de Geer et al. (2014) and Zhang and Zhang (2014).

The methods discussed in Berk et al. (2013) are based on the assumption that the true distribution is Gaussian and homoskedastic, and the authors consider only situations where linear models are fit to data. It is of substantial interest to generalize this approach, and to obtain generic methods for constructing confidence intervals post-model-selection that are widely applicable beyond the Gaussian homoskedastic model considered in Berk et al. (2013). We develop a general asymptotic theory for the construction of uniformly valid confidence sets post-model-selection. These results are applicable whenever the estimation error can be expanded as the sum of independent centered random vectors and a remainder term that is negligible relative to the variance of the leading term. Such a representation typically follows from standard first order linearization arguments, and can therefore be obtained in many situations.

Our confidence intervals can be based on either consistent estimators of the variance of the previously mentioned sum, or, more importantly, if such estimators are not available (which is usually the case when all working models are misspecified), can be based on variance estimators that consistently *overestimate* their targets. We also present results that allow one to obtain such estimators in general and demonstrate their construction in specific applications, where they often coincide with well known sandwich-type estimators. This overcomes another limitation present in Berk et al. (2013), namely the assumption that there exists an unbiased (and chi-square distributed) or uniformly consistent estimator of the variance of the observations (cf. the discussion in Remark 2.1 of Leeb et al. (2015) and in Section 2.2 in Bachoc et al. (2014)). The usage of variance estimators that overestimate their targets, while leading to more conservative inference, renders the approach applicable to the fully misspecified setting. Moreover, the suggested conservative estimators usually have the property that their bias vanishes if the selected model is correct (cf. Remark 2.7 and Subsection 3.1.1). Another important aspect of the results obtained is that they are valid uniformly over wide classes of potential underlying distributions, which is particularly important as this guarantees that the results provide a good description of finite sample properties (cf. Leeb and Pötscher (2003) and Tibshirani et al. (2015) for a discussion of related issues in a model selection context). Finally, our confidence intervals obtained in specific situations are particularly convenient for practitioners, because they are structurally very similar to the confidence sets one would use in practice following the naive (and invalid (see, e.g., Bachoc et al., 2014; Leeb et al., 2015)) approach that ignores that the model has been selected using the same data set. The main difference of our construction to the naive (and invalid) approach is the choice of a critical value: Quantiles from a standard normal or Student distribution are replaced by so-called POSI-constants.

The structure of the present article is as follows: We first develop a general asymptotic theory for the construction of uniformly valid confidence sets post-model-selection in Section 2. In Section 3, we apply our theoretical results to a number of examples: In Subsection 3.1 we consider a classical situation where homoskedastic linear models are fit to homoskedastic data, and in Subsection 3.2, we apply our methods to a situation where heteroskedastic linear models are fit to heteroskedastic data, leading to confidence intervals based on variance estimators in the spirit of Eicker (1967). Finally, in Subsection 3.3 we apply the general theory to the problem of fitting binary regression models to binary data, allowing also for a data driven choice of the link function. Of course, the selection of examples in Section 3 is by no means exhaustive. But besides covering three very important modeling frameworks, Section 3 serves as an illustration of how the general theory developed in Section 2 can be applied. In Section 4 we conclude and discuss possible extensions of the results obtained in this paper that are currently under investigation. The proofs are collected in Appendices A, B and C.

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

2.1. Framework and problem description

Consider a situation where we observe n independent $1 \times \ell$-dimensional random vectors $y_{1,n}, \ldots, y_{n,n}$ defined on a common probability space $(\Omega, \mathcal{A}, \mathbb{P})$. We denote the distribution of $y_n = (y'_{1,n}, \ldots, y'_{n,n})'$ on the Borel sets of $\mathbb{R}^{n\times\ell}$ by $\mathbb{P}_n$, which, denoting the distribution of $y_{i,n}$ by $\mathbb{P}_{i,n}$, then coincides with the product measure $\bigotimes_{i=1}^n \mathbb{P}_{i,n}$. Suppose further that we are given a set M_n consisting of d models, i.e., d nonempty sets of distributions $\mathbb{M}_{1,n}, \ldots, \mathbb{M}_{d,n}$ on the Borel sets of $\mathbb{R}^{n\times\ell}$, where d does not depend on n. We do *not* assume that the distribution of the data $\mathbb{P}_n$ is contained in one of those sets, i.e., the candidate set M_n might be *misspecified*. Therefore, one has to define for each model $\mathbb{M} \in \mathsf{M}_n$ a corresponding target of inference $\theta^*_{\mathbb{M},n} = \theta^*_{\mathbb{M},n}(\mathbb{P}_n)$, say, which we take as given throughout the present section. Furthermore we assume that for every $\mathbb{M}_{j,n} \in \mathsf{M}_n$ the target is an element of a Euclidean space of finite dimension $m(\mathbb{M}_{j,n})$ which does not depend on n. As an example, consider the situation where $\mathbb{M} \in \mathsf{M}_n$ is parameterized by a finite dimensional Euclidean space. Then $\theta^*_{\mathbb{M},n}$ would naturally be the value of the parameter that corresponds to the projection of $\mathbb{P}_n$ onto $\mathbb{M}$ w.r.t. some measure of closeness, e.g., the Kullback-Leibler divergence, or the Hellinger-distance. Note that in general such a projection might not uniquely exist, or might not exist at all, and that in each application additional conditions – on $\mathbb{P}_n$ and/or the candidate set M_n of models – need to be imposed to obtain well defined targets. Note also that the target is model-specific, i.e., it depends on $\mathbb{M}$. Lastly we emphasize that defining and working with (pseudo) targets of inference in potentially misspecified models has a long-standing tradition in statistics, dating back at least to Huber (1967), and we confer the reader to this strand of literature for further discussion.

Statistical inference in such a setting typically has to solve two problems: model selection, and statistical inference post-model-selection. We assume that we are given for every $\mathbb{M} \in \mathsf{M}_n$ an estimator $\hat{\theta}_{\mathbb{M},n}$ of the target $\theta^*_{\mathbb{M},n}$, i.e., a measurable map from the sample space $\mathbb{R}^{n\times\ell}$ to $\mathbb{R}^{m(\mathbb{M})}$. Secondly, we assume that there is available a model selection procedure, i.e., a measurable map $\hat{\mathbb{M}}_n : \mathbb{R}^{n\times\ell} \to \mathsf{M}_n$. Given those ingredients, inference proceeds in two steps: Firstly, the model selection procedure is applied to select a model $\hat{\mathbb{M}}_n$. The target of inference, which is data-dependent as well, then is $\theta^*_{\hat{\mathbb{M}}_n,n}$ and is estimated by $\hat{\theta}_{\hat{\mathbb{M}}_n,n}$. Often, in practice, it is naively ignored that the estimation step was preceded by a model selection step using the same data, which can then lead to invalid conclusions. In this article we shall introduce generic methods to obtain asymptotically valid confidence intervals in such a setup. More specifically, given $\alpha \in (0,1)$, we construct for every model $\mathbb{M} \in \mathsf{M}_n$ confidence intervals $\mathrm{CI}^{(j)}_{1-\alpha,\mathbb{M}}$ for $j = 1,\dots,m(\mathbb{M})$ for each coordinate $\theta^{*(j)}_{\mathbb{M},n}$ of the model-specific target based on the estimator $\hat{\theta}_{\mathbb{M},n}$, and we show that these confidence intervals satisfy for *any* (measurable) model selection procedure $\hat{\mathbb{M}}_n$ that

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

We work in a framework allowing for a triangular array of data, because in applications this then allows us to obtain uniform asymptotic results in the sense that for a suitable set $\mathbf{P}_n$ of feasible distributions for $\mathbb{P}_n$ we can verify that the confidence intervals post-model-selection satisfy for *any* (measurable) model selection procedure $\hat{\mathbb{M}}_n$ that

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

2.1.1. Notation

Before we proceed to our general theory and the corresponding basic assumption, we introduce some notation that is used throughout this article: A normal distribution with mean μ and (possibly singular) covariance matrix Σ is denoted by $N(\mu, \Sigma)$. For $\alpha \in (0,1)$ and a covariance matrix Γ we denote by $K_{1-\alpha}(\Gamma)$ the $1-\alpha$-quantile of the distribution of the supremum-norm $\|Z\|_\infty$ of $Z \sim N(0,\Gamma)$. The correlation matrix corresponding to a covariance matrix Σ is denoted by $\operatorname{corr}(\Sigma) = \operatorname{diag}(\Sigma)^{\dagger/2}\Sigma\operatorname{diag}(\Sigma)^{\dagger/2}$, where $\operatorname{diag}(\Sigma)$ denotes the diagonal matrix obtained from Σ by setting all off-diagonal elements equal to 0, $A^\dagger$ denotes the Moore-Penrose inverse of the quadratic matrix A, $A^{1/2}$ denotes the symmetric non-negative definite square root of the non-negative definite matrix A, and where we abbreviate $[A^\dagger]^{1/2}$ by $A^{\dagger/2}$. For a vector v with coordinates $v^{(1)}, \ldots, v^{(l)}$ we also use the symbol $\operatorname{diag}(v)$ to denote the diagonal matrix with first diagonal entry $v^{(1)}$, second $v^{(2)}$, and so on. The operator norm of a matrix A (w.r.t. the Euclidean norm) is denoted by $\|A\|$, and the Euclidean norm of a vector v is denoted by $\|v\|$. Furthermore, A_{ii}, the i-th diagonal element of a quadratic matrix A, is occasionally abbreviated as A_i. We also identify the indicator function $\mathbb{1}_B$ of a set B with the set B itself, whenever there is no risk of confusion. Weak convergence of a sequence of probability measures $\mathbb{Q}_n$ to $\mathbb{Q}$ is denoted by $\mathbb{Q}_n \Rightarrow \mathbb{Q}$. The image measure of a random variable (or vector) x on a probability space $(F, \mathcal{F}, \mathbb{Q})$ is denoted by $\mathbb{Q} \circ x$. If not stated otherwise, limits are taken as $n \to \infty$. The symbol d_w shall denote a metric generating the topology of weak convergence on the set of Borel probability measures on a Euclidean space of dimension indicated by the respective arguments; for specific examples see, e.g., the discussion in Dudley (2002) pp. 393. In the main part of the article (everywhere apart from the notation used in Appendix A) the expectation operator, the variance operator, and the variance-covariance operator w.r.t. $\mathbb{P}_n$ is denoted by $\mathbb{E}_n$, $\mathbb{V}_n$, and $\mathbb{VC}_n$, respectively; the expectation operator, the variance operator, and the variance-covariance operator w.r.t. $\mathbb{P}_{i,n}$ is denoted by $\mathbb{E}_{i,n}$, $\mathbb{V}_{i,n}$, and $\mathbb{VC}_{i,n}$, respectively; and the expectation operator, the variance operator, and the variance-covariance operator w.r.t. $\mathbb{P}$ is denoted by $\mathbb{E}$, $\mathbb{V}$, and $\mathbb{VC}$, respectively.

2.2. Assumptions

Our methods for constructing uniformly valid confidence intervals post-model-selection are developed under a high-level assumption imposed on the stacked vector of estimators $\hat{\theta}_n = (\hat{\theta}'_{\mathbb{M}_1,n}, \ldots, \hat{\theta}'_{\mathbb{M}_d,n})'$ centered at the corresponding stacked vector of targets $\theta^*_n = (\theta^{*'}_{\mathbb{M}_1,n}, \ldots, \theta^{*'}_{\mathbb{M}_d,n})'$. In this section we denote the dimension of $\hat{\theta}_n$ by

$$k \quad := \quad \sum_{j=1}^{d} m(\mathbb{M}_{j,n}),$$

which does not depend on n. The assumption is as follows:

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

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

where, writing $r_n(y_n) := \sum_{i=1}^n g_{i,n}(y_{i,n})$*, it holds for every* $i \in \{1, \ldots, n\}$ *and every* $j \in \{1, \ldots, k\}$ *that*

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

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

$$\mathbb{V}_n^{-1}\left(r_n^{(j)}\right)\sum_{i=1}^n \int_{\mathbb{R}^{1\times\ell}} \left[g_{i,n}^{(j)}\right]^2 \left\{|g_{i,n}^{(j)}| \geq \varepsilon \mathbb{V}_n^{\frac{1}{2}}(r_n^{(j)})\right\} d\mathbb{P}_{i,n} \to 0 \text{ for every } \varepsilon > 0, \tag{2.2}$$

and

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

Clearly, an expansion as in Equation (1) of Condition 1 is satisfied in many applications, and can typically be obtained by a standard linearization argument (cf. Subsection 3.3 for an example and further discussion). We emphasize that the two last assumptions in Condition 1 are formulated in terms of rescaled summands, which, in applications, can be exploited to circumvent restrictive compactness assumptions on moments of the distribution generating the data (e.g., in Subsections 3.1 and 3.2 we do not need to restrict variance parameters to a compact set - as opposed to the conditions used by Eicker (1967)).

Before we proceed to our main results we briefly highlight the main consequence of Condition 1 on which, besides the basic ideas in Berk et al. (2013), our methods of constructing confidence sets are based. Introduce, under Condition 1,

$$S_n(y_n) := \sum_{i=1}^n g_{i,n}(y_{i,n}) g_{i,n}'(y_{i,n}). \tag{2.3}$$

Then we can obtain the following result, which is established using tightness arguments, a result in Pollak (1972), and Raikov's theorem (cf. the statement in Gnedenko and Kolmogorov (1954) on p. 143, originally published in Raikov (1938)). See Section B.1 for the proof.

Lemma 2.1. *Under Condition 1, for* $\varepsilon > 0$ *we have*

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

and the statement remains valid upon replacing S_n *by* $\mathbb{VC}_n(r_n)$*.*

We emphasize that S_n is in general *not* an estimator of $\mathbb{VC}_n(r_n)$, since $g_{i,n}$ might, and typically will, depend on θ_n^*. In applications one therefore has to construct a "suitable" estimator for S_n, which can then be used, together with the results in Section 2.3, which are based on Lemma 2.1, to construct confidence intervals post-model-selection. The extent to which finding a suitable estimator is possible depends on the degree of misspecification of the candidate set of models, as will be discussed in the following subsection. In particular, we need to distinguish between the two cases where consistent estimators are available, and the more relevant case where estimators need to be used that, due to the presence of a non-negligible bias component, consistently overestimate their targets. An important part of the theory in Section 2.3 is that we show how such estimators can be constructed. We emphasize that in applications the variance estimators obtained will often coincide with Eicker-type 'sandwich' estimators (cf. Subsection 3.2), which is an aspect that is convenient for practitioners.

In light of Lemma 2.1 above, a remarkable aspect of Condition 1 is perhaps that we obtain a multivariate central limit theorem even though the condition *does not require* a joint Lindeberg-type condition concerning the random vectors $g_{i,n}$, but instead requires k separate Lindeberg conditions concerning the behavior of the marginals only. To verify that marginal Lindeberg conditions are sufficient for our theory to go through, we exploit a result due to Pollak (1972), showing that an infinitely-divisible distribution is normal if and only if each of its marginals is normal. This aspect can be convenient when applying our results given below, since in particular applications results on $\hat{\theta}_{\mathbb{M},n}$ as required in Condition 1 are likely to be available in the literature concerning asymptotic properties of estimators in misspecified models *without* a model selection procedure being applied before conducting inference. Note, however, that additional arguments might be needed to obtain asymptotic results that are uniform in the true distribution, which is one of our main objectives. We also emphasize the following alternative formulation of the Lindeberg condition appearing in Condition 1 above.

Remark 2.2. *Using, e.g., Gnedenko and Kolmogorov (1954) Theorem 3 in Paragraph 21, one obtains that Equation (1) in Condition 1 can be* equivalently *phrased as*

$$\mathbb{P}_n \circ \left(\frac{r_n^{(j)}}{\mathbb{V}_n^{1/2}(r_n^{(j)})} \right) \Rightarrow N(0,1) \quad \textit{and}$$

$$\max_{i=1,\ldots,n} \mathbb{P}_{i,n} \left(|g_{i,n}^{(j)}| \geq \mathbb{V}_n^{1/2}(r_n^{(j)})\varepsilon \right) \to 0 \ \ \textit{for every } \varepsilon > 0.$$

In some applications it might be easier to check these two conditions directly (for every j), in particular in case one can use existing results in the literature on misspecified models without model selection as indicated above.

2.3. Confidence intervals post-model-selection

In this subsection we shall now present our general asymptotic results for the construction of valid confidence intervals post-model-selection under Condition 1. We consider two different situations: (i) a situation where a consistent estimator of $\mathbb{VC}_n(r_n)$ is available; (ii) a situation where a consistent estimator of $\mathbb{VC}_n(r_n)$ is *not* available, but it is possible to construct estimators that "consistently overestimate" the diagonal entries of $\mathbb{VC}_n(r_n)$. Concrete examples of such consistent or "consistently overestimating" estimators are also provided, based on approximating the summands $g_{i,n}(y_{i,n})$ appearing in Condition 1. The main theoretical results in this section are Theorem 2.5 and Proposition 2.6.

Given $\mathbb{M} = \mathbb{M}_{i,n} \in \mathsf{M}_n$ and $j \in \{1, \dots, m(\mathbb{M})\}$ we abbreviate

$$j \star \mathbb{M} \quad := \quad \sum_{l=1}^{i-1} m(\mathbb{M}_{l,n}) + j,$$

where sums over an empty index set are to be interpreted as 0. For instance, $\theta_n^{*(j\star\mathbb{M})} = \theta_{\mathbb{M},n}^{*\ (j)}$.

2.3.1. Confidence intervals based on consistent estimators of $\mathbb{VC}_n(r_n)$

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

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

$$\mathbb{P}_n \left(\| \operatorname{corr}(\hat{S}_n) - \operatorname{corr}\left(\mathbb{VC}_n(r_n)\right) \| + \| \operatorname{diag}(\mathbb{VC}_n(r_n))^{-1} \operatorname{diag}(\hat{S}_n) - I_k \| \geq \varepsilon \right)$$

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

$$\mathbb{P}_n \left(\| \operatorname{corr}(\hat{S}_n) - \operatorname{corr}\left(S_n\right) \| + \| \operatorname{diag}(S_n)^{\dagger} \operatorname{diag}(\hat{S}_n) - I_k \| \geq \varepsilon \right) \to 0. \qquad (2.4)$$

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

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

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

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

Theorem 2.3 is based on the assumption that an estimator $\hat{S}_n$ is available that consistently estimates $\mathbb{VC}_n(r_n)$. Coming back to the discussion at the end of Subsection 2.2, the vectors $g_{i,n}(y_{i,n})$ appearing in the definition of S_n are typically *not* observable, because they will depend on the unknown target θ_n^*, i.e., they are, more explicitly, of the form $g_{i,n}(y_{i,n}, \theta_n^*)$. In such cases S_n is not a feasible candidate for $\hat{S}_n$ in the previous theorem, and therefore one will, in most cases, naturally try to obtain predictors $\hat{g}_{i,n}(y_n)$ for $g_{i,n}(y_{i,n})$ by replacing the unknown target by its estimator $\hat{\theta}_n$, i.e., by setting $\hat{g}_{i,n}(y_n) = g_{i,n}(y_{i,n}, \hat{\theta}_n(y_n))$. The subsequent proposition now provides conditions on predictors $\hat{g}_{i,n}(y_n)$, which, if satisfied, immediately allow the construction of a consistent estimator $\hat{S}_n$ of $\mathbb{VC}_n(r_n)$ by replacing each $g_{i,n}(y_{i,n})$ in Equation (2.2) by its predictor $\hat{g}_{i,n}(y_n)$. In the result the predictor $\hat{g}_{i,n}(y_n)$ might be of the form $g_{i,n}(y_{i,n}, \hat{\theta}_n(y_n))$ as discussed above, but the proposition is not restricted to that particular case. Again, the conditions are assumptions concerning the large sample behavior of the marginals only, which potentially facilitates their verification in practice.

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

$$\mathbb{P}\left(\frac{\sum_{i=1}^n \left(g_{i,n}^{(j)}(y_{i,n}) - \hat{g}_{i,n}^{(j)}(y_n)\right)^2}{\sum_{i=1}^n [g_{i,n}^{(j)}(y_{i,n})]^2} \geq \varepsilon\right) \to 0, \tag{2.5}$$

or equivalently that

$$\mathbb{P}\left(\frac{\sum_{i=1}^n \left(g_{i,n}^{(j)}(y_{i,n}) - \hat{g}_{i,n}^{(j)}(y_n)\right)^2}{\sum_{i=1}^n \mathbb{V}\left(g_{i,n}^{(j)}(y_{i,n})\right)} \geq \varepsilon\right) \to 0. \tag{2.6}$$

Then the convergence in (2.3) is satisfied for

$$\hat{S}_n(y_n) = \sum_{i=1}^n \hat{g}_{i,n}(y_n)\hat{g}_{i,n}'(y_n).$$

2.3.2. *Confidence intervals based on estimators that consistently overestimate the diagonal entries of $\mathbb{VC}_n(r_n)$*

Due to an asymptotically non-negligible bias term arising from misspecification of the model, it is typically difficult to obtain an estimator $\hat{S}_n$ satisfying the condition in Theorem 2.3 (see Remark 2.7 and Section 3.1.1 for details). Nevertheless, it is often still possible to construct estimators of the diagonal entries of the matrix $\mathbb{VC}_n(r_n)$ that, while possibly inconsistent, asymptotically *overestimate* their targets; for a corresponding constructive result see Proposition 2.6 below. Similarly, it is in general not difficult to find an estimator of $K_{1-\alpha}(\mathrm{corr}(S_n))$ that consistently overestimates that quantity, see the discussion and the result following Proposition 2.6 below concerning upper bounds on the function $K_{1-\alpha}(.)$ over the set of all correlation matrices (using this upper bound, although leading to wider confidence intervals, also leads to substantial computational advantages). Based on such estimators it is then possible to construct asymptotically valid confidence intervals post-model-selection, even though the candidate set of models might be (severely) misspecified. This is the content of the subsequent result, which, together with Proposition 2.6 below, is the main theoretical result in this section.

