Submitted to the Annals of Statistics

VALID POST-SELECTION INFERENCE

By Richard Berk, Lawrence Brown[*,†], Andreas Buja[*], Kai Zhang[*] and Linda Zhao[*]

The Wharton School, University of Pennsylvania

It is common practice in statistical data analysis to perform data-driven variable selection and derive statistical inference from the resulting model. Such inference enjoys none of the guarantees that classical statistical theory provides for tests and confidence intervals when the model has been chosen a priori. We propose to produce valid "post-selection inference" by reducing the problem to one of simultaneous inference. Simultaneity is required for all linear functions that arise as coefficient estimates in all submodels. By purchasing "simultaneity insurance" for all possible submodels, the resulting post-selection inference is rendered universally valid under all possible model selection procedures. This inference is therefore generally conservative for particular selection procedures, but it is always less conservative than full Scheffé protection. Importantly it does *not* depend on the truth of the selected submodel, and hence it produces valid inference even in wrong models. We describe the structure of the simultaneous inference problem and give some asymptotic results.

1. Introduction — The Problem with Statistical Inference after Model Selection. Classical statistical theory grants validity of statistical tests and confidence intervals assuming a wall of separation between the selection of a model and the analysis of the data being modeled. In practice, this separation rarely exists and more often a model is "found" by a data-driven selection process. As a consequence inferential guarantees derived from classical theory are invalidated. Among model selection methods that are problematic for classical inference, *variable selection* stands out because it is regularly taught, commonly practiced, and highly researched as a technology. Even though statisticians may have a general awareness that the data-driven selection of variables (predictors, covariates) must somehow affect subsequent classical inference from F- and t-based tests and confidence intervals, the practice is so pervasive that it appears in classical undergraduate textbooks on statistics such as Moore and McCabe (2003).

[*] Research supported in part by NSF Grant DMS-1007657.

[†] Corresponding Author, lbrown@wharton.upenn.edu.

AMS 2000 subject classifications: Primary 62J05, 62J15

Keywords and phrases: Linear Regression, Model Selection, Multiple Comparison, Family-wise Error, High-dimensional Inference, Sphere Packing

The reason for the invalidation of classical inference guarantees is that a data-driven variable selection process produces a model that is itself stochastic, and this stochastic aspect is not accounted for by classical theory. Models become stochastic when the stochastic component of the data is involved in the selection process. (In regression with fixed predictors the stochastic component is the response.) Models are stochastic in a well-defined way when they are the result of formal variable selection procedures such as stepwise or stagewise forward selection or backward elimination or all-subset searches driven by complexity penalties (such as C_p, AIC, BIC, risk-inflation, LASSO, ...) or prediction criteria such as cross-validation, or recent proposals such as LARS and the Dantzig selector (for an overview see, for example, Hastie, Tibshirani, and Friedman (2009)). Models are also stochastic but in an ill-defined way when they are informally selected through visual inspection of residual plots or normal quantile plots or generally through activities that may be characterized as "data snooping". Finally, models become stochastic in the most opaque way when their selection is affected by human intervention based on post hoc considerations such as "in retrospect only one of these two variables should be in the model" or "it turns out the predictive benefit of this variable is too weak to warrant the cost of collecting it." In practice, all three modes of variable selection may be excercised in the same data analysis: multiple runs of one or more formal search algorithms may be performed and compared, the parameters of the algorithms may be subjected to experimentation, and the results may be critiqued with graphical diagnostics; a round of fine-tuning based on substantive deliberations may finalize the analysis.

Posed so starkly, the problems with statistical inference after variable selection may well seem insurmountable. At a minimum, one would expect technical solutions to be possible only when a formal selection algorithm is (1) well-specified (1a) in advance and (1b) covering all eventualities, (2) strictly adhered to in the course of data analysis, and (3) not "improved" on by informal and post-hoc elements. It may, however, by unrealistic to expect this level of rigor in most data analysis contexts, with the exception of well-conducted clinical trials. The real challenge is therefore to devise statistical inference that is valid following any type of variable selection, be it formal, informal, post hoc, or a combination thereof. Meeting this challenge with a relatively simple proposal is the goal of this article. This proposal for valid ***Po***st-***S***election ***I***nference, or "***PoSI***" for short, consists of a large-scale family-wise error guarantee that can be shown to account for all types of variable selection, including those of the informal and post-hoc varieties. On the other hand, the proposal is no more conservative than necessary to account for selection, and in particular it can be shown to be less conservative than Scheffé's simultaneous inference.

The framework for our proposal is in outline as follows — details to be elaborated in subsequent sections: We consider linear regression with predictor variables whose values are considered fixed, and with a response variable that has normal and homoskedastic errors. The framework does not require that any of the eligible linear models is correct, not even the full model, as long as a valid error estimate is available. We assume that the selected model is the result of some procedure that makes use of the response, but the procedure does not need to be fully specified. A crucial aspect of the framework concerns the use and interpretation of the selected model: We assume that, after variable selection is completed, the selected predictor variables — and only they — will be relevant; all others will be eliminated from further consideration. This assumption, seemingly innocuous and natural, has critical consequences: It implies that statistical inference will be sought for the coefficients of the selected predictors only and in the context of the selected model only. Thus the appropriate targets of inference are the best linear coefficients within the selected model, where each coefficient is adjusted for the presence of all other included predictors but not those that were eliminated. Therefore the coefficient of an included predictor generally requires inference that is specific to the model in which it appears. Summarizing in a motto, a difference in adjustment implies a difference in parameters and hence in inference. The goal of the present proposal is therefore simultaneous inference for all coefficients in all submodels. Such inference can be shown to be valid following any variable selection procedure, be it formal, informal, post hoc, fully or only partly specified.

Problems associated with post-selection inference were recognized long ago, for example, by Buehler and Fedderson (1963), Brown (1967), Olshen (1973), Sen (1979), Sen and Saleh (1987), Dijkstra and Veldkamp (1988), Pötscher (1991), Kabaila (1998). More recently these problems have been the subject of incisive analyses by Leeb and Pötscher (2003; 2005; 2006a; 2006b; 2008a; 2008b), Kabaila and Leeb (2006), Leeb (2006), and Pötscher and Leeb (2009).

This article proceeds as follows: Section 2 starts by outlining some unsolvable difficulties of post-selection inference as they transpire from the work of Leeb and Pötscher cited above (Section 2.1); we then rethink the assumptions underlying their analyses and lay some groundwork by proposing new (or old) meanings for regression coefficients (Section 2.2); we conclude the section by discussing assumptions with a view towards valid inference in "wrong models" (Section 2.3). Section 3 is about estimation and its targets;

Section 4 develops the methodology for PoSI confidence intervals (CIs) and tests. After some structural results for the PoSI problem in Section 5 , we show in Section 6 that with increasing number of predictors p the width of PoSI CIs can range between the asymptotic rates $O(\sqrt{\log p})$ and $O(\sqrt{p})$. We give examples for both rates and, inspired by problems in sphere packing and covering, we give upper bounds for the limiting constant in the $O(\sqrt{p})$ case. In Section 7 we discuss computations that can be used to determine the required CI widths for any predictor matrix to satisfactory accuracy when $p \leq 20$, and we also provide computations for non-asymptotic upper bounds on these constants that work for larger p. We conclude with a discussion in Section 8. Some proofs are deferred to the appendix.

2. Model Selection Re-Interpreted.

2.1. *Post-Selection Inference for Full Model Parameters — a Dead End.* It is a natural intuition that model selection distorts inference by distorting sampling distributions of parameter estimates: One expects that estimates in selected models tend to generate more Type I errors than conventional theory would suggest because the typical selection procedure favors models with strong, hence highly significant predictors. This intuition correctly points toward a multiplicity problem which would tend to become more severe as the number of predictors subject to selection increases. This problem will be addressed here with a simultaneous inference approach.

A second problem with inference after model selection is pointed out by Leeb and Pötscher in the above referenced series of articles. The problem exists even in a two-predictor situation, as illustrated by Leeb and Pötscher (2005): They analyze a case with a predictor that is protected from selection and a covariate that is subject to selection, and they provide an explicit finite-sample formula for the sampling distribution of the coefficient estimate of the protected predictor, as the covariate is randomly selected/deselected according to a BIC-equivalent test to grant consistent model selection (ibid., p. 29). The analysis reveals in very graphic ways (ibid., Figure 2) that the sampling distribution depends critically on the unknown true coefficient of the covariate and the sample size, with egregious deviations from the fixed-model sampling distribution that ranges from bi-modality to approximate normality with inflated variance. Because the true covariate slope is not known, there is no way of determining whether the sample size places the sampling distribution in this realm of non-normality.

Generalizing to arbitrary linear models Leeb and Pötscher (2003; 2005; 2006a; 2006b; 2008a; 2008b) show that sampling distributions cannot be estimated after model selection, not even asymptotically. Ironically, the asymptotics that describe the devious finite sample behavior of sampling distributions best are those based on consistent model selection. They show that asymptotic normality is riddled with extreme non-uniformity of convergence and that risk functions behave erratically when telescoping true slopes to zero so as to approach submodels. Leeb and Pötscher (2005, p. 27) arrive at the following conclusion: "the post-model-selection estimator ... is nothing else than a variant of Hodges' so-called superefficient estimator."

It is little comfort that these problems are non-existent for perfectly orthogonal regression designs (Leeb and Pötscher 2005, p. 43f). In all except perfectly designed experiments with orthogonal predictors there is some degree of collinearity, and one of the purposes of model selection is to weed out predictor redundancies caused by partial collinearity. Leeb and Pötscher's analysis is compelling within their framework, but the intractable situation they expose suggests a need to renegotiate the assumptions that underlie their framework.

Leeb and Pötscher (ibid.), like many authors in this area, make the fundamental assumption that all estimation is in the full model. Thus, if a model selection procedure excludes a predictor, this is interpreted as forcing the estimate of its slope to zero. Consequently, a slope estimate $\hat{\beta}_j$ of a predictor is always defined, whether it is selected or not: $\hat{\beta}_j$ is the LS estimate in the selected submodel if the j^{th} predictor is included, and it is zero otherwise. Either way, the resulting value is interpreted as an estimate of β_j in the full model. A parallel consequence is that in this interpretation the coefficient of a predictor has always the same meaning as a full-model parameter, irrespective of which covariates are selected or excluded. It is under this framework that post-selection estimators can be interpreted as generalized superefficient Hodges estimators with the ensuing problems of non-uniformity (Leeb and Pötscher 2005). This problem can also be seen as an inferential analog of the "omitted variables bias" problem well-known in econometrics (see, for example, Angrist and Pischke 2009).

2.2. *The Meaning of Regression Coefficients.* Our solution to the inferential problem with the "omitted variables bias" is to assert that submodels have their own separate parameters, and it is these that are being estimated in the selected submodels. Our discussion starts with the following questions:

(1) When we select submodels in practice, do we think of excluded predictors as having a zero slope?
(2) Does the full model necessarily have special status?
(3) Can a slope estimate be interpreted as estimating the same target parameter, regardless of what the other predictors are?

The short answers are:

(1) The slopes of excluded predictors are not zero; they are not defined and therefore don't exist.
(2) The full model has no special status.
(3) The meaning of a slope depends on which predictors are in the selected model.

These answers call for some elaborations:

As for (1), assigning a zero value to predictors that are not in the model is an elegant technical device, but it is not something that describes how we think or even how we *should* think about slopes and their estimates. The PoSI framework we describe in Section 3 will not require zero slope fill-ins.

As for (2), the full model cannot be argued to have generally special status because there is generally a question of predictor redundancy. It is a common experience that models are proposed on theoretical grounds but found on empirical grounds to have their predictors hopelessly entangled by collinearities that permit little meaningful statistical inference. This is best illustrated with a concrete example (inspired by Mosteller and Tukey (1977), p. 326f): Consider a study of performance of students in a large school system. Interested in socio-economic factors, investigators wish to pin down the predictors that are most strongly associated with children's success in school: father's and mother's highest education levels, their high school GPAs, their SAT scores, their frequencies of intensive reading, the perceived importance they each assign to education, and so on. There should be little surprise that, if all these predictors are included in the model, the overall test rejects but none of the individual predictors is statistically significant. Informal model selection, however, may show that each predictor is highly statistically significant if retained alone in the model. The obvious reason is that these predictors measure essentially the same trait in parents, hence are highly collinear with each other. As a consequence, the full model is not viable in the first place. This situation is not limited to the social sciences: in gene expression studies it may well occur that numerous sites have a tendency to be expressed concurrently, hence as predictors in disease studies they will be hopelessly confounded. The bias in favor of full models may be particularly strong in econometrics where there is a "notion that a longer regression ... has a causal interpretation, while a shorter regression does not" (Angrist and Pischke 2009, p. 59). Even in causal models, however, there is a possibility that included adjustor variables will "adjust away" some of the causal variables of interest. In any creative observational study that delves into novel predictors it is difficult to determine whether the curse of collinearity will not force some rethinking. In conclusion, whenever predictor redundancy is a potential issue, it cannot a priori be claimed that the full model provides the parameters of primary interest.

As for (3), we do teach that the meaning of a slope depends on what other predictors are in the chosen model: "the slope is the average difference in the response for a unit difference in the predictor, *at fixed levels of all other predictors*." This last condition is sometimes rendered as "*adjusted for all other predictors*" and called the "Ceteris Paribus" clause (xxx need econometrics ref: Green 19xx). It is an essential part of the meaning of a slope. That there is a difference in meaning when there is a difference in covariates is most drastically evident when there is a case of Simpson's paradox. This is again best illustrated with a concrete example: A company is introducing a new high-tech device and conducts a consumer survey that includes a response for self-reported purchase likelihood (PL), as well as two predictors, Age and $Income$. We consider a model with Age alone and one with both Age and $Income$. [Note that the smaller model cannot be disregarded as "wrong". If $Income$ is difficult to measure, it may be useful to rely on the equation with Age alone. Further, if the variables have a jointly normal distribution, every linear submodel is "correct".] Now, it is sensibly anticipated that younger respondents will rate themselves with higher PL, but a regression of PL on Age alone produces a significantly positive slope estimate, indicating that older repondents have higher PL. On the other hand, a regression of PL on both Age and $Income$ yields a significantly negative slope estimate for Age, indicating that, *comparing only respondents at the same $Income$ level*, younger respondents have indeed higher PL. This instance of Simpson's paradox is enabled by a positive collinearity between Age and $Income$. Must we use the full model? Not if the improvement in R^2 is practically irrelevant even though $Income$ may be statistically significant (apart from the issue of availability of $Income$ measurements). Is the marginal slope of PL on Age an estimate of the $Income$-adjusted slope on Age? Certainly not — the two slopes answer very different questions, apart from having opposite signs. In conclusion, *differences in adjustment result in different parameters.*

From these considerations follows a framework in which the full model is no longer the sole provider of parameters, where much rather each submodel defines its own. The consequences of this view will be fully elaborated in Section 3.

2.3. *Assumptions, "Wrong Models", and Error Estimates.* We state assumptions for estimation and for the construction of valid tests and CIs. A major goal is to prepare the ground for valid statistical inference after model selection in "first order wrong models".

We consider a quantitative response vector $\mathbf{Y} \in \mathbb{R}^n$, assumed random, and a full predictor matrix $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p) \in \mathbb{R}^{n \times p}$, assumed fixed. We allow $\mathbf{X}$ to be of non-full rank, and n and p to be arbitrary. In particular, we allow $n < p$. Throughout the article we let

$$d \triangleq \operatorname{rank}(\mathbf{X}) = \dim(\operatorname{span}(\mathbf{X})), \qquad \text{hence} \quad d \leq \min(n, p). \tag{2.1}$$

Due to frequent reference we call $d = p$ $(\leq n)$ the "**classical case**".

It is common practice to assume the full model $\mathbf{Y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$ to be correct. In the present framework, however, first-order correctness, $\mathbf{E}[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}$, will not be assumed. By implication, first-order correctness of any submodel will not be assumed either. Effectively,

$$\boldsymbol{\mu} \triangleq \mathbf{E}[\mathbf{Y}] \in \mathbb{R}^n \tag{2.2}$$

is allowed to be unconstrained and, in particular, need not reside in the column space of $\mathbf{X}$. In other words, the model given by $\mathbf{X}$ is allowed to be "first-order wrong", and hence we are in a well-defined sense serious about G.E.P. Box' famous dictum. What he calls "wrong models", however, we prefer to call "approximations": All predictor matrices $\mathbf{X}$ provide approximations to $\boldsymbol{\mu}$, some better than others, but the degree of approximation plays no role in the clarification of statistical inference. We will echo Box as follows: all models are mere approximations, yet some are useful. The reason for elaborating this point here is that after model selection the case for "correct models" is even more questionable than for full models (Leeb and Pötscher 2003, p. 101), but even for full models it can be abandoned. As we proceed with estimation and inference guarantees in the absence of first-order correctness we will rely on assumptions as follows:

- For estimation (Section 3), we will only need the existence of $\boldsymbol{\mu} = \mathbf{E}[\mathbf{Y}]$.
- For testing and CI guarantees (Section 4), we will make conventional second order and distributional assumptions:

$$\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbf{I}). \tag{2.3}$$

The assumptions (2.3) of homoskedasticity and normality are as questionable as first order correctness, and we will report elsewhere on approaches that avoid them. In the present work, we choose to follow the large model selection literature that relies on the technical advantages of assuming homoskedastic and normal errors.

Accepting the assumption (2.3), we address the issue of estimating the error variance σ^2, because valid tests and CIs require a valid estimate $\hat{\sigma}^2$ of σ^2 that is independent of LS estimates. In the classical case, the most common way to assert such an estimate is to assume that the full model is first order correct, $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ in addition to (2.3), in which case the mean squared error (MSE), $\hat{\sigma}_F^2 = \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2/(n-p)$, of the full model will do. However, other possibilities for producing a valid estimate $\hat{\sigma}^2$ exist, and they may allow relaxing the assumption of first order correctness:

- Exact replications of the response obtained under identical conditions might be available in sufficient numbers. An estimate $\hat{\sigma}^2$ can be obtained as the MSE of the one-way ANOVA of the groups of replicates.
- In general, a larger linear model than the full model might be considered as correct, hence $\hat{\sigma}^2$ could be the MSE from this larger model.
- A different possibility is to use another dataset, similar to the one currently being analyzed, to produce an independent estimate $\hat{\sigma}^2$ by whatever valid estimation method.
- A special case of the preceding is a split-sample approach whereby one part of the data is reserved for producing $\hat{\sigma}^2$ and the other part for estimating coefficients, selecting models, and carrying out post-model selection inference.

The purpose of pointing out these possibilities is to separate at least in principle the issue of first-order model incorrectness from the issue of valid and independent error estimation under the assumption (2.3). This separation puts the case $n < p$ within our framework as the valid and independent estimation of σ^2 is a problem faced by all $n < p$ approaches.

3. Estimation and its Targets in Submodels.

Following Section 2.2, the meaning and numeric value of a regression coefficient depends on what the other predictors in the model are. This statement requires a qualification: it assumes that the predictors are non-orthogonal/partially collinear. If they are perfectly pairwise orthogonal, as in some designed experiments or in function fitting with orthogonal basis functions, a coefficient has the same identity across all submodels, both in meaning and in value, because adjustment of predictors for each other and the ceteris paribus clause become vacuous. This article is hence largely a story of (partial) collinearity.

3.1. *Multiplicity of Regression Coefficients.*

We will give meaning to LS estimators and their targets in the absence of any assumptions other than the existence of $\boldsymbol{\mu} = \mathbf{E}[\mathbf{Y}]$, which in turn is permitted to be entirely unconstrained in $\mathbb{R}^n$. Besides resolving the issue of estimation in "first order wrong models", the major point here is to follow up on the idea that the regression coefficient of a predictor generates different parameters in different submodels. As each predictor appears in 2^{p-1} submodels, the p regression coefficients of the full model generally proliferate into a plethora of as many as $p\,2^{p-1}$ distinct regression coefficients according to the submodels they appear in. We start with notation.

