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 and hence suitably widening conventional confidence and retention intervals. 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

[*]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

intervals, the practice is so pervasive that it appears in classical undergraduate textbooks on statistics such as Moore and McCabe (2003).

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 other regression diagnostics. Finally, models become stochastic in an 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 exercised 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, be 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 homoscedastic 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 within 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 specific problems have been the subject of incisive analyses and criticisms by the "Vienna School" of Pötscher, Leeb and Schneider; see, for example, Leeb and Pötscher (2003; 2005; 2006a; 2006b; 2008a; 2008b), Pötscher (2006), Leeb (2006), Pötscher and Leeb (2009), Pötscher and Schneider (2009, 2010, 2011), as well as Kabaila and Leeb (2006) and Kabaila (2009).

This article proceeds as follows: In Section 2 we first develop the "submodel view" of the targets of inference after model selection and contrast it with the "full model view" (Section 2.1); we then introduce assumptions with a view toward valid inference in "wrong models" (Section 2.2). Sec-

tion 3 is about estimation and its targets from the submodel point of view. 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. We conclude with a discussion in Section 7. Some proofs are deferred to the appendix, and some elaborations to the online appendix.

Computations will be described in a separate article. Simulation-based methods yield satisfactory accuracy specific to a design matrix up to $p \approx 20$, while non-asymptotic universal upper bounds can be computed for larger p.

2. Targets of Inference and Assumptions. It is a natural intuition that model selection distorts inference by distorting sampling distributions of parameter estimates: Estimates in selected models should tend to generate more Type I errors than conventional theory allows because the typical selection procedure favors models with strong, hence highly significant predictors. This intuition correctly points to a multiplicity problem that grows more severe as the number of predictors subject to selection increases. This is the problem we address in this article.

Model selection poses additional problems that are less obvious but no less fundamental: There exists an ambiguity as to the role and meaning of the parameters in submodels. On one view, the relevant parameters are always those of the full model, hence the selection of a submodel is interpreted as estimating the deselected parameters to be zero and estimating the selected parameters under a zero constraint on the deselected parameters. On another view, the submodel has its own parameters, and the deselected parameters are not zero but non-existent. These distinctions are not academic as they imply fundamentally different ideas regarding the targets of inference, the measurement of statistical performance, and the problem of post-selection inference. The two views derive from different purposes of equations:

- Underlying the full model view of parameters is the use of a full equation to describe a "data generating" mechanism for the response; the equation hence has a causal interpretation.
- Underlying the submodel view of parameters is the use of any equation to merely describe association between predictor and response variables; no data generating or causal claims are implied.

In this article we address the latter use of equations. Issues relating to the former use are discussed in the online appendix B.1.

2.1. *The Submodel Interpretation of Parameters.* In what follows we elaborate three points that set the submodel interpretation of coefficients apart from the full model interpretation, with important consequences for the rest of this article:

(1) The full model has no special status other than being the repository of available predictors.
(2) The coefficients of excluded predictors are not zero; they are not defined and therefore don't exist.
(3) The meaning of a predictor's coefficient depends on which other predictors are included in the selected model.

(1) The full model available to the statistician often cannot be argued to have special status because of inability to identify and measure all relevant predictors. Additionally, even when a large and potentially complete suite of predictors can be measured there is generally a question of predictor redundancy that may make it desirable to omit some of the measurable predictors from the final model. It is a common experience in the social sciences that models proposed on theoretical grounds are found on empirical grounds to have their predictors entangled by collinearities that permit little meaningful statistical inference. 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 strongly confounded. The emphasis on 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. Generally, in any creative observational study involving novel predictors it will be difficult a priori to exclude collinearities that might force a rethinking of the predictors. 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.

(2) In the submodel interpretation of parameters, claiming that the coefficients of deselected predictors are zero does not properly describe the role of predictors. Deselected predictors have no role in the submodel equation; they become no different than predictors that had never been considered. The selected submodel becomes the vehicle of substantive research irrespective of what the full model was. As such the submodel stands on its own. This view is especially appropriate if the statistician's task is to determine which predictors are to be measured in the future.

(3) The submodel interpretation of parameters is deeply seated in how we teach regression. We explain that the meaning of a regression coefficient depends on which of the other predictors are included in the model: "the slope is the average difference in the response for a unit difference in the predictor, *at fixed levels of all other predictors in the model*." This "ceteris paribus" clause is essential to 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. For example, if purchase likelihood of a high-tech gadget is predicted from *Age*, it might be found against expectations that younger people have lower purchase likelihood, whereas a regression on *Age* and *Income* might show that at fixed levels of income younger people have indeed higher purchase likelihood. This case of Simpson's paradox would be enabled by the expected positive collinearity between *Age* and *Income*. Thus the marginal slope on *Age* is distinct from the *Income*-adjusted slope on *Age* as the two slopes answer different questions, apart from having opposite signs. In summary, *different models result in different parameters with different meanings*.