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

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

satisfies

$$\mathbb{P}_n\left(\kappa_n \geq 1+\varepsilon\right) \to 0 \text{ for every } \varepsilon > 0, \tag{2.7}$$

(implicitly including that $\mathbb{P}_n(\kappa_n$ is well defined$) \to 1$) or, equivalently, that the condition in (2.5) holds with κ_n replaced by

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

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

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

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

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

In the important special case where $\hat{K}_n \geq K_{1-\alpha}(\mathrm{corr}(\mathbb{VC}_n(r_n)))$ holds eventually, the condition in Equation (2.5) is implied by the condition that for every $j = 1, \ldots, k$ it holds that

$$\mathbb{P}_n\left(\sqrt{\frac{[S_n]_j}{\hat{\nu}^2_{j,n}}} \geq 1+\varepsilon\right) \to 0 \text{ for every } \varepsilon > 0, \tag{2.8}$$

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

$$\mathbb{P}_n\left(\sqrt{\frac{[\mathbb{VC}_n(r_n)]_j}{\hat{\nu}^2_{j,n}}} \geq 1+\varepsilon\right) \to 0 \text{ for every } \varepsilon > 0. \tag{2.9}$$

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

Proposition 2.6. *Suppose Condition 1 is satisfied, and let $\tilde{g}_{i,n} : \mathbb{R}^{n\times\ell} \to \mathbb{R}^k$ and $\hat{g}_{i,n} : \mathbb{R}^{n\times\ell} \to \mathbb{R}^k$ be Borel measurable for $i = 1, \ldots, n$ for every n. Suppose that for every $j = 1, \ldots, k$ and for every $\varepsilon > 0$ the condition (2.4), or equivalently (2.4), is satisfied. Suppose further that there exist real numbers $a_{i,n}^{(j)}$ so that*

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

holds for every $n \in \mathbb{N}$, $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, k\}$. Then the statement in (2.5) is satisfied for

$$\hat{\nu}_{j,n}^2(y_n) \quad = \quad \sum_{i=1}^{n} \left[\tilde{g}_{i,n}^{(j)}(y_n)\right]^2 \qquad \text{for} \qquad j = 1, \ldots, k.$$

The previous proposition is based on the assumption that the vectors $g_{i,n}(y_{i,n})$ from Condition 1 can be well approximated in the sense of the condition appearing in Proposition 2.4 up to a non-stochastic additive component. The result is particularly geared towards the case where these non-stochastic additive components are non-negligible in the sense that

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

because in case the non-stochastic additive components are negligible in this sense, a consistent estimator of $\mathbb{VC}_n(r_n)$ in the sense of (2.3) can be constructed.

Remark 2.7. *Using the simple bound $(g_{i,n}^{(j)} - \tilde{g}_{i,n}^{(j)})^2 \leq 2(g_{i,n}^{(j)} - \hat{g}_{i,n}^{(j)})^2 + 2a_{i,n}^2$, it is easy to verify that if the non-stochastic additive components $a_{i,n}$ are negligible in the previously defined sense, then $\tilde{g}_{i,n}$ satisfies the assumptions of $\hat{g}_{i,n}$ appearing in Proposition 2.4. As a consequence the estimator*

$$\tilde{S}_n(y_n) = \sum_{i=1}^{n} \tilde{g}_{i,n}(y_n)\tilde{g}'_{i,n}(y_n)$$

satisfies (2.3), and one can construct confidence intervals based on this estimator as discussed in Theorem 2.3. Note that $\hat{\nu}_{j,n}^2(y_n) = [\tilde{S}_n(y_n)]_j$.

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

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

where $F_{Beta,1/2,(q-1)/2}$ is the cumulative distribution function of the Beta(1/2,$(q-1)/2$) distribution, and where G follows a chi-squared distribution with q degrees of freedom. The quantity $B_\alpha(q, N)$ corresponds to the quantity K_4 of Bachoc et al. (2014) in the known variance case (for a discussion of numerical algorithms for obtaining $B_\alpha(q, N)$ in practice we confer the reader to that reference). We have (Bachoc et al., 2014; Berk et al., 2013) that $B_\alpha(q, N)$ is larger than all the $1 - \alpha$ quantiles of random variables of the form $\max_{i=1,\ldots,N} |v_i' \epsilon|$, where $v_1, \ldots, v_N$ are column vectors of $\mathbb{R}^q$ with $||v_i||_2 \leq 1$ and where $\epsilon \sim N(0, I_q)$; also showing that for fixed α and N the function $B_\alpha(., N)$ is monotonically increasing.

Asymptotic approximations of $B_\alpha(q, N)$ for large q and N are provided in Bachoc et al. (2014), Berk et al. (2013) and Zhang (2015). In particular, as $q, N \to \infty$,

$$B_\alpha(q, N)/\sqrt{q\left(1 - N^{-2/(q-1)}\right)} \to 1,$$

from Proposition 2.10 in Bachoc et al. (2014), itself building on results from Berk et al. (2013) and Zhang (2015).

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

Lemma 2.8. *For every $\alpha \in (0, 1)$ and a $k \times k$ correlation matrix Γ we have*

$$K_{1-\alpha}(\Gamma) \leq B_\alpha(\mathrm{rank}(\Gamma), k).$$

In a particular application it might of course be possible to obtain better upper bounds by exploiting structural properties of the specific correlation matrix Γ at hand, cf. Subsection 3.1. Using the upper bound of Lemma 2.8 can also be very useful in situations where the computation of $K_{1-\alpha}(\Gamma)$ is infeasible.

3. Applications

In this section we now apply the general results obtained above to some important special cases that are frequently encountered in practice. In Subsection 3.1, we consider the case where fixed design homoskedastic linear models are fit to data, and we discuss the case where fixed design heteroskedastic linear models are fit to data in Subsection 3.2. Finally, in Subsection 3.3, we consider fitting binary regression models to binary data. In each of our applications we consider an asymptotic framework, where for each n the distribution of the data $\mathbb{P}_n$ is an element of a set of distributions $\mathbf{P}_n$ on the sample space equipped with its Borel σ-field, and construct confidence sets post-model-selection that are asymptotically valid *uniformly* over the set $\mathbf{P}_n$. To establish uniformity over the sets considered, we crucially rely on the fact that our general theory was developed for triangular arrays.

In this section we need some additional notation to define the feasible sets $\mathbf{P}_n$ of $\mathbb{P}_n$ we consider, and to define our candidate sets M_n of models: For n a natural number and T a Borel subset of $\mathbb{R}$, we shall denote by $M(T^n)$ the set of Borel probability measures supported on $T^n = \bigtimes_{i=1}^n T \subseteq \mathbb{R}^n$. The mean vector corresponding to an element Q of $M(T^n)$ is denoted by $\mu(Q)$. Given an element $Q \in M(T^1)$ and a real number $0 < q < \infty$, we shall use the symbol $m_q(Q)$ to denote the q-th absolute central moment of Q, assuming that $\mu(Q)$ exists. We shall also denote the set of product measures in $M(T^n)$ by $\bigotimes_{i=1}^n M(T)$. Given $Q \in \bigotimes_{i=1}^n M(T)$ we shall denote the corresponding marginals by Q_i, i.e., $Q = \bigotimes_{i=1}^n Q_i$.

3.1. Inference post-model-selection when fitting fixed design linear models to homoskedastic data

One important application of our general theory is the homoskedastic linear regression setting. The feasible sets for the true underlying distribution $\mathbb{P}_n$ we consider here depend on two parameters $\delta > 0$ and $\tau \geq 1$ and are defined as

$$\mathbf{P}_n^{(\mathrm{lm})}(\delta, \tau) := \left\{ Q \in \bigotimes_{i=1}^n M(\mathbb{R}) : \begin{array}{l} 0 < m_2(Q_1) = \ldots = m_2(Q_n) < \infty \\ \frac{\max_{i=1,\ldots,n} m_{2+\delta}(Q_i)^{\frac{2}{2+\delta}}}{m_2(Q_1)} \leq \tau \end{array} \right\}.$$

Note that $\mathbf{P}_n^{(\mathrm{lm})}(\delta, \tau)$ is empty for $\delta > 0$ and $\tau < 1$. Furthermore, observe that $\mathbf{P}_n^{(\mathrm{lm})}(\delta, \tau)$ contains the set of n-variate spherical normal distributions with unrestricted mean vector if

$$\Gamma\left(\frac{3+\delta}{2}\right) \leq (\tau/2)^{1+\delta/2} \sqrt{\pi}.$$

For such a pair (δ, τ) the set $\mathbf{P}_n^{(\mathrm{lm})}(\delta, \tau)$ contains the Gaussian model considered in Berk et al. (2013).

We are interested in a situation where one works with candidate sets consisting of homoskedastic linear models, i.e., a situation where one is interested in conducting inference on $\mu(\mathbb{P}_n)$, and it is *assumed* by the practitioner that $\mu(\mathbb{P}_n)$ is an element of $\operatorname{span}(X_n)$, the column span of a design matrix $X_n \in \mathbb{R}^{n \times p}$ with p not depending on n, or that $\mu(\mathbb{P}_n)$ is at least "well-approximated" by an element of that linear space; and where it is *known* (and taken into account in the construction of the confidence sets) that the observations have identical variances (for a situation where the observations are heteroskedastic see Subsection 3.2). In such a situation one often tries to decide in a data-driven way which regressors to use, i.e., one needs to solve a subset-selection problem. We assume that we are given a nonempty set $\mathcal{I} = \{M_1, \ldots, M_d\}$ of nonempty subsets of $\{1, 2, \ldots, p\}$. Given $M \in \mathcal{I}$ we shall denote by $X_n[M]$ the matrix obtained from X_n by striking all columns whose index is not an element of M. We then consider for each $j \in \{1, \ldots, d\}$ the sets

$$\mathbb{M}_{j,n} = \left\{ Q \in \bigotimes_{i=1}^{n} M(\mathbb{R}) : \begin{array}{l} 0 < m_2(Q_1) = \ldots = m_2(Q_n) < \infty \\ \mu(Q) \in \operatorname{span}(X_n[M_j]) \end{array} \right\},$$

and our candidate set of models is then given by

$$\mathsf{M}_n = \{\mathbb{M}_{j,n} : j = 1, \ldots, d\}.$$

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

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

$$\max_{i=1,\ldots,n} X_{i,n}[M] \left(X_n[M]' X_n[M]\right)^{-1} X_{i,n}[M]' \to 0.$$

Remark 3.1. *In a standard linear model framework, Condition X1 is necessary for asymptotic normality of the ordinary-least-squares estimator (see Arnold, 1980; Huber, 1973). It particularly holds if* $\operatorname{rank}(X_n) = p$*, eventually, and* $\max_{i=1,\ldots,n} X_{i,n}(X_n' X_n)^{-1} X_{i,n}' \to 0$*. Moreover, it also holds in case* $\|X_{i,n}\|$ *is bounded and* $\lambda_{\min}(\frac{1}{n} X_n' X_n)$ *is bounded away from* 0*, which is typically the case in sufficiently balanced factorial designs, but Condition X1 is obviously much more general. For example, it also covers the important cases of polynomial regressors, trigonometric regressors, or mixed polynomial and trigonometric regressors (cf. the discussion in Eicker (1967), pp. 64).*

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

$$\beta^*_{\mathbb{M},n} = \beta^*_{\mathbb{M},n}(\mathbb{P}_n) = \left(X_n[M]' X_n[M]\right)^{-1} X_n[M]' \mu(\mathbb{P}_n), \tag{3.1}$$

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

We shall now describe how asymptotically uniformly valid confidence sets can be constructed post-model-selection for the target defined in Equation (3.1) above: Given $\mathbb{M} \in \mathsf{M}_n$ with index set M, we estimate the corresponding target by the model-specific ordinary-least-squares estimator, i.e., by

$$\hat{\beta}_{\mathbb{M},n}(y) = \left(X_n[M]' X_n[M]\right)^{-1} X_n[M]' y; \tag{3.2}$$

let

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

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

$$\mathrm{CI}_{1-\alpha,\mathbb{M}}^{(j),\mathrm{lm}} = \hat{\beta}_{\mathbb{M},n}^{(j)} \pm \sqrt{\hat{\sigma}_{\mathbb{M},n}^2 \left[\left(X_n[M]' X_n[M] \right)^{-1} \right]_j} K_{1-\alpha}(\mathrm{corr}(\Gamma_n)), \qquad (3.3)$$

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

$$\left(X_n[M_s]' X_n[M_s] \right)^{-1} X_n[M_s]' X_n[M_t] \left(X_n[M_t]' X_n[M_t] \right)^{-1},$$

for $s,t \in \{1, \dots, d\}$. Essentially, the construction in (3.1) coincides with the confidence intervals of Berk et al. (2013). However, there are two major differences. First of all, we here do not assume that the data are Gaussian, which is why we resort to asymptotic results. This is also the reason why our constant $K_{1-\alpha}$, the so called POSI constant, is the quantile of a maximum of Gaussian rather than t-distributed random variables, as is the case in Berk et al. (2013). Furthermore, we simply use the usual variance estimator $\hat{\sigma}_{\mathbb{M},n}^2$ which, in general, is not unbiased or uniformly consistent (due to potential misspecification) as required in Berk et al. (2013), but we still obtain uniformly valid inference asymptotically. This shows that the somewhat restrictive assumption of Berk et al. (2013), that there exists an unbiased or a uniformly consistent estimator for σ_n^2 (cf. Lemma 3.4 below, as well as the discussion in Remark 2.1 of Leeb et al. (2015) and in Section 2.2 of Bachoc et al. (2014)), is not needed for uniform asymptotic validity. If the estimator $\hat{\sigma}_{\mathbb{M},n}^2$ is used in the construction of Berk et al. (2013), then their confidence intervals asymptotically coincide with our procedure. We also point out that the classical variance estimator used here adapts to misspecification in the sense that it is consistent for σ_n^2 if a first order correct model is selected and it otherwise overestimates the target in the sense of Section 2.3.2 (cf. Remark 2.7 and Subsection 3.1.1 for details).

It is also worth noting that up to the choice of the last multiplicative factor $K_{1-\alpha}(\mathrm{corr}(\Gamma_n))$ in the definition of the confidence intervals above, i.e., the POSI constant, this is just the usual confidence interval for the j-th coordinate of the coefficient vector one would typically use in practice working with homoskedastic linear models, and by following the naive way of ignoring the data-driven model selection step and the potential misspecification. The crucial difference, however, is that the naive approach is invalid (see, e.g., Bachoc et al., 2014; Leeb et al., 2015).

We now present the main result of this subsection, where we emphasize once more that the (measurable) model selection procedure $\hat{\mathbb{M}}_n$ is data-driven and unrestricted, and that some, or all of the candidate models in M_n may be misspecified, i.e., $\mathbf{P}_n^{(\mathrm{lm})}(\delta, \tau) \nsubseteq \bigcup \mathsf{M}_n$. Nevertheless it is possible to construct an asymptotically uniformly valid confidence set for the model-specific target vector $\beta^*_{\hat{\mathbb{M}}_n,n}$.

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

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

3.1.1. The POSI-intervals automatically adapt to misspecification

Let us for a moment forget about the model selection step and consider the classical construction of confidence intervals for $\beta^*_{\mathbb{M},n}$ based on the asymptotic normality of $\hat{\beta}_{\mathbb{M},n}$ for $\mathbb{M} \in \mathsf{M}_n$ fixed. It is well known that the usual variance estimator $\hat{\sigma}^2_{\mathbb{M},n}$ in model $\mathbb{M} \in \mathsf{M}_n$ with index set $M \in \mathcal{I}$ is consistent for the true error variance $\sigma_n^2 := m_2(\mathbb{P}_{1,n})$, if and only if, the model $\mathbb{M}$ is asymptotically first order correct. More specifically, the estimator $\hat{\sigma}^2_{\mathbb{M},n}$ is upward biased when $\mathbb{M}$ is mean-misspecified and the bias is given by $\|(I_n - P_{X_n[M]})\mu_n/\sigma_n\|^2/(n-|M|)$, where $P_{\dots}$ denotes the projection matrix corresponding to the column span of the matrix indicated in the subscript and where $\mu_n := \mu(\mathbb{P}_n)$. Consequently, using $\hat{\sigma}^2_{\mathbb{M},n}$ in the construction of confidence intervals for $\beta^*_{\mathbb{M},n}$ leads only to more 'conservative' inference in case $\mathbb{M}$ is misspecified, but the resulting inference is still valid. Moreover, if $\mathbb{M}$ is correct then the resulting inference on $\beta^*_{\mathbb{M},n}$ is also asymptotically efficient in the sense that the obtained intervals have minimal asymptotic length (they coincide with the infeasible intervals calculated with knowledge of σ_n^2).

If the working model is now selected by a data dependent selection procedure $\hat{\mathbb{M}}_n : \mathbb{R}^n \to \mathsf{M}_n$, the POSI-intervals suggested in (3.1) involve the post-model-selection estimator $\hat{\sigma}^2_{\hat{\mathbb{M}}_n,n}$ and we can again ask the question about its consistency properties. Clearly, consistency now depends on the existence of a correct model in the class M_n of candidate models as well as on the ability of $\hat{\mathbb{M}}_n$ to identify one such correct model. The following proposition provides a precise quantitative formulation of this claim (see Section C.1 for the proof). A similar statement is also discussed in Bachoc et al. (2014, Theorem 3.6 and Lemma B.2).

Proposition 3.3. *Fix* $\delta > 0$ *and* $\tau \geq 1$ *and suppose that eventually* $\mathrm{rank}(X_n) = p$*. For any sequence* $\mathbb{P}_n \in \mathbf{P}_n^{(\mathrm{lm})}(\delta,\tau)$*, the variance estimator* $\hat{\sigma}^2_{\mathbb{M},n}(y) = (n - |M|)^{-1} y'(I_n - P_{X_n[M]})y$ *satisfies*

$$\mathbb{P}_n \left(\left| \frac{\hat{\sigma}^2_{\mathbb{M},n}}{\sigma_n^2} \left(1 + \frac{\|(I_n - P_{X_n[M]})\mu_n/\sigma_n\|^2}{n-|M|} \right)^{-1} - 1 \right| > \varepsilon \right) \to 0,$$

for every $\varepsilon > 0$ *and for every* $M \in \mathcal{I}$*.*

In particular, since the index set $\mathcal{I}$ is finite, not depending on n, Proposition 3.3 shows that the post-model-selection estimator $\hat{\sigma}^2_{\hat{\mathbb{M}}_n,n}$ is consistent for

σ_n^2 (on a relative scale) along sequences $\mathbb{P}_n \in \mathbf{P}_n^{(\mathrm{lm})}(\delta, \tau)$ such that

$$\frac{\|(I_n - P_{X_n[\hat{M}_n]})\mu_n/\sigma_n\|^2}{n - |\hat{M}_n|} \xrightarrow[n\to\infty]{} 0, \tag{3.4}$$

in $\mathbb{P}_n$-probability. Of course, in general, this condition can not be verified in practice. But it still tells us that if the model selection procedure $\hat{\mathbb{M}}_n$ finds an approximately first order correct model in the sense of (3.1.1), then the POSI-intervals in (3.1) have the same asymptotic length as the infeasible intervals that use knowledge of σ_n^2. One can now raise the question if it is possible to construct a uniformly (over $\mathbf{P}_n^{(\mathrm{lm})}(\delta, \tau)$) consistent estimator for σ_n^2. However, in the framework we consider, in order to obtain valid confidence intervals, it is necessary to use an estimator that, for certain sequences $\mathbb{P}_n \in \mathbf{P}_n^{(\mathrm{lm})}(\delta, \tau)$, consistently overestimates the variance σ_n^2. More precisely, in Lemma 3.4 below (proved in Section C.3), we show that there does not exist a uniformly consistent estimator of the variance σ_n^2 if $\tau > 1$.

Lemma 3.4. *Let $\delta > 0$ and $\tau > 1$. There does not exist a sequence of measurable functions $(\hat{\sigma}_n^2)_{n\in\mathbb{N}}$ with $\hat{\sigma}_n^2 : \mathbb{R}^n \to [0, \infty)$, so that for every $\varepsilon > 0$*

$$\sup_{\mathbb{P}_n \in \mathbf{P}_n^{(\mathrm{lm})}(\delta,\tau)} \mathbb{P}_n\left(\left|\frac{\hat{\sigma}_n^2}{\sigma_n^2} - 1\right| > \varepsilon\right) \to 0. \tag{3.5}$$

3.2. Inference post-model-selection when fitting fixed design linear models to heteroskedastic data

The feasible sets for $\mathbb{P}_n$ we consider here again depend on two parameters $\delta > 0$ and $\tau \geq 1$ but, compared to the set $\mathbf{P}_n^{(\mathrm{lm})}(\delta, \tau)$ defined above, we now drop the requirement of homoskedasticity and let

$$\mathbf{P}_n^{(\mathrm{het})}(\delta, \tau) := \left\{ Q \in \bigotimes_{i=1}^n M(\mathbb{R}) : \begin{array}{l} 0 < m_2(Q_i) < \infty \text{ for } i = 1, \dots, n \\ \frac{\max_{i=1,\dots,n} m_{2+\delta}(Q_i)^{\frac{2}{2+\delta}}}{\min_{i=1,\dots,n} m_2(Q_i)} \leq \tau \end{array} \right\}.$$

Here, we consider a situation where one works with candidate sets consisting of heteroskedastic linear models, i.e., where similar as above one is interested in conducting inference on $\mu(\mathbb{P}_n)$, and it is *assumed* that $\mu(\mathbb{P}_n)$ of the true distribution $\mathbb{P}_n \in \mathbf{P}_n^{(\mathrm{het})}(\delta, \tau)$ is an element of $\mathrm{span}(X_n)$, the column span of a design matrices $X_n \in \mathbb{R}^{n\times p}$ with p fixed, or that $\mu(\mathbb{P}_n)$ is at least 'well-approximated' by an element of that linear space; but where it is taken into account that the observations might have different variances. We start with a set $\mathcal{I} = \{M_1, \dots, M_d\}$ as in Subsection 3.1, and we then define for each $j \in \{1, \dots, d\}$ the candidate sets

$$\mathbb{M}_{j,n} = \left\{ Q \in \bigotimes_{i=1}^n M(\mathbb{R}) : \begin{array}{l} 0 < m_2(Q_i) < \infty \text{ for } i = 1, \dots, n \\ \mu(Q) \in \mathrm{span}(X_n[M_j]) \end{array} \right\},$$

and hence our corresponding candidate set of models is then given by

$$\mathsf{M}_n = \{\mathbb{M}_{j,n} : j = 1, \ldots, d\}.$$

As above we assume that X_n satisfies Condition X1, and define our model-specific target of inference as in Equation (3.1). Again, we estimate the corresponding target by the model-specific ordinary-least-squares estimator in (3.1). For variance estimation we do no longer use the estimator as defined in Subsection 3.1, but now take into consideration, that the observations might be heteroskedastic. Therefore, we consider an approach based on estimators suggested by Eicker (1967). As in Subsection 3.1 the variance estimators used are not uniformly consistent due to potential model misspecification, but overestimate their targets in the sense of Subsection 2.3.2. Furthermore, in contrast to the construction made in Subsection 3.1 the construction of the confidence sets needs to incorporate an upper bound for the POSI constant $K_{1-\alpha}(\operatorname{corr}(\Gamma_n))$, because here $\operatorname{corr}(\Gamma_n)$ is unobserved and can not be estimated consistently due to potential misspecification. Define for every $\mathbb{M} \in \mathsf{M}_n$ with corresponding index set M the Eicker-estimator $\tilde{S}_{\mathbb{M},n}$ as

$$(X_n[M]'X_n[M])^{-1} X_n[M]' \operatorname{diag}\left(\hat{u}^2_{1,\mathbb{M}}, \ldots, \hat{u}^2_{n,\mathbb{M}}\right) X_n[M] \left(X_n[M]'X_n[M]\right)^{-1},$$

where we let $\hat{u}_{\mathbb{M}}(y) = (\hat{u}_{1,\mathbb{M}}(y), \ldots, \hat{u}_{n,\mathbb{M}}(y))' = y - X_n[M]\hat{\beta}_{\mathbb{M},n}(y)$, and denote the $j-$th diagonal entry $(j = 1, \ldots, m(\mathbb{M}))$ of $\tilde{S}_{\mathbb{M},n}$ by

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

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

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

with $k = \sum_{\mathbb{M}\in\mathsf{M}_n} m(\mathbb{M})$.