To denote a submodel we use the (non-empty) index set $\mathrm{M} = \{j_1, j_2, ..., j_m\} \subset \mathrm{M}_F = \{1, \ldots, p\}$ of the predictors $\mathbf{X}_{j_i}$ in the submodel; the size of the submodel is $m = |\mathrm{M}|$ and that of the full model is $p = |\mathrm{M}_F|$. Let $\mathbf{X}_\mathrm{M} = (\mathbf{X}_{j_1}, ..., \mathbf{X}_{j_m})$ denote the $n \times m$ submatrix of $\mathbf{X}$ with columns indexed by M. We will assume that only submodels M are considered for which $\mathbf{X}_\mathrm{M}$ is of full rank:

$$\mathrm{rank}(\mathbf{X}_\mathrm{M}) \;=\; m \;\leq\; d.$$

We let $\hat{\boldsymbol{\beta}}_\mathrm{M}$ be the unique least squares estimate in M:

$$\hat{\boldsymbol{\beta}}_\mathrm{M} = (\mathbf{X}_\mathrm{M}^T\mathbf{X}_\mathrm{M})^{-1}\mathbf{X}_\mathrm{M}^T\mathbf{Y}. \tag{3.1}$$

Now that $\hat{\boldsymbol{\beta}}_\mathrm{M}$ is an estimate, what is it estimating? A conclusion from Section 2.1 is that $\hat{\boldsymbol{\beta}}_\mathrm{M}$ does not estimate the coefficients in the full model. Because any larger model could have been the full model, we generalize by asserting that $\hat{\boldsymbol{\beta}}_\mathrm{M}$ does not estimate parameters in any other model than M itself. In M, it is natural to ask that $\hat{\boldsymbol{\beta}}_\mathrm{M}$ be an unbiased estimate of its target:

$$\begin{aligned} \boldsymbol{\beta}_\mathrm{M} \;\triangleq\; \mathbf{E}[\hat{\boldsymbol{\beta}}_\mathrm{M}] \;&=\; (\mathbf{X}_\mathrm{M}^T\mathbf{X}_\mathrm{M})^{-1}\mathbf{X}_\mathrm{M}^T\,\mathbf{E}[\mathbf{Y}] \\ &=\; \underset{\boldsymbol{\beta}'\in\mathbb{R}^m}{\mathrm{argmin}}\, \|\boldsymbol{\mu} - \mathbf{X}_\mathrm{M}\boldsymbol{\beta}'\|^2 \end{aligned} \tag{3.2}$$

This definition requires only the existence of $\boldsymbol{\mu} = \mathbf{E}[\mathbf{Y}]$ but no other assumptions. In particular there is no need to assume first order correctness of M or M_F. Nor does it matter to what degree M provides a good approximation to $\boldsymbol{\mu}$ in terms of approximation error $\|\boldsymbol{\mu} - \mathbf{X}_\mathrm{M}\boldsymbol{\beta}_\mathrm{M}\|^2$. Asserting that the model M is "correct" would mean $\boldsymbol{\mu} \in \mathrm{span}(\mathbf{X}_\mathrm{M})$ or equivalently the approximation error vanishes; in this case $\boldsymbol{\beta}_\mathrm{M}$ would be the "true" parameter.

In the classical case $d = p \leq n$, we can define the target of the full-model estimate $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ as a special case of (3.2) with $\mathrm{M} = \mathrm{M}_F$:

$$\boldsymbol{\beta} \;\triangleq\; \mathbf{E}[\hat{\boldsymbol{\beta}}] \;=\; (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{E}[\mathbf{Y}]. \tag{3.3}$$

In the general (non-classical) case, let $\boldsymbol{\beta}$ be any (possibly non-unique) minimizer of $\|\boldsymbol{\mu} - \mathbf{X}\boldsymbol{\beta}'\|^2$; the link between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_\mathrm{M}$ is as follows:

$$\boldsymbol{\beta}_\mathrm{M} \;=\; (\mathbf{X}_\mathrm{M}^T\mathbf{X}_\mathrm{M})^{-1}\mathbf{X}_\mathrm{M}^T\mathbf{X}\boldsymbol{\beta}. \tag{3.4}$$

Thus the target $\boldsymbol{\beta}_{\mathrm{M}}$ is an estimable linear function of $\boldsymbol{\beta}$, without any first-order assumptions. Equation (3.4) follows from $\operatorname{span}(\mathbf{X}_{\mathrm{M}}) \subset \operatorname{span}(\mathbf{X})$.

Notation: To distinguish regression coefficients as a function of the model they appear in, we write $\beta_{j\cdot\mathrm{M}} = \mathbf{E}[\hat{\beta}_{j\cdot\mathrm{M}}]$ for the components of $\boldsymbol{\beta}_{\mathrm{M}} = \mathbf{E}[\hat{\boldsymbol{\beta}}_{\mathrm{M}}]$ with $j \in \mathrm{M}$. An important convention we adopt throughout this article is that the index j of a coefficient refers to the coefficient's index in the original full model M_F: $\beta_{j\cdot\mathrm{M}}$ for $j \in \mathrm{M}$ refers not to the j'th coordinate of $\boldsymbol{\beta}_{\mathrm{M}}$, but to the coordinate of $\boldsymbol{\beta}_{\mathrm{M}}$ corresponding to the j'th predictor $\mathbf{X}_j$ in the full predictor matrix $\mathbf{X}$. We refer to this convention as "**full model indexing**".

3.2. *"Omitted Variables Bias".* By allowing each $\hat{\beta}_{j\cdot\mathrm{M}}$ to estimate its own target $\beta_{j\cdot\mathrm{M}}$ and thereby relieving $\hat{\beta}_{j\cdot\mathrm{M}}$ of the burden of estimating the parameter β_j in the full model, we eschew the problem of "omitted variables bias" and with it a major driver of the problems analyzed by Leeb and Pötscher (Section 2.1). In the present framework $\beta_j - \beta_{j\cdot\mathrm{M}}$ is not a bias as these are two different parameters that answer two different questions. Just the same, consider briefly the difference between β_j and $\beta_{j\cdot\mathrm{M}}$ in the classical case $d = p \leq n$. Compare the following two definitions:

$$\boldsymbol{\beta}_{\mathrm{M}} \triangleq \mathbf{E}[\hat{\boldsymbol{\beta}}_{\mathrm{M}}] \quad \text{and} \quad \boldsymbol{\beta}^{\mathrm{M}} \triangleq (\beta_j)_{j\in\mathrm{M}}, \tag{3.5}$$

the latter being the coefficients β_j from the full model M_F subsetted to the submodel M. While $\hat{\boldsymbol{\beta}}_{\mathrm{M}}$ estimates $\boldsymbol{\beta}_{\mathrm{M}}$, it does *not* generally estimate $\boldsymbol{\beta}^{\mathrm{M}}$. The difference $\boldsymbol{\beta}^{\mathrm{M}} - \boldsymbol{\beta}_{\mathrm{M}}$ is the vectorized "omitted variables bias".

In general, the definition of $\boldsymbol{\beta}_{\mathrm{M}}$ involves $\mathbf{X}$ and all of $\boldsymbol{\beta}$, not just $\boldsymbol{\beta}^{\mathrm{M}}$, through (3.4). A little algebra shows that $\boldsymbol{\beta}_{\mathrm{M}} = \boldsymbol{\beta}^{\mathrm{M}}$ if and only if

$$\mathbf{X}_{\mathrm{M}}^T \mathbf{X}_{\mathrm{M}^c} \boldsymbol{\beta}^{\mathrm{M}^c} = \mathbf{0}, \tag{3.6}$$

where M^c denotes the complement of M in the full model M_F. Special cases of (3.6) include: (1) the column space of $\mathbf{X}_{\mathrm{M}}$ is orthogonal to that of $\mathbf{X}_{\mathrm{M}^c}$, and (2) $\boldsymbol{\beta}^{\mathrm{M}^c} = \mathbf{0}$, meaning that the approximation to $\boldsymbol{\mu}$ in M_F is no better than in M, or if the full model M_F is first-order correct, so is the submodel M.

3.3. *Interpreting Regression Coefficients in First-Order Incorrect Models.* The regression coefficient $\beta_{j\cdot\mathrm{M}}$ is conventionally interpreted as the "average difference in the response for a unit difference in X_j, ceteris paribus in the model M". This interpretation no longer holds when the assumption of first order correctness is given up. Instead, the phrase "average difference in the response" should be replaced with the unwieldy but more correct phrase "average difference in the response approximated in the submodel M". The reason is that the fit in the submodel M is $\hat{\mathbf{Y}}_{\mathrm{M}} = \mathbf{H}_{\mathrm{M}}\mathbf{Y}$ ($\mathbf{H}_{\mathrm{M}} =$

$\mathbf{X}_{\mathrm{M}}(\mathbf{X}_{\mathrm{M}}^T\mathbf{X}_{\mathrm{M}})^{-1}\mathbf{X}_{\mathrm{M}}^T)$ whose target is $\boldsymbol{\mu}_{\mathrm{M}} = \mathbf{E}[\hat{\mathbf{Y}}_{\mathrm{M}}] = \mathbf{H}_{\mathrm{M}}\mathbf{E}[\mathbf{Y}] = \mathbf{H}_{\mathrm{M}}\boldsymbol{\mu}$. Thus in the submodel M we estimate not the true $\boldsymbol{\mu}$ but the LS approximation $\boldsymbol{\mu}_{\mathrm{M}}$ to $\boldsymbol{\mu}$ using $\mathbf{X}_{\mathrm{M}}$: $\boldsymbol{\mu}_{\mathrm{M}} = \mathbf{X}_{\mathrm{M}}\boldsymbol{\beta}_{\mathrm{M}}$, where $\boldsymbol{\beta}_{\mathrm{M}} = \mathrm{argmin}_{\boldsymbol{\beta}'} \|\boldsymbol{\mu} - \mathbf{X}_{\mathrm{M}}\boldsymbol{\beta}'\|^2$.

A second interpretation of regression coefficients is in terms of adjusted predictors: For $j \in \mathrm{M}$ define the M-adjusted predictor $\mathbf{X}_{j\cdot\mathrm{M}}$ as the residual vector of the regression of $\mathbf{X}_j$ on all other predictors in M. Multiple regression coefficients, both estimates $\hat{\beta}_{j\cdot\mathrm{M}}$ and parameters $\beta_{j\cdot\mathrm{M}}$, can be expressed as simple regression coefficients with regard to the M-adjusted predictor:

$$\hat{\beta}_{j\cdot\mathrm{M}} \;=\; \frac{\mathbf{X}_{j\cdot\mathrm{M}}^T\mathbf{Y}}{\|\mathbf{X}_{j\cdot\mathrm{M}}\|^2}, \qquad \beta_{j\cdot\mathrm{M}} \;=\; \frac{\mathbf{X}_{j\cdot\mathrm{M}}^T\,\boldsymbol{\mu}}{\|\mathbf{X}_{j\cdot\mathrm{M}}\|^2}. \tag{3.7}$$

The left hand formula lends itself to an interpretation of $\hat{\beta}_{j\cdot\mathrm{M}}$ in terms of the well-known leverage plot which shows Y plotted against $\mathbf{X}_{j\cdot\mathrm{M}}$ and the line with slope $\hat{\beta}_{j\cdot\mathrm{M}}$. This plot is valid without first-order correctness assumption.

A third interpretation can be derived from the second: For notational reasons let $\mathbf{x} = (x_i)_{i=1\ldots n}$ be any adjusted predictor $\mathbf{X}_{j\cdot\mathrm{M}}$, so that $\hat{\beta} = \mathbf{x}^T\mathbf{Y}/\|\mathbf{x}\|^2$ and $\beta = \mathbf{x}^T\boldsymbol{\mu}/\|\mathbf{x}\|^2$ are the corresponding $\hat{\beta}_{j\cdot\mathrm{M}}$ and $\beta_{j\cdot\mathrm{M}}$. Introduce case-wise slopes through the origin, both as estimates $\hat{\beta}_{(i)} = Y_i/x_i$ and as parameters $\beta_{(i)} = \mu_i/x_i$, as well as case-wise weights $w_{(i)} = x_i^2/\sum_{i'=1\ldots n} x_{i'}^2$. Equations (3.7) are then equivalent to the following:

$$\hat{\beta} \;=\; \sum_i w_{(i)}\hat{\beta}_{(i)}, \qquad \beta \;=\; \sum_i w_{(i)}\beta_{(i)}.$$

Hence regression coefficients are weighted averages of case-wise slopes, and this interpretation holds without first-order assumptions.

4. Universally Valid Post-Selection Confidence Intervals.

4.1. *Test Statistics with One Error Estimate for All Submodels.* After defining $\boldsymbol{\beta}_{\mathrm{M}}$ as the target of the estimate $\hat{\boldsymbol{\beta}}_{\mathrm{M}}$, we consider inference for it in terms of test statistics. Following Section 2.3 we require a normal homoskedastic model for $\mathbf{Y}$, but we leave its mean $\boldsymbol{\mu} = \mathbf{E}[\mathbf{Y}]$ unspecified: $\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbf{I})$. We then have equivalently

$$\hat{\boldsymbol{\beta}}_{\mathrm{M}} \sim \mathcal{N}(\boldsymbol{\beta}_{\mathrm{M}}, \sigma^2(\mathbf{X}_{\mathrm{M}}^T\mathbf{X}_{\mathrm{M}})^{-1}) \quad \text{and} \quad \hat{\beta}_{j\cdot\mathrm{M}} \sim \mathcal{N}(\beta_{j\cdot\mathrm{M}}, \sigma^2/\|\mathbf{X}_{j\cdot\mathrm{M}}\|^2).$$

Again following Section 2.3 we assume the availability of a valid estimate $\hat{\sigma}^2$ of σ^2 that is independent of all estimates $\hat{\beta}_{j\cdot\mathrm{M}}$, and we further assume $\hat{\sigma}^2 \sim \sigma^2\chi_r^2/r$ for r degrees of freedom. If the full model is assumed correct,

$n > p$ and $\hat{\sigma}^2 = \hat{\sigma}_F^2$, then $r = n - p$. In the limit $r \to \infty$ we obtain $\hat{\sigma} = \sigma$, the case of known σ, which will be used starting with Section 6.

Let $t_{j\cdot\mathrm{M}}$ denote a t-ratio for $\beta_{j\cdot\mathrm{M}}$ that uses $\hat{\sigma}$ irrespective of the submodel M:

$$t_{j\cdot\mathrm{M}} \triangleq \frac{\hat{\beta}_{j\cdot\mathrm{M}} - \beta_{j\cdot\mathrm{M}}}{\left((\mathbf{X}_\mathrm{M}^T\mathbf{X}_\mathrm{M})^{-1}\right)_{jj}^{\frac{1}{2}} \hat{\sigma}} = \frac{\hat{\beta}_{j\cdot\mathrm{M}} - \beta_{j\cdot\mathrm{M}}}{\hat{\sigma}/\|\mathbf{X}_{j\cdot\mathrm{M}}\|} = \frac{(\mathbf{Y} - \boldsymbol{\mu})^T\mathbf{X}_{j\cdot\mathrm{M}}}{\hat{\sigma}\|\mathbf{X}_{j\cdot\mathrm{M}}\|}. \tag{4.1}$$

[According to full model indexing, $(...)_{jj}$ refers to the diagonal element corresponding to $\mathbf{X}_j$.] The quantity $t_{j\cdot\mathrm{M}} = t_{j\cdot\mathrm{M}}(\mathbf{Y})$ has a central t-distribution with r degrees of freedom. Essential is that the standard error estimate in the denominator of (4.1) does *not* involve the MSE $\hat{\sigma}_\mathrm{M}$ from the submodel M, for two reasons:

- We do not assume that the submodel M is first-order correct; therefore each MSE $\hat{\sigma}_\mathrm{M}^2$ could have a distribution that is a multiple of a non-central χ^2 distribution with unknown non-centrality parameter.
- More disconcertingly, the MSE would be the result of selection: $\hat{\sigma}_{\hat{\mathrm{M}}}^2$. Nothing is known about its distribution.

These problems are avoided by using one valid estimate $\hat{\sigma}^2$ that is independent of all submodels.

With this choice of $\hat{\sigma}$, a marginal $1-\alpha$ confidence interval for $\beta_{j\cdot\mathrm{M}}$ is

$$\begin{aligned} \mathrm{CI}_{j\cdot\mathrm{M}}(K) &\triangleq \left[\hat{\beta}_{j\cdot\mathrm{M}} \pm K \left[(\mathbf{X}_\mathrm{M}^T\mathbf{X}_\mathrm{M})^{-1}\right]_{jj}^{\frac{1}{2}} \hat{\sigma}\right] \\ &= \left[\hat{\beta}_{j\cdot\mathrm{M}} \pm K\,\hat{\sigma}/\|\mathbf{X}_{j\cdot\mathrm{M}}\|\right]. \end{aligned} \tag{4.2}$$

where $K = t_{r,1-\alpha/2}$ is the $1 - \alpha/2$ quantile of a t-distribution with r degrees of freedom. This interval is valid, that is,

$$\mathbf{P}[\beta_{j\cdot\mathrm{M}} \in \mathrm{CI}_{j\cdot\mathrm{M}}(K)] \;\geq\; 1 - \alpha,$$

under the assumption that the submodel M is chosen independently of the response $\mathbf{Y}$.

4.2. *Model Selection and Its Implications for Parameters.* In practice, the model M tends to be the result of some form of model selection that makes use of the stochastic component of the data, which is the response vector $\mathbf{Y}$ in the present context (Section 2.3). This model should therefore be expressed as $\hat{\mathrm{M}} = \hat{\mathrm{M}}(\mathbf{Y})$. In general we allow a model selection procedure to be any (measurable) map

$$\hat{\mathrm{M}}: \quad \mathbf{Y} \mapsto \hat{\mathrm{M}}(\mathbf{Y}), \qquad \mathbb{R}^n \to \mathcal{M}_{\mathrm{all}}, \tag{4.3}$$

where

$$\mathcal{M}_{\rm all} \;\triangleq\; \{\mathrm{M}\,|\,\mathrm{M} \subset \{1,2,...,p\},\ \mathrm{rank}(\mathbf{X}_{\mathrm{M}}) = |\mathrm{M}|\,\} \tag{4.4}$$

is the set of all full-rank submodels. Thus $\hat{\mathrm{M}}$ divides $\mathbb{R}^n$ into as many as $|\mathcal{M}|$ different regions with shared outcome of model selection.

Data dependence of the selected model $\hat{\mathrm{M}}$ has strong consequences:

- Most fundamentally, the selected model $\hat{\mathrm{M}} = \hat{\mathrm{M}}(\mathbf{Y})$ is now random. Whether the model has been selected by an algorithm or by human choice, if the response $\mathbf{Y}$ has been involved in the selection, the resulting model is a random object because it could have been different for a different realization of the random vector $\mathbf{Y}$.
- Associated with the random model $\hat{\mathrm{M}}(\mathbf{Y})$ is the parameter vector of coefficients $\boldsymbol{\beta}_{\hat{\mathrm{M}}(\mathbf{Y})}$, which is now randomly chosen also: (1) It has a random dimension, $\boldsymbol{\beta}_{\hat{\mathrm{M}}(\mathbf{Y})} \in \mathbb{R}^{m(\mathbf{Y})}$ for $m(\mathbf{Y}) = |\hat{\mathrm{M}}(\mathbf{Y})|$; (2) for any fixed j, it may or may not be the case that $j \in \hat{\mathrm{M}}(\mathbf{Y})$; (3) conditional on $j \in \hat{\mathrm{M}}(\mathbf{Y})$, the parameter $\beta_{j\cdot\hat{\mathrm{M}}(\mathbf{Y})}$ changes randomly as the adjustor covariates in $\hat{\mathrm{M}}(\mathbf{Y})$ vary randomly.

Thus the set of parameters for which inference is sought is random also.