Must we use the full model with both predictors? Not if *Income* data is difficult to obtain or if it provides little improvement in R^2 beyond *Age*. The model based on *Age* alone cannot be said to be a priori "wrong". If, for example, the predictor and response variables have jointly multivariate normal distributions, then every linear submodel is "correct". These considerations drive home, once again, that sometimes no model has special status.

In summary, a range of applications call for a framework in which the full model is not the sole provider of parameters, where rather each submodel defines its own. The consequences of this view will be developed in Section 3.

2.2. *Assumptions, Models as Approximations, and Error Estimates.* We state assumptions for estimation and for the construction of valid tests and CIs when fitting arbitrary linear equations. The main goal is to prepare the ground for valid statistical inference after model selection — *not* assuming that selected models are correct.

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

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

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 $\mathbf{X}$. That is, 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 quote. What he calls "wrong models" 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. The main reason for elaborating this point is as follows: after model selection the case for "correct models" is clearly questionable, even for "consistent model selection procedures" (Leeb and Pötscher 2003, p. 101); but if correctness of submodels is not assumed, it is only natural to abandon this assumption for the full model also, in line with the idea that the full model has no special status. 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 homoscedasticity and normality are as questionable as first order correctness, and we will report elsewhere on approaches that avoid them. For now we follow the vast model selection literature that relies on the technical advantages of assuming homoscedastic and normal errors.

Accepting the assumption (2.3), we address the issue of estimating the error variance σ^2, because the valid tests and CIs we construct 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 residual (MSR) $\hat{\sigma}^2_F = \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2/(n-p)$ 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 MSR 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 MSR 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 random 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.
- A different type of estimates $\hat{\sigma}^2$ may be based on considerations borrowed from non-parametric function estimation (Hall and Carroll 1989).

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 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 2.1, the value and meaning of a regression coefficient depends on what the other predictors in the model are. An exception occurs, of course, when the predictors are perfectly orthogonal, as in some designed experiments or in function fitting with orthogonal basis functions. In this case a coefficient has the same value and meaning across all submodels. This article is hence 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 purpose here is to elaborate the idea that the slope 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. To describe the situation 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 in-

dexed by M. We will only allow submodels M 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? Following Section 2.1, we will not interpret $\hat{\boldsymbol{\beta}}_\mathrm{M}$ as estimates of the full model coefficients and, more generally, of any model other than M. Thus it is natural to ask that $\hat{\boldsymbol{\beta}}_\mathrm{M}$ define its own target through the requirement of unbiasedness:

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

This definition requires no other assumption than the existence of $\boldsymbol{\mu}=\mathbf{E}[\mathbf{Y}]$. 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 $\|\boldsymbol{\mu} - \mathbf{X}_\mathrm{M}\boldsymbol{\beta}_\mathrm{M}\|^2$.

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 (including the 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 first-order correctness assumptions. Equation (3.4) follows from $\mathrm{span}(\mathbf{X}_\mathrm{M}) \subset \mathrm{span}(\mathbf{X})$.

Notation: To distinguish the regression coefficients of the predictor $\mathbf{X}_j$ relative to the submodel it appears 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 is that indexes are always elements of the full model, $j \in \{1, 2, ..., p\} = \mathrm{M}_F$, for what we call "full model indexing".

3.2. *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 phrase "average difference in the response approximated in the submodel M". The reason is that the target of the fit $\hat{\mathbf{Y}}_{\mathrm{M}} = \mathbf{X}_{\mathrm{M}}\hat{\boldsymbol{\beta}}_{\mathrm{M}}$ in the submodel M is $\boldsymbol{\mu}_{\mathrm{M}} = \mathbf{X}_{\mathrm{M}}\boldsymbol{\beta}_{\mathrm{M}}$, hence in M we estimate unbiasedly not the true $\boldsymbol{\mu}$ but its LS approximation $\boldsymbol{\mu}_{\mathrm{M}}$.