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

Theorem 3.5. *Let $\alpha \in (0,1)$, $\delta > 0$ and $\tau \geq 1$, suppose Condition X1 holds, and let $\hat{\mathbb{M}}_n$ be a (measurable) model selection procedure, i.e., a measurable map from the sample space to M_n. Then*

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

3.3. Inference post-model-selection when fitting binary regression models to binary data

The feasible sets for $\mathbb{P}_n$ we consider here depend on a parameter $\tau > 0$ and are defined as

$$\mathbf{P}_n^{(\mathrm{bin})}(\tau) := \left\{ Q \in \bigotimes_{i=1}^{n} M(\{0,1\}) : Q_i(\{0\}) Q_i(\{1\}) \geq \tau \ \forall i = 1, \ldots, n \right\}.$$

We consider a situation where binary regression models are fit to binary data generated under one of the elements $\mathbb{P}_n \in \mathbf{P}_n^{(\mathrm{bin})}(\tau)$. In binary regression the maintained *modeling assumption* is that the probability of a success on the i-th observation ($y_{i,n} = 1$) is given by $Q_i(\{1\}) = h(X_{i,n}\beta)$, for some $\beta \in \mathbb{R}^p$, some response function $h : \mathbb{R} \to (0,1)$ and where $X_{i,n}$ is the i-th row of a design matrix $X_n \in \mathbb{R}^{n \times p}$. Usually, when h is invertible, h^{-1} is called the link function. Thus, unlike the previous two examples, here we also have to make a choice for the response function h, in addition to selecting variables from X_n. Classical choices are the logit and the probit functions, but we allow also for other choices of response functions h, as long as they belong to a finite set $\mathcal{H} = \{h_1, \ldots, h_{d_1}\}$ of potential candidates. Together with the collection $\mathcal{I} = \{M_1, \ldots, M_{d_2}\} \subseteq 2^{\{1,\ldots,p\}} \setminus \varnothing$ of candidate regressors, we can define a candidate model by

$$\mathbb{M}_{(j_1,j_2),n} = \left\{ Q \in \bigotimes_{i=1}^{n} M(\{0,1\}) : \begin{array}{l} \exists \beta \in \mathbb{R}^{|M_{j_2}|} : \forall i = 1, \ldots, n : \\ Q_i(\{1\}) = h_{j_1}(X_{i,n}[M_{j_2}]\beta) \end{array} \right\},$$

for $j_1 \in \{1, \ldots, d_1\}$ and $j_2 \in \{1, \ldots, d_2\}$. Thus, our candidate set of size $d = d_1 \cdot d_2$ is given by

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

We need to impose some regularity conditions on the possible response functions $h \in \mathcal{H}$ and the design X_n.

Condition X2. *Let $C > 0$ be fixed. Eventually, we have*

(i) $\mathrm{rank}(X_n) = p$;
(ii) $\max_{i=1,\ldots,n} X_{i,n}[M] \left(X_n[M]' X_n[M] \right)^{-1} X_{i,n}[M]' \leq C/n$, *for every* $M \in \mathcal{I}$;
(iii) $\lambda_{max}(X_n' X_n) / \lambda_{min}(X_n' X_n) \leq C$;

Condition H. *The elements $h \in \mathcal{H}$ have the following properties:*

(i) $h : \mathbb{R} \to (0,1)$ *is a continuous cumulative distribution function;*
(ii) *The functions* $\phi_1(\gamma) := \log(h(\gamma))$ *and* $\phi_2(\gamma) := \log(1 - h(\gamma))$ *are strictly concave on* $\mathbb{R}$*;*

(iii) *h is twice continuously differentiable and ϕ_1 and ϕ_2 have strictly negative second derivatives on $\mathbb{R}$;*

(iv) *The derivative $\dot{h}$ of h is strictly positive on $\mathbb{R}$;*

Remark 3.6. *Condition X2 is a strengthened version of Condition X1. It is still satisfied if $\|X_{i,n}\|$ is bounded and $\lambda_{min}(\frac{1}{n}X_n'X_n)$ is bounded away from 0, as is typically the case in factorial designs.*

Remark 3.7. *The Conditions H(i) and H(iv) are rather natural and essential for parameter identification. Condition H(iii) is also classical and used to ensure continuity of the Hessian of the log-likelihood (cf. Fahrmeir, 1990; Fahrmeir and Kaufmann, 1985). Finally, Condition H(ii), which is implied by Condition H(iii), ensures strict concavity of the log-likelihood, which, in turn, guarantees uniqueness of pseudo parameters and the MLE (see Lemma 3.8 and Lemma 3.9 below). It is easy to see that Condition H is satisfied, e.g., for response functions corresponding to the classical logit, probit, log-log and complementary log-log link functions discussed in McCullagh and Nelder (1989, p.108).*

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

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

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

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

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

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

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

As our target of inference we take the model dependent vector $\beta^*_{\mathbb{M},n} \in \mathbb{R}^{|M|}$ that maximizes the expected log-likelihood $\beta \mapsto \mathbb{E}_n[\ell_{\mathbb{M},n}(\cdot,\beta)]$ under the true data generating distribution $\mathbb{P}_n \in \mathbf{P}_n^{(\mathrm{bin})}(\tau)$. If $\beta^*_{\mathbb{M}}$ exists, then it is easy to see that it also minimizes the Kullback-Leibler divergence between the true data generating distribution $\mathbb{P}_n$ and the class of distributions specified by the working model $\mathbb{M} \in \mathsf{M}_n$. Focusing on the Kullback-Leibler minimizer has a longstanding tradition in the misspecification literature dating back at least to Huber (1967) (see also White (1982) and the references given therein). For references more specific to generalized linear models see Fahrmeir (1990) and Lv and Liu (2014). That this target uniquely exists in the present context of binary regression is the subject of the following lemma (see Section C.5 for the proof).

Lemma 3.8. *Suppose that* $\operatorname{rank}(X_n) = p$ *and H(i,ii) hold. Then, for every* $\mathbb{M} \in \mathsf{M}_n$ *and for every* $\mathbb{P} \in \bigcup_{\delta>0} \mathbf{P}_n^{(\mathrm{bin})}(\delta)$*, there exists a unique vector* $\beta^*_{\mathbb{M},n} = \beta^*_{\mathbb{M},n}(\mathbb{P}) \in \mathbb{R}^{m(\mathbb{M})}$*, such that*

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

Furthermore, it is well known that for some points in the sample space $\{0,1\}^n$ the MLE in the binary regression model does not exist (see, e.g., Wedderburn, 1976). But those samples have vanishing asymptotic probability. The following lemma establishes this asymptotic existence of the MLE $\hat\beta_{\mathbb{M},n}$ in the present setting, along with uniform consistency. Its proof is deferred to Section C.7.

Lemma 3.9. *Suppose that Conditions X2(i,ii) and H(i,ii,iii) hold and fix* $\tau > 0$*. Then, for every* $n \in \mathbb{N}$*, every* $\mathbb{M} \in \mathsf{M}_n$ *and every* $\mathbb{P} \in \mathbf{P}_n^{(\mathrm{bin})}(\tau)$*, there exists a function* $\hat\beta_{\mathbb{M},n} : \{0,1\}^n \to \mathbb{R}^{m(\mathbb{M})}$ *(depending only on* n *and* $\mathbb{M}$*) and a set* $E_{\mathbb{M},\mathbb{P},n} \subseteq \{0,1\}^n$*, such that*

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

and

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

Moreover, for the pseudo parameter $\beta^*_{\mathbb{M},n} \in \mathbb{R}^{m(\mathbb{M})}$ *of Lemma 3.8, we have*

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

as $\delta \to \infty$*.*

To construct asymptotically valid confidence intervals for the components of $\beta^*_{\mathbb{M},n}$, we need an estimate of the asymptotic covariance matrix of $\hat\beta_{\mathbb{M},n}$. In the misspecified setting it is usually not possible to obtain a consistent estimator. We here follow the suggestion of Fahrmeir (1990, p. 491) who proposed a sandwich-type estimator for misspecified generalized linear models. This estimator fits with the general idea of Section 2.3.2. For $\mathbb{M} \in \mathsf{M}_n$, $\mathbb{M} \triangleq (h, M)$, define

$$\tilde S_{\mathbb{M},n} = \hat H^{-1}_{\mathbb{M},n} X_n[M]' \operatorname{diag}\left(\hat u^2_{1,\mathbb{M}}, \dots, \hat u^2_{n,\mathbb{M}}\right) X_n[M] \hat H^{-1}_{\mathbb{M},n},$$

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

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

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

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

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

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

with $k = \sum_{\mathbb{M}\in\mathsf{M}_n} m(\mathbb{M})$.

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

Theorem 3.10. *Let $\alpha \in (0,1)$ and $\tau > 0$, suppose Conditions X2 and H hold, and let $\hat{\mathbb{M}}_n$ be a model selection procedure, i.e., a map from the sample space $\{0,1\}^n$ to M_n. Then*

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

Remark 3.11. *It is important to note that if one decides a priori to use only the canonical link function, which, in the present case of binary regression, corresponds to the logistic response function $h^{(c)}(\gamma) := e^\gamma/(1+e^\gamma)$, then Theorem 3.10 holds with $B_\alpha(\min(k,n),k)$ decreased to $B_\alpha(\min(k,p),k)$. See Corollary C.3 in Section C.9.*

Remark 3.12. *We point out that similar principles used to derive Theorem 3.10 can also be employed to treat other quasi-maximum likelihood or general M-, and Z-estimation problems (see, e.g., Fahrmeir, 1990, for a more general treatment of generalized linear models). The general theory of M- and Z-estimation as presented, e.g., in van der Vaart and Wellner (1996, Sections 3.2 and 3.3), usually also leads to expansions of the form required in Conditon 1 (cf. van der Vaart and Wellner, 1996, Theorem 3.2.16 and Theorem 3.3.1). These results are stated in a pointwise fashion but can be made uniform over large classes of data generating processes by using ideas from Section 2.8 of the same reference. However, in more specific examples such as the present binary regression setting, conditions can be directly imposed on the design and the link functions and can be optimized for this setup.*

Finally, the main difference compared to the previous two examples is that here we fit non-linear regression models to binary observations. Therefore, a data driven model selection procedure $\hat{\mathbb{M}}_n$ may not only select variables among the p candidate regressors in X_n, but may also result in a choice of a response function h from some pre-specified class $\mathcal{H}$. In practice, $\mathcal{H}$ often contains certain classical candidates such as, e.g., the response functions corresponding to the logit, probit or complementary log-log link function. A working model could then be selected, for instance, by minimizing some penalized likelihood criterion over all possible choices of $(h, M) \in \mathcal{H} \times \mathcal{I}$. However, we emphasize once more, that the specifics of the possibly data driven model selection procedure $\hat{\mathbb{M}}_n$ are completely inconsequential for the validity of our proposed confidence intervals and could also involve graphical inspection of the data and subjective preferences.

4. Conclusion

We have presented a general theory for the construction of asymptotically valid confidence sets post-model-selection. Our methods can be used in a wide number of situations, because they are only based on a standard representation that can often be obtained by simple linearization arguments. We have also applied our theory to construct valid confidence sets after selecting and fitting fixed design linear models to (possibly non-Gaussian) homoskedastic or heteroskedastic data. Moreover, we have investigated the practically very important case when binary regression models are fit to binary data. In this case, in addition to selecting variables from a given design matrix, also the choice of an appropriate link function can be made in a data driven way. The general theory and the proposed methods are applicable irrespective of whether any of the candidate models under consideration is correctly specified, leading to more or less conservative inference depending on the severity of misspecification (see Remark 2.7). This feature is also present in the applications of Section 3.1 (see Subsection 3.1.1), 3.2 and 3.3.

Open questions that go beyond the scope of this article, but are currently under investigation, include the extension of the approach discussed here to dependent data; the applicability and performance of bootstrap procedures; and the development of procedures in the spirit of Berk et al. (2013) in the challenging situation when the number of models fitted can grow with sample size. In ongoing work we apply our methods to real data and investigate if they can prevent spurious findings while detecting true reproducible effects.

Acknowledgements

David Preinerstorfer was supported, in part, by the Austrian Science Fund (FWF): P27398 and by the Danish National Research Foundation (grant DNRF 78, CREATES), and Lukas Steinberger was supported by the Austrian Science Fund (FWF): P28233, which is greatly acknowledged. Results related to the present article were presented in the Statistics and Econometrics Research Seminar at the Department of Statistics and Operations Research at the University of Vienna, and we would like to thank the participants, in particular Hannes Leeb, Benedikt M. Pötscher and Ulrike Schneider, for helpful comments and suggestions.

References

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

Bachoc, F., Leeb, H. and Pötscher, B. M. (2014). Valid confidence intervals for post-model-selection predictors. *arXiv:1412.4605*.

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

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

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

Billingsley, P. (1968). *Convergence of probability measures.* John Wiley & Sons.

Dudley, R. M. (2002). *Real analysis and probability.* Cambridge University Press.

Eicker, F. (1967). Limit theorems for regressions with unequal and dependent errors. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1. 59–82.

Fahrmeir, L. (1990). Maximum likelihood estimation in misspecified generalized linear models. *Statistics*, **21** 487–502.

Fahrmeir, L. and Kaufmann, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *Ann. Statist*, **13** 342–368.

Fithian, W., Sun, D. and Taylor, J. (2015). Optimal inference after model selection. *arXiv:1410.2597*.

Gnedenko, B. and Kolmogorov, A. N. (1954). *Limit distributions for sums of independent random variables.* Addison-Wesley.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability.* 1, 221–233.

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and monte carlo. *Ann. Statist.*, **1** 799–821.

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

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

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

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

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

Leeb, H. and Pötscher, B. M. (2006). Performance limits for estimators of the risk or distribution of shrinkage-type estimators, and some general lower risk-bound results. *Econometric Theory*, **22** 69–97.

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

Leeb, H., Pötscher, B. M. and Ewald, K. (2015). On various confidence intervals post-model-selection. *Statist. Sci.*, **30** 216–227.

Lv, J. and Liu, J. S. (2014). Model selection principles in misspecified models. *J. R. Statist. Soc. B*, **76** 141–167.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models.* 2nd ed. London, New York: Chapman & Hall.

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

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

Raikov, D. (1938). On a connection between the central limit-law of the theory of probability and the law of great numbers. *Izvestiya Rossiiskoi Akademii Nauk. Seriya Matematicheskaya*, **2** 323–338.

Tibshirani, R. J., Rinaldo, A., Tibshirani, R. and Wasserman, L. (2015). Uniform asymptotic inference and the bootstrap after model selection. *arXiv:1506.06266v2*.

Tibshirani, R. J., Taylor, J., Lockhart, R. and Tibshirani, R. (2014). Exact post-selection inference for sequential regression procedures. *arXiv:1401.3889*.

van de Geer, S., Bühlmann, P., Ritov, Y. and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.*, **42** 1166–1202.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics.* Springer.

Wedderburn, R. W. M. (1976). On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. *Biometrika*, **63** 27–32.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, **50** 1–25.

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

Zhang, K. (2015). Spherical cap packing asymptotics and rank-extreme detection. *arXiv:1511.06198*.

Appendix A: Auxiliary results

In this section, for every $n \in \mathbb{N}$ and for some $k \in \mathbb{N}$, not depending on n, let $z_{1,n}, \ldots, z_{n,n}$ be independent k-variate random vectors defined on a probability space $(\Omega_n, \mathcal{A}_n, \mathbb{P}_n)$. Denote

$$r_n \quad = \quad (r_n^{(1)}, \ldots, r_n^{(k)})' \quad = \quad \sum_{i=1}^{n} (z_{i,n}^{(1)}, \ldots, z_{i,n}^{(k)})',$$

and let

$$S_n \quad = \quad \sum_{i=1}^{n} z_{i,n} z_{i,n}'.$$

In this section, the expectation operator, the variance-covariance operator, and the variance operator w.r.t. $\mathbb{P}_n$ is denoted by $\mathbb{E}_n$, $\mathbb{VC}_n$, and $\mathbb{V}_n$, respectively.

Condition 2. *For every $n \in \mathbb{N}$, every $j \in \{1, \ldots, k\}$ and every $i \in \{1, \ldots, n\}$*

$$\mathbb{E}_n(z_{i,n}^{(j)}) = 0 \quad \textit{and} \quad \mathbb{V}_n(r_n^{(j)}) = 1. \tag{A.1}$$

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

$$\mathbb{P}_n \circ r_n^{(j)} \quad \Rightarrow \quad N(0, 1), \tag{A.2}$$

and for every $\varepsilon > 0$

$$\max_{1 \le i \le n} \mathbb{P}_n \left(|z_{i,n}^{(j)}| \ge \varepsilon \right) \to 0. \tag{A.3}$$

The first statement in the subsequent lemma is essentially Corollary 2 in Pollak (1972) combined with a tightness argument. The second statement is obtained via an application of Raikov's theorem (Raikov (1938), cf. the statement given in Gnedenko and Kolmogorov (1954) on p. 143).

Lemma A.1. *Let Condition 2 hold. Then*

$$d_w\left(\mathbb{P}_n \circ r_n, N(0, \mathbb{VC}_n(r_n))\right) \to 0.$$

Furthermore, for every $\varepsilon > 0$ it holds that

$$\mathbb{P}_n \left(\left\| S_n - \mathbb{VC}_n(r_n) \right\| \ge \varepsilon \right) \to 0,$$

and hence that

$$\mathbb{P}_n \left(d_w\left(\mathbb{P}_n \circ r_n, N(0, S_n)\right) \ge \varepsilon \right) \to 0. \tag{A.4}$$

Proof. For the first claim, let n' be an arbitrary subsequence. From Equation (2) we see that $\mathbb{E}_n(r_n) = 0$ and that $\mathbb{VC}_n(r_n)$ is norm-bounded and, hence, that $\mathbb{P}_n \circ r_n$ is tight. Therefore, there exists a subsequence n'' of n' along which $\mathbb{VC}_n(r_n)$ converges to Σ, say, and along which $\mathbb{P}_n \circ r_n$ converges weakly. We now need to show that $d_w(\mathbb{P}_{n''} \circ r_{n''}, N(0,\Sigma)) \to 0$, which then proves the statement in view of the triangle inequality, continuity of $\Gamma \mapsto N(0,\Gamma)$ w.r.t. d_w, and the fact that n' was arbitrary. That the weak limit of $\mathbb{P}_{n''} \circ r_{n''}$ must be normal follows from Equations (2) and (2), applying Corollary 2 in Pollak (1972). The mean vector of the limiting distribution of $\mathbb{P}_{n''} \circ r_{n''}$ is 0 from (2). It remains to verify that the covariance matrix of the limiting distribution is Σ. From Equations (2) and (2) and, e.g., Theorem 5.4 in Billingsley (1968) it follows that $[r_n^{(j)}]^2$ is uniformly integrable for $j = 1, \ldots, k$. The inequality $ab \leq \frac{1}{2}(a^2 + b^2)$, together with the fact that the sum of two uniformly integrable sequences is uniformly integrable, then shows uniform integrability of $r_n^{(s)} r_n^{(t)}$, and hence (e.g., again Theorem 5.4 in Billingsley (1968) together with weak convergence of $\mathbb{P}_{n''} \circ r_{n''}$ and the continuous mapping theorem) that the covariance matrix of the limiting distribution of $\mathbb{P}_{n''} \circ r_{n''}$ coincides with Σ.

To prove the second claim, we start with the observation that it suffices to verify that for every $\varepsilon > 0$ and every $\gamma \in \mathbb{R}^k$ it holds that

$$\mathbb{P}_n \left(|\gamma' \left(S_n - \mathbb{VC}_n(r_n) \right) \gamma| \geq \varepsilon \right) \to 0.$$

To see this, it suffices to first take γ equal to the elements of the standard basis in $\mathbb{R}^k$ in order to show that the diagonal entries converge to zero. Then, taking γ equal to $(1,1,0,\ldots,0)'$, $(0,1,1,0,\ldots,0)'$, etc., and using symmetry shows that also the entries above and below the main diagonal converge. Continuing this process with vectors containing exactly three, four, five, etc., consecutive ones, establishes the claim. Next, to verify the statement in the previous display, let $\gamma \in \mathbb{R}^k$, $\varepsilon > 0$, and let n' be an arbitrary subsequence. Choose n'' a subsequence of n' along which $\mathbb{P}_n \circ r_n \Rightarrow N(0,\Sigma)$ - such a subsequence exists because of the already established part of the lemma. We also already know from the uniform integrability argument above, that then $\mathbb{VC}_{n''}(r_{n''}) \to \Sigma$, and hence that $\sigma^2_{n''} := \gamma' \mathbb{VC}_{n''}(r_{n''})\gamma \to \gamma'\Sigma\gamma =: \sigma^2$. Now, if $\sigma^2 = 0$, then eventually $\sigma^2_{n''} < \varepsilon/2$, and by Markov's inequality

$$\mathbb{P}_{n''} \left(\left|\gamma' \left(S_{n''} - \mathbb{VC}_{n''}(r_{n''}) \right) \gamma\right| \geq \varepsilon \right) \leq \mathbb{P}_{n''} \left(\gamma' S_{n''} \gamma \geq \frac{\varepsilon}{2} \right) \leq \frac{2\sigma^2_{n''}}{\varepsilon} \to 0.$$

Suppose next that $\sigma^2 > 0$. Then, we can assume without loss of generality that $0 < \delta_1 \leq \sigma^2_{n''} \leq \delta$ for some $\delta, \delta_1 \in \mathbb{R}$, and it remains to verify that

$$\mathbb{P}_{n''} \left(\left| \frac{\gamma' S_{n''} \gamma}{\sigma^2_{n''}} - 1 \right| \geq \frac{\varepsilon}{\delta} \right) \to 0. \tag{A.5}$$

To that end, define

$$\xi_{i,n''} \quad = \quad \frac{\gamma' z_{i,n''}}{\sigma_{n''}}$$

where $\sigma_{n''}$ denotes the positive square root of $\sigma^2_{n''}$, and note that by Equation (2) we have $\mathbb{E}_{n''}(\xi_{i,n''}) = 0$, that by construction $\sigma^{-2}_{n''} \gamma' S_{n''} \gamma = \sum_{i=1}^{n''} \xi^2_{i,n''}$, and that $\mathbb{V}_{n''}(\sum_{i=1}^{n''} \xi_{i,n''}) = 1$. Note also that it follows from Equation (2) that for every $\bar{\delta} > 0$ we have

$$\max_{1 \le i \le n''} \mathbb{P}_{n''} \left(|\xi_{i,n''}| \ge \bar{\delta} \right) \to 0 \text{ as } n'' \to \infty.$$

Furthermore, since $\mathbb{P}_{n''} \circ r_{n''} \Rightarrow N(0, \Sigma)$, we see that $\sum_{i=1}^{n''} \xi_{i,n''}$ is asymptotically normal with mean 0 and variance 1. But then Equation (A) follows from Raikov's theorem (Gnedenko and Kolmogorov (1954), p. 143, Theorem 4). Since n' was arbitrary, this proves the second statement.