4.3. *Valid Post-Selection Confidence Intervals.* Unless a predictor is forced to be in the selected model, it is not meaningul to ask for marginal probability guarantees for $\mathrm{CI}_{j\cdot\hat{\mathrm{M}}}$ for a fixed j because the guarantee requires $j\in\hat{\mathrm{M}}(\mathbf{Y})$ whereas the probability $\mathbf{P}[j\in\hat{\mathrm{M}}(\mathbf{Y})]$ all by itself can be easily less than $1-\alpha$ even for a strong predictor. One may therefore be tempted to look for guarantees in terms of conditional probabilities given $j\in\hat{\mathrm{M}}$, but nothing is known about such events and the associated conditional distribution of $|t_{j\cdot\mathrm{M}}|$ for common selection methods. However, a solution in terms of marginal rather than conditional probability can be found by binding j with a quantifier and requiring a simultaneous guarantee in terms of $\mathbf{P}[\beta_{j\cdot\hat{\mathrm{M}}} \in \mathrm{CI}_{j\cdot\hat{\mathrm{M}}}(K)\ \forall j\in\hat{\mathrm{M}}]$. For this mathematically well-defined probabiliy there exists in principle a confidence guarantee through suitable choice of the constant K such that

$$\mathbf{P}\left[\beta_{j\cdot\hat{\mathrm{M}}} \in \mathrm{CI}_{j\cdot\hat{\mathrm{M}}}(K)\ \forall j \in \hat{\mathrm{M}}\right] \;\geq\; 1-\alpha. \tag{4.5}$$

Thus the impossibility of a marginal guarantee for any particular $j\in\hat{\mathrm{M}}$ implies that only a simultaneous guarantee for all $j\in\hat{\mathrm{M}}$ can be given.

4.4. *Universal Validity for all Selection Procedures.* The difficulty with the guarantee (4.5) is that the constant would be specific to the model selection procedure $\hat{\mathrm{M}}$: $K = K(\hat{\mathrm{M}})$. Finding procedure-specific constants may be a challenge, and this is not what we attempt to do in this article. Rather, the "PoSI" procedure proposed here produces a constant K that provides universally valid post-selection inference *for all model selection procedures* $\hat{\mathrm{M}}$:

$$\mathbf{P}\left[\beta_{j\cdot\hat{\mathrm{M}}} \in \mathrm{CI}_{j\cdot\hat{\mathrm{M}}}(K)\ \forall j \in \hat{\mathrm{M}}\right] \;\geq\; 1-\alpha \quad \forall\,\hat{\mathrm{M}}. \tag{4.6}$$

Universal validity irrespective of the model selection procedure $\hat{\mathrm{M}}$ is a strong property that raises questions of whether the approach is too conservative. There are, however, some arguments in its favor:

(1) Universal validity may be desirable or even essential for applications in which the model selection procedure is not specified in advance or for which the analysis involves some ad hoc elements that cannot be accurately prespecified. Even so, we should think of the actually chosen model as part of a "procedure" $\mathbf{Y} \mapsto \hat{\mathrm{M}}(\mathbf{Y})$, and though the ad hoc steps are not specified for $\mathbf{Y}$ other than the observed one, this is not a problem because our protection is irrespective of what a specification might have been. This view also allows data analysts to change their minds, to improvise and informally decide in favor of a model other than that produced by a formal selection procedure, or to experiment with multiple selection procedures.

(2) There exists a model selection procedure that requires the full strength of universally valid PoSI, and this procedure may not be entirely unrealistic as an approximation to some types of data analytic activities: "significance hunting", that is, selecting that model which contains the statistically most significant coefficient; see Section 4.8.

(3) There is a general question about the wisdom of proposing ever tighter confidence and retention intervals for practical use when in fact these intervals are valid only under tightly controlled conditions. It might be reasonable to suppose that much applied work involves more data peeking than is reported in published articles. With inference that is universally valid after any model selection procedure we have a way to establish which rejections are safe, irrespective of unreported data peeking as part of selecting a model.

4.5. *Restricted Model Selection.* The concerns over PoSI's conservative nature can be alleviated somewhat by introducing a degree of flexibility to the PoSI problem with regard to the universe of models being searched. Such flexibility is additionally called for from a practical point of view because it is not true that *all* submodels in $\mathcal{M}_{\mathrm{all}}$ (4.4) are being searched all the time. Rather, in many applications the search is limited in a way that can be specified a priori, without involvement of $\mathbf{Y}$. For example, a predictor of interest may be forced into the submodels, or there may be a restriction on the size of the submodels. Indeed, if p is large, a restriction to a manageable set of submodels is a computational necessity. In much of what follows we can allow the universe $\mathcal{M}$ of submodels to be an (almost) arbitrary but pre-specified non-empty subset of $\mathcal{M}_{\text{all}}$; the only requirement is that every predictor is used in at least one model:

$$\bigcup_{\mathrm{M}\in\mathcal{M}} \mathrm{M} \;=\; \{1, 2, ..., p\}. \tag{4.7}$$

Because we allow only non-singular submodels (see (4.4)) we have $|\mathrm{M}| \le d$ $\forall \mathrm{M} \in \mathcal{M}$, where as always $d = \operatorname{rank}(\mathbf{X})$. — Selection procedures are now maps

$$\hat{\mathrm{M}} : \;\mathbf{Y} \mapsto \hat{\mathrm{M}}(\mathbf{Y}), \qquad \mathbb{R}^n \to \mathcal{M}. \tag{4.8}$$

The following are examples of model universes with practical relevance (see also Leeb and Pötscher (2008a), Section 1.1, Example 1).

(1) Submodels that contain the first p' predictors ($1 \le p' \le p$):
$\mathcal{M}_1 = \{\mathrm{M} \in \mathcal{M}_{\text{all}} \,|\, \{1, 2, ..., p'\} \subset \mathrm{M}\}$.
Classical: $|\mathcal{M}_1| = 2^{p-p'}$. Example: forcing an intercept into all models.

(2) Submodels of size m' or less ("sparsity option"):
$\mathcal{M}_2 = \{\mathrm{M} \in \mathcal{M}_{\text{all}} \,|\, |\mathrm{M}| \le m'\}$. Classical: $|\mathcal{M}_2| = \binom{p}{1} + ... + \binom{p}{m'}$.

(3) Submodels with fewer than m' predictors dropped from the full model:
$\mathcal{M}_3 = \{\mathrm{M} \in \mathcal{M}_{\text{all}} \,|\, |\mathrm{M}| > p - m'\}$. Classical: $|\mathcal{M}_3| = |\mathcal{M}_2|$.

(4) Nested models: $\mathcal{M}_4 = \{\{1, ..., j\} \,|\, j \in \{1, ..., p\}\}$. $|\mathcal{M}_4| = p$.
Example: selecting the degree up to $p-1$ in a polynomial regression.

(5) Models dictated by an ANOVA hierarchy of main effects and interactions in a factorial design.

This list is just an indication of possibilities. In general, the smaller the set $\tilde{\mathcal{M}} = \{(j, \mathrm{M}) \,|\, j \in \mathrm{M} \in \mathcal{M}\}$ is, the less conservative the PoSI approach is, and the more computationally manageable the problem becomes. With sufficiently strong restrictions, in particular using the sparsity option (2) and assuming the availability of an independent valid estimate $\hat{\sigma}$, it is possible to apply PoSI in certain non-classical $p > n$ situations.

Further reduction of the PoSI problem is possible by pre-screening adjusted predictors *without the response* $\mathbf{Y}$. In a fixed-design regression, any variable selection procedure that does *not* involve $\mathbf{Y}$ does *not* invalidate statistical inference. For example, one may decide not to seek inference for predictors in submodels that impart a "Variance Inflation Factor" (*VIF*) above a user-chosen threshold: $\mathit{VIF}_{j\cdot\mathrm{M}} = \|\mathbf{X}_j\|^2 / \|\mathbf{X}_{j\cdot\mathrm{M}}\|^2$ if $\mathbf{X}_j$ is centered, hence does not make use of $\mathbf{Y}$, and elimination according to $\mathit{VIF}_{j\cdot \mathrm{M}} > c$ does not invalidate inference.

4.6. *Reduction of Universally Valid Post-Selection Inference to Simultaneous Inference.* We show that universally valid post-selection inference (4.6) follows from simultaneous inference in the form of family-wise error control for all parameters in all submodels. The argument depends on the following lemma that may fall into the category of the "trivial but not immediately obvious".

Lemma 4.1. *("Significant Triviality Bound") For any model selection procedure $\hat{\mathrm{M}} : \mathbb{R}^n \to \mathcal{M}$, the following inequality holds for all $\mathbf{Y} \in \mathbb{R}^n$:*

$$\max_{j \in \hat{\mathrm{M}}(\mathbf{Y})} |t_{j\cdot \hat{\mathrm{M}}(\mathbf{Y})}(\mathbf{Y})| \;\leq\; \max_{\mathrm{M} \in \mathcal{M}} \max_{j \in \mathrm{M}} |t_{j\cdot \mathrm{M}}(\mathbf{Y})|$$

Proof: This is a special case of the triviality $f(\hat{\mathrm{M}}(\mathbf{Y})) \leq \max_{\mathrm{M}} f(\mathrm{M})$, where $f(\mathrm{M}) = \max_{j \in \mathrm{M}} |t_{j\cdot \mathrm{M}}(\mathbf{Y})|$. □

For a model selection procedure $\hat{\mathrm{M}}$ that attains the right hand bound of the lemma, see Section 4.8.

Theorem 4.1. *Let K satisfy*

$$\mathbf{P}\left[\max_{\mathrm{M} \in \mathcal{M}} \max_{j \in \mathrm{M}} |t_{j\cdot \mathrm{M}}| \leq K\right] \;\geq\; 1 - \alpha. \tag{4.9}$$

Then the following holds for all model selection procedures $\hat{\mathrm{M}} : \mathbb{R}^n \to \mathcal{M}$:

$$\mathbf{P}\left[\max_{j \in \hat{\mathrm{M}}} |t_{j\cdot \hat{\mathrm{M}}}| \leq K\right] \;\geq\; 1 - \alpha. \tag{4.10}$$

Proof: This follows immediately from Lemma 4.1. □

Although mathematically trivial we give the above the status of a theorem as it is the central statement of the reduction of universal post-selection inference to simultaneous inference.

Let K be the minimal constant satisfying (4.9). By definition K does not depend on the selection procedure $\hat{\mathrm{M}}$, but it does depend on the full predictor matrix $\mathbf{X}$, the set of submodels $\mathcal{M}$, the required coverage $1 - \alpha$, and the degrees of freedom r in $\hat{\sigma}$. We will ignore the dependence on $\mathcal{M}$ if it is understood that $\mathcal{M} = \mathcal{M}_{\text{all}}$ and we will variously write

$$K = K(\mathbf{X}, \mathcal{M}, \alpha, r), \quad K(\mathbf{X}), \quad K(\mathbf{X}, p), \quad K(\mathbf{X}, \alpha, p, r), \tag{4.11}$$

the last two being useful in the classical case ($d = p \leq n$) for asymptotics as $p \to \infty$. We call $K(\mathbf{X})$ the "PoSI constant", and for M and $j \in \mathrm{M}$ we call $\mathrm{CI}_{j\cdot\mathrm{M}}(K(\mathbf{X}))$ the "PoSI simultaneous confidence interval" or simply "PoSI CI". From (4.10) follows the desired coverage guarantee:

Theorem 4.2. *"Simultaneous Post-Selection Confidence Guarantees" hold for any model selection procedure* $\hat{\mathrm{M}}\colon \mathbb{R}^n \to \mathcal{M}$*:*

$$\mathbf{P}[\,\beta_{j\cdot\hat{\mathrm{M}}} \in \mathrm{CI}_{j\cdot\hat{\mathrm{M}}}(K(\mathbf{X}, \alpha))\;\forall j \in \hat{\mathrm{M}}\,] \;\geq\; 1 - \alpha.$$

Simultaneous inference provides strong family-wise error control, which in turn translates to strong error control following model selection.

Theorem 4.3. *"Strong Post-Selection Error Control" holds for any model selection procedure* $\hat{\mathrm{M}}\colon \mathbb{R}^n \to \mathcal{M}$*:*

$$\mathbf{P}[\,\forall j \in \hat{\mathrm{M}} : |t^{(0)}_{j\cdot\hat{\mathrm{M}}}| > K(\mathbf{X}, \alpha) \;\Rightarrow\; \beta_{j\cdot\hat{\mathrm{M}}} \neq 0\,] \;\geq\; 1 - \alpha,$$

where $t^{(0)}_{j\cdot\mathrm{M}}$ *is the* t*-statistic for the null hypothesis* $\beta_{j\cdot\mathrm{M}} = 0$*.*

The proof is in the Appendix. The theorem states that, with probability $1 - \alpha$, in a selected model *all* PoSI-significant rejections have detected true alternatives.

4.7. *Scheffé Protection.* Realizing the idea that the LS estimators in different submodels generally estimate different parameters, we generated a simultaneous inference problem involving up to $p\,2^{p-1}$ linear contrasts $\beta_{j\cdot\mathrm{M}}$. In view of the enormous number of linear combinations for which simultaneous inference is sought, one should wonder whether the problem is not best solved by Scheffé's method (1953; 1959) which provides simultaneous inference for *all* linear combinations. To accommodate rank-deficient $\mathbf{X}$, we cast Scheffé's result in terms of t-statistics for arbitrary non-zero $\mathbf{x} \in \mathrm{span}(\mathbf{X})$:

$$t_{\mathbf{x}} \;\triangleq\; \frac{(\mathbf{Y} - \boldsymbol{\mu})^T \mathbf{x}}{\hat{\sigma}\|\mathbf{x}\|}. \tag{4.12}$$

The t-statistics in (4.1) are obtained for $\mathbf{x} = \mathbf{X}_{j\cdot\mathrm{M}}$. Scheffé's guarantee is

$$\mathbf{P}\left[\sup_{\mathbf{x}\in\mathrm{span}(\mathbf{X})} |t_{\mathbf{x}}| \leq K_{\mathrm{Sch}}\right] \;=\; 1-\alpha, \tag{4.13}$$

where the Scheffé constant is

$$K_{\mathrm{Sch}} \;=\; K_{\mathrm{Sch}}(\alpha, d, r) \;=\; \sqrt{d\mathrm{F}_{d,r,1-\alpha}}. \tag{4.14}$$

It provides an upper bound for *all* PoSI constants:

Proposition 4.1. $K(\mathbf{X}, \mathcal{M}, \alpha, r) \;\leq\; K_{\mathrm{Sch}}(\alpha, d, r) \;\; \forall\, \mathbf{X},\, \mathcal{M},\, d{=}\mathrm{rank}(\mathbf{X})$.

Thus parameter estimates $\hat{\beta}_{j\cdot\mathrm{M}}$ whose t-ratios exceed K_{Sch} in magnitude are universally safe from invalidation due to model selection. The universality of the Scheffé constant is a tip-off that it may be too loose for some predictor matrices $\mathbf{X}$, and obtaining the sharper constant $K(\mathbf{X})$ may be worth the effort. An indication is given by the following comparison as $r \to \infty$:

- For the Scheffé constant it holds $K_{\mathrm{Sch}} \sim \sqrt{d}$.
- For orthogonal designs it holds $K_{\mathrm{orth}} \sim \sqrt{2\log d}$.

(For orthogonal designs see Section 5.6.) Thus the PoSI constant K_{orth} is much smaller than K_{Sch}. The big gap between the two suggests that the Scheffé constant may be too conservative at least in some cases. We will study the case of certain non-orthogonal designs for which the PoSI constant is $O(\sqrt{\log(d)})$ in Section 6.1. On the other hand, the PoSI constant can approach the order $O(\sqrt{d})$ of the Scheffé constant K_{Sch} as well, and we will study one such case in Section 6.2.

Even though in this article we will give asymptotic results for $d = p \to \infty$ and $r \to \infty$ only, we mention another kind of asymptotics whereby r is held constant while $d = p \to \infty$: In this case K_{Sch} is in the order of the product of $\sqrt{d}$ and the $1-\alpha$ quantile of the inverse-chi-square distribution with r degrees of freedom. In a similar way, the constant K_{orth} for orthogonal designs is in the order of the product of $\sqrt{2\log d}$ and the $1-\alpha$ quantile of the inverse-chi-square distribution with r degrees of freedom.

4.8. *PoSI-Sharp Model Selection — "SPAR" and "SPAR1".* There exists a model selection procedure that requires the full protection of the simultaneous inference procedure (4.9). It is the "significance hunting" procedure that selects the model containing the most significant "effect":

$$\hat{\mathrm{M}}_{\mathrm{SPAR}}(\mathbf{Y}) \;\triangleq\; \operatorname*{argmax}_{\mathrm{M}\in\mathcal{M}} \max_{j\in\mathrm{M}} |t_{j\cdot\mathrm{M}}(\mathbf{Y})|.$$

We name this procedure "SPAR" for "*Single Predictor Adjusted Regression.*" It achieves equality with the "significant triviality bound" in Lemma 4.1 and is therefore the worst case procedure for the PoSI problem. In the selected submodel $\hat{\mathrm{M}}_{\mathrm{SPAR}}(\mathbf{Y})$ the less significant predictors matter only in so far as they boost the significance of the winning predictor by adjusting it accordingly. This procedure ignores the quality of the fit to $\mathbf{Y}$ provided by the model. While our present purpose is to point out the existence of a selection procedure that requires full PoSI protection, SPAR could be of practical interest when the analysis is centered on strength of "effects", not quality of model fit.

Practically of greater interest is a restricted version of SPAR whereby a predictor of interest is determined a priori and the search is for adjustment that optimizes this predictor's "effect". We name the resulting procedure "SPAR1". If the predictor of interest is $\mathbf{X}_p$, say, then the model universe is $\mathcal{M}_{\mathrm{SPAR1}} = \{\mathrm{M} \in \mathcal{M}_{\mathrm{all}} \,|\, p \in \mathrm{M}\,\}$ and the model selection procedure is

$$\hat{\mathrm{M}}_{\mathrm{SPAR1}}(\mathbf{Y}) \;\triangleq\; \underset{\mathrm{M}\in\mathcal{M}_{\mathrm{SPAR1}}}{\operatorname{argmax}} \; |t_{p\cdot\mathrm{M}}(\mathbf{Y})|.$$

Importantly, the SPAR1 guarantee that we seek is not for all coefficients in the models $\mathrm{M} \in \mathcal{M}_{\mathrm{SPAR1}}$ but only for the X_p-coefficient $\beta_{p\cdot\mathrm{M}}$:

$$\mathbf{P}\left[\max_{\mathrm{M}\in\mathcal{M}_{\mathrm{SPAR1}}} |t_{p\cdot\mathrm{M}}| \le K_{\mathrm{SPAR1}}\right] \;\ge\; 1-\alpha,$$

where K_{SPAR1} is the minimal constant satisfying this condition. As $\mathcal{M}_{\mathrm{SPAR1}} \subset \mathcal{M}_{\mathrm{all}}$ and SPAR1 inference is for $j = p$ only, the unrestricted PoSI constant dominates the SPAR1 constant: $K(\mathbf{X}, \mathcal{M}_{\mathrm{all}}) \ge K_{\mathrm{SPAR1}}(\mathbf{X})$. Even so, we will construct in Section 6.2 an example where the SPAR1 constant increases at the Scheffé rate and is asymptotically more than 63% of K_{Sch}. This is the technical reason for introducing SPAR1.

4.9. *PoSI P-Value Adjustment for Model Selection.* Statistical inference for regression coefficients is more often carried out in terms of p-values than confidence intervals. The usual p-values are for null hypotheses $\beta_{j\cdot\mathrm{M}} = 0$, hence the test statistics are

$$t^{(0)}_{j\cdot\mathrm{M}} \;=\; \hat{\beta}_{j\cdot\mathrm{M}}/(\hat{\sigma}/\|\mathbf{X}_{j\cdot\mathrm{M}}\|), \qquad t^{(0)}_{\max} \;=\; \max_{\mathrm{M}\in\mathcal{M}}\max_{j\in\mathrm{M}} |t^{(0)}_{j\cdot\mathrm{M}}|.$$

To define marginal and adjusted p-values we introduce two c.d.f.s:

$$F_{j\cdot\mathrm{M}}(t) = \mathbf{P}[\,|t^{(0)}_{j\cdot\mathrm{M}}| < t\,], \qquad F_{\max}(t) \;=\; \mathbf{P}[t^{(0)}_{\max} < t]. \tag{4.15}$$

The former measures marginal null coverage of a two-sided retention interval $[-t, +t]$, while the latter measures simultaneous coverage of a retention cube $[-t, +t]^k$ where $k = |\{(j, \mathrm{M}) \,|\, j \in \mathrm{M} \in \mathcal{M}\}|$ is the number of tests performed, which can be as many as $p\,2^{p-1}$ in the classical case $d = p \leq n$ for $\mathcal{M} = \mathcal{M}_{\mathrm{all}}$. (See Section 7.1 for the approximate computation of $F_{\max}(t)$.) Denoting by $t^{obs}_{j\cdot\mathrm{M}}$ and $t^{obs}_{\max}$ the observed values of $t^{(0)}_{j\cdot\mathrm{M}}$ and $t^{(0)}_{\max}$, respectively, the following p-values can be defined:

(1) Marginal: $\mathrm{pval}_{j\cdot\mathrm{M}} = 1 - F_{j\cdot\mathrm{M}}(\,|t^{obs}_{j\cdot\mathrm{M}}|\,)$

(2) Global adjusted: $\mathrm{pval}^{PoSI}_{j\cdot\mathrm{M}} = 1 - F_{\max}(t^{obs}_{\max})$

(3) Individual adjusted: $\mathrm{pval}^{PoSI}_{j\cdot\mathrm{M}} = 1 - F_{\max}(|t^{obs}_{j\cdot\mathrm{M}}|)$

Comments:

(1) The marginal p-value ignores the fact that k tests are being performed.

(2) The global adjusted p-value establishes whether at least the strongest "effect" is statistically significant, and it is therefore an overall test similar to, but more specific than, the overall F-test. Because the latter is derived from Scheffé protection, the global adjusted PoSI p-value is more powerful and still protects against any model selection in the model universe $\mathcal{M}$.

(3) The individual adjusted p-value adjusts each $|t_{j\cdot\mathrm{M}}|$ as if it were a max statistic, hence results in an over-adjustment for all but $t_{\max}$. A sharper method than this "one-step adjustment" would be a simulation-based "step-down" method, but the computational expense may be prohibitive and the gain in statistical efficiency may be small.

The adjusted p-values are recommended because they account universally for any model selection in the model universe $\mathcal{M}$.
[Note on terminology: "adjustment of a p-value for simultaneity" and "adjustment of a predictor for other predictors" are two concepts that share nothing except the partial homonym.]

5. The Structure of the PoSI Problem.

5.1. *Canonical Coordinates.* We can reduce the dimensionality of the PoSI problem from $n \times p$ to $d \times p$, where $d = \mathrm{rank}(X) \leq \min(n, p)$, by introducing Scheffé's canonical coordinates. This reduction is important both geometrically and computationally because the PoSI coverage problem really takes place in the column space of $\mathbf{X}$.

DEFINITION: *Let* $\mathbf{Q} = (\mathbf{q}_1, ..., \mathbf{q}_d) \in \mathbb{R}^{n \times d}$ *be any orthonormal basis of the column space of* $\mathbf{X}$. *Note that* $\hat{\mathbf{Y}} = \mathbf{Q}\mathbf{Q}^T\mathbf{Y}$ *is the orthogonal projection* *of* $\mathbf{Y}$ *onto the column space of* $\mathbf{X}$ *even if* $\mathbf{X}$ *is not of full rank. We call* $\tilde{\mathbf{X}} = \mathbf{Q}^T\mathbf{X} \in \mathbb{R}^{d\times p}$ *and* $\tilde{\mathbf{Y}} = \mathbf{Q}^T\hat{\mathbf{Y}} \in \mathbb{R}^d$ *canonical coordinates of* $\mathbf{X}$ *and* $\hat{\mathbf{Y}}$*.*

We extend the notation $\mathbf{X}_{\mathrm{M}}$ for extraction of subsets of columns to canonical coordinates $\tilde{\mathbf{X}}_{\mathrm{M}}$. Accordingly slopes obtained from canonical coordinates will be denoted by $\hat{\boldsymbol{\beta}}_{\mathrm{M}}(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}) = (\tilde{\mathbf{X}}_{\mathrm{M}}^T\tilde{\mathbf{X}}_{\mathrm{M}})^{-1}\tilde{\mathbf{X}}_{\mathrm{M}}^T\tilde{\mathbf{Y}}$ to distinguish them from the slopes obtained from the original data $\hat{\boldsymbol{\beta}}_{\mathrm{M}}(\mathbf{X}, \mathbf{Y}) = (\mathbf{X}_{\mathrm{M}}^T\mathbf{X}_{\mathrm{M}})^{-1}\mathbf{X}_{\mathrm{M}}^T\mathbf{Y}$, if only to state in the following proposition that they are identical.

Proposition 5.1. *Properties of canonical coordinates:*

(1) $\tilde{\mathbf{Y}} = \mathbf{Q}^T\mathbf{Y}$.

(2) $\tilde{\mathbf{X}}_{\mathrm{M}}^T\tilde{\mathbf{X}}_{\mathrm{M}} = \mathbf{X}_{\mathrm{M}}^T\mathbf{X}_{\mathrm{M}}$ *and* $\tilde{\mathbf{X}}_{\mathrm{M}}^T\tilde{\mathbf{Y}} = \mathbf{X}_{\mathrm{M}}^T\mathbf{Y}$.

(3) $\hat{\boldsymbol{\beta}}_{\mathrm{M}}(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}) = \hat{\boldsymbol{\beta}}_{\mathrm{M}}(\mathbf{X}, \mathbf{Y})$ *for all submodels* M.

(4) $\tilde{\mathbf{Y}} \sim \mathcal{N}(\tilde{\boldsymbol{\mu}}, \sigma^2\mathbf{I}_d)$, *where* $\tilde{\boldsymbol{\mu}} = \mathbf{Q}^T\boldsymbol{\mu}$.

(5) $\tilde{\mathbf{X}}_{j\cdot\mathrm{M}} = \mathbf{Q}^T\mathbf{X}_{j\cdot\mathrm{M}}$, *where* $j \in \mathrm{M}$ *and* $\tilde{\mathbf{X}}_{j\cdot\mathrm{M}} \in \mathbb{R}^d$ *is the residual vector of the regression of* $\tilde{\mathbf{X}}_j$ *onto the other columns of* $\tilde{\mathbf{X}}_{\mathrm{M}}$.

(6) $t_{j\cdot M} = (\hat{\beta}_{j\cdot M}(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}) - \beta_{j\cdot M})/(\hat{\sigma}/\|\tilde{\mathbf{X}}_{j\cdot\mathrm{M}}\|)$.

(7) In the classical case $d=p$, $\tilde{\mathbf{X}}$ *can be chosen to be an upper triangular or a symmetric matrix.*

The proofs of *(1)-(6)* are elementary. As for *(7)*, an upper triangular $\tilde{\mathbf{X}}$ can be obtained from a QR-decomposition based on a Gram-Schmidt procedure: $\mathbf{X} = \mathbf{QR}$, $\tilde{\mathbf{X}} = \mathbf{R}$. A symmetric $\tilde{\mathbf{X}}$ is obtained from a singular value decomposition: $\mathbf{X} = \mathbf{UDV}^T$, $\mathbf{Q} = \mathbf{UV}^T$, $\tilde{\mathbf{X}} = \mathbf{VDV}^T$.

Canonical coordinates allow us to analyze the PoSI coverage problem in $\mathbb{R}^d$. In what follows we will freely assume that all objects are rendered in canonical coordinates and write $\mathbf{X}$ and $\mathbf{Y}$ for $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$, implying that the predictor matrix is of size $d \times p$ and the response is of size $d \times 1$.

5.2. *PoSI Coefficient Vectors in Canonical Coordinates.* The PoSI coverage problem (4.9) can be simplified as follows: Due to pivotality of t-statistics, the problem is invariant under translation of $\boldsymbol{\beta}$ and rescaling of σ, and hence it suffices to solve coverage problems for $\boldsymbol{\beta} = \mathbf{0}$ and $\sigma = 1$. In canonical coordinates this implies $\mathbf{E}[\tilde{\mathbf{Y}}] = \mathbf{0}_d$, hence $\tilde{\mathbf{Y}} \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$. For this reason we write $\mathbf{Z}$ instead of $\tilde{\mathbf{Y}}$, so that $\mathbf{Z}/\hat{\sigma}$ has a d-dimensional t-distribution with r degrees of freedom and any linear combination $\mathbf{u}^T\mathbf{Z}/\hat{\sigma}$ with a unit vector $\mathbf{u}$ has a 1-dimensional t-distribution. Letting $\mathbf{X}_{j\cdot\mathrm{M}}$ be the adjusted predictors in canonical coordinates, the estimates (3.7) and their t-statistics (4.1) simplify to

$$\hat{\beta}_{j\cdot\mathrm{M}} = \frac{\mathbf{X}_{j\cdot\mathrm{M}}^T\mathbf{Z}}{\|\mathbf{X}_{j\cdot\mathrm{M}}\|^2} = \boldsymbol{l}_{j\cdot\mathrm{M}}^T\mathbf{Z}, \qquad t_{j\cdot\mathrm{M}} = \frac{\mathbf{X}_{j\cdot\mathrm{M}}^T\mathbf{Z}}{\|\mathbf{X}_{j\cdot\mathrm{M}}\|\hat{\sigma}} = \bar{\boldsymbol{l}}_{j\cdot\mathrm{M}}^T\mathbf{Z}/\hat{\sigma}, \tag{5.1}$$

where we took advantage of the fact that these are linear functions of $\mathbf{Z}$ and $\mathbf{Z}/\hat{\sigma}$, respectively, with "PoSI coefficient vectors" $\boldsymbol{l}_{j\cdot\mathrm{M}}$ and $\bar{\boldsymbol{l}}_{j\cdot\mathrm{M}}$ that equal $\mathbf{X}_{j\cdot\mathrm{M}}$ up to scale:

$$\text{(5.2)} \qquad \boldsymbol{l}_{j\cdot\mathrm{M}} \;\triangleq\; \frac{\mathbf{X}_{j\cdot\mathrm{M}}}{\|\mathbf{X}_{j\cdot\mathrm{M}}\|^2}, \qquad \bar{\boldsymbol{l}}_{j\cdot\mathrm{M}} \;\triangleq\; \frac{\boldsymbol{l}_{j\cdot\mathrm{M}}}{\|\boldsymbol{l}_{j\cdot\mathrm{M}}\|} \;=\; \frac{\mathbf{X}_{j\cdot\mathrm{M}}}{\|\mathbf{X}_{j\cdot\mathrm{M}}\|}.$$

As we now operate in canonical coordinates we have $\boldsymbol{l}_{j\cdot\mathrm{M}} \in \mathbb{R}^d$ and $\bar{\boldsymbol{l}}_{j\cdot\mathrm{M}} \in S^{d-1}$, where S^{d-1} is the unit sphere in $\mathbb{R}^d$. To complete the structural description of the PoSI problem we let

$$\text{(5.3)} \qquad \mathcal{L}(\mathbf{X},\mathcal{M}) \;\triangleq\; \{\bar{\boldsymbol{l}}_{j\cdot\mathrm{M}} \,|\, j\in\mathrm{M}\in\mathcal{M}\,\} \;\subset\; S^{d-1},$$

If $\mathcal{M}=\mathcal{M}_{\mathrm{all}}$ we omit the second argument and write $\mathcal{L}(\mathbf{X})$.

Proposition 5.2. *The PoSI problem* (4.9) *is equivalent to a d-dimensional coverage problem for linear functions of the multivariate t-vector $\mathbf{Z}/\hat{\sigma}$:*

$$\text{(5.4)} \quad \mathbf{P}\left[\max_{\mathrm{M}\in\mathcal{M}}\max_{j\in\mathrm{M}} |t_{j\cdot\mathrm{M}}| \le K\right] \;=\; \mathbf{P}\left[\max_{\bar{l}\in\mathcal{L}(\mathbf{X},\mathcal{M})} |\bar{\boldsymbol{l}}^T\mathbf{Z}/\hat{\sigma}| \le K\right] \;\overset{(\geq)}{=}\; 1-\alpha.$$

5.3. *Orthogonalities of PoSI Coefficient Vectors.* The set $\mathcal{L}(\mathbf{X},\mathcal{M})$ of unit vectors $\bar{\boldsymbol{l}}_{j\cdot\mathrm{M}}$ has intrinsically interesting geometric structure, which is the subject of this and the following subsections. The next proposition (proof in Appendix A.2) elaborates in so many ways the fact that $\bar{\boldsymbol{l}}_{j\cdot\mathrm{M}}$ is essentially the predictor vector $\mathbf{X}_j$ orthogonalized with regard to the other predictors in the model M. In what follows vectors are always assumed in canonical coordinates and hence d-dimensional.

Proposition 5.3. *Orthogonalities in $\mathcal{L}(\mathbf{X},\mathcal{M})$: The following statements hold assuming that the models referred to are in $\mathcal{M}$ (hence are of full rank).*

1. *Adjustment properties:*
   $\bar{\boldsymbol{l}}_{j\cdot\mathrm{M}} \in \mathrm{span}\{\,\mathbf{X}_j |\, j\in\mathrm{M}\}$ *and* $\bar{\boldsymbol{l}}_{j\cdot\mathrm{M}} \perp \mathbf{X}_{j'}$ *for $j \neq j'$ both* $\in \mathrm{M}$.
2. *The following vectors form an orthonormal "Gram-Schmidt" series:*
   $$\{\bar{\boldsymbol{l}}_{1\cdot\{1\}},\; \bar{\boldsymbol{l}}_{2\cdot\{1,2\}},\; \bar{\boldsymbol{l}}_{3\cdot\{1,2,3\}},\; \ldots,\; \bar{\boldsymbol{l}}_{d\cdot\{1,2,\ldots,d\}}\}$$
   Other series are obtained using $(j_1, j_2, \ldots, j_d)$ *in place of* $(1, 2, \ldots, d)$.
3. *Vectors* $\bar{\boldsymbol{l}}_{j\cdot\mathrm{M}}$ *and* $\bar{\boldsymbol{l}}_{j'\cdot\mathrm{M}'}$ *are orthogonal if* $\mathrm{M}\subset\mathrm{M}'$, $j\in\mathrm{M}$ *and* $j'\in\mathrm{M}'\setminus\mathrm{M}$.

4. Classical case $d = p$ and $\mathcal{M} = \mathcal{M}_{\mathrm{all}}$: Each vector $\bar{\boldsymbol{l}}_{j\cdot\mathrm{M}}$ is orthogonal to $(p-1)\,2^{p-2}$ vectors $\bar{\boldsymbol{l}}_{j'\cdot\mathrm{M}'}$ (not all of which may be distinct).

The cardinality of orthogonalities in the classical case and $\mathcal{M} = \mathcal{M}_{\mathrm{all}}$ is as follows: If the predictor vectors $\mathbf{X}_j$ have no orthogonal pairs among them, then $|\mathcal{L}(\mathbf{X})| = p\,2^{p-1}$. If there exist orthogonal pairs, then $|\mathcal{L}(\mathbf{X})|$ is less. For example, if there exists exactly one orthogonal pair, then $|\mathcal{L}(\mathbf{X})| = (p-1)\,2^{p-1}$. When $\mathbf{X}$ is a fully orthogonal design, then $|\mathcal{L}(\mathbf{X})| = p$.

5.4. *The PoSI Process.* An alternative way of looking at the PoSI problem is in terms of a stochastic process indexed by (j, M) for $j \in \mathrm{M}$. We mention this view because it is the basis of some software implementations used to solve simultaneous inference and coverage problems, even though in this case it does not result in a practicable approach. In the PoSI problem the obvious process is $\mathbf{W} = (t_{j\cdot\mathrm{M}})_{j\in\mathrm{M}\in\mathcal{M}}$, which is a t-process for finite degrees of freedom r in $\hat{\sigma}$ and a Gaussian process in the limit $r \to \infty$.

The covariance structure of $\mathbf{W}$ exists for $r > 2$ and is proportional (by a factor $r/(r-2)$) to the correlation matrix

$$\boldsymbol{\Sigma} \;=\; (\Sigma_{j\cdot\mathrm{M};\,j'\cdot\mathrm{M}'}), \qquad \Sigma_{j\cdot\mathrm{M};\,j'\cdot\mathrm{M}'} \;\triangleq\; \bar{\boldsymbol{l}}_{j\cdot\mathrm{M}}^T \bar{\boldsymbol{l}}_{j'\cdot\mathrm{M}'}. \tag{5.5}$$

The coverage problem (5.4) can be written as $\mathbf{P}[\|\mathbf{W}\|_\infty \le K] = 1 - \alpha$. Software that computes such coverages (for example, Genz et al. (2010)) allows users to specify a structure such as $\boldsymbol{\Sigma}$, intervals such as $[-K, +K]$ for the components, and error degrees of freedom r. In our experiments this approach worked in the classical case $d = p$ and $\mathcal{M} = \mathcal{M}_{\mathrm{all}}$ for up to $p = 7$, the limiting factor being the space requirement $p\,2^{p-1} \times p\,2^{p-1}$ for the matrix $\boldsymbol{\Sigma}$. By comparison the approach described in Section 7 works for up to $p \approx 20$.

Proposition 5.3 above implies that there exist certain necessary orthogonalities in $\mathcal{L}(\mathbf{X}, \mathcal{M})$. In terms of the correlation structure $\boldsymbol{\Sigma}$, orthogonalities in $\mathcal{L}(\mathbf{X}, \mathcal{M})$ correspond to zero correlations in $\boldsymbol{\Sigma}$. Part *4.* of the proposition states that in the classical case and $\mathcal{M} = \mathcal{M}_{\mathrm{all}}$ each "row" of $\boldsymbol{\Sigma}$ has $(p-1)\,2^{p-2}$ zeros out of $p\,2^{p-1}$ entries, amounting to a fraction $(p-1)/(2p) \to 0.5$, implying that the overall fraction of zeros in $\boldsymbol{\Sigma}$ approaches half for increasing p. Thus $\boldsymbol{\Sigma}$, though not sparse, is rich in zeros. It can be much sparser in the presence of exact orthogonalities among the predictors.

5.5. *The PoSI Polytope.* Coverage problems can be framed geometrically in terms of probability coverage of polytopes in $\mathbb{R}^d$. For the PoSI problem the polytope with half-width K is defined by

$$\boldsymbol{\Pi}_K \;=\; \boldsymbol{\Pi}_K(\mathbf{X}, \mathcal{M}) \;\triangleq\; \{\mathbf{z} \in \mathbb{R}^d \,|\; |\bar{\boldsymbol{l}}^T \mathbf{z}| \le K,\; \forall \bar{\boldsymbol{l}} \in \mathcal{L}(\mathbf{X}, \mathcal{M})\,\}, \tag{5.6}$$

henceforth called the "PoSI polytope". The PoSI coverage problem (5.4) is equivalent to calibrating K such that

$$\mathbf{P}[\mathbf{Z}/\hat{\sigma} \in \mathbf{\Pi}_K] \;=\; 1-\alpha.$$

The simplest case of a PoSI polytope, for $d{=}p{=}2$, is illustrated in Figure 1. More general polytopes are obtained for arbitrary sets $\mathcal{L}$ of unit vectors, that is, subsets $\mathcal{L} \subset S^{d-1}$ of the unit sphere in $\mathbb{R}^d$. For the special case $\mathcal{L} = S^{d-1}$ the "polytope" is the "Scheffé ball" with coverage $\sqrt{d\mathrm{F}_{d,r}} \to \sqrt{\chi^2_d}$ as $r \to \infty$:

$$\mathbf{B}_K \;\triangleq\; \{\mathbf{z} \in \mathbb{R}^d |\; \|\mathbf{z}\| \le K\,\}, \qquad \mathbf{P}[\mathbf{Z}/\hat{\sigma} \in \mathbf{B}_K] \;=\; F_{\mathrm{F}_{d,r}}(K^2/d).$$

Many properties of the polytopes $\mathbf{\Pi}_K$ are not specific to PoSI because they hold for polytopes (5.6) generated by simultaneous inference problems for linear functions with arbitrary sets $\mathcal{L}$ of unit vectors. These polytopes ...

1. ... form scale families of geometrically similar bodies: $\mathbf{\Pi}_K = K\mathbf{\Pi}_1$.
2. ... are point symmetric about the origin: $\mathbf{\Pi}_K = -\mathbf{\Pi}_K$.
3. ... contain the Scheffé ball: $\mathbf{B}_K \subset \mathbf{\Pi}_K$.
4. ... are intersections of "slabs" of width $2K$:
$$\mathbf{\Pi}_K \;=\; \bigcap_{\bar{\boldsymbol{l}} \in \mathcal{L}} \{\mathbf{z} \in \mathbb{R}^d |\; |\mathbf{z}^T \bar{\boldsymbol{l}}\,| \;\le K\,\}.$$
5. ... have $2\,|\mathcal{L}|$ faces (assuming $\mathcal{L} \cap -\mathcal{L} = \emptyset$), and each face is tangent to the Scheffé ball $\mathbf{B}_K$ with tangency points $\pm K\bar{\boldsymbol{l}}$ ($\bar{\boldsymbol{l}} \in \mathcal{L}$).