The statement in Equation (A.1) is an immediate consequence of the triangle inequality and the first two statements. ☐

Condition 3. *For every $n \in \mathbb{N}$, every $j \in \{1, \dots, k\}$ and every $i \in \{1, \dots, n\}$ we have*

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

Furthermore, setting $r_{n,*}^{(j)} = \frac{r_n^{(j)}}{\sqrt{\mathbb{V}_n(r_n^{(j)})}}$ and $z_{i,n,*}^{(j)} = \frac{z_{i,n}^{(j)}}{\sqrt{\mathbb{V}_n(r_n^{(j)})}}$, for every $j \in \{1, \dots, k\}$ we have

$$\mathbb{P}_n \circ r_{n,*}^{(j)} \Rightarrow N(0, 1),$$

and for every $\varepsilon > 0$

$$\max_{1 \le i \le n} \mathbb{P}_n \left(|z_{i,n,*}^{(j)}| \ge \varepsilon \right) \to 0 \text{ as } n \to \infty.$$

Lemma A.2. *Let Condition 3 hold. Then for every $\varepsilon > 0$ it holds that*

$$\mathbb{P}_n \left(\| \operatorname{diag}(\mathbb{VC}_n(r_n))^{-1} \operatorname{diag}(S_n) - I_k \| \ge \varepsilon \right) \to 0 \tag{A.6}$$

and

$$\mathbb{P}_n \left(\| \operatorname{corr}(S_n) - \operatorname{corr}(\mathbb{VC}_n(r_n)) \| \ge \varepsilon \right) \to 0. \tag{A.7}$$

Furthermore, for every $\varepsilon > 0$, $\hat{r}_{n,*} = \operatorname{diag}(S_n)^{\dagger/2} \, r_n$ satisfies

$$\mathbb{P}_n \left(d_w \left(\mathbb{P}_n \circ \hat{r}_{n,*}, N(0, \operatorname{corr}(S_n)) \right) \ge \varepsilon \right) \to 0.$$

Proof. For the statements in Equations (A.2) and (A.2) we first note that the triangular array $z_{i,n,*}^{(j)}$, and the corresponding quantities $r_{n,*}$ and $r_{n,*}^{(j)}$, satisfy Condition 2. Hence, Lemma A.1 is applicable, and shows, in particular, for every $\varepsilon > 0$ and with the abbreviation $S_{n,*} = \sum_{i=1}^{n} z_{i,n,*} z_{i,n,*}'$, that

$$\mathbb{P}_n \left(\| S_{n,*} - \mathbb{VC}_n(r_{n,*}) \| \ge \varepsilon \right) \to 0 \quad \text{as } n \to \infty.$$

Noting that the diagonal entries of $\mathbb{VC}_n(r_{n,*})$ are all equal to 1 (in fact $\mathbb{VC}_n(r_{n,*}) = \operatorname{corr}(\mathbb{VC}_n(r_n))$) and that $\operatorname{diag}(S_{n,*}) = \operatorname{diag}(\mathbb{VC}_n(r_n))^{-1} \operatorname{diag}(S_n)$ holds, establishes the claimed convergence in (A.2). But this together with the preceding display then establishes the convergence in (A.2), because, using the abbreviation $A_n = \operatorname{diag}(S_n)^{\dagger/2} \operatorname{diag}(\mathbb{VC}_n(r_n))^{1/2}$, we have $A_n \to I_k$ in $\mathbb{P}_n$-probability and

$$\begin{aligned}\operatorname{corr}(S_n) - \operatorname{corr}(\mathbb{VC}_n(r_n)) &= \operatorname{diag}(S_n)^{\dagger/2} S_n \operatorname{diag}(S_n)^{\dagger/2} - \mathbb{VC}_n(r_{n,*}) \\ &= A_n(S_{n,*} - \mathbb{VC}_n(r_{n,*}))A_n' \\ &\quad + A_n \mathbb{VC}_n(r_{n,*})A_n' - \mathbb{VC}_n(r_{n,*}),\end{aligned}$$

which converges to zero in $\mathbb{P}_n$-probability. The last part is an application of the statements already established, together with

$$\mathbb{P}_n \left(d_w \left(\mathbb{P}_n \circ r_{n,*}, N(0, S_{n,*}) \right) \geq \varepsilon \right) \to 0 \text{ for every } \varepsilon > 0,$$

which we obtain (as above) from Lemma A.1. □

Lemma A.3. *For every $\alpha \in (0,1)$ the map $\Gamma \mapsto K_{1-\alpha}(\Gamma)$ is continuous on the subset of $k \times k$-dimensional covariance matrices of $\mathbb{R}^{k\times k}$.*

Proof. Let Γ_n be a sequence of covariance matrices converging to Γ. By definition, $K_{1-\alpha}(\Gamma_n)$ is the $1-\alpha$-quantile of the distribution of $\|Z_n\|_\infty$, where Z_n is a Gaussian random vector with mean 0 and covariance matrix Γ_n. By the continuous mapping theorem, $\|Z_n\|_\infty$ converges weakly to $\|Z\|_\infty$, where Z is a Gaussian random vector with mean 0 and covariance matrix Γ. In case $\Gamma \neq 0$ it is easy to see that the distribution function of $\|Z\|_\infty$ is everywhere continuous and strictly increasing on $[0,\infty)$, and the result then follows, because weak convergence of distribution functions is equivalent to weak convergence of the corresponding quantile functions. Consider now the case where $\Gamma = 0$. Fix $0 < \varepsilon < 1$. Let z be a random variable taking values in $[0,\infty)$, with continuous and strictly increasing (on $[0,\infty)$) distribution function and $1-\alpha$-quantile equal to ε. Clearly, $\|Z_n\|_\infty + z$ converges weakly to z. Hence K_n, say, the $1-\alpha$ quantile of $\|Z_n\|_\infty + z$ converges to ε. From $K_{1-\alpha}(\Gamma_n) \leq K_n$ it then follows that

$$0 \quad \leq \quad \limsup_{n\to\infty} K_{1-\alpha}(\Gamma_n) \quad \leq \quad \varepsilon.$$

Therefore, $K_{1-\alpha}(\Gamma_n) \to 0 = K_{1-\alpha}(0)$. □

Lemma A.4. *For $n \in \mathbb{N}$, for every $i = 1,\ldots,n$, and for $j = 1,2$, let $a_{i,n}(j)$ and $b_{i,n}(j)$ be random variables on a probability space $(\Omega_n, \mathcal{A}_n, \mathbb{P}_n)$.*

1. *If $\mathbb{P}_n(\sum_{i=1}^n a_{i,n}^2(1) = 0) \to 0$ holds, and if $\frac{\sum_{i=1}^n (a_{i,n}(1)-b_{i,n}(1))^2}{\sum_{i=1}^n a_{i,n}^2(1)} = o_{\mathbb{P}_n}(1)$, then*

$$\mathbb{P}_n \left(\left| \frac{\sum_{i=1}^n b_{i,n}^2(1)}{\sum_{i=1}^n a_{i,n}^2(1)} - 1 \right| \geq \varepsilon \right) \to 0 \quad \textit{for every } \varepsilon > 0.$$

2. *If* $\mathbb{P}_n(\sum_{i=1}^n a_{i,n}^2(j) = 0) \to 0$ *and* $\frac{\sum_{i=1}^n (a_{i,n}(j) - b_{i,n}(j))^2}{\sum_{i=1}^n a_{i,n}^2(j)} = o_{\mathbb{P}_n}(1)$ *holds for* $j = 1, 2$*, then for every* $\varepsilon > 0$

$$\mathbb{P}_n\Bigg(\Bigg|\frac{\sum_{i=1}^n a_{i,n}(1)a_{i,n}(2)}{\sqrt{\sum_{i=1}^n a_{i,n}^2(1)}\sqrt{\sum_{i=1}^n a_{i,n}^2(2)}} - \frac{\sum_{i=1}^n b_{i,n}(1)b_{i,n}(2)}{\sqrt{\sum_{i=1}^n b_{i,n}^2(1)}\sqrt{\sum_{i=1}^n b_{i,n}^2(2)}}\Bigg| \geq \varepsilon\Bigg) \to 0.$$

3. *Suppose that* $\mathbb{V}_n(\sum_{i=1}^n a_{i,n}(1)) > 0$ *holds eventually. Assume that for every* n *the random variables* $a_{i,n}(1)$ *for* $i = 1, \dots, n$ *have mean* 0 *and are uncorrelated, and that it holds that*

$$\mathbb{P}_n\left(\left|\frac{\sum_{i=1}^n a_{i,n}^2(1)}{\mathbb{V}_n(\sum_{i=1}^n a_{i,n}(1))} - 1\right| > \varepsilon\right) \to 0 \text{ for every } \varepsilon > 0, \tag{A.8}$$

that

$$\frac{\max_{i=1,\dots,n} \mathbb{V}_n(a_{i,n}(1))}{\mathbb{V}_n(\sum_{i=1}^n a_{i,n}(1))} \to 0, \tag{A.9}$$

and that $\frac{\sum_{i=1}^n (a_{i,n}(1) - b_{i,n}(1))^2}{\sum_{i=1}^n a_{i,n}^2(1)} = o_{\mathbb{P}_n}(1)$*. Then for every array of real numbers* $c_{i,n}$ *we have*

$$\mathbb{P}_n\left(\frac{\sum_{i=1}^n (b_{i,n}(1) + c_{i,n})^2}{\mathbb{V}_n(\sum_{i=1}^n a_{i,n}(1))} \leq 1 - \varepsilon\right) \to 0 \quad \text{for every } \varepsilon > 0.$$

Proof. For the first part note that the quotient under consideration is well defined with probability converging to one, that

$$\frac{\sum_{i=1}^n b_{i,n}^2(1)}{\sum_{i=1}^n a_{i,n}^2(1)} - 1 = o_{\mathbb{P}_n}(1) + 2\frac{\sum_{i=1}^n a_{i,n}(1)[b_{i,n}(1) - a_{i,n}(1)]}{\sum_{i=1}^n a_{i,n}^2(1)},$$

and that by the Cauchy-Schwarz inequality

$$\left|\frac{\sum_{i=1}^n a_{i,n}(1)[b_{i,n}(1) - a_{i,n}(1)]}{\sum_{i=1}^n a_{i,n}^2(1)}\right| \leq \sqrt{\frac{\sum_{i=1}^n (a_{i,n}(1) - b_{i,n}(1))^2}{\sum_{i=1}^n a_{i,n}^2(1)}} = o_{\mathbb{P}_n}(1).$$

For the second part note that the quotients are well defined with probability converging to 1 (by applying Part 1), and write

$$\frac{\sum_{i=1}^n a_{i,n}(1)a_{i,n}(2)}{\sqrt{\sum_{i=1}^n a_{i,n}^2(1)}\sqrt{\sum_{i=1}^n a_{i,n}^2(2)}} - \frac{\sum_{i=1}^n b_{i,n}(1)b_{i,n}(2)}{\sqrt{\sum_{i=1}^n b_{i,n}^2(1)}\sqrt{\sum_{i=1}^n b_{i,n}^2(2)}}$$

as the sum of

$$A_n := \frac{\sum_{i=1}^n \big(a_{i,n}(1) - b_{i,n}(1)\big)\big(a_{i,n}(2) - b_{i,n}(2)\big)}{\sqrt{\sum_{i=1}^n a_{i,n}^2(1)}\sqrt{\sum_{i=1}^n a_{i,n}^2(2)}}$$

$$B_n := \frac{\sum_{i=1}^n b_{i,n}(2)\big(a_{i,n}(1) - b_{i,n}(1)\big)}{\sqrt{\sum_{i=1}^n a_{i,n}^2(1)}\sqrt{\sum_{i=1}^n a_{i,n}^2(2)}}$$

$$C_n := \frac{\sum_{i=1}^n b_{i,n}(1)\big(a_{i,n}(2) - b_{i,n}(2)\big)}{\sqrt{\sum_{i=1}^n a_{i,n}^2(1)}\sqrt{\sum_{i=1}^n a_{i,n}^2(2)}}$$

and

$$D_n := \frac{\sum_{i=1}^n b_{i,n}(1)b_{i,n}(2)}{\sqrt{\sum_{i=1}^n a_{i,n}^2(1)}\sqrt{\sum_{i=1}^n a_{i,n}^2(2)}} - \frac{\sum_{i=1}^n b_{i,n}(1)b_{i,n}(2)}{\sqrt{\sum_{i=1}^n b_{i,n}^2(1)}\sqrt{\sum_{i=1}^n b_{i,n}^2(2)}}.$$

Using Cauchy-Schwarz inequality, the assumptions, and the first part of the lemma, we now see that A_n, B_n, C_n, and D_n are $o_{\mathbb{P}_n}(1)$.

For the third part we abbreviate $a_{i,n} = a_{i,n}(1)$ and $b_{i,n} = b_{i,n}(1)$ throughout, and note that (eventually)

$$\frac{\sum_{i=1}^n (b_{i,n} + c_{i,n})^2}{\mathbb{V}_n(\sum_{i=1}^n a_{i,n})} - 1 \quad = \quad F_n + G_n + H_n,$$

where

$$\begin{aligned} F_n &:= \frac{\sum_{i=1}^n (b_{i,n} - a_{i,n})^2}{\mathbb{V}_n(\sum_{i=1}^n a_{i,n})} = o_{\mathbb{P}_n}(1) \\ G_n &:= \frac{\sum_{i=1}^n (a_{i,n} + c_{i,n})^2}{\mathbb{V}_n(\sum_{i=1}^n a_{i,n})} - 1 = \frac{\sum_{i=1}^n c_{i,n}(2a_{i,n} + c_{i,n})}{\mathbb{V}_n(\sum_{i=1}^n a_{i,n})} + o_{\mathbb{P}_n}(1) \\ H_n &:= 2\frac{\sum_{i=1}^n (b_{i,n} - a_{i,n})(a_{i,n} + c_{i,n})}{\mathbb{V}_n(\sum_{i=1}^n a_{i,n})} \end{aligned}$$

where the second equality for F_n follows from the last assumption appearing in Part 3 together with (3), and the second equality for G_n follows from (3). By the Cauchy-Schwarz inequality and the last assumption appearing in Part 3

$$|H_n| \le o_{\mathbb{P}_n}(1)\sqrt{G_n + 1}. \tag{A.10}$$

Now, define

$$\kappa_n \quad = \quad \frac{\sum_{i=1}^n c_{i,n}(2a_{i,n} + c_{i,n})}{\mathbb{V}_n(\sum_{i=1}^n a_{i,n})}$$

and note that, since by assumption $\mathbb{E}_n(a_{i,n}) = 0$, we have

$$\mathbb{E}_n(\kappa_n) = \frac{\sum_{i=1}^n c_{i,n}^2}{\mathbb{V}_n(\sum_{i=1}^n a_{i,n})} =: d_n \geq 0,$$

and it holds, using uncorrelatedness of $a_{i,n}$ for $i = 1, \ldots, n$, that

$$\mathbb{V}_n(\kappa_n) = 4\frac{\sum_{i=1}^n c_{i,n}^2 \mathbb{V}_n(a_{i,n})}{\mathbb{V}_n^2(\sum_{i=1}^n a_{i,n})} \leq 4\frac{\max_{i=1,\ldots,n} \mathbb{V}_n(a_{i,n})}{\mathbb{V}_n(\sum_{i=1}^n a_{i,n})}\ d_n. \tag{A.11}$$

We need to verify that for every $\varepsilon > 0$ it holds that

$$\mathbb{P}_n\left(F_n + G_n + H_n \leq -\varepsilon\right) \to 0.$$

We argue by contradiction: Suppose there exists an $\varepsilon > 0$ so that the convergence in the previous display does not hold. Then, by compactness of the Cartesian product of the extended real line with the unit interval, there exists a subsequence n' along which d_n converges to a $c \in [0, \infty]$ and along which the probability in the previous display converges to a $\gamma \in (0, 1]$. Suppose first that $0 \leq c < \infty$. Then, from Equation (A) and Assumption (3), we see that $\kappa_{n'}$, and hence $G_{n'}$, converges to c in $\mathbb{P}_{n'}$-probability, and, by Equation (A), that $H_{n'}$ converges to 0 in $\mathbb{P}_{n'}$-probability, showing that $F_{n'} + G_{n'} + H_{n'}$ converges to $c \geq 0$ in $\mathbb{P}_{n'}$-probability, and hence that the sequence in the previous display converges along n' to $0 < \gamma$, a contradiction. Assume next that $c = \infty$, and assume then, without loss of generality, that $d_{n'} > 0$ holds. We show that $(F_{n'} + G_{n'} + H_{n'})/d_{n'}$ converges to 1 in $\mathbb{P}_{n'}$-probability, which then again contradicts $\gamma \in (0, 1]$. Note that it suffices to verify that $\kappa_{n'}/d_{n'}$ converges to 1 in $\mathbb{P}_{n'}$-probability. But this follows, because the expectation of $\kappa_{n'}/d_{n'}$ is 1, and because, by relation (A), its variance is bounded from above by

$$4d_{n'}^{-1}\frac{\max_{i=1,\ldots,n'} \mathbb{V}_{n'}(a_{i,n'})}{\mathbb{V}_{n'}(\sum_{i=1}^{n'} a_{i,n'})} \to 0,$$

where we used $d_{n'} \to \infty$ and Assumption (3) to obtain the limit. □

Appendix B: Proofs for Section 2

B.1. *Proof of Lemma 2.1*

We actually prove the following more detailed statement.

Lemma B.1. *Under Condition 1, for $\varepsilon > 0$ we have*

$$\mathbb{P}_n\left(\|\operatorname{diag}(\mathbb{VC}_n(r_n))^{-1}\operatorname{diag}(S_n) - I_k\| \geq \varepsilon\right) \to 0,$$

$$\mathbb{P}_n\left(\|\operatorname{corr}(S_n) - \operatorname{corr}(\mathbb{VC}_n(r_n))\| \geq \varepsilon\right) \to 0 \quad \textit{and}$$

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

The last statement remains valid upon replacing S_n by $\mathbb{VC}_n(r_n)$.

Proof. Lemma A.2 applied to the array $z_{i,n} = g_{i,n} \circ \pi_{i,n}$ defined on the space $(\mathbb{R}^{n\times\ell}, \mathcal{B}(\mathbb{R}^{n\times\ell}), \mathbb{P}_n)$, where $\pi_{i,n} : \mathbb{R}^{n\times\ell} \to \mathbb{R}^{1\times\ell}$ extracts the $i - th$ row of an $n\times\ell$ matrix (to verify Condition 3 we use Condition 1 and replace the Lindeberg condition as discussed in Remark 2.2), shows that for every $\varepsilon > 0$ we have

$$\begin{aligned} &\mathbb{P}_n \left(\| \operatorname{diag}(\mathbb{VC}_n(r_n))^{-1} \operatorname{diag}(S_n) - I_k\| \geq \varepsilon\right) \to 0, \\ &\mathbb{P}_n \left(\| \operatorname{corr}(S_n) - \operatorname{corr}(\mathbb{VC}_n(r_n)) \| \geq \varepsilon\right) \to 0 \end{aligned} \tag{B.1}$$

and

$$\mathbb{P}_n \left(d_w \left(\mathbb{P}_n \circ \hat{r}_{n,*}, N(0, \operatorname{corr}(S_n))\right) \geq \varepsilon\right) \to 0,$$

where $\hat{r}_{n,*} = \operatorname{diag}(S_n)^{\dagger/2} r_n$. The last part of Condition 1 together with Equation (B.1) now shows that $\operatorname{diag}(S_n)^{\dagger/2}\Delta_n \to 0$ w.r.t. $\mathbb{P}_n$, so that

$$\hat{r}_{n,*} = \operatorname{diag}(S_n)^{\dagger/2} \left(\hat{\theta}_n - \theta_n^*\right) + o_{\mathbb{P}_n}(1),$$

which then proves the claim. □

B.2. Proof of Theorem 2.3

For any (measurable) model selection procedure $\hat{\mathbb{M}}_n$ we have

$$\begin{aligned} &\mathbb{P}_n \left(\theta^{*(j)}_{\hat{\mathbb{M}}_n,n} \in \mathrm{CI}^{(j),\mathrm{est}}_{1-\alpha,\hat{\mathbb{M}}_n} \text{ for all } j = 1, \ldots, m(\hat{\mathbb{M}}_n)\right) \\ \geq\ &\mathbb{P}_n \left(\theta^{*(j)}_{\mathbb{M},n} \in \mathrm{CI}^{(j),\mathrm{est}}_{1-\alpha,\mathbb{M}} \text{ for all } \mathbb{M} \in \mathsf{M}_n \text{ and } j \in \{1, \ldots, m(\mathbb{M})\}\right). \end{aligned}$$

It hence suffices to verify that the lower bound converges to $1 - \alpha$. And for that (cf. Equation (B.2) below, and Condition 1) it suffices to verify that the following quantity converges to $1 - \alpha$:

$$\mathbb{P}_n \left(\| \operatorname{diag}(\hat{S}_n)^{\dagger/2}(\hat{\theta}_n - \theta_n^*)\|_\infty \leq K_{1-\alpha}(\operatorname{corr}(\hat{S}_n))\right).$$

Lemma B.1 shows that for every $\varepsilon > 0$ we have

$$\begin{aligned} &\mathbb{P}_n \left(\| \operatorname{diag}(\mathbb{VC}_n(r_n))^{-1} \operatorname{diag}(S_n) - I_k\| \geq \varepsilon\right) \to 0 \\ &\mathbb{P}_n \left(\| \operatorname{corr}(S_n) - \operatorname{corr}(\mathbb{VC}_n(r_n)) \| \geq \varepsilon\right) \to 0 \\ &\mathbb{P}_n \left(d_w \left(\mathbb{P}_n \circ \left[\operatorname{diag}(S_n)^{\dagger/2}(\hat{\theta}_n - \theta_n^*)\right], N(0, \operatorname{corr}(S_n))\right) \geq \varepsilon\right) \to 0. \end{aligned}$$

This also shows that the two conditions imposed on $\hat{S}_n$ in the statement of the theorem are indeed equivalent. Furthermore, we immediately see that from any of these two assumptions, together with the previous display, it follows that for every $\varepsilon > 0$ we have

$$\mathbb{P}_n \left(\| \operatorname{diag}(\mathbb{VC}_n(r_n))^{-1} \operatorname{diag}(\hat{S}_n) - I_k\| \geq \varepsilon\right) \to 0 \tag{B.2}$$

$$\mathbb{P}_n \left(\| \operatorname{corr}\left(\hat{S}_n\right) - \operatorname{corr}(\mathbb{VC}_n(r_n)) \| \geq \varepsilon\right) \to 0 \tag{B.3}$$

$$\mathbb{P}_n \left(d_w \left(\mathbb{P}_n \circ \bar{r}_{n,*}, N(0, \operatorname{corr}(\hat{S}_n))\right) \geq \varepsilon\right) \to 0,$$

where $\bar{r}_{n,*} = \operatorname{diag}(\hat{S}_n)^{\dagger/2}(\hat{\theta}_n - \theta_n^*)$. Next, let n' be an arbitrary subsequence of n, and let n'' be a subsequence of n' along which the norm-bounded sequence $\operatorname{corr}(\mathbb{VC}_n(r_n))$ converges to $\bar{\Sigma}$, say. By (B.2) it holds that $\operatorname{corr}(\hat{S}_{n''})$ converges to $\bar{\Sigma}$ in $\mathbb{P}_{n''}$-probability, and from the previous display it follows that $\mathbb{P}_{n''} \circ \bar{r}_{n'',*} \Rightarrow N(0, \bar{\Sigma})$. Combining these two statements, it then follows that

$$\mathbb{P}_{n''} \circ (\bar{r}_{n'',*}, \operatorname{corr}(\hat{S}_{n''})) \Rightarrow Q_{\bar{\Sigma}} \otimes \delta_{\bar{\Sigma}},$$

where $Q_{\bar{\Sigma}} := N(0, \bar{\Sigma})$, and where $\delta_{\bar{\Sigma}}$ denotes point mass at $\bar{\Sigma} \in \mathbb{R}^{k \times k}$. Now, define the map $F : \mathbb{R}^k \times \mathbb{R}^{k \times k}_{s, \geq 0} \to \mathbb{R}$ via $(z, \Sigma) \mapsto \|z\|_\infty - K_{1-\alpha}(\Sigma)$, where $\mathbb{R}^{k \times k}_{s, \geq 0}$ denotes the set of real, symmetric and nonnegative definite $k \times k$ dimensional matrices, and note that the map F is continuous everywhere (cf. Lemma A.3). It follows from the continuous mapping theorem together with the previous display that

$$\mathbb{P}_{n''} \circ F(\bar{r}_{n'',*}, \operatorname{corr}(\hat{S}_{n''})) \Rightarrow Q_{\bar{\Sigma}} \circ \left(\|.\|_\infty - K_{1-\alpha}(\bar{\Sigma})\right).$$