Specific to PoSI are the orthogonalities described in Proposition 5.3.

5.6. *PoSI Optimality of Orthogonal Designs.* In orthogonal designs, adjustment has no effect: $\mathbf{X}_{j\cdot \mathrm{M}} = \mathbf{X}_j$ for all $j \in \mathrm{M}$, hence $\bar{\boldsymbol{l}}_{j\cdot\mathrm{M}} = \mathbf{X}_j/\|\mathbf{X}_j\|$ and $\mathcal{L}(\mathbf{X},\mathcal{M}) = \{\mathbf{X}_1/\|\mathbf{X}_1\|, ..., \mathbf{X}_p/\|\mathbf{X}_p\|\}$. The polytope $\mathbf{\Pi}_K$ is therefore a hypercube. This simple observation implies an optimality property of orthogonal designs if the submodel universes $\mathcal{M}$ are sufficiently rich to force $\mathcal{L}(\mathbf{X},\mathcal{M})$ to contain an orthonormal basis of $\mathbb{R}^d$: The polytope generated by an orthonormal basis is a hypercube, hence the polytope $\mathbf{\Pi}_K(\mathbf{X},\mathcal{M})$ is contained in this hypercube; thus $\mathbf{\Pi}_K(\mathbf{X},\mathcal{M})$ has maximal extent iff it is equal to this hypercube, which is the case iff $\mathcal{L}(\mathbf{X},\mathcal{M})$ is this orthonormal basis and nothing more, that is, $\mathbf{X}$ is an orthogonal design. — A simple sufficient condition for $\mathcal{M}$ to grant the existence of an orthonormal basis in $\mathcal{L}(\mathbf{X},\mathcal{M})$ is the existence of a maximal nested sequence of submodels such as $\{1\}$, $\{1,2\}$,...,$\{1,2,...,d\}$ in $\mathcal{M}$. It follows according to item 2. in Proposition 5.3 that there exists an orthonormal Gram-Schmidt basis in $\mathcal{L}(\mathbf{X},\mathcal{M})$. We summarize:

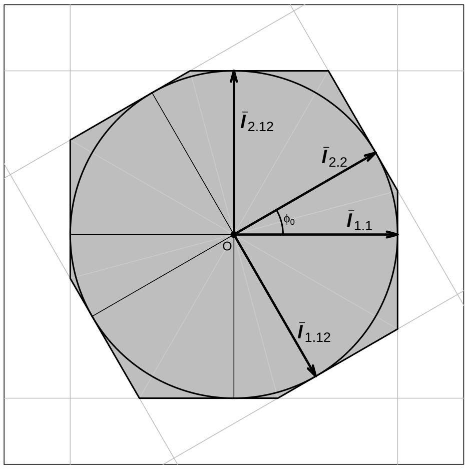

FIG 1. *The PoSI polytope $\mathbf{\Pi}_{K=1}$ tangent to the Scheffé disk (2-D ball) $\mathbf{B}_{K=1}$ for $d = p = 2$: The normalized raw predictor vectors are $\bar{\boldsymbol{l}}_{1\cdot\{1\}} \sim \mathbf{X}_1$ and $\bar{\boldsymbol{l}}_{2\cdot\{2\}} \sim \mathbf{X}_2$, and the normalized adjusted versions are $\bar{\boldsymbol{l}}_{1\cdot\{1,2\}}$ and $\bar{\boldsymbol{l}}_{2\cdot\{1,2\}}$. Shown in gray outline are the two squares (2-D cubes) generated by the o.n. bases $(\bar{\boldsymbol{l}}_{1\cdot\{1\}}, \bar{\boldsymbol{l}}_{2\cdot\{1,2\}})$ and $(\bar{\boldsymbol{l}}_{2\cdot\{2\}}, \bar{\boldsymbol{l}}_{1\cdot\{1,2\}})$, respectively. The PoSI polytope is the intersection of the two squares.*

PROPOSITION 5.4. *Among predictor matrices with $\mathrm{rank}(\mathbf{X}){=}d$ and model universes $\mathcal{M}$ that contain at least one maximal nested sequence of submodels, orthogonal designs with $p{=}d$ columns yield*

- *the maximal coverage probability $\mathbf{P}[\mathbf{Z}/\hat{\sigma} \in \mathbf{\Pi}_K]$ for fixed K, and*
- *the minimal PoSI constant K satisfying $\mathbf{P}[\mathbf{Z}/\hat{\sigma} \in \mathbf{\Pi}_K] = 1 - \alpha$ for fixed α: $\inf_{\mathrm{rank}(X)=d} K(\mathbf{X}, \mathcal{M}, \alpha, r) = K_{\mathrm{orth}}(\alpha, d, r)$.*

The proposition holds not only for multivariate t-vectors and their Gaussian limits but for arbitrary spherically symmetric distributions. — Optimality of orthogonal designs translates to optimal asymptotic behavior of their constant $K(\mathbf{X}, \alpha)$ for large d:

PROPOSITION 5.5. *Consider the Gaussian limit $r \to \infty$. For $\mathbf{X}$ and $\mathcal{M}$ as in Proposition 5.4, the asymptotic lower bound for the constant K as $d \to \infty$ is attained for orthogonal designs for which the asymptotic rate is*

$$\inf_{\mathrm{rank}(\mathbf{X})=d} K(\mathbf{X}, \mathcal{M}, \alpha) \;=\; K_{\mathrm{orth}}(d, \alpha) \;=\; \sqrt{2 \log d} + o(d).$$

The above facts show that the PoSI problem is bounded on one side by orthogonal designs: $\inf_{\mathrm{rank}(\mathbf{X})=d} K(\mathbf{X},\alpha,r) = K_{\mathrm{orth}}(\alpha,d,r)$, for all α, d and r. On the other side, the Scheffé ball yields a loose upper bound: $\sup_{\mathrm{rank}(\mathbf{X})=d,\mathcal{M}} K(\mathbf{X},\mathcal{M},\alpha,r) < K_{\mathrm{Sch}}(\alpha,d,r)$. The question of how close to the Scheffé bound $\sup_{\mathrm{rank}(\mathbf{X})=d,\mathcal{M}} K(\mathbf{X},\mathcal{M},\alpha,r)$ can get for $r \to \infty$ will occupy us in Section 6.2. Unlike the infimum problem, the supremum problem does not appear to have a unique optimizing design $\mathbf{X}$ uniformly in α, d and r.

5.7. *A Duality Property of PoSI Vectors.* In the classical case $d{=}p$ and $\mathcal{M}=\mathcal{M}_{\mathrm{all}}$ there exists a duality for PoSI vectors $\mathcal{L}(\mathbf{X})$ which we will use in Section 6.1 below but which is also of independent interest. We require some preliminaries: Letting $\mathrm{M}_F = \{1,2,...,p\}$ be the full model, we observe that the (unnormalized) PoSI vectors $\boldsymbol{l}_{j\cdot\mathrm{M}_F} = \mathbf{X}_{j\cdot\mathrm{M}_F}/\|\mathbf{X}_{j\cdot\mathrm{M}_F}\|^2$ form the rows of the matrix $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ because $\hat{\boldsymbol{\beta}}_F = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$. In a change of perspective, we interpret the transpose matrix

$$\mathbf{X}^* \;=\; \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}$$

as a predictor matrix as well, to be called the "dual design" of $\mathbf{X}$. It is also of size $p \times p$ in canonical coordinates, and its columns are the PoSI vectors $\boldsymbol{l}_{j\cdot\mathrm{M}_F}$. It turns out that $\mathbf{X}^*$ and $\mathbf{X}$ pose identical PoSI problems if $\mathcal{M}{=}\mathcal{M}_{\mathrm{all}}$:

THEOREM 5.1. $\mathcal{L}(\mathbf{X}^*) = \mathcal{L}(\mathbf{X})$, $\;\boldsymbol{\Pi}_K(\mathbf{X}^*) = \boldsymbol{\Pi}_K(\mathbf{X})$, $\;K(\mathbf{X}^*) = K(\mathbf{X})$.

Recall that $\mathcal{L}(\mathbf{X})$ and $\mathcal{L}(\mathbf{X}^*)$ contain the normalized versions of the respective adjusted predictor vectors. The theorem follows from the following lemma which establishes the identities of vectors between $\mathcal{L}(\mathbf{X}^*)$ and $\mathcal{L}(\mathbf{X})$. We extend obvious notations from $\mathbf{X}$ to $\mathbf{X}^*$ as follows:

$$\mathbf{X}^*_j \;=\; \boldsymbol{l}^*_{j\cdot\{j\}} \;=\; \boldsymbol{l}_{j\cdot\mathrm{M}_F}\,.$$

Submodels for $\mathbf{X}^*$ will be denoted M^*, but they, too, will be given as subsets of $\{1,2,...,p\}$ which, however, refer to columns of $\mathbf{X}^*$. Finally, the normalized version of $\boldsymbol{l}^*_{j\cdot\mathrm{M}^*}$ will be written as $\bar{\boldsymbol{l}}^*_{j\cdot\mathrm{M}^*}$.

LEMMA 5.1. *For two submodels* M *and* M^* *that satisfy* $\mathrm{M}\cap\mathrm{M}^* = \{j\}$ *and* $\mathrm{M}\cup\mathrm{M}^* = \mathrm{M}_F$*, we have*

$$\bar{\boldsymbol{l}}^*_{j\cdot\mathrm{M}^*} \;=\; \bar{\boldsymbol{l}}_{j\cdot\mathrm{M}}\,, \qquad \|\boldsymbol{l}^*_{j\cdot\mathrm{M}^*}\|\,\|\boldsymbol{l}_{j\cdot\mathrm{M}}\| \;=\; 1$$

The proof is in Appendix A.3. The assertion about norms is really only needed to exclude collapse of $\boldsymbol{l}^*_{j\cdot \mathrm{M}^*}$ to zero.

A special case arises when the predictor matrix (in canonical coordinates) is chosen to be symmetric according to Proposition 5.1 (*7.*): if $\mathbf{X}^T = \mathbf{X}$, then $\mathbf{X}^* = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} = \mathbf{X}^{-1}$, and hence:

Corollary 5.1. *If* $\mathbf{X}$ *is symmetric in canonical coordinates, then*

$$\mathcal{L}(\mathbf{X}^{-1}) = \mathcal{L}(\mathbf{X}), \quad \mathbf{\Pi}_K(\mathbf{X}^{-1}) = \mathbf{\Pi}_K(\mathbf{X}), \quad \textit{and} \quad K(\mathbf{X}^{-1}) = K(\mathbf{X})$$

6. Illustrative Examples and Asymptotic Results.

In this section we consider examples in the classical case $d = p$ and $\mathcal{M} = \mathcal{M}_{\mathrm{all}}$. Also, we work with the Gaussian limit $r \to \infty$ assuming $\sigma = 1$ is known.

6.1. *Example 1: The PoSI Problem for Exchangeable Designs.*

In exchangeable designs $\mathbf{X}$ all pairs of predictor vectors enclose the same angle. In canonical coordinates a convenient parametrization of a family of symmetric exchangeable design is

$$\mathbf{X} \;=\; \mathbf{I}_p + a\mathbf{E}_{p\times p}, \tag{6.1}$$

where $-1/p < a < \infty$, and $\mathbf{E}$ is a matrix with all entries equal to 1. The range restriction on a assures that $\mathbf{X}$ is positive definite. Writing $\mathbf{X} = \mathbf{X}(a)$ when the parameter a matters, we will make use of the fact that

$$\mathbf{X}(a)^{-1} = \mathbf{X}(-a/(1+pa))$$

is also an exchangeable design. The function $c_p(a) = -a/(1+pa)$ maps the interval $(-1/p, \infty)$ onto itself, and it holds $c_p(0) = 0$, $c_p(a) \downarrow -1/p$ as $a \uparrow +\infty$, and vice versa. Exchangeable designs include orthogonal designs for $a = 0$, and they extend to two types of strict collinearities: for $a \uparrow \infty$ the predictor vectors collapse to a single dimension span$(\mathbf{1})$, and for $a \downarrow -1/p$ they collapse to a subspace span$(\mathbf{1})^\perp$ of dimension $(p-1)$, where $\mathbf{1} = (1, 1, ..., 1)^T$.

As non-orthogonality/collinearity drives the fracturing of the regression coefficients into model-dependent quantities $\beta_{j\cdot \mathrm{M}}$, it is of interest to analyze $K(\mathbf{X})$ as $\mathbf{X} = \mathbf{X}(a)$ moves from orthogonality at $a = 0$ toward either of the two types of collinearity. Here is what we find:

- Unguided intuition might suggest that the collapse to rank 1 calls for larger $K(\mathbf{X})$ than the collapse to rank $(p-1)$. This turns out to be entirely wrong: collapse to rank 1 or rank $p-1$ has identical effects on $K(\mathbf{X})$. The reason is duality (Section 5.7): for exchangeable

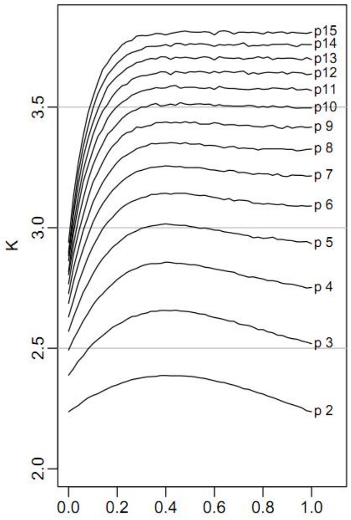

Fig 2. *The PoSI constant* $K(\mathbf{X}, \alpha = 0.05)$ *for exchangeable designs* $\mathbf{X} = \mathbf{I} + a\mathbf{E}$ *for* $a \in [0, \infty)$. *The horizontal axis shows* $a/(1+a)$, *hence the locations 0, 0.5 and 1.0 represent* $a = 0,\ 1,\ \infty$, *respectively. Surprisingly, the largest* $K(\mathbf{X})$ *is not attained at* $a = \infty$, *the point of perfect collinearity, at least not for dimensions up to* $p = 10$. *The graph is based on 10,000 random samples in* p *dimensions for* $p = 2, ..., 15$.

designs, $\mathbf{X}(a)$ collapses to rank 1 iff $\mathbf{X}(a)^* = \mathbf{X}(a)^{-1} = \mathbf{X}(-a/(1+pa))$ collapses to rank $p-1$, and vice versa, while $K(\mathbf{X}(a)^{-1}) = K(\mathbf{X}(a))$ according to Corollary 5.1.

- A more basic intuition would suggest that $K(\mathbf{X})$ increases as $\mathbf{X}$ moves away from orthogonality and approaches collinearity. Even this intuition is not fully born out: In Figure 2 we depict numerical approximations to $K(\mathbf{X}(a), \alpha = 0.05)$ for $a \in [0, \infty)$ ($a \in (-1/p, 0]$ being redundant due to duality). As the traces show, $K(\mathbf{X}(a))$ increases as $\mathbf{X}(a)$ moves away from orthogonality, up to a point, whereafter it descends as it approaches collinarity, at least for dimensions $p \leq 10$.

In summary, the dependence of $K(\mathbf{X})$ on the design $\mathbf{X}$ is not a simple matter. While duality provides some insights, there are no simple intuitions for inferring from $\mathbf{X}$ the geometry of the sets of unit vectors $\mathcal{L}(\mathbf{X})$, their polytopes $\mathbf{\Pi}_K$, their coverage probabilities and PoSI constants $K(\mathbf{X})$.

We next address the asymptotic behavior of $K = K(\mathbf{X}, \alpha, p)$ for increasing p. As noted in Section 4.7, there is a wide gap between orthogonal designs with $K_{\text{orth}} \sim \sqrt{2 \log p}$ and the full Scheffé protection with $K_{\text{Sch}} \sim \sqrt{p}$. The following theorem shows how exchangeable designs fall into this gap:

THEOREM 6.1. *PoSI constants of exchangeable design matrices* $\mathbf{X}(a)$ *have the following limiting behavior:*

$$\lim_{p\to\infty} \sup_{a\in(-1/p,\infty)} \frac{K(\mathbf{X}(a),\alpha,p)}{\sqrt{2\log p}} \;=\; 2.$$

The proof can be found in Appendix A.4. The theorem shows that for exchangeable designs the PoSI constant remains much closer to the orthogonal case than the Scheffé case. Thus, for this family of designs it is possible to improve on the Scheffé constant by a considerable margin.

The following detail of geometry for exchangeable designs has a bearing on the behavior of their PoSI constant: The angle between pairs of predictor vectors as a function of a is $\cos(\mathbf{X}_j(a),\mathbf{X}_{j'}(a)) = a(2+pa)/(pa^2+4a+2)$. In particular, as the vectors fall into the rank-$(p-1)$ collinearity at $a = -1/p$, the cosine becomes $-1/(2p-3)$, which converges to zero as $p\to\infty$. Thus, with increasing dimension, exchangeable designs approach orthogonal designs even at their most collinear extreme.

We finish with a geometric depiction of the limiting polytope $\mathbf{\Pi}_K$ as $\mathbf{X}(a)$ approaches either collinearity: For $a\uparrow\infty$, the predictor vectors fall into the 1-D subspace $\mathrm{span}(\mathbf{1})$, and for $a\downarrow -1/p$ they fall into $\mathrm{span}(\mathbf{1})^\perp$. With duality in mind and considering the permutation symmetry of exchangeable designs, it follows that the limiting polytope is a prismatic polytope with a p-simplex as its base in $\mathrm{span}(\mathbf{1})^\perp$. In Figure 3 we show this prism for $p=3$. The unit vectors $\bar{\boldsymbol{l}}_{1\cdot\{1\}}\sim\mathbf{X}_1$, $\bar{\boldsymbol{l}}_{2\cdot\{2\}}\sim\mathbf{X}_2$ and $\bar{\boldsymbol{l}}_{3\cdot\{3\}}\sim\mathbf{X}_3$ form an equilateral triangle. The plane $\mathrm{span}(\mathbf{1})^\perp$ also contains the six once-adjusted vectors $\bar{\boldsymbol{l}}_{j\cdot\{j,j'\}}$ $(j'\neq j)$, while the three fully adjusted vectors $\bar{\boldsymbol{l}}_{j\cdot\{1,2,3\}}$ collapse to $\mathbf{1}/\sqrt{p}$, turning the polytope into a prism.

6.2. *Example 2: Where* $K(\mathbf{X})$ *is close to the Scheffé Bound.* We describe a situation in which the asymptotic upper bound for $K(\mathbf{X},\alpha,p)$ is $O(\sqrt{p})$, hence close to the Scheffé constant K_{Sch} in terms of the asymptotic rate. We consider SPAR1 (Section 4.8) whereby a predictor of interest has been chosen, $\mathbf{X}_p$, say. The goal of model selection with SPAR1 is to "boost the effect" of $\mathbf{X}_p$ by adjusting it for optimally chosen predictors $\mathbf{X}_j$ $(j<p)$. The search is over the 2^{p-1} models that contain $\mathbf{X}_p$, but inference is sought only for the adjusted coefficient $\beta_{p\cdot\mathrm{M}}$.

The task is to construct a design for which simultaneous inference for all adjusted coefficients $\beta_{p\cdot\mathrm{M}}$ requires the constant $K_{\mathrm{SPAR1}}(\mathbf{X})$ to be in the order of $\sqrt{p}$. To this end consider the following upper triangular $p\times p$ design matrix in canonical coordinates:

$$\mathbf{X} = (\mathbf{e}_1,\ldots,\mathbf{e}_{p-1},\mathbf{1}_p)\,, \tag{6.2}$$

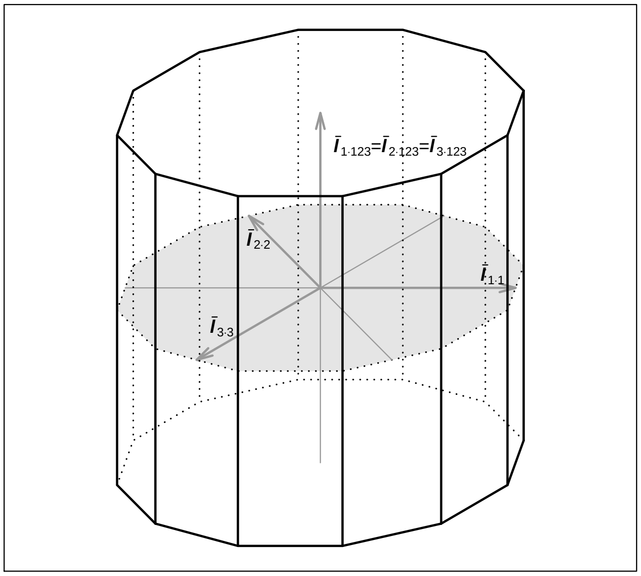
$\bar{l}_{1\cdot123}=\bar{l}_{2\cdot123}=\bar{l}_{3\cdot123}$
$\bar{l}_{2\cdot2}$
$\bar{l}_{1\cdot1}$
$\bar{l}_{3\cdot3}$

FIG 3. *Exchangeable Designs: The geometry of the limiting PoSI polytope for* $p = 3$ *as* $a \downarrow -1/p$ *or* $a \uparrow +\infty$ *in* (6.1).

where $\mathbf{e}_j$ are the canonical basis vectors, $(\mathbf{e}_j)_i = \delta_{ij}$, and $\mathbf{1}_p = (1, ..., 1)^T \in \mathbb{R}^p$. We have the following theorem:

THEOREM 6.2. *The designs* (6.2) *have SPAR1 simultaneous* $1 - \alpha$ *confidence intervals for* $\mathbf{X}_p$ *of the form* $\left[\hat{\beta}_p \pm K_{\text{SPAR1}}(\mathbf{X})\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{pp}}\right]$ *where*

$$\lim_{p\to\infty} \frac{K_{\text{SPAR1}}(\mathbf{X})}{\sqrt{p}} \;=\; 0.6363....$$

A (partial) proof is in Appendix A.5 where we show the $\geq$ part. As always, we consider the case of "large r," that is, σ known; for small r the constant is larger. The theorem shows that even if we restrict consideration to a single predictor $\mathbf{X}_p$ and its adjustments, the constant K_{SPAR1} to reach valid simultaneous inference against all submodels containing that coefficient can be much greater than the $O(1)$ t-quantiles used in common practice. Also, since for the unrestricted PoSI constant $K(\mathbf{X})$ we have $K(\mathbf{X}) \geq K_{\text{SPAR1}}(\mathbf{X})$, the theorem shows that there exist predictor matrices for which the PoSI constants are of the asymptotic order of the Scheffé constants.

6.3. *Bounding Away from Scheffé.*

We provide a rough asymptotic upper bound on all PoSI constants $K(\mathbf{X}, \mathcal{M}, \alpha, d)$. It is strictly smaller than the Scheffé constant but not by much. The bound, however, is loose because it is based on letting go of the rich structure of the sets $\mathcal{L}(\mathbf{X}, \mathcal{M})$ (Section 5.3) and only using their cardinality $|\mathcal{L}|$ $(= p\, 2^{p-1}$ in the classical case $d = p$ and $\mathcal{M} = \mathcal{M}_{\text{all}})$.

THEOREM 6.3. *Denote by $\mathcal{L}_d$ arbitrary finite sets of d-dimensional unit vectors, $\mathcal{L}_d \subset S^{d-1}$, such that $|\mathcal{L}_d| \leq a_d$ where $a_d^{1/d} \to a$ (> 0). Denote by $K(\mathcal{L}_d)$ the $(1-\alpha)$-quantile of $\sup_{\bar{\boldsymbol{l}} \in \mathcal{L}_d} |\bar{\boldsymbol{l}}^T \mathbf{Z}|$. Then the following describes an asymptotic worst-case bound for $K(\mathcal{L}_d)$ and its attainment:*

$$\lim_{d \to \infty} \sup_{|\mathcal{L}_d| \leq a_d} \frac{K(\mathcal{L}_d)}{\sqrt{d}} = \left(1 - \frac{1}{a^2}\right)^{1/2}.$$

The proof of Theorem 6.3 (see the Appendix A.6) is an adaptation of Wyner's (1967) techniques for sphere packing and sphere covering. The worst-case bound ($\leq$) is based on a surprisingly crude Bonferroni-style inequality for caps on spheres. Attainment of the bound ($\geq$) makes use of the artifice of picking the vectors $\bar{\boldsymbol{l}} \in \mathcal{L}$ randomly and independently. — Applying the theorem to PoSI sets $\mathcal{L} = \mathcal{L}(\mathbf{X}_{n \times p}, \mathcal{M}_{\text{all}})$ in the classical case $d = p$, we have $|\mathcal{L}| = p\, 2^{p-1} = a_p$, hence $a_p^{1/p} \to 2$, so the theorem applies with $a = 2$:

COROLLARY 6.1. *In the classical case $d = p$ a universal asymptotic upper bound for the PoSI constant $K(\mathbf{X}_{n \times p}, \mathcal{M}_{\text{all}})$ is*

$$\lim_{p \to \infty} \sup_{\mathbf{X}_{n \times p}} \frac{K(\mathbf{X}_{n \times p}, \mathcal{M}_{\text{all}})}{\sqrt{p}} \leq \frac{\sqrt{3}}{2} = 0.866\ldots.$$

The corollary shows that the asymptotic rate of the PoSI constant is strictly below that of the Scheffé constant, but possibly not by much. We do not know whether there exist designs for which the bound of the corollary is attained, but the theorem implies the bound is sharp for unstructured sets $\mathcal{L}$.

7. Computations.

The results in this section are for fixed finite r and p, and hence for t- rather than z-statistics. The computations of the design-specific constant $K = K(\mathbf{X}_{n\times p}, \mathcal{M}, \alpha, r)$ are MC-based, while those of the universal upper bound $K = K_{univ}(\alpha, d, r)$ derive from an analytical formula for a lower bound on the coverage probability inspired by Theorem 6.3.

7.1. *Computation of PoSI Constants $K = K(\mathbf{X}_{n\times p}, \alpha, r)$.*

For the derivation of an MC algorithm for PoSI problems, we continue from Proposition 5.2 in Section 5.2 which shows that the problem consists of estimating and calibrating a coverage probability $\mathbf{P}[\max_{\bar{\boldsymbol{l}}\in\mathcal{L}(\mathbf{X},\mathcal{M})} |\bar{\boldsymbol{l}}^T\mathbf{Z}/\hat{\sigma}| \leq K]$. As the algorithm discussed here is general and makes no use of the special structure of the set $\mathcal{L}(\mathbf{X},\mathcal{M})$ of PoSI coeffcient vectors, we write $\mathcal{L}$ from now on. As a reminder, in a PoSI problem the vectors in $\mathcal{L}$ are the adjusted predictor vectors $\mathbf{X}_{j\cdot\mathrm{M}}$ normalized to unit length and represented in canonical coordinates (Section 5.1), so that $\mathcal{L} \subset S^{d-1}$ according to (5.2) and (5.3).

The obvious MC algorithm would be to sample $\mathbf{Z}^{(1)}$,..., $\mathbf{Z}^{(I)}$ i.i.d. from $\mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)$ and independently $\hat{\sigma}^{(1)}$,..., $\hat{\sigma}^{(I)}$ i.i.d. from $\sqrt{\chi^2_r/r}$, and calibrate an MC estimate of the coverage probability with a bisection search for K:

$$\mathbf{P}\left[\max_{\bar{\boldsymbol{l}}\in\mathcal{L}} |\bar{\boldsymbol{l}}^T\mathbf{Z}|/\hat{\sigma} \leq K\right] \;\approx\; \frac{1}{I}|\{i\,|\,\max_{\bar{\boldsymbol{l}}\in\mathcal{L}} |\bar{\boldsymbol{l}}^T\mathbf{Z}^{(i)}|/\hat{\sigma}^{(i)} \leq K\}| \;=\; 1-\alpha.$$

It is, however, possible to increase precision by removing some of the variance from this MC simulation by integrating out the radial component of $\mathbf{Z}/\hat{\sigma}$. As the distribution of $\mathbf{Z}/\hat{\sigma}$ is spherically symmetric about $\mathbf{0}_d$, it can be decomposed into a radial and an angular component (see also the proof of Theorem 6.3, Appendix A.6):

$$\mathbf{Z}/\hat{\sigma} = R\mathbf{U}, \quad R^2/d \sim \mathrm{F}_{d,r}\,, \quad \mathbf{U} \sim \mathrm{Unif}(S^{d-1}), \quad R, \mathbf{U} \text{ independent.}$$

The maximal t-statistic becomes

$$\max_{\bar{\boldsymbol{l}}\in\mathcal{L}} |\bar{\boldsymbol{l}}^T(R\mathbf{U})| \;=\; R\max_{\bar{\boldsymbol{l}}\in\mathcal{L}} |\bar{\boldsymbol{l}}^T\mathbf{U}| \;=\; R\,C, \qquad C \triangleq \max_{\bar{\boldsymbol{l}}\in\mathcal{L}} |\bar{\boldsymbol{l}}^T\mathbf{U}|,$$

where C is the random variable that measures the maximal magnitude of the cosine between $\mathbf{U}$ and the vectors in $\mathcal{L}$. We integrate the radial component:

$$\begin{aligned}\mathbf{P}\left[\max_{\bar{\boldsymbol{l}}\in\mathcal{L}} |\bar{\boldsymbol{l}}^T\mathbf{Z}|/\hat{\sigma} \leq K\right] &= \mathbf{P}[\,RC \leq K\,]\\ &= \mathbf{E}[\,\mathbf{P}[\,R^2/d \leq (K/C)^2/d\,|\,C\,]\,]\\ &= \mathbf{E}[\,F_{\mathrm{F}_{d,r}}((K/C)^2/d)\,]\,,\end{aligned}$$

where $F_{\mathrm{F}_{d,r}}(...)$ denotes the c.d.f. of the F-distribution.

PROPOSITION 7.1. *If* $\mathbf{U} \sim \mathrm{Unif}(S^{p-1})$ *and* $C = \max_{\bar{\boldsymbol{l}} \in \mathcal{L}} |\bar{\boldsymbol{l}}^T \mathbf{U}|$, *then*

$$\mathbf{P}\left[\max_{\bar{\boldsymbol{l}} \in \mathcal{L}} |\bar{\boldsymbol{l}}^T \mathbf{Z}|/\hat{\sigma} \leq K\right] \;=\; \mathbf{E}[\, F_{\mathrm{F}_{d,r}}((K/C)^2/d)\,] \tag{7.1}$$

The expectation on the right hand side refers to the random variable C. The connection of this formula to Scheffé protection is as follows: The Scheffé case arises for $\mathcal{L} = S^{d-1}$ and hence $C \equiv 1$, in which case calibration of the right hand side of (7.1) requires $F_{\mathrm{F}_{d,r}}(K^2/d) = 1 - \alpha$, which reproduces the Scheffé constant $K_{\mathrm{Sch}} = \sqrt{d\, F^{-1}_{\mathrm{F}_{d,r}}(1-\alpha)}$. To gain on Scheffé, one needs $C < 1$ in distribution, which *is* the case for any finite $\mathcal{L}$. — The PoSI constant can now be approximated by calibrating an MC estimate of (7.1):

PoSI ALGORITHM: Sample i.i.d. unit vectors $\mathbf{U}^{(1)}, ..., \mathbf{U}^{(I)} \sim \mathrm{Unif}(S^{d-1})$. *Calculate their maximal cosine magnitudes:* $C^{(i)} = \max_{\bar{\boldsymbol{l}} \in \mathcal{L}} |\bar{\boldsymbol{l}}^T \mathbf{U}^{(i)}|$. *Calibrate the MC estimate of PoSI coverage with a bisection search for* K*:*

$$\frac{1}{I} \sum_{i=1,\ldots,I} F_{\mathrm{F}_{d,r}}((K/C^{(i)})^2/d) \;=\; 1 - \alpha\,. \tag{7.2}$$

The computational expense is in calculating the values $C^{(i)}$, which involves the inner product of $\mathbf{U}^{(i)}$ with the $|\mathcal{L}|$ (or up to $p\,2^{p-1}$) PoSI vectors in $\mathcal{L}$. Currently we completely enumerate the set $\mathcal{L}$, and this is why the computations are limited to about $p \leq 20$. For $p = 20$ ($p\,2^{p-1} = 10,485,760$), elapsed time on 2012 computing equipment is about one hour. Once the values $C^{(i)}$ are calculated (for $I = 1,000$, say), they are reused in (7.2) for all values of K that are tried in the bisection search, which therefore takes up negligible time. — An implementation in *R*-code (`http://www.r-project.org/`) can be obtained from the authors' web pages.

7.2. *Computation of Universal Upper Bounds* $K_{univ}(\mathcal{L}, \alpha, r)$. The same technique that produced the universal asymptotic upper bound in Section 6.3 (see the proof in Appendix A.6) can be used to compute a universal finite-r/finite-p upper bound: $K_{univ}(\alpha, p, r) \geq K(\mathbf{X}, \alpha, p, r)\ \forall \mathbf{X}$, yet strictly less than the Scheffé bound, $K_{univ}(\alpha, p, r) < K_{\mathrm{Sch}}(\alpha, p, r)$. This exercise has two purposes: (a) providing a computational method, and (b) showing that the asymptotic bound 0.866 of Theorem 6.3 is somewhat optimistic because for $p \to \infty$ it is approached from above.

As in Section 6.3 we ignore all structure in $\mathcal{L}$ except for its cardinality, hence the technique really works for arbitrary coverage problems involving many linear estimable functions. We start from the expression (7.1) for the coverage probability and observe that for any random variable C' that dominates C stochastically, $C' \overset{\mathcal{D}}{\geq} C$, we have

$$\mathbf{E}[F_{p,r}((K/C')^2/p)] \;\leq\; \mathbf{E}[F_{p,r}((K/C)^2/p)],$$

because $c \mapsto F_{p,r}((K/c)^2/p)$ is monotone decreasing for $c > 0$. Thus calibrating the left hand side results in a more conservative constant $K' \geq K$. We can create C' following the lead of Theorem 6.3 by using its proof's Bonferroni-style bound (A.15) and relying on the fact that every $(\bar{\boldsymbol{l}}^T \mathbf{U})^2$ has a Beta$(1/2,(p-1)/2)$-distribution (A.16):

$$\mathbf{P}[C > c] \;=\; \mathbf{P}[\max_{\bar{\boldsymbol{l}} \in \mathcal{L}} |\bar{\boldsymbol{l}}^T \mathbf{U}| > c] \;\leq\; |\mathcal{L}|(1 - F_{Beta,1/2,(p-1)/2}(c^2))$$

The r.h.s. depends on $\mathcal{L}$ (and hence on $\mathbf{X}$) only through the cardinality $|\mathcal{L}|$. We define the c.d.f. of C' in terms of a capped version of the r.h.s.:

Proposition 7.2. *Let C' be a random variable with the following c.d.f.:*

$$F_{C'}(c) = 1 - \min(1, |\mathcal{L}|(1 - F_{Beta,1/2,(p-1)/2}(c^2))).$$

Then a universal lower bound on PoSI coverage probabilities for all $\mathbf{X}$ is:

$$\mathbf{E}[F_{p,r}((K'/C')^2/p)] \;\leq\; \mathbf{P}\left[\max_{\mathrm{M} \in \mathcal{M}} \max_{j \in \mathrm{M}} |t_{j \cdot \mathrm{M}}| \leq K'\right]. \tag{7.3}$$

Thus, if the l.h.s. in (7.3) is calibrated with a search over K' so that

$$\mathbf{E}[F_{p,r}((K'/C')^2/p)] \;=\; 1 - \alpha,$$

we obtain a constant $K' = K_{univ}(\alpha,p,r)$ that satisfies for all $\mathbf{X}$

$$K(\mathbf{X},\alpha,p,r) \;\leq\; K_{univ}(\alpha,p,r) \;<\; K_{\mathrm{Sch}}(\alpha,p,r).$$

(For the last inequality note $\mathbf{P}[C' < 1] = 1$ but $C \equiv 1$ for Scheffé.)

Good approximations of the l.h.s. in (7.3) for arbitrary K' are obtained by calculating a grid of equi-probability quantiles for the distribution of C' once for all, and re-use it in the bisection search for K'. For a grid of length I, the grid points can be obtained by solving $F_{C'}(c_i) = i/(I+1)$ or, equivalently but more conveniently, $1 - F_{C'}(c_i) = i/(I+1)$:

$$c_i \;=\; F^{-1}_{Beta,1/2,(p-1)/2}\left(1 - \frac{i}{(I+1)\,|\mathcal{L}|}\right)^{1/2}.$$

Numerically this works for up to about $p = 40$ when $\mathcal{M} = \mathcal{M}_{\text{all}}$; for larger p the "Bonferroni" denominator $|\mathcal{L}| = p\,2^{p-1}$ creates quantiles that are too extreme for conventional numerical routines of Beta quantiles. If it works, the following approximation to the l.h.s. of (7.3) can be used for calibration:

$$\mathbf{E}[F_{p,r}((K'/C')^2/p)] \;\approx\; \frac{1}{I}\sum_{i=1\ldots I} F_{p,r}((K'/c_i)^2/p) \;=\; 1-\alpha.$$

A comparison of the distribution of the variable C' with the values $C \equiv 1$ (Scheffé) and $C = \sqrt{3}/2 = 0.866...$ (asymptotic bound) is of interest because it shows (a) to what degree simultaneous inference problems involving $|\mathcal{L}|$ estimable functions are necessarily less stringent than Scheffé, and (b) to what degree the asymptotic bound is approximated by C'. Such comparisons are given in Figure 4: all distributions are strictly below 1, beating Scheffé as they obviously should, but the asymptotic bound 0.866 is approached from above, which means that this value should be enlarged somewhat. In view of Figure 4, a good rough and ready rule for practice might be using a fraction 0.9 of the Scheffé constant as an approximate universal PoSI constant.

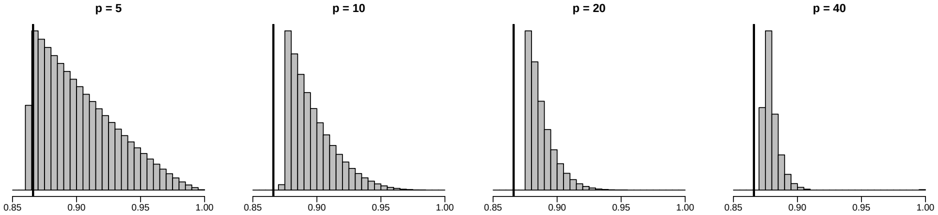

FIG 4. *Distribution of the variable C' of Section 7.2 used for calculating the universal upper bound on the PoSI constants. The vertical bold line on the left shows the asymptotic bound $\sqrt{3}/2$. The distributions approach the asymptotic value from above, thereby showing that universal upper bounds for finite p tend to be somewhat larger than the asymptotic limit, yet less than 1 and hence less than the Scheffé constant.*

known σ or error degrees of freedom $r \to \infty$)

8. Summary and Discussion.

We investigated the Post-Selection Inference or "PoSI" problem for linear models whereby valid statistical tests and confidence intervals are sought after variable selection, that is, after selecting a subset of the predictors in a data-driven way. We adopted a framework that does *not* assume any of the linear models under consideration to be correct. We allowed the response vector to be centered at an arbitrary mean vector but with homoscedastic and Gaussian errors. We further allowed the full predictor matrix $\mathbf{X}_{n\times p}$ to be rank-deficient, $d = \text{rank}(\mathbf{X}) < p$, and we also allowed the set $\mathcal{M}$ of models M under consideration to be largely arbitrary. In this framework we showed that valid post-selection inference is possible via simultaneous inference. An important enabling factor is the principle that the regression coefficient of a given predictor as distinct when it occurs in different submodels: $\beta_{j\cdot\mathrm{M}}$ and $\beta_{j\cdot\mathrm{M}'}$ are generally different parameters if $\mathrm{M} \neq \mathrm{M}'$. We showed that simultaneity protection for all parameters $\beta_{j\cdot\mathrm{M}}$ provides valid post-selection inference. In practice this means enlarging the constant $t_{1-\alpha/2,r}$ used in conventional inference to a constant $K(\mathbf{X}_{n\times p}, \alpha, r)$ that provides simultaneity protection for up to $p\,2^{p-1}$ parameters $\beta_{j\cdot\mathrm{M}}$. We showed that the constant depends strongly on the predictor matrix $\mathbf{X}$ as the asymptotic bound for $K(\mathbf{X}, \alpha, r)$ with $d = \text{rank}(\mathbf{X})$ ranges between the minimum of $\sqrt{2\log d}$ achieved for orthogonal designs on the one hand, and a large fraction of the Scheffé bound $\sqrt{d}$ on the other hand. This wide asymptotic range suggests that computation is critical for problems with large numbers of predictors. In the classical case $d = p$ our current computational methods are feasible up to about $p \approx 20$.