Since the diagonal elements of $\bar{\Sigma}$ are ones (by its definition together with Condition 1), one can easily show that the $Q_{\bar{\Sigma}}$-probability of $\|.\|_\infty - K_{1-\alpha}(\bar{\Sigma})$ being equal to 0 is 0. It hence follows from the Portmanteau theorem, together with the definition of $K_{1-\alpha}(\bar{\Sigma})$ and the previous display, that

$$\begin{aligned}
&\mathbb{P}_{n''}\left(\|\operatorname{diag}(\hat{S}_{n''})^{\dagger/2}(\hat{\theta}_{n''} - \theta^*_{n''})\|_\infty \leq K_{1-\alpha}(\operatorname{corr}(\hat{S}_{n''}))\right) \\
&= \mathbb{P}_{n''} \circ F(\bar{r}_{n'',*}, \operatorname{corr}(\hat{S}_{n''}))\,((-\infty, 0]) \\
&\to Q_{\bar{\Sigma}}\left(y \in \mathbb{R}^k : \|y\|_\infty \leq K_{1-\alpha}(\bar{\Sigma})\right) = 1 - \alpha.
\end{aligned}$$

This finishes the proof. □

B.3. Proof of Proposition 2.4

As in the proof of Lemma B.1, Lemma A.2 applied to the array $z_{i,n} := g_{i,n} \circ \pi_{i,n}$ defined on $(\mathbb{R}^{n \times \ell}, \mathcal{B}(\mathbb{R}^{n \times \ell}), \mathbb{P}_n)$ shows (in particular) that for every $\varepsilon > 0$

$$\mathbb{P}_n\left(\|\operatorname{diag}(\mathbb{VC}_n(r_n))^{-1} \operatorname{diag}(S_n) - I_k\| \geq \varepsilon\right) \to 0.$$

This shows that the two conditions given in the statement of the proposition are indeed equivalent, and, together with Condition 1, it also shows that for every $j = 1, \ldots, k$ we have $\mathbb{P}_n(\sum_{i=1}^n [z_{i,n}^{(j)}]^2 = 0) \to 0$. Now, for $j = 1, \ldots, k$, we apply the first part of Lemma A.4 (with $a_{i,n}(1) = g_{i,n}^{(j)} \circ \pi_{i,n}$ and $b_{i,n} = \hat{g}_{i,n}^{(j)} \circ \pi_{i,n}$) to obtain for every $\varepsilon > 0$ that

$$\mathbb{P}_n\left(\|\operatorname{diag}(S_n)^\dagger \operatorname{diag}(\hat{S}_n) - I_k\| \geq \varepsilon\right) \to 0.$$

Next, we can, in a similar way, apply the second part of Lemma A.4 to obtain

$$\mathbb{P}_n\left(\|\operatorname{corr}\left(\hat{S}_n\right) - \operatorname{corr}\left(S_n\right)\| \geq \varepsilon\right) \to 0.$$

This finishes the proof. □

B.4. Proof of Theorem 2.5

Similarly as in the proof of Theorem 2.3 we now need to verify that

$$\liminf_{n\to\infty} \mathbb{P}_n \left(\| \operatorname{diag}(\hat{\nu}_n^2)^{\dagger/2} \left[r_n + \Delta_n \right] \|_\infty \leq \hat{K}_n \right) \geq 1 - \alpha. \tag{B.4}$$

We make the following preparatory observation: Denote the event on which κ_n is well defined by A_n (recall that $\mathbb{P}_n(A_n) \to 1$), and let $\varepsilon > 0$. Observe that the limit inferior in Equation (B.4) is not smaller than

$$\liminf_{n\to\infty} \mathbb{P}_n \left(\kappa_n \| \operatorname{diag}(\mathbb{VC}_n(r_n))^{-1/2} \left[r_n + \Delta_n \right] \|_\infty \leq K_{1-\alpha}(\operatorname{corr}(\mathbb{VC}_n(r_n))), A_n \right),$$

which, in turn, is bounded from below (using that κ_n is positive on A_n, and Equation (2.5)) by

$$\liminf_{n\to\infty} \mathbb{P}_n \left(\| \operatorname{diag}(\mathbb{VC}_n(r_n))^{-1/2} \left[r_n + \Delta_n \right] \|_\infty \leq \frac{K_{1-\alpha}(\operatorname{corr}(\mathbb{VC}_n(r_n)))}{1+\varepsilon} \right). \tag{B.5}$$

Now, we argue by contradiction, and suppose that (B.4) is false: Then there exists a $\delta > 0$, that can be chosen independently of ε, so that the limit inferior in (B.4) is an element of $[0, 1-\alpha-\delta)$. Next, let $n'(\varepsilon)$ denote a subsequence along which (B.4) is attained. Arguing as in the proof of Theorem 2.3 (borrowing some of its notation) we can obtain a subsequence $n''(\varepsilon)$ of $n'(\varepsilon)$ along which the sequence of probabilities in the preceding display converges to

$$Q_{\bar{\Sigma}(\varepsilon)} \left(y \in \mathbb{R}^k : \|y\|_\infty \leq \frac{K_{1-\alpha}(\bar{\Sigma}(\varepsilon))}{1+\varepsilon} \right) \in [0, 1-\alpha-\delta).$$

Note that $\varepsilon > 0$ was arbitrary, and let $\varepsilon_m > 0$ converge to 0. Assume (otherwise pass to a subsequence) that the sequence of correlation matrices $\bar{\Sigma}(\varepsilon_m)$ (with diagonal entries equal to 1) converges to $\bar{\Sigma}$, say. It is then not difficult to obtain (by a weak convergence argument involving Portmanteau theorem) the contradiction

$$\begin{aligned} &Q_{\bar{\Sigma}(\varepsilon_m)} \left(y \in \mathbb{R}^k : \|y\|_\infty \leq \frac{K_{1-\alpha}(\bar{\Sigma}(\varepsilon_m))}{1+\varepsilon_m} \right) \\ \xrightarrow[m\to\infty]{} \quad &Q_{\bar{\Sigma}} \left(y \in \mathbb{R}^k : \|y\|_\infty \leq K_{1-\alpha}(\bar{\Sigma}) \right) = 1-\alpha. \end{aligned}$$

The remaining part follows immediately from what we have already established. □

B.5. Proof of Proposition 2.6

Fix j and note that with the same notation and argumentation as in the beginning of the proof of Proposition 2.4, for every $\varepsilon > 0$, it holds that

$$\mathbb{P}_n \left(\left| \frac{\sum_{i=1}^n [z_{i,n}^{(j)}]^2}{\sum_{i=1}^n \mathbb{V}_n(z_{i,n}^{(j)})} - 1 \right| \geq \varepsilon \right) \to 0. \tag{B.6}$$

Obviously, setting $\tilde{z}_{i,n} = \tilde{g}_{i,n} \circ \pi_{i,n}$ we obtain

$$\tilde{z}_{i,n}^{(j)} = \left(\hat{z}_{i,n}^{(j)} - z_{i,n}^{(j)}\right) + z_{i,n}^{(j)} + a_{i,n}^{(j)}.$$

Equation (2.4), or equivalently (equivalence being due to Equation (B.5) above) Equation (2.4), together with Part 3 of Lemma A.4 (applied with: $a_{i,n}(1) = z_{i,n}^{(j)}$, $b_{i,n}(1) = \left(\hat{z}_{i,n}^{(j)} - z_{i,n}^{(j)}\right) + z_{i,n}^{(j)}$ and $c_{i,n} = a_{i,n}^{(j)}$) now shows that

$$\mathbb{P}_n\left(\frac{\sum_{i=1}^n [\tilde{z}_{i,n}^{(j)}]^2}{\sum_{i=1}^n \mathbb{V}_n(z_{i,n}^{(j)})} \leq 1 - \varepsilon\right) \to 0 \text{ for every } \varepsilon > 0,$$

implying the claimed statement. Note that Equation (3) in Lemma A.4 is satisfied here because Condition 1 (in particular the Lindeberg condition in Equation (1)) implies the corresponding Feller condition

$$\frac{\max_{i=1,\ldots,n} \mathbb{V}_n(z_{i,n}^{(j)})}{\sum_{i=1}^n \mathbb{V}_n(z_{i,n}^{(j)})} \to 0.$$

All remaining assumptions in Part 3 of Lemma A.4 can be easily checked using Condition 1, (2.4) and (B.5). □

B.6. Proof of Lemma 2.8

Let $\omega := \operatorname{rank}(\Gamma)$ and let $Z \sim N(0, \Gamma)$. Since Γ is a correlation matrix of rank ω, by the spectral decomposition, we can find a $k \times \omega$-dimensional matrix V so that $VV' = \Gamma$. In particular if $\varepsilon \sim N(0, I_\omega)$ it holds that $V\varepsilon \sim N(0, \Gamma)$, and hence the $1-\alpha$-quantiles of the distributions of $\|Z\|_\infty$ and of $\|V\varepsilon\|_\infty = \max_{i=1,\ldots,k} |v_i \varepsilon|$ coincide, v_i denoting the i-th row of V. Since Γ is a correlation matrix it furthermore holds that each row v_i of V has Euclidean norm less than or equal to 1. From the discussion after the definition of B_α it then follows that $K_{1-\alpha}(\Gamma)$, the $1-\alpha$-quantile of the distributions of $\|Z\|_\infty$, is not greater than $B_\alpha(\omega, k)$. □

Appendix C: Proofs for Section 3

C.1. Proof of Proposition 3.3

Fix $\delta > 0$, $\tau \geq 1$, $n \in \mathbb{N}$, $\mathbb{P}_n \in \mathbf{P}_n^{(\mathrm{lm})}(\delta, \tau)$ and $\mathbb{M} \in \mathsf{M}_n$ with corresponding index set $M \in \mathcal{I}$. Abbreviate $\mu_n = \mu(\mathbb{P}_n)$, $\sigma_n^2 = m_2(\mathbb{P}_{1,n})$, $m = |M|$, $u_n = u_n(y) = y - \mu_n$ and $H_M = I_n - P_{X_n[M]}$. The mean of $\hat{\sigma}_{\mathbb{M},n}^2(y) = (n-m)^{-1} y' H_M y$ is easily seen to be $\mathbb{E}_n[\hat{\sigma}^2_{\mathbb{M},n}] = \mu_n' H_M \mu_n/(n-m) + \sigma_n^2$, eventually. Now consider

$$\begin{aligned}
\left|\frac{\hat{\sigma}^2_{\mathbb{M},n}}{\mathbb{E}_n[\hat{\sigma}^2_{\mathbb{M},n}]} - 1\right| &= \left|\frac{2\mu_n H_M u_n/(n-m) + u_n' H_M u_n/(n-m) - \sigma_n^2}{\mu_n' H_M \mu_n/(n-m) + \sigma_n^2}\right| \\
&= \left|\frac{2(\mu_n/\sigma_n) H_M (u_n/\sigma_n)/(n-m) + (u_n/\sigma_n)' H_M (u_n/\sigma_n)/(n-m) - 1}{(\mu_n/\sigma_n)' H_M (\mu_n/\sigma_n)/(n-m) + 1}\right| \\
&\leq 2\left|\frac{(\mu_n/\sigma_n) H_M (u_n/\sigma_n)}{(\mu_n/\sigma_n)' H_M (\mu_n/\sigma_n) + (n-m)}\right| + \left|\frac{(u_n/\sigma_n)' H_M (u_n/\sigma_n)}{n-m} - 1\right|.
\end{aligned}$$

The first fraction on the last line of the previous display converges to zero in $\mathbb{P}_n$-probability, because its mean is 0 and its variance is upper bounded by $\|H_M \mu_n/\sigma_n\|^2/(\|H_M\mu_n/\sigma_n\|^4 + (n-m)^2)$, which converges to 0, as is seen by maximizing it with respect to $\|H_M\mu_n/\sigma_n\|^2$. To show that the second fraction converges to one, abbreviate the random n-vector $v_n = u_n/\sigma_n$, and note that v_n has independent standardized components under $\mathbb{P}_n$. Now decompose the quadratic form as

$$\frac{v_n' H_M v_n}{n-m} = \frac{1}{n-m}\sum_{i=1}^n (H_M)_{ii} v_{i,n}^2 + \frac{1}{n-m}\sum_{i\neq j}(H_M)_{ij} v_{i,n} v_{j,n},$$

and note that $(n-m)^{-1}\sum_{i\neq j}(H_M)_{ij} v_{i,n} v_{j,n}$ has mean zero and variance equal to $(n-m)^{-2}\sum_{i\neq j}[(H_M)_{ij}]^2 \leq (n-m)^{-2}\operatorname{trace}(H_M^2) = (n-m)^{-1} \to 0$. To show that $(n-m)^{-1}\sum_{i=1}^n (H_M)_{ii} v_{i,n}^2$ converges to one, we use a standard truncation argument. For $K > 0$ define $\tilde{v}_{i,n} = v_{i,n}\{|v_{i,n}| \leq K\}$, $S_n = \frac{1}{n-m}\sum_{i=1}^n (H_M)_{ii} v_{i,n}^2$, $\tilde{S}_n = \frac{1}{n-m}\sum_{i=1}^n (H_M)_{ii}\tilde{v}_{i,n}^2$ and

$$D_n := S_n - \tilde{S}_n = \frac{1}{n-m}\sum_{i=1}^n (H_M)_{ii} v_{i,n}^2\{|v_{i,n}| > K\} \ \geq \ 0.$$

Using Hölder's inequality and Markov's inequality, the mean of D_n can be bounded by

$$\begin{aligned}
\mathbb{E}_n[D_n] &\leq \frac{1}{n-m}\sum_{i=1}^n (H_M)_{ii}\left(\mathbb{E}_n[|v_{i,n}|^{2+\delta}]\right)^{2/(2+\delta)} \mathbb{P}_n(|v_{i,n}| > K)^{\delta/(2+\delta)} \\
&\leq \tau K^{-\frac{2\delta}{2+\delta}}.
\end{aligned}$$

Now for $\varepsilon > 0$,

$$\begin{aligned}
&\mathbb{P}_n(|S_n - 1| > \varepsilon) \\
&\quad \leq \mathbb{P}_n(|S_n - \tilde{S}_n| > \varepsilon/2) + \mathbb{P}_n(|\tilde{S}_n - \mathbb{E}_n[\tilde{S}_n] + \mathbb{E}_n[\tilde{S}_n] - 1| > \varepsilon/2) \\
&\quad \leq 2\tau K^{-\frac{2\delta}{2+\delta}}/\varepsilon + 16\mathbb{V}_n[\tilde{S}_n]/\varepsilon^2 + \mathbb{P}_n(\tau K^{-\frac{2\delta}{2+\delta}} > \varepsilon/4).
\end{aligned}$$

Since the variance of $\tilde{S}_n$ clearly converges to zero as $n \to \infty$, for every $K > 0$, the limit superior of $\mathbb{P}_n(|S_n - 1| > \varepsilon)$ is bounded by a quantity that approaches zero as $K \to \infty$. Thus, we have established the convergence $\hat{\sigma}^2_{\mathbb{M},n}/\mathbb{E}_n[\hat{\sigma}^2_{M,n}] \to 1$, in $\mathbb{P}_n$-probability. Therefore, we can write

$$\frac{\hat{\sigma}^2_{\mathbb{M},n}}{\sigma_n^2} = \frac{\hat{\sigma}^2_{\mathbb{M},n}}{\mathbb{E}_n[\hat{\sigma}^2_{\mathbb{M},n}]} \frac{\mathbb{E}_n[\hat{\sigma}^2_{\mathbb{M},n}]}{\sigma_n^2} = (1 + o_{\mathbb{P}_n}(1)) \left(\frac{(\mu_n/\sigma_n)' H_M (\mu_n/\sigma_n)}{n-m} + 1 \right),$$

which finishes the proof. □

C.2. Proof of Theorem 3.2

Let $\mathbb{P}_n \in \mathbf{P}_n^{(\mathrm{lm})}(\delta, \tau)$ for every $n \in \mathbb{N}$. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space that carries a triangular array of random variables $y_{1,n}, \ldots, y_{n,n}$ for $n \in \mathbb{N}$ so that $y_{1,n}, \ldots, y_{n,n}$ are independent for every $n \in \mathbb{N}$, and so that the distribution of $y_n = (y_{1,n}, \ldots, y_{n,n})'$ on the Borel sets of $\mathbb{R}^n$ coincides with $\mathbb{P}_n$. The expectation, variance and variance-covariance matrix operators w.r.t. $\mathbb{P}$ are denoted by $\mathbb{E}$, $\mathbb{V}$ and $\mathbb{VC}$, respectively. The expectation, variance and variance-covariance matrix operators w.r.t. $\mathbb{P}_n$ are denoted by $\mathbb{E}_n$ and $\mathbb{V}_n$ and $\mathbb{VC}_n$, respectively, and the expectation operator w.r.t. $\mathbb{P}_{i,n}$ is denoted by $\mathbb{E}_{i,n}$. Now, from $\mathbb{P}_n \in \mathbf{P}_n^{(\mathrm{lm})}(\delta, \tau)$ we conclude that: $\mathbb{E}(y_{i,n}) =: \mu_{i,n}$ exists, furthermore we let $\mu_n = (\mu_{1,n}, \ldots, \mu_{n,n})'$; $\mathbb{V}(y_{i,n})$ exists, is positive, does not depend on i, and will be denoted by σ_n^2; we also let $u_{i,n}(y_{i,n}) := y_{i,n} - \mu_{i,n}$, set $u_n(y_n) = (u_{1,n}(y_{1,n}), \ldots, u_{n,n}(y_{n,n}))'$, and note that

$$\frac{\max_{i=1,\ldots,n} \mathbb{E}_{i,n}(|u_{i,n}|^{2+\delta})^{\frac{2}{2+\delta}}}{\sigma_n^2} \leq \tau. \tag{C.1}$$

We now verify Condition 1 (with $\Delta_n \equiv 0$): Let M_l be the index set corresponding to model $\mathbb{M}_{l,n} \in \mathsf{M}_n$, $l = 1, \ldots, d$. Define the stacked vector of estimators $\hat{\theta}_n = (\hat{\beta}'_{\mathbb{M}_1,n}, \ldots, \hat{\beta}'_{\mathbb{M}_d,n})'$ and define the corresponding stacked vector of targets $\theta_n^* = (\beta^{*'}_{\mathbb{M}_1,n}, \ldots, \beta^{*'}_{\mathbb{M}_d,n})'$, where we recall that for every $l = 1, \ldots, d$

$$\beta^*_{\mathbb{M}_l,n} = (X_n[M_l]' X_n[M_l])^{-1} X_n[M_l]' \mu_n = U(l,n)^{-1} X_n[M_l]' \mu_n,$$

for $U(l,n) := (X_n[M_l]' X_n[M_l])$. Let $j \in \{1, \ldots, k\}$ be arbitrary. The j-th coordinate of the vector

$$r_n(y_n) \ := \ \hat{\theta}_n(y_n) - \theta_n^*$$

can be written as

$$\sum_{i=1}^n e'_{|M_l|}(h) U(l,n)^{-1} X_{i,n}[M_l]' u_{i,n} =: \sum_{i=1}^n w_{l,h,i,n} u_{i,n}.$$

for an obvious choice of $l = l(j) \in \{1, \ldots, d\}$ and $h = h(j) \in \{1, \ldots, |M_{l(j)}|\}$, where $e_m(h)$ denotes the h-th element of the canonical basis of $\mathbb{R}^m$. Next, define the function $g_{i,n} : \mathbb{R}^{1\times 1} \to \mathbb{R}^k$ via

$$g_{i,n}^{(j)}(y) = w_{l(j),h(j),i,n}(y - \mu_{i,n}), \quad j = 1, \ldots, k.$$

By definition $\mathbb{E}_{i,n}\left(g_{i,n}^{(j)}\right) = 0$ and $0 < \mathbb{V}_n\left(r_n^{(j)}\right) < \infty$ holds (eventually), the latter following from

$$\mathbb{V}_n\left(r_n^{(j)}\right) = \sigma_n^2 e'_{|M_{l(j)}|}(h(j))U(l(j),n)^{-1}e_{|M_{l(j)}|}(h(j)) \tag{C.2}$$

together with Condition X1. We now verify that for every $\varepsilon > 0$ it holds that

$$\mathbb{V}_n^{-1}\left(r_n^{(j)}\right)\sum_{i=1}^{n}\int_{\mathbb{R}}\left[g_{i,n}^{(j)}\right]^2\left\{|g_{i,n}^{(j)}| \geq \varepsilon\mathbb{V}_n^{\frac{1}{2}}\left(r_n^{(j)}\right)\right\}d\mathbb{P}_{i,n} \to 0. \tag{C.3}$$

An application of Hölder's inequality (with $p = \frac{2+\delta}{2}$ and $q = \frac{2+\delta}{\delta}$) shows that the quantity to the left in the previous display is bounded from above by

$$\frac{\max_{i=1,\dots,n}\mathbb{E}_{i,n}(|u_{i,n}|^{2+\delta})^{\frac{2}{2+\delta}}}{\sigma_n^2}\left(\max_{i=1,\dots,n}\mathbb{P}_{i,n}\left(|g_{i,n}^{(j)}| \geq \varepsilon\mathbb{V}_n^{\frac{1}{2}}\left(r_n^{(j)}\right)\right)\right)^{\frac{\delta}{2+\delta}},$$

which, using the bound (C.2) above and Markov's inequality, does not exceed

$$\tau\left(\frac{\max_{i=1,\dots,n} w_{l(j),h(j),i,n}^2}{\sum_{i=1}^{n} w_{l(j),h(j),i,n}^2}\right)^{\frac{\delta}{2+\delta}}\varepsilon^{-\frac{2\delta}{2+\delta}}. \tag{C.4}$$

Finally, since the term within brackets coincides with

$$\max_{i=1,\dots,n}\left(X_{i,n}[M_{l(j)}]U(l(j),n)^{-\frac{1}{2}}\frac{W_{j,n}}{\|W_{j,n}\|}U(l(j),n)^{-\frac{1}{2}}X_{i,n}[M_{l(j)}]'\right),$$

for

$$W_{j,n} = U(l(j),n)^{-\frac{1}{2}}e_{|M_{l(j)}|}(h(j))e'_{|M_{l(j)}|}(h(j))U(l(j),n)^{-\frac{1}{2}},$$

and since this quantity is not greater than

$$\max_{i=1,\dots,n} X_{i,n}[M_{l(j)}]U(l(j),n)^{-1}X_{i,n}[M_{l(j)}]' \to 0,$$

where convergence holds by Condition X1, the statement in (C.2) follows. Since j was arbitrary, we have verified Condition 1.