We carried out post-selection inference in a limited framework. Several problems remain open, and many natural extensions are desirable:

- Among open problems is the quest for the largest fraction of the asymptotic Scheffé rate $\sqrt{d}$ attained by PoSI constants. So far we know this fraction to be at least 0.6363 but no more than 0.8660... in the classical case $d = p$.
- Computations for $p > 20$ are a challenge. Straight enumeration of the set of up to $p\,2^{p-1}$ linear combinations should be replaced with heuristic shortcuts that yield practically useful upper bounds on $K(\mathbf{X}_{n\times p}, \mathcal{M}, \alpha, r)$ that are specific to $\mathbf{X}$ and the set of submodels $\mathcal{M}$, unlike the 0.8660 fraction of the Scheffé bound which is universal.
- The methodology is easily adapted to practically useful variations by suitable choice of the set of models $\mathcal{M}$: (1) Data analysts might be interested only in small submodels, $|\mathrm{M}| \leq 5$, say, when p is large. (2) We introduced SPAR ("Single Predictor Adjusted Regression", Section 4.8) defined as "significance hunting" or the search for the strongest adjusted "effect" in any predictor. In practice one might be more interested in SPAR1 or the search for strong adjusted effects in one predetermined focal predictor. — Any limitation to a lesser number of submodels or regression coefficients to be searched increases the computationally accessible number of predictors.
- Among models to which the PoSI framework should be extended next are generalized linear models and mixed effects models.

R code for computing the PoSI constant for up to $p = 20$ can be obtained from the authors' webpages.

Acknowledgments. We thank L. Dicker, M. Freiman, E. George, E. Pitkin, L. Shepp, N. Sloane, P. Shaman and M. Traskin for very helpful discussions. The acronym "SPAR" is due to M. Freiman.

APPENDIX A: PROOFS

A.1. Proof of Theorem 4.3. We start with the statement of strong family-wise error control by defining the true null hypotheses and true alternatives for the true $\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\mu}$, as well as the sets of insignificant and significant tests for the observed $\mathbf{Y}$:

$$\begin{aligned}
H_0 &\triangleq \{ (j, \mathrm{M}) \mid \beta_{j\cdot\mathrm{M}} = 0,\ j \in \mathrm{M} \in \mathcal{M} \}, \\
H_1 &\triangleq \{ (j, \mathrm{M}) \mid \beta_{j\cdot\mathrm{M}} \neq 0,\ j \in \mathrm{M} \in \mathcal{M} \}, \\
\hat{H}_0 &\triangleq \{ (j, \mathrm{M}) \mid |t^{(0)}_{j\cdot\mathrm{M}}| \leq K(\mathbf{X}, \alpha),\ j \in \mathrm{M} \in \mathcal{M} \}, \\
\hat{H}_1 &\triangleq \{ (j, \mathrm{M}) \mid |t^{(0)}_{j\cdot\mathrm{M}}| > K(\mathbf{X}, \alpha),\ j \in \mathrm{M} \in \mathcal{M} \}.
\end{aligned}$$

where $t^{(0)}_{j\cdot\mathrm{M}} \triangleq \hat{\beta}_{j\cdot\mathrm{M}}/(\hat{\sigma}/\|\mathbf{X}_{j\cdot\mathbf{M}}\|)$ has the parameter set to $\beta_{j\cdot\mathrm{M}} = 0$.

Lemma A.1. *"Strong Family-Wise Error Control" holds for $K(\mathbf{X}, \alpha)$:*

$$\mathbf{P}[H_0 \subset \hat{H}_0] \;=\; \mathbf{P}[H_1 \supset \hat{H}_1] \;\geq\; 1 - \alpha.$$

Proof: Standard; just the same: $H_0 \subset \hat{H}_0 \;\Leftrightarrow\; H_1 \supset \hat{H}_1$ implies the equality of the two probabilities. Further, using $t^{(0)}_{j\cdot\mathrm{M}} = t_{j\cdot\mathrm{M}} \Leftrightarrow (j, \mathrm{M}) \in H_0$,

$$\begin{aligned}
\mathbf{P}[\, H_0 \subset \hat{H}_0 \,] &= \mathbf{P}[\max_{(j,\mathrm{M}) \in H_0} |t_{j\cdot\mathrm{M}}| \leq K(\mathbf{X}, \alpha)\,] \\
&\geq \mathbf{P}[\max_{\mathrm{M} \in \mathcal{M}} \max_{j \in \mathrm{M}} |t_{j\cdot\mathrm{M}}| \leq K(\mathbf{X}, \alpha)\,] \;\geq\; 1 - \alpha
\end{aligned}$$

by the definition of $K(\mathbf{X}, \alpha)$ (4.11). □

THEOREM 4.3. *"Strong Post-Selection Error Control" holds for any model selection procedure $\hat{\mathrm{M}} : \mathbb{R}^n \to \mathcal{M}$:*

$$\mathbf{P}[\forall j \in \hat{\mathrm{M}} : |t^{(0)}_{j\cdot\hat{\mathrm{M}}}| > K(\mathbf{X}, \alpha) \;\Rightarrow\; \beta_{j\cdot\hat{\mathrm{M}}} \neq 0\,] \;\geq\; 1 - \alpha.$$

Proof: Define $\hat{\mathrm{M}}' \triangleq \{(j, \hat{\mathrm{M}}) \mid j \in \hat{\mathrm{M}}\}$. The event $H_1 \supset \hat{H}_1$ implies the event $H_1 \cap \hat{\mathrm{M}}' \supset \hat{H}_1 \cap \hat{\mathrm{M}}'$, hence, using Lemma A.1:

$$1 - \alpha \;\leq\; \mathbf{P}[\, H_1 \supset \hat{H}_1 \,] \;\leq\; \mathbf{P}[\, H_1 \cup \hat{\mathrm{M}}' \supset \hat{H}_1 \cup \hat{\mathrm{M}}' \,]. \qquad \square$$

A.2. Proof of Proposition 5.3.

1. The matrix $\mathbf{X}^*_{\mathrm{M}} = \mathbf{X}_{\mathrm{M}}(\mathbf{X}^T_{\mathrm{M}}\mathbf{X}_{\mathrm{M}})^{-1}$ has the vectors $\boldsymbol{l}_{j\cdot\mathrm{M}}$ as its columns. Thus $\boldsymbol{l}_{j\cdot\mathrm{M}} \in \mathrm{span}(\mathbf{X}_j : j \in \mathrm{M})$. Orthogonality $\boldsymbol{l}_{j\cdot\mathrm{M}} \perp \mathbf{X}_{j'}$ for $j' \neq j$ follows from $\mathbf{X}^T_{\mathrm{M}}\mathbf{X}^*_{\mathrm{M}} = \mathbf{I}_p$. The same properties hold for the normalized vectors $\bar{\boldsymbol{l}}_{j\cdot\mathrm{M}}$.
2. The vectors $\{\bar{\boldsymbol{l}}_{1\cdot\{1\}}, \ \bar{\boldsymbol{l}}_{2\cdot\{1,2\}}, \ \bar{\boldsymbol{l}}_{3\cdot\{1,2,3\}}, \ \ldots, \ \bar{\boldsymbol{l}}_{p\cdot\{1,2,\ldots,p\}}\}$ form a Gram-Schmidt series with normalization, hence they are an o.n. basis of $\mathbb{R}^p$.
3. For $\mathrm{M} \subset \mathrm{M}'$, $j \in \mathrm{M}$, $j' \in \mathrm{M}' \setminus \mathrm{M}$, we have $\bar{\boldsymbol{l}}_{j\cdot\mathrm{M}} \perp \bar{\boldsymbol{l}}_{j'\cdot\mathrm{M}}$ because they can be embedded in an o.n. basis by first enumerating M and subsequently $\mathrm{M}' \setminus \mathrm{M}$, with j being last in the enumeration of M and j' last in the enumeration of $\mathrm{M}' \setminus \mathrm{M}$.
4. For any (j_0, M_0), $j_0 \in \mathrm{M}_0$, there are $(p-1)\,2^{p-2}$ ways to choose a partner (j_1, M_1) such that either $j_1 \in \mathrm{M}_1 \subset \mathrm{M}_0 \setminus j_0$ or $\mathrm{M}_0 \subset \mathrm{M}_1 \setminus j_1$, both of which result in $\bar{\boldsymbol{l}}_{j_0\cdot\mathrm{M}_0} \perp \bar{\boldsymbol{l}}_{j_1\cdot\mathrm{M}_1}$ by the previous part.

A.3. Proof of Duality: Lemma 5.1 and Theorem5.1. The proof relies on a careful analysis of orthogonalities as described in Proposition 5.3, part *3*. In what follows we write $[\mathbf{A}]$ for the column space of a matrix $\mathbf{A}$, and $[\mathbf{A}]^\perp$ for its orthogonal complement. We show first that, for $\mathrm{M} \cap \mathrm{M}^* = \{j\}$, $\mathrm{M} \cup \mathrm{M}^* = \mathrm{M}_F$, the vectors $\bar{\boldsymbol{l}}^*_{j\cdot\mathrm{M}^*}$ and $\bar{\boldsymbol{l}}_{j\cdot\mathrm{M}}$ are in the same one-dimensional subspace, hence are a multiple of each other. To this end we observe:

$$\bar{\boldsymbol{l}}_{j\cdot\mathrm{M}} \in [\mathbf{X}_{\mathrm{M}}], \qquad \bar{\boldsymbol{l}}_{j\cdot\mathrm{M}} \in [\mathbf{X}_{\mathrm{M}\setminus j}]^\perp, \tag{A.1}$$

$$\bar{\boldsymbol{l}}^*_{j\cdot\mathrm{M}^*} \in [\mathbf{X}^*_{\mathrm{M}^*}], \qquad \bar{\boldsymbol{l}}^*_{j\cdot\mathrm{M}^*} \in [\mathbf{X}^*_{\mathrm{M}^*\setminus j}]^\perp, \tag{A.2}$$

$$[\mathbf{X}^*_{\mathrm{M}^*}] = [\mathbf{X}_{\mathrm{M}\setminus j}]^\perp, \qquad [\mathbf{X}^*_{\mathrm{M}^*\setminus j}]^\perp = [\mathbf{X}_{\mathrm{M}}]. \tag{A.3}$$

The first two lines state that $\bar{\boldsymbol{l}}_{j\cdot\mathrm{M}}$ and $\bar{\boldsymbol{l}}^*_{j\cdot\mathrm{M}^*}$ are in the respective column spaces of their models, but orthogonalized with regard to all other predictors in these models. The last line, which can also be obtained from the orthogonalities implied by $\mathbf{X}^T\mathbf{X}^* = \mathbf{I}_p$, establishes that the two vectors fall in the same one-dimensional subspace:

$$\bar{\boldsymbol{l}}_{j\cdot\mathrm{M}} \in [\mathbf{X}_{\mathrm{M}}] \cap [\mathbf{X}_{\mathrm{M}\setminus j}]^\perp \ = \ [\mathbf{X}^*_{\mathrm{M}^*}] \cap [\mathbf{X}^*_{\mathrm{M}^*\setminus j}]^\perp \ni \bar{\boldsymbol{l}}^*_{j\cdot\mathrm{M}^*}.$$

Since they are normalized, it follows $\bar{\boldsymbol{l}}^*_{j\cdot\mathrm{M}^*} = \pm\bar{\boldsymbol{l}}_{j\cdot\mathrm{M}}$. This result is sufficient to imply all of Theorem 5.1. The lemma, however, makes a slightly stronger statement involving lengths which we now prove. In order to express $\boldsymbol{l}_{j\cdot\mathrm{M}}$ and $\boldsymbol{l}^*_{j\cdot\mathrm{M}^*}$ according to (5.2), we use $\mathbf{P}_{\mathrm{M}\setminus j}$ as before and we write $\mathbf{P}^*_{\mathrm{M}^*\setminus j}$ for the analogous projection onto the space spanned by the columns $\mathrm{M}^* \setminus j$ of

$\mathbf{X}^*$. The method of proof is to evaluate $\boldsymbol{l}_{j\cdot\mathrm{M}}^T\, \boldsymbol{l}_{j\cdot\mathrm{M}^*}^*$. The main argument is based on

$$\mathbf{X}_j^T(\mathbf{I}-\mathbf{P}_{\mathrm{M}\setminus j})(\mathbf{I}-\mathbf{P}^*_{\mathrm{M}^*\setminus j})\mathbf{X}_j^* \;=\; 1, \tag{A.4}$$

which follows from these facts:

$$\mathbf{P}_{\mathrm{M}\setminus j}\mathbf{P}^*_{\mathrm{M}^*\setminus j}=\mathbf{0},\quad \mathbf{P}_{\mathrm{M}\setminus j}\mathbf{X}_j^*=\mathbf{0},\quad \mathbf{P}^*_{\mathrm{M}^*\setminus j}\mathbf{X}_j=\mathbf{0},\quad \mathbf{X}_j^T\mathbf{X}_j^*=1,$$

which in turn are consequences of (A.3) and $\mathbf{X}^T\mathbf{X}^* = \mathbf{I}_p$. We also know from (5.2) that

$$\|\boldsymbol{l}_{j\cdot\mathrm{M}}\| \;=\; 1/\|(\mathbf{I}-\mathbf{P}_{\mathrm{M}\setminus j})\mathbf{X}_j\|\,, \qquad \|\boldsymbol{l}^*_{j\cdot\mathrm{M}^*}\| \;=\; 1/\|(\mathbf{I}-\mathbf{P}^*_{\mathrm{M}^*\setminus j})\mathbf{X}^*_j\|\,. \tag{A.5}$$

Putting together (A.4), (A.5), and (5.2), we obtain

$$\boldsymbol{l}_{j\cdot\mathrm{M}}^T\,\boldsymbol{l}^*_{j\cdot\mathrm{M}^*} \;=\; \|\boldsymbol{l}_{j\cdot\mathrm{M}}\|^2\,\|\boldsymbol{l}^*_{j\cdot\mathrm{M}^*}\|^2 \;>\; 0. \tag{A.6}$$

Because the two vectors are scalar multiples of each other, we also know that

$$\boldsymbol{l}_{j\cdot\mathrm{M}}^T\,\boldsymbol{l}^*_{j\cdot\mathrm{M}^*} \;=\; \pm\,\|\boldsymbol{l}_{j\cdot\mathrm{M}}\|\,\|\boldsymbol{l}^*_{j\cdot\mathrm{M}^*}\|. \tag{A.7}$$

Putting together (A.6) and (A.7) we conclude

$$\|\boldsymbol{l}_{j\cdot\mathrm{M}}\|\,\|\boldsymbol{l}^*_{j\cdot\mathrm{M}^*}\| \;=\; 1, \qquad \bar{\boldsymbol{l}}^*_{j\cdot\mathrm{M}^*} \;=\; \bar{\boldsymbol{l}}_{j\cdot\mathrm{M}},$$

This proves the lemma and the theorem. □

A.4. Proof of Theorem 6.1. The parameter a can range from $-1/p$ to ∞, but because of duality there is no loss of generality in considering only the case in which $a \geq 0$, and we do so in the following. Let $\mathrm{M} \subset \{1,\ldots,p\}$ and $j \in \mathrm{M}$. If $\mathrm{M} = \{j\}$ then $\boldsymbol{l}_{j\cdot\mathrm{M}} = \mathbf{X}_j$, the j-th column of $\mathbf{X}$, and $\bar{\boldsymbol{l}}_{j\cdot\mathrm{M}} = \boldsymbol{l}_{j\cdot\mathrm{M}}/\sqrt{pa^2+2a+1}$. It follows that for $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}_p, \mathbf{I}_p)$,

$$|\bar{\boldsymbol{l}}_{j\cdot\mathbf{M}}^T\mathbf{Z}| \leq |\sum_{k\neq j} Z_k|/\sqrt{p} + |Z_j| \leq \sqrt{2\log p}(1+o_p(1)) \tag{A.8}$$

because $\|\mathbf{Z}\|_\infty = (1+o_p(1))\sqrt{2\log p}$.

Because of (A.8) we now need only consider model selection sets, M, that contain at least two indices. For notational convenience, consider the case that $j = 1$ and $\mathrm{M} = \{1,\ldots,m\}$ with $2 \leq m \leq p$. The following results can then be applied to arbitrary j and M by permuting coordinates.

When $m \geq 2$ the projection of $\mathbf{X}_1$ on the space spanned by $\mathbf{X}_2, \dots, \mathbf{X}_m$ must be of the form

$$\mathrm{Proj} \;=\; \frac{c}{m-1}\sum_{k=2}^{m}\mathbf{X}_k \;=\; \left(ca, \underbrace{ca + \frac{c}{m-1}, \dots, ca + \frac{c}{m-1}}_{m-1}, \underbrace{ca, \dots, ca}_{p-m} \right)$$

where the constant c satisfies $\boldsymbol{l}_{1\cdot\mathrm{M}} = (\mathbf{X}_1 - \mathrm{Proj}) \perp \mathrm{Proj}$. This follows from symmetry; no calculation of projection matrices is needed to verify this. Let $d = 1 - c$. Then

$$(\boldsymbol{l}_{1\cdot\mathrm{M}})_k = \begin{cases} 1 + da & k = 1 \\ da - \frac{1-d}{m-1} & 2 \leq k \leq m \\ da & k \geq m+1 \end{cases}.$$

Some algebra starting from $\boldsymbol{l}_{1\cdot\mathrm{M}}^T \mathbf{X}_2 = 0$ yields

$$d \;=\; \frac{1/(m-1)}{pa^2 + 2a + 1/(m-1)}.$$

The term $d = d(a)$ is a simple rational function of a, and it is easy to check when $m \geq 2$ that $0 \leq da < 1/(2\sqrt{p})$.

Note also that $\|\boldsymbol{l}_{1\cdot\mathrm{M}}\| \geq 1$. Hence $\bar{\boldsymbol{l}}_{1\cdot\mathrm{M}} = \boldsymbol{l}_{1\cdot\mathrm{M}}/\|\boldsymbol{l}_{1\cdot\mathrm{M}}\|$ satisfies

$$|\bar{\boldsymbol{l}}_{1\cdot\mathrm{M}}^T \mathbf{Z}| \leq |Z_1| + |\frac{1}{m-1}\sum_{j=2}^{m} Z_j| + |\frac{1}{2\sqrt{p}}\sum_{j=1}^{p} Z_j| \leq 2\sqrt{2\log p}(1 + o_p(1)) + O_p(1).$$

This verifies that

$$\limsup_{p\to\infty} \frac{\sup_{a\in(-1/p,\infty)} K(\mathbf{X}(a))}{\sqrt{2\log p}} \leq 2 \quad \text{in probability.} \tag{A.9}$$

It remains to prove that equality holds in (A.9). Let $Z_{(1)} < Z_{(2)} < \ldots < Z_{(p)}$ denote the order statistics of $\mathbf{Z}$. Fix m. It is well-known that, in probability,

$$\lim_{p\to\infty} \frac{Z_{(1)}}{\sqrt{2\log p}} = -1 \quad \text{and} \quad \lim_{p\to\infty} \frac{Z_{(j)}}{\sqrt{2\log p}} = 1 \quad \forall j :\; p - m \;\leq\; j \;\leq\; p.$$

Note that

$$\lim_{a\to\infty} da = 0 \quad \text{and} \quad \lim_{a\to\infty} \|\boldsymbol{l}_{1\cdot\mathrm{M}}\|^2 = 1 + (m-1)^{-1}.$$

For any given $\mathbf{Z}$ one may choose to look at $\boldsymbol{l}_{j^*\cdot \mathrm{M}^*}$, with j^* being the index of $Z_{(1)}$ and $\mathrm{M}^* = \{j^*\} \cup \{j \,|\, Z_j = Z_{(k)}, p-m+2 \leq k \leq p\}$. The above then yields, in probability,

$$\lim_{p\to\infty, a\to\infty} \frac{|\bar{\boldsymbol{l}}_{j^*\cdot \mathrm{M}^*}^T \mathbf{Z}|}{\sqrt{2\log p}} \geq \frac{2}{\sqrt{1+(m-1)^{-1}}}.$$

Choosing m arbitrarily large and combining this with (A.9) yields the desired conclusion.

A.5. Partial Proof of Theorem 6.2. While the theorem is correct as stated, of interest is only the inequality

$$K_{\text{SPAR1}}(\mathbf{X}) \;\geq\; (0.6363...)\,\sqrt{p}\,(1+o_P(1)), \tag{A.10}$$

the point being that a non-zero fraction of the Scheffé rate $\sqrt{p}$ can be attained by SPAR1 constants. As the proof of the reverse inequality is lengthy, we provide here only the more straightforward proof of inequality (A.10) and indicate below the missing part which can be obtained from the authors on request. The following preparations are required for both inequalities.

To find $\bar{\boldsymbol{l}}_{p\cdot \mathrm{M}}$, we need to adjust the predictor of interest $\mathbf{X}_p = \mathbf{1}_p$ for other predictors $\mathbf{X}_j = \mathbf{e}_j$ $(j < p)$ in the model M. In this case adjusting means zeroing out the components of $\mathbf{X}_p$ for $j \in \mathrm{M}$ with the exception of $j = p$, hence the z-statistic (5.1) for the predictor of interest are

$$z_{p\cdot \mathrm{M}} \;=\; \bar{\boldsymbol{l}}_{p\cdot \mathrm{M}}^T \mathbf{Z} \;=\; \frac{Z_p + \sum_{j\notin \mathrm{M}} Z_j}{\left(1+\sum_{j\notin \mathrm{M}} 1\right)^{1/2}}\,.$$

We will consider only the one-sided problem based on $\max_{\mathrm{M}(\ni p)} z_{p\cdot \mathrm{M}}$ as the two-sided criterion is the larger of the one-sided criteria for $\mathbf{Z}$ and $-\mathbf{Z}$, which are asymptotically the same. We also simplify the problem by dropping the terms Z_p and 1 which are asymptotically irrelevant:

$$\max_{\mathrm{M}\,(\ni p)} \frac{z_{p\cdot \mathrm{M}}}{\sqrt{p}} \;=\; \max_{\mathrm{M}:\, p\in \mathrm{M},\, |\mathrm{M}|>1} \frac{z'_{p\cdot \mathrm{M}}}{\sqrt{p}} \;+\; o_P(1)\,,$$

where

$$z'_{p\cdot \mathrm{M}} \;=\; \frac{\sum_{j\notin \mathrm{M}} Z_j}{\left(\sum_{j\notin \mathrm{M}} 1\right)^{1/2}}\,.$$

Next we observe that for a fixed model size $|\mathrm{M}| = \sum_{j\in \mathrm{M}} 1 = m$ (> 1) and a given $\mathbf{Z}$ the maximizing model has to include the predictors j for which Z_j

is smallest, hence is of the form $\mathrm{M}_B = \{\, j \,|\, Z_j < B \,\} \cup \{p\}$, where B is chosen such that $|\mathrm{M}| = m$. It is therefore sufficient to consider only models of the form M_B. Furthermore, we can limit the search to $B \geq 0$ because adding j with $Z_j < 0$ to the model increases the numerator of the above ratio and makes it positive, and also it decreases the numerator, thereby increasing the ratio again:

$$\max_{\mathrm{M}:\, p \in \mathrm{M},\ |\mathrm{M}|>1} z'_{p\cdot\mathrm{M}} \;=\; \max_{B \geq 0} z'_{p\cdot\mathrm{M}_B} \;=\; \max_{B \geq 0} \frac{\sum_{j<p} Z_j\, \mathbf{1}_{Z_j > B}}{\left(\sum_{j<p} \mathbf{1}_{Z_j > B}\right)^{1/2}} .$$

The asymptotic properties of the right hand ratio is provided by the following lemma:

Lemma A.2. *Define $A(B) = \phi(B)/\sqrt{1 - \Phi(B)}$. Then it holds uniformly in $B \geq 0$:*

$$\frac{z'_{p\cdot\mathrm{M}_B}}{\sqrt{p}} \;=\; A(B) + o_P(1).$$

It is the uniformity in the statement of the lemma that is needed to prove the reverse inequality of (A.10). We provide here only the simple proof of the pointwise statement, which is sufficient for (A.10): Because $\mathbf{E}[Z_j\, \mathbf{1}_{Z_j > B}] = \phi(B)$, we have for $p \to \infty$

$$\frac{1}{p-1} \sum_{j<p} Z_j\, \mathbf{1}_{Z_j > B} \;\overset{P}{\to}\; \phi(B), \qquad \frac{1}{p-1} \sum_{j<p} \mathbf{1}_{Z_j > B} \;\overset{P}{\to}\; 1 - \Phi(B).$$

The pointwise assertion of the lemma follows. □

Continuing with the proof of the theorem, we look for $B \geq 0$ that makes $z_{p\cdot\mathrm{M}_B}$ asymptotically the largest. Following the lead of the lemma we obtain the maximum of $A(B)$ as $A_{\max} = \max_{B \geq 0} \phi(B)/\sqrt{1 - \Phi(B)} = 0.6363...$, attained at $B_{max} \approx 0.6121$. (The graph of $A(B)$ is shown in Figure 5.) We have therefore $z_{p\cdot\mathrm{M}_{B_{max}}} = (0.6363... + o_P(1))\sqrt{p}$. Since $\max_B z_{p\cdot\mathrm{M}_B} \geq z_{p\cdot\mathrm{M}_{B_{max}}}$, the SPAR1 constant is lower-bounded by $K_{\text{SPAR1}} \geq 0.6363...\sqrt{p}\,(1 + o_p(1))$. □

A.6. Proof of Theorem 6.3. We will show that if $a_p^{1/p} \to a\ (> 0)$, we have

- a uniform asymptotic worst-case bound:

$$\lim_{p\to\infty} \sup_{|\mathcal{L}_p| \leq a_p} \max_{\bar{\boldsymbol{l}} \in \mathcal{L}_p} |\bar{\boldsymbol{l}}^T \mathbf{Z}|/\sqrt{p} \;\overset{\mathbf{P}}{\leq}\; \sqrt{1 - 1/a^2};$$

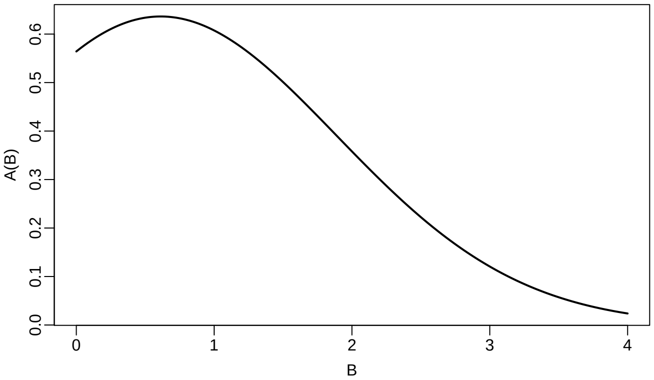

FIG 5. *The function* $A(B) = \frac{\phi(B)}{\sqrt{1-\Phi(B)}}$ *from the proof of Theorem 6.2 in Appendix A.5.*

- attainment of the bound when $|\mathcal{L}_p| = a_p$ and $\bar{\boldsymbol{l}} \in \mathcal{L}_p$ are i.i.d. Unif(S^{p-1}) independent of $\mathbf{Z}$:
$$\lim_{p\to\infty} \max_{\bar{\boldsymbol{l}}\in\mathcal{L}_p} |\bar{\boldsymbol{l}}^T\mathbf{Z}|/\sqrt{p} \overset{\mathbf{P}}{\geq} \sqrt{1-1/a^2}.$$

These facts imply the assertions about $(1-\alpha)$-quantiles $K(\mathcal{L}_p)$ of $\max_{\bar{\boldsymbol{l}}\in\mathcal{L}_p} |\bar{\boldsymbol{l}}^T\mathbf{Z}|$ in Theorem 6.3. We decompose $\mathbf{Z} = R\mathbf{U}$ where $R^2 = \|\mathbf{Z}\|^2 \sim \chi_p^2$ and $\mathbf{U} = \mathbf{Z}/\|\mathbf{Z}\| \sim$ Unif(S^{p-1}) are independent. Due to $R/\sqrt{p} \xrightarrow{\mathbf{P}} 1$ it is sufficient to show the following:

- uniform asymptotic worst-case bound:
$$\lim_{p\to\infty} \sup_{|\mathcal{L}_p|\leq a_p} \max_{\bar{\boldsymbol{l}}\in\mathcal{L}_p} |\bar{\boldsymbol{l}}^T\mathbf{U}| \overset{\mathbf{P}}{\leq} \sqrt{1-1/a^2}\,; \tag{A.11}$$

- attainment of the bound when $|\mathcal{L}_p| = a_p$ and $\bar{\boldsymbol{l}} \in \mathcal{L}_p$ are i.i.d. Unif(S^{p-1}) independent of $\mathbf{U}$:
$$\lim_{p\to\infty} \max_{\bar{\boldsymbol{l}}\in\mathcal{L}_p} |\bar{\boldsymbol{l}}^T\mathbf{U}| \overset{\mathbf{P}}{\geq} \sqrt{1-1/a^2}\,. \tag{A.12}$$

To show (A.11), we upper-bound the non-coverage probability and show that it converges to zero for $K' > \sqrt{1-1/a^2}$. To this end we start with a

Bonferroni-style bound, as in Wyner (1967):

$$\mathbf{P}[\max_{\bar{\boldsymbol{l}}\in\mathcal{L}} |\bar{\boldsymbol{l}}^T\mathbf{U}| > K'] = \mathbf{P}\bigcup_{\bar{\boldsymbol{l}}\in\mathcal{L}}[\,|\bar{\boldsymbol{l}}^T\mathbf{U}| > K'] \tag{A.13}$$

$$\leq \sum_{\bar{\boldsymbol{l}}\in\mathcal{L}} \mathbf{P}[\,|\bar{\boldsymbol{l}}^T\mathbf{U}| > K'] \tag{A.14}$$

$$= |\mathcal{L}_p|\,\mathbf{P}[\,|U| > K'], \tag{A.15}$$

where U is any coordinate of $\mathbf{U}$ or projection of $\mathbf{U}$ onto a unit vector. We will show that the bound (A.15) converges to zero. We use the fact that $U^2 \sim \text{Beta}(1/2, (p-1)/2)$, hence

$$\mathbf{P}[\,|U| > K'] = \frac{1}{\mathrm{B}(1/2,(p-1)/2)} \int_{K'^2}^{1} x^{-1/2}(1-x)^{(p-3)/2}dx \tag{A.16}$$

We bound the Beta function and the integral separately:

$$\frac{1}{\mathrm{B}(1/2,(p-1)/2)} = \frac{\Gamma(p/2)}{\Gamma(1/2)\Gamma((p-1)/2)} < \sqrt{\frac{(p-1)/2}{\pi}}\,,$$

where we used $\Gamma(x+1/2)/\Gamma(x) < \sqrt{x}$ (a good approximation, really) and $\Gamma(1/2) = \sqrt{\pi}$.

$$\int_{K'^2}^{1} x^{-1/2}(1-x)^{(p-3)/2}dx \leq \frac{1}{K'}\frac{1}{(p-1)/2}(1-K'^2)^{(p-1)/2},$$

where we used $x^{-1/2} \leq 1/K'$ on the integration interval. Continuing with the chain of bounds from (A.15) we have:

$$|\mathcal{L}_p|\mathbf{P}[\,|U| > K'] \leq \frac{1}{K'}\left(\frac{2}{(p-1)\pi}\right)^{1/2}\left(|\mathcal{L}_p|^{1/(p-1)}\sqrt{1-K'^2}\right)^{p-1}.$$

Since $|\mathcal{L}_p|^{1/(p-1)} \to a\ (>0)$ by assumption, the right hand side converges to zero at geometric speed if $a\sqrt{1-K'^2} < 1$, that is, if $K' > \sqrt{1-1/a^2}$. This proves (A.11).

To show (A.12), we upper-bound the coverage probability and show that it converges to zero for $K' < \sqrt{1-1/a^2}$. We make use of independence of $\bar{\boldsymbol{l}} \in \mathcal{L}_p$, as in Wyner (1967):

$$\mathbf{P}[\max_{\bar{\boldsymbol{l}}\in\mathcal{L}_p} |\bar{\boldsymbol{l}}^T\mathbf{U}| \leq K'] = \prod_{\bar{\boldsymbol{l}}\in\mathcal{L}_p} \mathbf{P}[\,|\bar{\boldsymbol{l}}^T\mathbf{U}| \leq K'] = \mathbf{P}[\,|U| \leq K']^{|\mathcal{L}_p|} \tag{A.17}$$

$$= \left(1-\mathbf{P}[\,|U| > K']\right)^{|\mathcal{L}_p|} \tag{A.18}$$

$$\leq \exp\left(-|\mathcal{L}_p|\,\mathbf{P}[\,|U| > K']\right). \tag{A.19}$$

We will lower-bound the probability $\mathbf{P}[\,|U| > K']$ recalling (A.16) and again deal with the Beta function and the integral separately:

$$\frac{1}{\mathrm{B}(1/2,(p-1)/2)} \;=\; \frac{\Gamma(p/2)}{\Gamma(1/2)\Gamma((p-1)/2)} \;>\; \sqrt{\frac{p/2-3/4}{\pi}}\,,$$

where we used $\Gamma(x{+}1)/\Gamma(x{+}1/2) > \sqrt{x+1/4}$ (again, a good approximation really).

$$\int_{K'^2}^{1} x^{-1/2}(1-x)^{(p-3)/2}dx \;\geq\; \frac{1}{(p-1)/2}(1-K'^2)^{(p-1)/2},$$

where we used $x^{-1/2} \geq 1$. Putting it all together we bound the exponent in (A.19):

$$|\mathcal{L}_p|\,\mathbf{P}[\,|U| > K'] \;\geq\; \frac{\sqrt{p/2-3/4}}{\sqrt{\pi}\,(p-1)/2}\left(|\mathcal{L}_p|^{1/(p-1)}\sqrt{1-K'^2}\right)^{p-1}.$$

Since $|\mathcal{L}_p|^{1/(p-1)} \to a$ (> 0) by assumption, the right hand side converges to $+\infty$ at nearly geometric speed if $a\sqrt{1-K'^2} > 1$, that is, if $K' < \sqrt{1-1/a^2}$. This proves (A.12).

REFERENCES

[1] Angrist, X. and Pischke, X. (2009). *Mostly Harmless Econometrics*, ... xxxxx

[2] Berk, R., Brown, L. D. and Zhao, L. (2010). Statistical Inference after Model Selection. *Journal of Quantitative Criminology* **26**, 217–236.

[3] Brown, L. D. (1967). The Conditional Level of Student's t-Test. *The Annals of Mathematical Statistics* **38**, 1068–1071.

[4] Brown, L. D. (1990). An Ancillarity Paradox which Appears in Multiple Linear Regression. *The Annals of Statistics* **18**, 471–493.

[5] Buehler, R. J. and Fedderson, A. P. (1963). Note on a conditional property of Student's t *The Annals of Mathematical Statistics* **34**, 1098–1100.

[6] Dijkstra, T. K. and Veldkamp, J. H. (1988). Data-driven Selection of Regressors and the Bootstrap, in *On Model Uncertainty and Its Statistical Implications* (T. K. Dijkstra, ed.), 17–38, Berlin: Springer.

[7] Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning — Data Mining, Inference, and Prediction*, 2nd ed. Corr. 3rd printing. Springer Series in Statistics, New York: Springer.

[8] Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., Bornkamp, B. and Hothorn, T. (2010). *mtvnorm: Multivariate Normal and t Distributions*, http://cran.r-project.org/web/packages/mvtnorm/

[9] Kabaila, P. (1998). Valid Confidence Intervals in Regression after Variable Selection. , –.

[10] Kabaila, P. and Leeb, H. (2006). On the Large-Sample Minimal Coverage Probability of Confidence Intervals after Model Selection. *Journal of the American Statistical Association* **101** (474), 619–629.

[11] Leeb, H. (2006). The Distribution of a Linear Predictor after Model Selection: Unconditional Finaite-Sample Distributions and Asymptotic Approximations. *IMS Lecture Notes - Monograph Series* **49**, 291–311.

[12] Leeb, H. and Pötscher, B. M. (2003). The Finite-Sample Distributions of Post-Model-Selection Estimators and Uniform versus Nonuniform Approximations. *Econometric Theory* **19**, 100–142.

[13] Leeb, H. and Pötscher, B. M. (2005). Model Selection and Inference: Facts and Fiction, *Econometric Theory* **21**, 21–59.

[14] 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.

[15] Leeb, H. and Pötscher, B. M. (2006). Can One Estimate the Conditional Distribution of Post-Model-Selection Estimators? *The Annals of Statistics* **34**, 2554–2591.

[16] Leeb, H. and Pötscher, B. M. (2008a). Model Selection, in *The Handbook of Financial Time Series* (T. G. Anderson, R. A. Davis, J. -P. Kreib, and T. Mikosch, eds), 785–821, New York: Springer.

[17] Leeb, H. and Pötscher, B. M. (2008b). Can One Estimate the Unconditional Distribution of Post-Model-Selection Estimators? *Econometric Theory* **24** (2), 338–376.

[18] Leeb, H. and Pötscher, B. M. (2008c). Sparse Estimators and the Oracle Property, or the Return of Hodges' Estimator. *Journal of Econometrics* **142**, 201–211.

[19] Moore, D. S. and McCabe, G. P. (2003). *Introduction to the Practice of Statistics*, 4th ed., New York: W. H. Freeman and Company.

[20] Mosteller, F. and Tukey, J. W. (1977). *Data Analysis and Regression: A Second Course in Statistics*, New York: Addison-Wesley.

[21] Olshen, R. A. (1973). The Conditional Level of the F-Test. *Journal of the American Statistical Association* **68**, 692–698.

[22] Pacifico, M. P., Genovese, C., Verdinelli, I. and Wasserman, L. (2004). False Discovery Control for Random Fields. *Journal of the American Statistical Association*, **99** (468), 1002–1014.

[23] Pötscher, B. M. (1991). Effects of Model Selection on Inference. *Econometric Theory* **7**, 163–185.

[24] Pötscher, B. M. and Leeb, H. (2009). On the Distribution of Penalized Maximum Likelihood Estimators: The LASSO, SCAD, and Thresholding. *Journal of Multivariate Analysis* **100**, 2065–2082.

[25] Scheffé, H. (1953). A Method for Judging All Contrasts in the Analysis of Variance. *Biometrika* **40**, 87–104.

[26] Scheffé, H. (1959). *The Analysis of Variance*, New York: John Wiley & Sons.

[27] Sen, P. K. (1979). Asymptotic Properties of Maximum Likelihood Estimators Based on Conditional Specification. *The Annals of Statistics*, **7**, 742–755.

[28] Sen, P. K. and Saleh, A. K. M. E. (1987). On Preliminary Test and Shrinkage M-Estimation in Linear Models. *The Annals of Statistics*, **15**, 1580–1592.

[29] Wyner, A. D. (1967). Random Packings and Coverings of the Unit n-Sphere. *Bell System Technical Journal*, **46**, 2111–2118.

Statistics Department, The Wharton School, University of Pennsylvania,
471 Jon M. Huntsman Hall, Philadelphia, PA 19104-6340.
Office: (215) 898-8222, Fax: (215) 898-1280.
E-mail: berk@wharton.upenn.edu, lbrown@wharton.upenn.edu, buja.at.wharton@gmail.com,
zhangk@wharton.upenn.edu, lzhao@wharton.upenn.edu