Now, for $j \in \{1,\dots,k\}$, set $\hat{\nu}_{j,n}^2 = \hat{\sigma}_{\mathbb{M}_{l(j)},n}^2[(X_n[M_{l(j)}]'X_n[M_{l(j)}])^{-1}]_{h(j)} = \hat{\sigma}_{\mathbb{M}_{l(j)},n}^2 e'_{|M_{l(j)}|}(h(j))U(l(j),n)^{-1}e_{|M_{l(j)}|}(h(j))$. Thus, using (C.2) and Proposition 3.3, we see that

$$\mathbb{P}_n\left(\sqrt{\frac{[\mathbb{VC}_n(r_n)]_j}{\hat{\nu}_{j,n}^2}} \geq 1+\varepsilon\right) = \mathbb{P}_n\left(\sqrt{\frac{\sigma_n^2}{\hat{\sigma}_{\mathbb{M}_{l(j)},n}^2}} \geq 1+\varepsilon\right) \to 0.$$

Finally, note that

$$\mathbb{VC}_n(r_n) = \sigma_n^2\Gamma_n,$$

and that $K_{1-\alpha}(\mathrm{corr}(\Gamma_n)) = K_{1-\alpha}(\mathrm{corr}(\mathbb{VC}_n(r_n))) > 0$. It now follows from Theorem 2.5 (the special case with $\hat{K}_n = K_{1-\alpha}(\mathrm{corr}(\Gamma_n))$, that for any (measurable) model selection procedure $\hat{\mathbb{M}}_n$ it holds that

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

The theorem now follows because the selection $\mathbb{P}_n$ we started with was arbitrary. $\square$

C.3. Proof of Lemma 3.4

We argue by contradiction, and assume existence of a sequence of measurable functions $(\hat{\sigma}^2_n)_{n\in\mathbb{N}}$ with $\hat{\sigma}^2_n : \mathbb{R}^n \to [0, \infty)$ and so that for every $\varepsilon > 0$ Equation (3.4) holds. First, we define for every $x \in \mathbb{R}^n$ and every $\rho > 0$ the measure

$$Q_n(x, \rho) = \bigotimes_{i=1}^{n} \Big(Q(x_i) * N(0, \rho) \Big),$$

where $Q(x_i)$ puts mass $1/2$ to $x_i + 1$ and to $x_i - 1$, respectively, and $Q(x_i) * N(0, \rho)$ denotes the convolution of $Q(x_i)$ and $N(0, \rho)$, where we interpret $N(0, 0)$ as point mass at 0, i.e., $Q(x_i) * N(0, 0) = Q(x_i)$. We note that $Q_n(x, 0) \in \mathbf{P}^{(\mathrm{lm})}_n(\delta, \tau)$, because $m_{2+h}(Q(x_i))^{\frac{2}{2+h}} = 1$ for every $h \geq 0$, and since $\tau > 1$ holds by assumption. It is easy to verify that for every $h \geq 0$ we have $m_{2+h}(Q(x_i) * N(0, \rho))^{\frac{2}{2+h}} \to 1$ as $\rho \to 0$. Hence there exists a $\rho^* > 0$ so that for every $0 \leq \rho \leq \rho^*$ and every $x \in \mathbb{R}^n$ we have $Q_n(x, \rho) \in \mathbf{P}^{(\mathrm{lm})}_n(\delta, \tau)$. Therefore, by our assumption, it holds for every $\varepsilon > 0$ that

$$\sup_{x\in\mathbb{R}^n} \sup_{0\leq\rho\leq\rho^*} Q_n(x, \rho) \left(\left| \frac{\hat{\sigma}^2_n}{1 + \rho} - 1 \right| > \varepsilon \right) \to 0. \tag{C.5}$$

Next, let $(F, \mathcal{F}, \mathbb{Q})$ be a probability space on which, for every n, there are defined two independent random n-vectors $X^{(1)}_n$ and $X^{(2)}_n$, so that $\mathbb{Q} \circ X^{(1)}_n \sim Q_n(0, 0)$ and $\mathbb{Q} \circ X^{(2)}_n \sim N(0, \rho^* I_n)$, and hence the distribution of $Y_n := X^{(1)}_n + X^{(2)}_n$ is $Q_n(0, \rho^*)$. Let $\varepsilon > 0$ be fixed. From the previous display it follows that $\mathbb{Q}(|\hat{\sigma}^2_n(Y_n) - (1 + \rho^*)| > \varepsilon)$ converges to 0. Since the conditoinal distribution of Y_n given $X^{(2)}_n$ is $Q_n(X^{(2)}_n, 0)$, it furthermore holds that

$$\mathbb{Q}(|\hat{\sigma}^2_n(Y_n) - (1 + \rho^*)| > \varepsilon) = \mathbb{E}_{\mathbb{Q}} \left(Q_n(X^{(2)}_n, 0)(|\hat{\sigma}^2_n - (1 + \rho^*)| > \varepsilon) \right) \to 0,$$

from which it now follows that $I_n(\omega) = Q_n(X^{(2)}_n(\omega), 0)(|\hat{\sigma}^2_n - (1 + \rho^*)| > \varepsilon) \to 0$ in $\mathbb{Q}$-probability as $n \to \infty$. Thus, there exists a subsequence n' so that $I_{n'} \to 0$, $\mathbb{Q}$-almost surely. As a consequence, there exists $\bar{\omega} \in \Omega$ for which $I_{n'}(\bar{\omega}) \to 0$ as $n' \to \infty$. But this now means that for $x_{n'} = X^{(2)}_{n'}(\bar{\omega})$ it holds that $Q_{n'}(x_{n'}, 0)(|\hat{\sigma}^2_{n'} - (1 + \rho^*)| > \varepsilon) \to 0$. Since $\rho^* > 0$, this contradicts Equation (C.3) which implies $Q_{n'}(x_{n'}, 0)(|\hat{\sigma}^2_{n'} - 1| > \varepsilon) \to 0$. $\square$

C.4. *Proof of Theorem 3.5*

We proceed as in the proof of Theorem 3.2, noting that now $\mathbb{V}(y_{i,n})$ exists and is positive, but might depend on i and is hence denoted by $\sigma^2_{i,n}$. The bound (C.2) in the proof of Theorem 3.2 is now replaced by

$$\frac{\max_{i=1,\ldots,n} \mathbb{E}_{i,n}(|u_{i,n}|^{2+\delta})^{\frac{2}{2+\delta}}}{\min_{i=1,\ldots,n} \sigma^2_{i,n}} \leq \tau. \tag{C.6}$$

To verify Condition 1 we replace (C.2) in the proof of Theorem 3.2 by

$$\begin{aligned}\mathbb{V}_n\left(r_n^{(j)}\right) = e'_{|M_{l(j)}|}(h(j))U(l(j),n)^{-1} &\sum_{i=1}^{n} \sigma^2_{i,n} X_{i,n}[M_{l(j)}]' X_{i,n}[M_{l(j)}] \\ \times \qquad & U(l(j),n)^{-1} e_{|M_{l(j)}|}(h(j))\end{aligned}$$

which, replacing each $\sigma^2_{i,n}$ by $\min_{i=1,\ldots,n}\sigma^2_{i,n} > 0$, is seen to be eventually positive by Condition X1. For the verification of the Lindeberg condition (C.2) we use essentially the same argument as in the proof of Theorem 3.2, now using (C.4) above. Hence, Condition 1 holds and we then obtain from Lemma B.1,

$$\mathbb{P}_n\left(\| \operatorname{diag}(\mathbb{VC}_n(r_n))^{-1} \operatorname{diag}(S_n) - I_k\| \geq \varepsilon\right) \to 0 \text{ for every } \varepsilon > 0,$$

where

$$[S_n]_{s,t} = \sum_{i=1}^{n} w_{l(s),h(s),i,n} w_{l(t),h(t),i,n} u^2_{i,n}.$$

Now, for $l \in \{1,\ldots,d\}$ we define $\hat{u}_{i,l,n} = y_{i,n} - X_{i,n}[M_l]\hat{\beta}_{\mathbb{M}_l,n}(y_n)$, and note that for every $h \in \{1,\ldots,|M_l|\}$ we have

$$\begin{aligned}&\frac{\sum_{i=1}^{n} w^2_{l,h,i,n}\left(X_{i,n}[M_l]U^{-1}(l,n)X_n[M_l]'u_n\right)^2}{\mathbb{V}_n\left(\sum_{i=1}^{n} w_{l,h,i,n}u_{i,n}\right)} \\ \leq \quad & \frac{\max_{i=1,\ldots,n} w^2_{l,h,i,n}}{\sum_{i=1}^{n} w^2_{l,h,i,n}} \times \frac{u_n' X_n[M_l]U^{-1}(l,n)X_n[M_l]'u_n}{\min_{i=1,\ldots,n}\sigma^2_{i,n}},\end{aligned} \tag{C.7}$$

that the first ratio in the upper bound converges to 0 by an argument as in the proof of Theorem 3.2 based on Condition X1 (cf. (C.2)), and that the second ratio is non-negative, with (using Equation (C.4)) an expectation bounded from above by

$$\frac{|M_l| \max_{i=1,\ldots,n}\sigma^2_{i,n}}{\min_{i=1,\ldots,n}\sigma^2_{i,n}} \leq |M_l|\tau.$$

Therefore, the quantity in the first line of Display (C.4) converges to 0 in $\mathbb{P}_n$-probability, and we can apply Part 3 of Lemma A.4 (with $a_{i,n}(1) = w_{l,h,i,n}u_{i,n}$, $b_{i,n}(1) = w_{l,h,i,n}\left(u_{i,n} - X_{i,n}[M_l]U^{-1}(l,n)X_n[M_l]'u_n\right)$, and $c_{i,n} = w_{l,h,i,n}(\mu_{i,n} -$

$X_{i,n}[M_l]\beta^*_{\mathbb{M}_l,n})))$ to obtain that for every $l \in \{1,\dots,d\}$, every $h \in \{1,\dots,|M_l|\}$, and every $\varepsilon > 0$ we have

$$\mathbb{P}_n\left(\sqrt{\frac{\mathbb{V}_n(\sum_{i=1}^n w_{l,h,i,n}u_{i,n})}{\sum_{i=1}^n w^2_{l,h,i,n}\hat{u}^2_{i,l,n}}} \geq 1+\varepsilon\right) \to 0.$$

Note that $\mathbb{V}_n(\sum_{i=1}^n w_{l(j),h(j),i,n}u_{i,n}) = [\mathbb{VC}_n(r_n)]_j$ and $\sum_{i=1}^n w^2_{l,h,i,n}\hat{u}^2_{i,l,n} = \hat{\sigma}^2_{h,\mathbb{M}_l,n}$. By Lemma 2.8 we have that $B_\alpha(\min(k,p),k) \geq K_{1-\alpha}(\mathrm{corr}(\mathbb{VC}_n(r_n))) > 0$ (it is easy to see that the rank of the $k\times k$-dimensional matrix $\mathbb{VC}_n(r_n)$ can not exceed p), so we can now apply the special case discussed in Theorem 2.5 to conclude that

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

which proves the claim as the family $\mathbb{P}_n \in \mathbf{P}^{(\mathrm{het})}_n(\delta,\tau)$ was arbitrary. □

C.5. Proof of Lemma 3.8

Fix $\mathbb{M} \in \mathsf{M}_n$ and $\mathbb{P} \in \bigcup_{\delta>0}\mathbf{P}^{(\mathrm{bin})}_n(\delta)$ and recall that $\mathbb{M} \triangleq (h,M) \in \mathcal{H}\times\mathcal{I}$. For $i=1,\dots,n$, we abbreviate $\pi_i = \mathbb{P}_i(\{1\})$ and for $\gamma\in\mathbb{R}$, $\phi_1(\gamma) = \log(h(\gamma))$, $\phi_2(\gamma) = \log(1-h(\gamma))$, and we note that $\pi_i \in (0,1)$. Thus, the expected log-likelihood function can be expressed as

$$\beta \mapsto \int_{\mathbb{R}^n} \ell_{\mathbb{M},n}(y,\beta)\,d\mathbb{P}(y) = \sum_{i=1}^n \left[\pi_i\phi_1(X_{i,n}[M]\beta) + (1-\pi_i)\phi_2(X_{i,n}[M]\beta)\right]. \tag{C.8}$$

The function in the previous display is continuous on its domain $\mathbb{R}^{|M|}$, by Condition H(i). To see that it also has a maximizer on $\mathbb{R}^{|M|}$, consider an arbitrary sequence $\beta_k \in \mathbb{R}^{|M|}$ such that $\|\beta_k\| \to \infty$ as $k\to\infty$. Then $\|X_n[M]\beta_k\|^2 \geq \|\beta_k\|^2\lambda_{\min}(X_n[M]'X_n[M]) \to \infty$ as $k\to\infty$, by Condition X2(i), so that at least for a sequence $i_k \in \{1,\dots,n\}$, we must have $|X_{i_k,n}[M]\beta_k| \to \infty$ as $k\to\infty$. Therefore, using the fact that $\pi_i\in(0,1)$ for $i=1,\dots,n$, it is easy to see that the sequence of summands in (C.5) corresponding to the indices i_k, $k\in\mathbb{N}$, with β replaced by β_k, converges to $-\infty$ as $k\to\infty$. Since for each k the remaining summands in (C.5) are non-positive, we see that the expected likelihood diverges to $-\infty$ along (β_k), and thus, by continuity, attains its maximum at some $\beta^*_{\mathbb{M},n}(\mathbb{P},X_n[M]) \in \mathbb{R}^{|M|}$.

For uniqueness, we show that the function in (C.5) is strictly concave. Take $\beta_1,\beta_2\in\mathbb{R}^{|M|}$, $\beta_1\neq\beta_2$ and $\alpha\in(0,1)$, and note that because $X_n[M]$ is of full rank $|M|$, we must have $X_n[M]\beta_1 \neq X_n[M]\beta_2$. Thus, there is at least one $i_0 \in \{1,\dots,n\}$ such that $X_{i_0,n}[M]'\beta_1 \neq X_{i_0,n}[M]'\beta_2$, and, by strict concavity (Condition H(ii)), $\phi_j(\alpha X_{i_0,n}[M]'\beta_1 + (1-\alpha)X_{i_0,n}[M]'\beta_2) > \alpha\phi_j(X_{i_0,n}[M]'\beta_1) + (1-\alpha)\phi_j(X_{i_0,n}[M]'\beta_2)$, for $j=1,2$. For the remaining indices $i\neq i_0$, the same inequalities hold, but are possibly not strict. Therefore, the expected log-likelihood in (C.5) is strictly concave and has a unique maximizer $\beta^*_{\mathbb{M},n}(\mathbb{P}, X_n[M]) \in \mathbb{R}^{|M|}$. $\square$

C.6. Auxiliary results for Section 3.3

Lemma C.1. *Suppose that Conditions X2(i,ii) and H(i,ii) hold and fix $\tau > 0$. There exists a finite positive constant $K^*(\tau, C)$, depending only on τ and the constant C from Condition X2(ii), such that eventually*

$$\max_{i=1,\dots,n} |X_{i,n}[M]\beta^*_{\mathbb{M},n}(\mathbb{P})| \;\leq\; K^*(\tau, C)$$

for all $\mathbb{P} \in \mathbf{P}_n^{(\mathrm{bin})}(\tau)$ and all $\mathbb{M} \in \mathsf{M}_n$. Here, $\beta^*_{\mathbb{M},n}(\mathbb{P})$ is the pseudo parameter from Lemma 3.8.

Proof. We begin by establishing the following preliminary result. For every pair $(h, M) \in \mathcal{H} \times \mathcal{I}$, there exists a bounded set $B_{h,M}(\tau, C) \subseteq \mathbb{R}^{|M|}$, such that eventually $U_{M,n}^{1/2}\beta^*_{\mathbb{M},n}(\mathbb{P}) \in B_{h,M}(\tau, C)$, for all $\mathbb{P} \in \mathbf{P}_n^{(\mathrm{bin})}(\tau)$, where $U_{M,n} = X_n[M]'X_n[M]/n$. Here, $C > 0$ is the constant from Condition X2. Fix $(h, M) \in \mathcal{H} \times \mathcal{I}$ and $t > 0$, and define $g_h(t) = \sup_{|\gamma|>t} \min\{\phi_1(\gamma), \phi_2(\gamma)\}$ and the set $B_{h,M}(\tau, C)$ by

$$B_{h,M}(\tau, C) = \{v \in \mathbb{R}^{|M|} : g_h(\|v\|/\sqrt{2}) \geq 2C[\phi_1(0) + \phi_2(0)]/\tau\},$$

where ϕ_1 and ϕ_2 are as in Condition H(ii). Note that, indeed, $B_{h,M}(\tau, C)$ is bounded, because $g_h(t) \to -\infty$ as $t \to \infty$, in view of Condition H(i). Next, fix n large enough, such that Conditions X2(i,ii) hold. Then the pseudo parameter $\beta^*_{\mathbb{M},n}$ of Lemma 3.8 uniquely exists. For $\beta \in \mathbb{R}^{|M|}$ and $\xi > 0$, define $R_{\beta,n}(\xi) = \{i \leq n : |X_{i,n}[M]\beta| \geq \xi\|U_{M,n}^{1/2}\beta\|\}$. Now fix $\xi > 0$ and $\beta \in \mathbb{R}^{|M|}$ such that $\|U_{M,n}^{1/2}\beta\| = 1$, and observe that

$$\begin{aligned}
1 \;=\; \beta'X_n[M]'X_n[M]\beta/n \;&=\; \frac{1}{n}\sum_{i=1}^{n}(X_{i,n}[M]U_{M,n}^{-1/2}U_{M,n}^{1/2}\beta)^2 \\
&\leq\; \xi^2 + \frac{1}{n}\sum_{i\in R_{\beta,n}(\xi)} \|U_{M,n}^{-1/2}X_{i,n}[M]'\|^2 \;\leq\; \frac{1}{n}|R_{\beta,n}(\xi)|C \;+\; \xi^2,
\end{aligned}$$

which implies that $\inf_{\beta\in\mathbb{R}^{|M|}} |R_{\beta,n}(\xi)| = \inf_{\beta:\|U_{M,n}^{1/2}\beta\|=1} |R_{\beta,n}(\xi)| \geq n(1-\xi^2)/C$. Since ϕ_1 and ϕ_2 are negative, we get for every $\mathbb{P} \in \mathbf{P}_n^{(\mathrm{bin})}(\tau)$, every $\beta \in \mathbb{R}^{|M|}$ and for $\xi = 1/\sqrt{2}$, that

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}}[\ell_{\mathbb{M},n}(\cdot, \beta)] \;&\leq\; \sum_{i\in R_{\beta,n}(\xi)} \mathbb{P}_i(\{1\})\mathbb{P}_i(\{0\})\Big(\phi_1(X_{i,n}[M]\beta) + \phi_2(X_{i,n}[M]\beta)\Big) \\
&\leq\; \tau\,|R_{\beta,n}(\xi)|\,g_h(\xi\|U_{M,n}^{1/2}\beta\|) \;\leq\; n\frac{\tau}{2C}g_h(\|U_{M,n}^{1/2}\beta\|/\sqrt{2}),
\end{aligned}$$

where $\mathbb{M} \triangleq (h, M)$. Therefore, we have $n[\phi_1(0) + \phi_2(0)] \leq \mathbb{E}_{\mathbb{P}}[\ell_{\mathbb{M},n}(\cdot, 0)] \leq \mathbb{E}_{\mathbb{P}}[\ell_{\mathbb{M},n}(\cdot, \beta^*_{\mathbb{M},n}(\mathbb{P})] \leq n\tau g_h(\|U^{1/2}_{M,n}\beta^*_{\mathbb{M},n}\|/\sqrt{2})/(2C)$, which yields

$$g_h(\|U^{1/2}_{M,n}\beta^*_{\mathbb{M},n}\|/\sqrt{2}) \ \geq \ \frac{[\phi_1(0) + \phi_2(0)]2C}{\tau},$$

i.e., $U^{1/2}_{M,n}\beta^*_{\mathbb{M},n}(\mathbb{P}) \in B_{h,M}(\tau, C)$. So we have established the preliminary result. Since the bounded set $B_{h,M}(\tau, C)$ depends only on the indicated quantities, there exists a finite positive constant $K^*_{h,M}(\tau, C)$, depending on the same quantities, such that for all large n, for all $\mathbb{P} \in \mathbf{P}^{(\mathrm{bin})}_n(\tau)$ and for all $\mathbb{M} \in \mathsf{M}_n$,

$$\begin{aligned}\max_i |X_{i,n}[M]\beta^*_{\mathbb{M},n}(\mathbb{P})| &\leq \max_i \|U^{-1/2}_{M,n} X_{i,n}[M]'\| \|U^{1/2}_{M,n}\beta^*_{\mathbb{M},n}(\mathbb{P})\| \\ &\leq \sqrt{C}K^*_{h,M}(\tau, C) \leq \max_{(h,M)\in\mathcal{H}\times\mathcal{I}} \sqrt{C}K^*_{h,M}(\tau, C) =: K^*(\tau, C).\end{aligned}$$

This finishes the proof. □

Lemma C.2. *Suppose that Conditions X2(i,ii) and H(i,ii,iii) hold and fix $\tau > 0$.*

(i) There exist positive constants $\overline{K}(\tau, C)$ and $\underline{K}(\tau, C)$, depending only on τ and the constant C from Condition X2(ii), such that for all sufficiently large $n \in \mathbb{N}$, for all $\mathbb{M} \in \mathsf{M}_n$, all $\mathbb{P} \in \mathbf{P}^{(\mathrm{bin})}_n(\tau)$ and all $y \in \{0, 1\}^n$,

$$\begin{aligned}\overline{K}(\tau, C) &\geq \lambda_{max}\left((X_n[M]'X_n[M])^{-1/2} H^*_{\mathbb{M},n}(y)(X_n[M]'X_n[M])^{-1/2}\right) \\ &\geq \lambda_{min}\left((X_n[M]'X_n[M])^{-1/2} H^*_{\mathbb{M},n}(y)(X_n[M]'X_n[M])^{-1/2}\right) \geq \underline{K}(\tau, C),\end{aligned}$$

where $H^*_{\mathbb{M},n}(y) := H_{\mathbb{M},n}(y, \beta^*_{\mathbb{M},n}(\mathbb{P}))$, $H_{\mathbb{M},n}(y, \beta) = -\frac{\partial^2 \ell_{\mathbb{M},n}(y,\beta)}{\partial\beta\partial\beta'}$ with $\beta \in \mathbb{R}^{m(\mathbb{M})}$ and $\beta^*_{\mathbb{M},n}(\mathbb{P})$ is the pseudo parameter of Lemma 3.8.

(ii) For $\delta > 0$ and n sufficiently large, such that the pseudo parameter $\beta^*_{\mathbb{M},n}$ of Lemma 3.8 exists, define

$$N_{\mathbb{M},\mathbb{P},n}(\delta) = \left\{\beta \in \mathbb{R}^{m(\mathbb{M})} : \left\|(X_n[M]'X_n[M])^{1/2}(\beta - \beta^*_{\mathbb{M},n}(\mathbb{P}))\right\| \leq \delta\right\}.$$

Then, for every $\delta > 0$,

$$\sup_{\substack{\mathbb{M}\in\mathsf{M}_n \\ \mathbb{P}\in\mathbf{P}^{(\mathrm{bin})}_n(\tau)}} \ \sup_{y\in\{0,1\}^n} \ \sup_{\beta\in N_{\mathbb{M},\mathbb{P},n}(\delta)} \left\| H^*_{\mathbb{M},n}(y)^{-1/2} H_{\mathbb{M},n}(y, \beta) H^*_{\mathbb{M},n}(y)^{-1/2} - I_{m(\mathbb{M})}\right\|$$

converges to zero as $n \to \infty$.

(iii) Suppose that, in addition, also Condition X2(iii) holds. Then there exists a positive finite constant $K(\tau, C)$, depending only on τ and the constant C from Condition X2, such that eventually

$$\sup_{\mathbb{M}\in\mathsf{M}_n} \ \sup_{\mathbb{P}\in\mathbf{P}^{(\mathrm{bin})}_n(\tau)} \frac{\lambda_{max}(\mathbb{E}_{\mathbb{P}}[H^*_{\mathbb{M},n}])}{\lambda_{min}(\mathbb{E}_{\mathbb{P}}[H^*_{\mathbb{M},n}])} \ \leq \ K(\tau, C).$$

Proof. First, fix n large enough, such that the bound of Condition X2(ii) holds, the pseudo parameter of Lemma 3.8 exists and the bound of Lemma C.1 applies. Fix $\mathbb{M} \in \mathsf{M}_n$ and $\mathbb{P} \in \mathbf{P}_n^{(\mathrm{bin})}(\tau)$. Since $H_{\mathbb{M},n}(y,\beta) = X_n[M]' D_{\mathbb{M},n}(y,\beta) X_n[M]$, for a diagonal matrix $D_{\mathbb{M},n}(y,\beta)$ whose i-th diagonal entry is given by

$$-y_i \ddot{\phi}_1(X_{i,n}[M]\beta) - (1-y_i)\ddot{\phi}_2(X_{i,n}[M]\beta) \ > \ 0,$$

in view of Conditions H(i,ii,iii), we see that $H^*_{\mathbb{M},n}(y)$ is positive definite. Moreover, from continuity and positivity of $-\ddot{\phi}_j$, $j = 1,2$, Lemma C.1 and finiteness of $\mathcal{H}$, we conclude that the diagonal entries of $D_{\mathbb{M},n}(y, \beta^*_{\mathbb{M},n}(\mathbb{P}))$ are lower and upper bounded by positive constants that depend only on τ and C. This finishes the claim in (i). Consider now

$$\begin{aligned}
&\sup_{\beta \in N_{\mathbb{M},\mathbb{P},n}(\delta)} \|H^*_{\mathbb{M},n}(y)^{-1/2} H_{\mathbb{M},n}(y,\beta) H^*_{\mathbb{M},n}(y)^{-1/2} - I_{m(\mathbb{M})}\| \\
&\quad \le \ \|X_n[M] H^*_{\mathbb{M},n}(y)^{-1/2}\|^2 \sup_{\beta \in N_{\mathbb{M},\mathbb{P},n}(\delta)} \|D_{\mathbb{M},n}(y,\beta) - D_{\mathbb{M},n}(y, \beta^*_{\mathbb{M},n}(\mathbb{P}))\| \\
&\quad \le \ \frac{\|X_n[M](X_n[M]'X_n[M])^{-1} X_n[M]'\|}{\min_{i=1,\dots,n}(-y_i\ddot{\phi}_1(X_{i,n}[M]\beta^*_{\mathbb{M},n}) - (1-y_i)\ddot{\phi}_2(X_{i,n}[M]\beta^*_{\mathbb{M},n}))} \times \\
&\qquad\qquad \sup_{\beta \in N_{\mathbb{M},\mathbb{P},n}(\delta)} \max_{\substack{i=1,\dots,n \\ j=1,2}} |\ddot{\phi}_j(X_{i,n}[M]\beta) - \ddot{\phi}_j(X_{i,n}[M]\beta^*_{\mathbb{M},n})|. \qquad \text{(C.9)}
\end{aligned}$$

We have just seen that the minimum on the far right side of the previous display is lower bounded by a positive constant that depends only on τ and C. To finish the proof, note that for $\beta \in N_{\mathbb{M},\mathbb{P},n}(\delta)$, we have

$$\begin{aligned}
&|X_{i,n}[M]\beta - X_{i,n}[M]\beta^*_{\mathbb{M},n}| \\
&\le \|X_{i,n}[M](X_n[M]'X_n[M])^{-1/2}\| \|(X_n[M]'X_n[M])^{1/2}(\beta - \beta^*_{\mathbb{M},n})\| \le \delta\sqrt{C/n}
\end{aligned}$$

and that $|X_{i,n}[M]\beta^*_{\mathbb{M},n}| \le K^*(\tau, C)$, by Lemma C.1. Therefore, by uniform continuity of $\ddot{\phi}_j$ on the compact interval $[-K^*(\tau,C) - \delta\sqrt{C}, K^*(\tau,C) + \delta\sqrt{C}]$, for every $\eta > 0$, there exists $n_0 = n_0(\eta, \tau, \delta, C, h)$, such that the supremum in (C.6) is bounded by η, for all $n \ge n_0$. Since $\mathcal{H}$ is finite, the proof of (ii) is finished. For part (iii), simply combine part (i) and Condition X2(iii). ☐

C.7. Proof of Lemma 3.9

The proof is a variation of the consistency part of the proof of Theorem 4 in Fahrmeir (1990). Fix n large enough, such that the bound of Condition X2(ii) holds, the pseudo parameter of Lemma 3.8 exists and the bound of Lemma C.1 applies. Fix $\mathbb{M} \in \mathsf{M}_n$ and $\mathbb{P} \in \mathbf{P}_n^{(\mathrm{bin})}(\tau)$, write $\beta^*_{\mathbb{M},n} = \beta^*_{\mathbb{M},n}(\mathbb{P})$ and note that by Condition H(iii), the function $\beta \mapsto \ell_{\mathbb{M},n}(y,\beta)$ is twice continuously differentiable on $\mathbb{R}^{m(\mathbb{M})}$ and thus, for every $y \in \{0,1\}^n$, admits the expansion

$$\begin{aligned}\ell_{\mathbb{M},n}(y,\beta) \quad &= \quad \ell_{\mathbb{M},n}(y,\beta^*_{\mathbb{M},n}) \;+\; (\beta-\beta^*_{\mathbb{M},n})'U_{M,n}^{1/2}U_{M,n}^{-1/2}s^*_{\mathbb{M},n}(y)\\ &\quad - \;\frac{1}{2}(\beta-\beta^*_{\mathbb{M},n})'U_{M,n}^{1/2}U_{M,n}^{-1/2}H_{\mathbb{M},n}(y,\tilde\beta_n)U_{M,n}^{-1/2}U_{M,n}^{1/2}(\beta-\beta^*_{\mathbb{M},n}),\end{aligned}$$

for some $\tilde\beta \in \{a\beta+(1-a)\beta^*_{\mathbb{M},n} : a\in[0,1]\}$, and where $s^*_{\mathbb{M},n}(y) = \frac{\partial \ell_{\mathbb{M},n}(y,\beta)}{\partial\beta}\Big|_{\beta=\beta^*_{\mathbb{M},n}}$, $H_{\mathbb{M},n}(y,\beta) = -\frac{\partial\ell_{\mathbb{M},n}(y,\beta)}{\partial\beta\partial\beta'}$, and $U_{M,n} = X_n[M]'X_n[M]/n$. For $\delta>0$, define $\lambda_n = \sqrt{n}U_{M,n}^{1/2}(\beta-\beta^*_{\mathbb{M}.n})/\delta$ to rewrite the previous equation as

$$\begin{aligned}\ell_{\mathbb{M},n}(y,\beta) \;-\; \ell_{\mathbb{M},n}(y,\beta^*_{\mathbb{M},n}) \;&=\; \delta\lambda_n' U_{M,n}^{-1/2}s^*_{\mathbb{M},n}(y)/\sqrt{n}\\ &\quad -\;\frac{1}{2}\delta^2\lambda_n'U_{M,n}^{-1/2}(H_{\mathbb{M},n}(y,\tilde\beta_n)/n)U_{M,n}^{-1/2}\lambda_n,\end{aligned}$$

for all $\beta\in\mathbb{R}^{m(\mathbb{M})}$ and all $y\in\{0,1\}^n$. For $N_n(\delta) := N_{\mathbb{M},\mathbb{P},n}(\delta)$ as in Lemma C.2(ii), define $L_n(y,\delta) := \inf_{\beta\in N_n(\delta)}\lambda_{\min}\left(U_{M,n}^{-1/2}(H_{\mathbb{M},n}(y,\beta)/n)U_{M,n}^{-1/2}\right)$, take $\beta\in\partial N_n(\delta) = \{\beta\in\mathbb{R}^{m(\mathbb{M})} : \|\sqrt{n}U_{M,n}^{1/2}(\beta-\beta^*_{\mathbb{M},n})\| = \delta\}$ and observe that now $\|\lambda_n\| = 1$, $\delta\lambda_n'U_{M,n}^{-1/2}s^*_{\mathbb{M},n}(y)/\sqrt{n} \le \delta\|U_{M,n}^{-1/2}s^*_{\mathbb{M},n}(y)/\sqrt{n}\|$ and

$$\begin{aligned}\frac{1}{2}\delta^2 L_n(y,\delta) \;&\le\; \frac{1}{2}\delta^2\lambda_{\min}\left(U_{M,n}^{-1/2}(H_{\mathbb{M},n}(y,\tilde\beta)/n)U_{M,n}^{-1/2}\right)\\ &\le\; \frac{1}{2}\delta^2\lambda_n'U_{M,n}^{-1/2}(H_{\mathbb{M},n}(y,\tilde\beta)/n)U_{M,n}^{-1/2}\lambda_n,\end{aligned}$$

for all $y\in\{0,1\}^n$. Therefore, we have the inclusion

$$\begin{aligned}E_{\mathbb{M},\mathbb{P},n}(\delta) &:= \{y\in\{0,1\}^n : \delta\|U_{M,n}^{-1/2}s^*_{\mathbb{M},n}(y)/\sqrt{n}\| < \delta^2 L_n(y,\delta)/2\}\\ &\subseteq \{y\in\{0,1\}^n : \forall\beta\in\partial N_n(\delta) : \ell_{\mathbb{M},n}(y,\beta) < \ell_{\mathbb{M},n}(y,\beta^*_{\mathbb{M},n})\} =: F_n(\delta).\end{aligned}$$

As a consequence, for every $y\in E_{\mathbb{M},\mathbb{P},n}(\delta)$, the function $\beta\mapsto\ell_{\mathbb{M},n}(y,\beta)$ has a local maximum $\hat\beta_{\mathbb{M},n}(y)$ on the interior of $N_n(\delta)$. By strict concavity (Conditions H(ii) and X2(i)), this is a unique global maximum. Moreover, we have $F_n(\delta)\subseteq\{y : \hat\beta_{\mathbb{M},n}(y)\in N_n(\delta)\} = \{y : \|\sqrt{n}U_{M,n}^{1/2}(\hat\beta_{\mathbb{M},n}(y)-\beta^*_{\mathbb{M},n})\|\le\delta\}$. Hence,

$$\mathbb{P}(\|\sqrt{n}U_{M,n}^{1/2}(\hat\beta_{\mathbb{M},n}(y)-\beta^*_{\mathbb{M},n})\| > \delta) \;\le\; \mathbb{P}(E_{\mathbb{M},\mathbb{P},n}(\delta)^c).$$

It remains to verify that $\mathbb{P}(E_{\mathbb{M},\mathbb{P},n}(\delta)^c)$ is small for large n, uniformly in $\mathbb{M}$ and $\mathbb{P}$. Take $\varepsilon>0$ and note that

$$\begin{aligned}\mathbb{P}(E_{\mathbb{M},\mathbb{P},n}(\delta)^c) \;&\le\; \mathbb{P}(4\|U_{M,n}^{-1/2}s^*_{\mathbb{M},n}(\cdot)/\sqrt{n}\|^2 \ge \delta^2L_n^2(\cdot,\delta), L_n^2(\cdot,\delta)\ge\varepsilon^2)\\ &\quad +\;\mathbb{P}(L_n^2(\cdot,\delta)<\varepsilon^2) \qquad\qquad (C.10)\\ &\le\; \mathbb{P}(4\|U_{M,n}^{-1/2}s^*_{\mathbb{M},n}(\cdot)/\sqrt{n}\|^2\ge\delta^2\varepsilon^2) \;+\; \mathbb{P}(L_n^2(\cdot,\delta)<\varepsilon^2)\\ &\le\; 4\frac{\mathrm{trace}(U_{M,n}^{-1/2}\mathbb{V}\mathbb{C}_{\mathbb{P}}(s^*_{\mathbb{M},n}/\sqrt{n})U_{M,n}^{-1/2})}{\delta^2\varepsilon^2} \;+\; \mathbb{P}(L_n(\cdot,\delta)<\varepsilon),\end{aligned}$$

in view of Markov's inequality and since $\mathbb{E}_{\mathbb{P}}[s^*_{\mathbb{M},n}] = 0$. Note that $\mathbb{VC}_{\mathbb{P}}(s^*_{\mathbb{M},n}) = X_n[M]' V^*_{\mathbb{M},\mathbb{P},n} X_n[M]$, for a diagonal matrix $V^*_{\mathbb{M},\mathbb{P},n}$ whose diagonal entries satisfy

$$\begin{aligned}[V^*_{\mathbb{M},\mathbb{P},n}]_{ii} &= \mathbb{P}_i(\{1\})\mathbb{P}_i(\{0\})(\dot{\phi}_1(X_{i,n}\beta^*_{\mathbb{M},n}) - \dot{\phi}_2(X_{i,n}\beta^*_{\mathbb{M},n}))^2 \\ &\leq \sup_{|\gamma| \leq K^*(\tau,C)} (\dot{\phi}_1(\gamma) - \dot{\phi}_2(\gamma))^2,\end{aligned}$$

for the constant $K^*(\tau, C)$ of Lemma C.1. Thus, the trace on the last line of display (C.7) is bounded by $p \max_{h\in\mathcal{H}} \sup_{|\gamma|\leq K^*(\tau,C)} (\dot{\phi}_1(\gamma) - \dot{\phi}_2(\gamma))^2$, which does not depend on n, $\mathbb{M}$ or $\mathbb{P}$. Finally, to bound the remaining probability, note that $L_n(y, \delta)$ is lower bounded by the product of $\lambda_{\min}(U^{-1/2}_{M,n} H^*_{\mathbb{M},n}(y) U^{-1/2}_{M,n})/n$ and $\inf_{\beta\in N_n(\delta)} \lambda_{\min}(H^*_{\mathbb{M},n}(y)^{-1/2} H_{\mathbb{M},n}(y, \beta) H^*_{\mathbb{M},n}(y)^{-1/2})$. The first factor is itself lower bounded by the positive constant $\underline{K}(\tau, C)$ from Lemma C.2(i). Thus, $\mathbb{P}(L_n(\cdot, \delta) < \varepsilon)$ is upper bounded by

$$\mathbb{P}\left(\sup_{\beta\in N_n(\delta)} \left\| H^*_{\mathbb{M},n}(y)^{-1/2} H_{\mathbb{M},n}(y,\beta) H^*_{\mathbb{M},n}(y)^{-1/2} - I_{m(\mathbb{M})}\right\| > 1 - \frac{\varepsilon}{\underline{K}(\tau, C)}\right).$$

Choosing $\varepsilon = \underline{K}(\tau, C)/2$ and using Lemma C.2(ii), we conclude that for every $\delta > 0$,

$$\begin{aligned}&\sup_{\mathbb{M}\in\mathsf{M}_n} \sup_{\mathbb{P}\in\mathbf{P}^{(\mathrm{bin})}_n(\tau)} \mathbb{P}\left(\left\|(X_n[M]'X_n[M])^{1/2}(\hat{\beta}_{\mathbb{M},n} - \beta^*_{\mathbb{M},n}(\mathbb{P}))\right\| > \delta\right) \quad \text{(C.11)} \\ &\leq \sup_{\mathbb{M}\in\mathsf{M}_n} \sup_{\mathbb{P}\in\mathbf{P}^{(\mathrm{bin})}_n(\tau)} \mathbb{P}(E_{\mathbb{M},\mathbb{P},n}(\delta)^c) \\ &\leq 16 \frac{p \max_{h\in\mathcal{H}} \sup_{|\gamma|\leq K^*(\tau,C)} (\dot{\phi}_1(\gamma) - \dot{\phi}_2(\gamma))^2}{\delta^2 \underline{K}(\tau, C)^2} + o(1),\end{aligned}$$

where the $o(1)$ term refers to convergence as $n \to \infty$. Now, to establish the asymptotic existence of the MLE, we simply take $E_{\mathbb{M},\mathbb{P},n} := E_{\mathbb{M},\mathbb{P},n}(\delta_n)$, for $\delta_n \to \infty$ sufficiently slowly as $n \to \infty$. For the uniform consistency part, note that the limit superior as $n \to \infty$ of the expression in (C.7) is bounded by a quantity that converges to zero as $\delta \to \infty$. ☐

C.8. Proof of Theorem 3.10

Fix $n \in \mathbb{N}$, a candidate model $\mathbb{M} \in \mathsf{M}_n$ and $\mathbb{P}_n \in \mathbf{P}^{(\mathrm{bin})}_n(\tau)$, and let $\mathbb{E}_n$ and $\mathbb{V}_n$ denote the expectation and variance operators with respect to $\mathbb{P}_n$ on $\{0, 1\}^n$. Define $s_{\mathbb{M},n}(y, \beta) := \partial\ell_{\mathbb{M},n}(y, \beta)/\partial\beta$ and note that by assumption $\beta \mapsto s_{\mathbb{M},n}(y, \beta)$ is continuously differentiable on $\mathbb{R}^{m(\mathbb{M})}$, for all $y \in \{0, 1\}^n$. Therefore, we can expand $s_{\mathbb{M},n}$ around $\beta_0 \in \mathbb{R}^{m(\mathbb{M})}$ as follows,

$$s_{\mathbb{M},n}(y, \beta) - s_{\mathbb{M},n}(y, \beta_0) = \int_0^1 H_{\mathbb{M},n}(y, t\beta + (1 - t)\beta_0) dt \cdot (\beta_0 - \beta).$$

For n sufficiently large, such that $\beta^*_{\mathbb{M},n} = \beta^*_{\mathbb{M},n}(\mathbb{P}_n)$ of Lemma 3.8 exists, define $\tilde{H}_{\mathbb{M},n}(y) := \int_0^1 H_{\mathbb{M},n}(y, t\beta^*_{\mathbb{M},n} + (1-t)\hat{\beta}_{\mathbb{M},n}(y))\,dt$ and note that with this we have

$$s^*_{\mathbb{M},n}(y) \;:=\; s_{\mathbb{M},n}(y, \beta^*_{\mathbb{M},n}) \;=\; \tilde{H}_{\mathbb{M},n}(y) \cdot (\hat{\beta}_{\mathbb{M},n}(y) - \beta^*_{\mathbb{M},n}), \tag{C.12}$$

for $y \in E_{\mathbb{M},n} := E_{\mathbb{M},\mathbb{P}_n,n}$, the set defined in Lemma 3.9. Moreover, since $H_{\mathbb{M},n}(y,\beta)$ is positive definite under Conditions H(iii) and X2(i), so are $\tilde{H}_{\mathbb{M},n}(y)$ and $H^*_{\mathbb{M},n}(y) := H_{\mathbb{M},n}(y, \beta^*_{\mathbb{M},n})$. Thus, if we set

$$\begin{aligned} \psi^*_{i,n,\mathbb{M}}(y_i) \;&:=\; y_i \dot{\phi}_1(X_{i,n}[M]\beta^*_{\mathbb{M},n}) + (1-y_i)\dot{\phi}_2(X_{i,n}[M]\beta^*_{\mathbb{M},n}), \\ g_{i,n,\mathbb{M}}(y_i) \;&:=\; \mathbb{E}_n[H^*_{\mathbb{M},n}]^{-1} X_{i,n}[M]' \left(\psi^*_{i,n,\mathbb{M}}(y_i) - \mathbb{E}_n[\psi^*_{i,n,\mathbb{M}}]\right), \end{aligned}$$

and

$$\Delta_{n,\mathbb{M}}(y) \;:=\; \hat{\beta}_{\mathbb{M},n}(y) - \beta^*_{\mathbb{M},n} - \sum_{i=1}^n g_{i,n,\mathbb{M}}(y_i),$$

we see that (1) is satisfied, that $\mathbb{E}_n[g_{i,n,\mathbb{M}}] = 0$, $r_{n,\mathbb{M}}(y) := \sum_{i=1}^n g_{i,n,\mathbb{M}}(y_i) = \mathbb{E}_n[H^*_{\mathbb{M},n}]^{-1} s^*_{\mathbb{M},n}(y)$, because $\mathbb{E}_n[s^*_{\mathbb{M},n}] = 0$, and that $\mathbb{V}_n[r^{(j)}_{n,\mathbb{M}}]$ is given by

$$\sum_{i=1}^n \left(e_{m(\mathbb{M})}(j)' \mathbb{E}_n[H^*_{\mathbb{M},n}]^{-1} X_{i,n}[M]'\right)^2 \mathbb{V}_n[\psi^*_{i,n,\mathbb{M}}],$$

where $e_m(j)$ is the j-th element of the canonical basis in $\mathbb{R}^m$ and $j \in \{1, \dots, m(\mathbb{M})\}$. Note that by Lemma C.1, Conditions H(i,iii,iv) and the finiteness of $\mathcal{H}$, there exists a positive constant $\underline{K'}(\tau, C)$, depending only on $\tau > 0$ and C from Condition X2, such that for all large n,

$$\infty > \mathbb{V}_n[\psi^*_{i,n,\mathbb{M}}] \geq \tau \left(\frac{\dot{h}(X_{i,n}[M]\beta^*_{\mathbb{M},n})}{h(X_{i,n}[M]\beta^*_{\mathbb{M},n})} + \frac{\dot{h}(X_{i,n}[M]\beta^*_{\mathbb{M},n})}{1 - h(X_{i,n}[M]\beta^*_{\mathbb{M},n})} \right)^2 \geq \underline{K'}(\tau, C) > 0.$$

In particular, for such n, we have $0 < \mathbb{V}_n[r^{(j)}_{n,\mathbb{M}}] < \infty$. Furthermore, by a similar argument, we obtain the upper bound $|\psi^*_{i,n,\mathbb{M}}(y_i)|^2 \leq \overline{K'}(\tau, C)$ and, in turn,

$$\frac{|g^{(j)}_{i,n,\mathbb{M}}(y_i)|^2}{\mathbb{V}_n[r^{(j)}_{n,\mathbb{M}}]} \leq 2 \frac{\overline{K'}(\tau, C)}{\underline{K'}(\tau, C)} \frac{\left(e_{m(\mathbb{M})}(j)' \mathbb{E}_n[H^*_{\mathbb{M},n}]^{-1} X_{i,n}[M]'\right)^2}{\sum_{i=1}^n \left(e_{m(\mathbb{M})}(j)' \mathbb{E}_n[H^*_{\mathbb{M},n}]^{-1} X_{i,n}[M]'\right)^2}. \tag{C.13}$$

But the numerator of the second fraction on the right of the previous display can be bounded by

$$\begin{aligned} &\|(X_n[M]'X_n[M])^{1/2} \mathbb{E}_n[H^*_{\mathbb{M},n}]^{-1} e_{m(\mathbb{M})}(j)\|^2 \|(X_n[M]'X_n[M])^{-1/2} X_{i,n}[M]'\|^2 \\ &\quad \leq \|X_n[M] \mathbb{E}_n[H^*_{\mathbb{M},n}]^{-1} e_{m(\mathbb{M})}(j)\|^2 \cdot C/n, \end{aligned}$$

in view of Condition X2(ii), whereas the denominator of that same fraction coincides with $\|X_n[M] \mathbb{E}_n[H^*_{\mathbb{M},n}]^{-1} e_{m(\mathbb{M})}(j)\|^2$. Thus, we conclude that also (1) is satisfied. Finally, for asymptotic negligibility of $\mathbb{V}_n[r_{n,\mathbb{M}}^{(j)}]^{-1/2}\Delta_{n,\mathbb{M}}^{(j)}(y)$, first note that using (C.8), for $y \in E_{\mathbb{M},n}$, we have

$$\begin{aligned}
\Delta_{n,\mathbb{M}}(y) &= \tilde{H}_{\mathbb{M},n}(y)^{-1}\tilde{H}_{\mathbb{M},n}(y)(\hat{\beta}_{\mathbb{M},n}(y) - \beta^*_{\mathbb{M},n}) - \mathbb{E}_n[H^*_{\mathbb{M},n}]^{-1}s^*_{\mathbb{M},n}(y) \\
&= \left(\tilde{H}_{\mathbb{M},n}(y)^{-1}\mathbb{E}_n[H^*_{\mathbb{M},n}] - I_{m(\mathbb{M})}\right)\mathbb{E}_n[H^*_{\mathbb{M},n}]^{-1}s^*_{\mathbb{M},n}(y) \\
&= \mathbb{E}_n[H^*_{\mathbb{M},n}]^{-1/2}\left(\mathbb{E}_n[H^*_{\mathbb{M},n}]^{1/2}\tilde{H}_{\mathbb{M},n}(y)^{-1}\mathbb{E}_n[H^*_{\mathbb{M},n}]^{1/2} - I_{m(\mathbb{M})}\right)\mathbb{E}_n[H^*_{\mathbb{M},n}]^{1/2} \\
&\quad \times R_{n,\mathbb{M}}^{1/2}R_{n,\mathbb{M}}^{-1/2}r_{n,\mathbb{M}}(y),
\end{aligned}$$

where $R_{n,\mathbb{M}} = \operatorname{diag}(\mathbb{V}_n[r_{n,\mathbb{M}}^{(1)}], \dots, \mathbb{V}_n[r_{n,\mathbb{M}}^{(m(\mathbb{M}))}])$. Therefore,

$$\begin{aligned}
&\|R_{n,\mathbb{M}}^{-1/2}\Delta_{n,\mathbb{M}}(y)\| \\
&\quad \le \|\mathbb{E}_n[H^*_{\mathbb{M},n}]^{1/2}\tilde{H}_{\mathbb{M},n}(y)^{-1}\mathbb{E}_n[H^*_{\mathbb{M},n}]^{1/2} - I_{m(\mathbb{M})}\|\frac{\lambda_{\max}(\mathbb{E}_n[H^*_{\mathbb{M},n}]^{1/2})}{\lambda_{\min}(\mathbb{E}_n[H^*_{\mathbb{M},n}]^{1/2})} \\
&\quad\quad \times \frac{\max_j \mathbb{V}_n^{1/2}[r_{n,\mathbb{M}}^{(j)}]}{\min_j \mathbb{V}_n^{1/2}[r_{n,\mathbb{M}}^{(j)}]}\|R_{n,\mathbb{M}}^{-1/2}r_{n,\mathbb{M}}(y)\|.
\end{aligned}$$

Here, $R_{n,\mathbb{M}}^{-1/2}r_{n,\mathbb{M}}(y)$ has mean zero and covariance matrix with ones on the main diagonal, and consequently its norm is bounded in probability. The ratio of the largest and smallest variance component of $r_{n,\mathbb{M}}$ is bounded by the condition number of the matrix $\mathbb{E}_n[H^*_{\mathbb{M},n}]^{-1}X_n[M]'X_n[M]\mathbb{E}_n[H^*_{\mathbb{M},n}]^{-1}$ times a constant that depends only on τ and C, because of the previously derived upper and lower bounds on $|\psi^*_{i,n,\mathbb{M}}(y_i)|^2$ and $\mathbb{V}_n[\psi^*_{i,n,\mathbb{M}}]$, respectively. But this condition number is eventually bounded by a finite constant that depends only on τ and C from Condition X2, in view of Lemma C.2(iii). In particular, this lemma shows that the condition number of $\mathbb{E}_n[H^*_{\mathbb{M},n}]$ is bounded, eventually. Therefore, since $\mathbb{P}_n(E_{\mathbb{M},n}) \to 1$, as $n \to \infty$, it remains to show that $\mathbb{E}_n[H^*_{\mathbb{M},n}]^{1/2}\tilde{H}_{\mathbb{M},n}(y)^{-1}\mathbb{E}_n[H^*_{\mathbb{M},n}]^{1/2} \to I_{m(\mathbb{M})}$, in $\mathbb{P}_n$-probability. The result follows if we can show that the eigenvalues of $H^*_{\mathbb{M},n}(y)^{-1/2}\tilde{H}_{\mathbb{M},n}(y)H^*_{\mathbb{M},n}(y)^{-1/2}$ and of $\mathbb{E}_n[H^*_{\mathbb{M},n}]^{-1/2}H^*_{\mathbb{M},n}(y)\mathbb{E}_n[H^*_{\mathbb{M},n}]^{-1/2}$ converge to 1, in $\mathbb{P}_n$-probability, because for $A = \mathbb{E}_n[H^*_{\mathbb{M},n}]$, $B = \tilde{H}_{\mathbb{M},n}(y)$ and $C = H^*_{\mathbb{M},n}(y)$, we have

$$\begin{aligned}
&\|A^{1/2}B^{-1}A^{1/2} - I_{m(\mathbb{M})}\| \\
&= \|A^{1/2}C^{-1/2}C^{1/2}B^{-1}C^{1/2}C^{-1/2}A^{1/2} - I_{m(\mathbb{M})}\| \\
&\le \|A^{1/2}C^{-1}A^{1/2}\|\|C^{1/2}B^{-1}C^{1/2} - I_{m(\mathbb{M})}\| + \|A^{1/2}C^{-1}A^{1/2} - I_{m(\mathbb{M})}\|.
\end{aligned}$$

For the first of these two, let $N_{\mathbb{M},n}(\delta) := N_{\mathbb{M},\mathbb{P}_n,n}(\delta)$ be as in Lemma C.2(ii), which is a convex set, and note that for every $\varepsilon > 0$ and every $\delta > 0$,

$$\begin{aligned}
&\mathbb{P}_n\left(\|H^*_{\mathbb{M},n}(y)^{-1/2}\tilde{H}_{\mathbb{M},n}(y)H^*_{\mathbb{M},n}(y)^{-1/2} - I_{m(\mathbb{M})}\| > \varepsilon\right) \\
&\quad\le \mathbb{P}_n\Big(\sup_{\beta\in N_{\mathbb{M},n}(\delta)}\|H^*_{\mathbb{M},n}(y)^{-1/2}H_{\mathbb{M},n}(y,\beta)H^*_{\mathbb{M},n}(y)^{-1/2} - I_{m(\mathbb{M})}\| > \varepsilon, \\
&\qquad\qquad \hat{\beta}_{\mathbb{M},n} \in N_{\mathbb{M},n}(\delta)\Big) + \mathbb{P}_n\left(\hat{\beta}_{\mathbb{M},n} \notin N_{\mathbb{M},n}(\delta)\right) \\
&\quad= o(1) + \mathbb{P}_n\left(\sqrt{n}\|U^{1/2}_{M,n}(\hat{\beta}_{\mathbb{M},n} - \beta^*_{\mathbb{M},n})\| > \delta\right),
\end{aligned}$$

where $U_{M,n} = X_n[M]'X_n[M]/n$. Since $\delta > 0$ was arbitrary Lemma 3.9 shows that the probability on the far left-hand-side of the previous display converges to zero as $n \to \infty$. Finally, for $v_1, v_2 \in \mathbb{R}^{m(\mathbb{M})}$ with $\|v_1\| = \|v_2\| = 1$, write

$$\begin{aligned}
&v_1'(\mathbb{E}_n[H^*_{\mathbb{M},n}]^{-1/2}H^*_{\mathbb{M},n}(y)\mathbb{E}_n[H^*_{\mathbb{M},n}]^{-1/2} - I_{m(\mathbb{M})})v_2 \\
&= \sum_{i=1}^n v_1'\mathbb{E}_n[H^*_{\mathbb{M},n}]^{-1/2}X_{i,n}[M]'X_{i,n}[M]\mathbb{E}_n[H^*_{\mathbb{M},n}]^{-1/2}v_2\left(D^*_{i,\mathbb{M},n}(y_i) - \mathbb{E}_n[D^*_{i,\mathbb{M},n}]\right),
\end{aligned}$$

where $D^*_{i,\mathbb{M},n}(y_i) = -y_i\ddot{\phi}_1(X_{i,n}[M]\beta^*_{\mathbb{M},n}) - (1-y_i)\ddot{\phi}_2(X_{i,n}[M]\beta^*_{\mathbb{M},n})$ and $\mathbb{V}_n[D^*_{i,\mathbb{M},n}]$ is bounded by a constant that depends only on τ and C. The mean of the expression in the previous display is clearly equal to zero, while its variance is bounded by $\sum_{i=1}^n (X_{i,n}[M]\mathbb{E}_n[H^*_{\mathbb{M},n}]^{-1}X_{i,n}[M]')^2$ times a constant that depends only on τ and C. In view of Lemma C.2(i) and Condition X2(ii), the latter sum is itself bounded by $\sum_{i=1}^n (X_{i,n}[M](X_n[M]'X_n[M])^{-1}X_{i,n}[M]')^2 \le nC^2/n^2 \to 0$, where we have omitted another constant that depends only on τ and C. We have thus verified Condition 1.

To show that the proposed estimators $\hat{\sigma}^2_{j,\mathbb{M},n}$ of (3.3) consistently overestimate the asymptotic variances of the MLE, we verify the assumptions of Proposition 2.6 with

$$\tilde{g}_{i,n,\mathbb{M}}(y) = \hat{H}_{\mathbb{M},n}(y)^{-1}X_{i,n}[M]'\hat{\psi}_{i,n,\mathbb{M}}(y),$$

and $a_{i,n,\mathbb{M}} = \mathbb{E}_n[H^*_{\mathbb{M},n}]^{-1}X_{i,n}[M]'\mathbb{E}_n[\psi^*_{i,n,\mathbb{M}}]$, where $\hat{H}_{\mathbb{M},n}(y) = H_{\mathbb{M},n}(y, \hat{\beta}_{\mathbb{M},n}(y))$ and $\hat{\psi}_{i,n,\mathbb{M}}(y) = y_i\dot{\phi}_1(X_{i,n}[M]\hat{\beta}_{\mathbb{M},n}(y)) + (1-y_i)\dot{\phi}_2(X_{i,n}[M]\hat{\beta}_{\mathbb{M},n}(y))$. In particular, $\hat{g}_{i,n,\mathbb{M}}(y) = \tilde{g}_{i,n,\mathbb{M}}(y) - a_{i,n,\mathbb{M}}$. First note that for any $\gamma \in \mathbb{R}$,

$$y_i\dot{\phi}_1(\gamma) + (1-y_i)\dot{\phi}_2(\gamma) = \frac{\dot{h}(\gamma)}{h(\gamma)(1-h(\gamma))}(y_i - h(\gamma)),$$

so that $\hat{\psi}_{i,n,\mathbb{M}}(y) = \hat{u}_{i,\mathbb{M}}(y)$ and the diagonal entries of $\tilde{S}_{\mathbb{M},n}$ can, indeed, be represented as

$$\hat{\sigma}^2_{j,\mathbb{M},n}(y) = \sum_{i=1}^n \left[\tilde{g}^{(j)}_{i,n,\mathbb{M}}(y)\right]^2,$$

for $j = 1, \ldots, m(\mathbb{M})$, as required for the application of Proposition 2.6. Next, consider

$$
\begin{aligned}
|g^{(j)}_{i,n,\mathbb{M}}(y_i) - \hat{g}^{(j)}_{i,n,\mathbb{M}}(y)|^2 &= |e_{m(\mathbb{M})}(j)'\Big(\mathbb{E}_n[H^*_{\mathbb{M},n}]^{-1}X_{i,n}[M]'\psi^*_{i,n,\mathbb{M}}(y_i) \\
&\quad - \hat{H}_{\mathbb{M},n}(y)^{-1}X_{i,n}[M]'\hat{\psi}_{i,n,\mathbb{M}}(y)\Big)|^2 \\
&\leq 2|e_{m(\mathbb{M})}(j)'(\mathbb{E}_n[H^*_{\mathbb{M},n}]^{-1} - \hat{H}_{\mathbb{M},n}(y)^{-1})X_{i,n}[M]'\psi^*_{i,n,\mathbb{M}}(y_i)|^2 \\
&\quad + 2|e_{m(\mathbb{M})}(j)'\hat{H}_{\mathbb{M},n}(y)^{-1}X_{i,n}[M]'(\psi^*_{i,n,\mathbb{M}}(y_i) - \hat{\psi}_{i,n,\mathbb{M}}(y))|^2 \\
&\leq 2|e_{m(\mathbb{M})}(j)'(\mathbb{E}_n[H^*_{\mathbb{M},n}]^{-1} - \hat{H}_{\mathbb{M},n}(y)^{-1})X_{i,n}[M]'|^2\overline{K'}(\tau, C) \\
&\quad + 2|e_{m(\mathbb{M})}(j)'\hat{H}_{\mathbb{M},n}(y)^{-1}X_{i,n}[M]'(\psi^*_{i,n,\mathbb{M}}(y_i) - \hat{\psi}_{i,n,\mathbb{M}}(y))|^2.
\end{aligned}
$$

We want to show that

$$
\frac{\sum_{i=1}^n \left[g^{(j)}_{i,n,\mathbb{M}}(y_i) - \hat{g}^{(j)}_{i,n,\mathbb{M}}(y)\right]^2}{\mathbb{V}_n[r^{(j)}_{n,\mathbb{M}}]} \leq \frac{\sum_{i=1}^n \left[g^{(j)}_{i,n,\mathbb{M}}(y_i) - \hat{g}^{(j)}_{i,n,\mathbb{M}}(y)\right]^2}{\underline{K'}(\tau, C)\|X_n[M]\mathbb{E}_n[H^*_{\mathbb{M},n}]^{-1}e_{m(\mathbb{M})}(j)\|^2},
$$

converges to zero in $\mathbb{P}_n$-probability, as in (2.4), where the inequality follows by the same argument as in (C.8). Hence, it suffices to show that

$$
\frac{\|X_n[M](\mathbb{E}_n[H^*_{\mathbb{M},n}]^{-1} - \hat{H}_{\mathbb{M},n}(y)^{-1})e_{m(\mathbb{M})}(j)\|^2}{\|X_n[M]\mathbb{E}_n[H^*_{\mathbb{M},n}]^{-1}e_{m(\mathbb{M})}(j)\|^2} \tag{C.14}
$$

and

$$
\frac{\|X_n[M]\hat{H}_{\mathbb{M},n}(y)^{-1}e_{m(\mathbb{M})}(j)\|^2}{\|X_n[M]\mathbb{E}_n[H^*_{\mathbb{M},n}]^{-1}e_{m(\mathbb{M})}(j)\|^2} \max_{i=1,\ldots,n}(\psi^*_{i,n,\mathbb{M}}(y_i) - \hat{\psi}_{i,n,\mathbb{M}}(y))^2, \tag{C.15}
$$

both converge to zero in $\mathbb{P}_n$-probability. For (C.8), simply note that this expression is bounded by

$$
\|I_{m(\mathbb{M})} - \mathbb{E}_n[H^*_{\mathbb{M},n}]\hat{H}_{\mathbb{M},n}(y)^{-1}\|^2
$$

times the condition number of the matrix $\mathbb{E}_n[H^*_{\mathbb{M},n}]^{-1}X_n[M]'X_n[M]\mathbb{E}_n[H^*_{\mathbb{M},n}]^{-1}$. We have already seen above that the latter is bounded by a constant that depends only on τ and C. To see that $\mathbb{E}_n[H^*_{\mathbb{M},n}]\hat{H}_{\mathbb{M},n}(y)^{-1}$ converges to $I_{m(\mathbb{M})}$ in $\mathbb{P}_n$-probability, note that

$$
\begin{aligned}
&\|I_{m(\mathbb{M})} - \mathbb{E}_n[H^*_{\mathbb{M},n}]\hat{H}_{\mathbb{M},n}(y)^{-1}\|^2 \\
&\quad \leq \|I_{m(\mathbb{M})} - \mathbb{E}_n[H^*_{\mathbb{M},n}]^{1/2}\hat{H}_{\mathbb{M},n}(y)^{-1}\mathbb{E}_n[H^*_{\mathbb{M},n}]^{1/2}\|^2\frac{\lambda_{\max}(\mathbb{E}_n[H^*_{\mathbb{M},n}])}{\lambda_{\min}(\mathbb{E}_n[H^*_{\mathbb{M},n}])} \\
&\quad \leq \|I_{m(\mathbb{M})} - \mathbb{E}_n[H^*_{\mathbb{M},n}]^{1/2}\hat{H}_{\mathbb{M},n}(y)^{-1}\mathbb{E}_n[H^*_{\mathbb{M},n}]^{1/2}\|^2 K(\tau, C),
\end{aligned}
$$

by Lemma C.2(iii). Furthermore, for every $\varepsilon > 0$ and $\delta > 0$,

$$
\begin{aligned}
&\mathbb{P}_n\left(\|\mathbb{E}_n[H^*_{\mathbb{M},n}]^{-1/2}\hat{H}_{\mathbb{M},n}(y)\mathbb{E}_n[H^*_{\mathbb{M},n}]^{-1/2} - I_{m(\mathbb{M})}\| > \varepsilon\right) \\
&\quad \leq \mathbb{P}_n\Big(\sup_{\beta \in N_{\mathbb{M},n}(\delta)} \|\mathbb{E}_n[H^*_{\mathbb{M},n}]^{-1/2}H_{\mathbb{M},n}(y, \beta)\mathbb{E}_n[H^*_{\mathbb{M},n}]^{-1/2} - I_{m(\mathbb{M})}\| > \varepsilon\Big) \\
&\qquad + \mathbb{P}_n\Big(\hat{\beta}_{\mathbb{M},n} \notin N_{\mathbb{M},n}(\delta)\Big),
\end{aligned}
$$

and we have already seen before that this entails convergence to zero of the probability on the left-hand-side of the previous display. We conclude that (C.8) does converge to zero in $\mathbb{P}_n$-probability. To establish the same convergence also for (C.8), first note that it follows from the previous arguments that the fraction in that display is bounded in $\mathbb{P}_n$-probability. Finally, we have to establish the desired convergence for the maximum in that display. But this follows from the continuity of $\dot{\phi}_1$ and $\dot{\phi}_2$ on $\mathbb{R}$, the bound on $|X_{i,n}[M]\beta^*_{\mathbb{M},n}|$ from Lemma C.1 and the consistency of Lemma 3.9. Therefore, Proposition 2.6 shows that (2.5) is satisfied. Note that $\mathbb{VC}_n(r_n)$ has rank no larger than $\min(k,n)$, where $r_n = (r_{n,\mathbb{M}})_{\mathbb{M}\in\mathsf{M}_n}$. Hence, Theorem 2.5, together with Lemma 2.8, finishes the proof. □

C.9. Canonical link function

Corollary C.3. *In the setting of Theorem 3.10, if $\mathcal{H}$ contains only the canonical link function $h^{(c)}(\gamma) = e^\gamma/(1+e^\gamma)$, then the confidence intervals*

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

satisfy

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

Proof. From the first few lines of the proof of Theorem 3.10, we see that the $m(\mathbb{M})$-dimensional sub-vector $r_{n,\mathbb{M}}(y)$ of $r_n(y)$ that corresponds to the model $\mathbb{M} \triangleq (h,M) \in \{h^{(c)}\}\times\mathcal{I}$, is given by

$$r_{n,\mathbb{M}}(y) = \mathbb{E}_n[H^*_{\mathbb{M},n}]^{-1}s^*_{\mathbb{M},n}(y),$$

where $H^*_{\mathbb{M},n}(y) := H_{\mathbb{M},n}(y,\beta^*_{\mathbb{M},n})$, $s^*_{\mathbb{M},n}(y) := s_{\mathbb{M},n}(y,\beta^*_{\mathbb{M},n})$ and $s_{\mathbb{M},n}(y,\beta) := \partial\ell_{\mathbb{M},n}(y,\beta)/\partial\beta = X_n[M]'C_{\mathbb{M},n}(y,\beta)$, where $C_{\mathbb{M},n}(y,\beta)$ is an $n\times 1$ vector with i-th entry given by $y_i\dot{\phi}_1(X_{i,n}[M]\beta) + (1-y_i)\dot{\phi}_2(X_{i,n}[M]\beta)$. But it is easy to see that for $h = h^{(c)}$, $\phi_1(\gamma)-\phi_2(\gamma) = \gamma$ and thus $\dot{\phi}_1(\gamma)-\dot{\phi}_2(\gamma) = 1$, so that in this case the matrix $\mathbb{VC}_n(r_n)$ reduces to

$$\mathbb{VC}_n(r_n) = \left(\mathbb{E}_n[H^*_{\mathbb{M}_s,n}]^{-1}X_n[M_s]'\mathbb{VC}(y_n)X_n[M_t]\mathbb{E}_n[H^*_{\mathbb{M}_t,n}]^{-1}\right)^d_{s,t=1}.$$

The rank of this matrix is not larger than $\min(k,p)$, so by Lemma 2.8 we obtain the smaller bound $K_{1-\alpha}(\mathrm{corr}(\mathbb{VC}_n(r_n))) \leq B_\alpha(\min(k,p),k)$. □