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 predictors:

$$\hat{\beta}_{j\cdot\mathrm{M}} = \frac{\mathbf{X}_{j\cdot\mathrm{M}}^T\mathbf{Y}}{\|\mathbf{X}_{j\cdot\mathrm{M}}\|^2}, \qquad \beta_{j\cdot\mathrm{M}} = \frac{\mathbf{X}_{j\cdot\mathrm{M}}^T\boldsymbol{\mu}}{\|\mathbf{X}_{j\cdot\mathrm{M}}\|^2}. \tag{3.5}$$

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 $\mathbf{Y}$ 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: To unclutter notation 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 (1) case-wise slopes through the origin, both as estimates $\hat{\beta}_{(i)} = Y_i/x_i$ and as parameters $\beta_{(i)} = \mu_i/x_i$, and (2) case-wise weights $w_{(i)} = x_i^2/\sum_{i'=1\ldots n} x_{i'}^2$. Equations (3.5) 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.* We consider inference for $\hat{\boldsymbol{\beta}}_{\mathrm{M}}$ and its target $\boldsymbol{\beta}_{\mathrm{M}}$. Following Section 2.2 we require a normal homoscedastic model for $\mathbf{Y}$, but we leave its mean $\boldsymbol{\mu} = \mathbf{E}[\mathbf{Y}]$ entirely 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 2.2 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 M:

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

where $(...)_{jj}$ 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 MSR $\hat{\sigma}_\mathrm{M}$ from the submodel M, for two reasons:

- We do not assume that the submodel M is first-order correct, hence $\hat{\sigma}_\mathrm{M}^2$ would in general have a distribution that is a multiple of a non-central χ^2 distribution with unknown non-centrality parameter.
- More disconcertingly, $\hat{\sigma}_\mathrm{M}^2$ would be the result of selection: $\hat{\sigma}_{\hat{\mathrm{M}}}^2$ (see Section 4.2). Not much of real use is known about its distribution (see, for example, Brown 1967 and Olshen 1973).

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

With this choice of $\hat{\sigma}$, confidence intervals for $\beta_{j\cdot\mathrm{M}}$ take the form

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

If $K = t_{r,1-\alpha/2}$ is the $1-\alpha/2$ quantile of a t-distribution with r degrees of freedom, then the interval is marginally valid with a $1-\alpha$ coverage guarantee:

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

This holds if the submodel M is *not* the result of variable selection.

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 $\mathbf{Y}$ ($\mathbf{X}$ being fixed, Section 2.2). This model should therefore be expressed as $\hat{\mathrm{M}} = \hat{\mathrm{M}}(\mathbf{Y})$. In general we allow a variable 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 $\mathcal{M}_{\text{all}}$ is the set of all full-rank submodels:

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

Thus the procedure $\hat{\mathrm{M}}$ is a discrete map that divides $\mathbb{R}^n$ into as many as $|\mathcal{M}_{all}|$ 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:
  - It has a random dimension $m(\mathbf{Y}) = |\hat{\mathrm{M}}(\mathbf{Y})|$: $\boldsymbol{\beta}_{\hat{\mathrm{M}}(\mathbf{Y})} \in \mathbb{R}^{m(\mathbf{Y})}$.
  - For any fixed j, it may or may not be the case that $j \in \hat{\mathrm{M}}(\mathbf{Y})$.
  - 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. *Post-Selection Coverage Guarantees for Confidence Intervals.* With randomness of the selected model and its parameters in mind, what is a desirable form of post-selection coverage guarantee for confidence intervals? A natural requirement would be a $1{-}\alpha$ confidence guarantee for the coefficients of the predictors that are selected into the model:

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

Several points should be noted:

- The guarantee is family-wise for all selected predictors $j \in \hat{\mathrm{M}}$, though the sense of "family-wise" is unusual because $\hat{\mathrm{M}} = \hat{\mathrm{M}}(\mathbf{Y})$ is random.
- The guarantee has nothing to say about predictors $j \notin \hat{\mathrm{M}}$ that have been deselected, regardless of the substantive interest they might have. Predictors of overarching interest should be protected from variable selection, and for these one can use a modification of the PoSI approach which we call "PoSI1"; see Section 4.10.
- Because predictor selection is random, $\hat{\mathrm{M}} = \hat{\mathrm{M}}(\mathbf{Y})$, two realized samples $\mathbf{y}^{(1)},\ \mathbf{y}^{(2)} \in \mathbb{R}^n$ from $\mathbf{Y}$ may result in different sets of selected

predictors, $\hat{\mathrm{M}}(\mathbf{y}^{(1)}) \neq \hat{\mathrm{M}}(\mathbf{y}^{(2)})$. It would be a fundamental misunderstanding to wonder whether the guarantee holds for both realizations. Instead, the guarantee (4.5) is about the *procedure*

$$\mathbf{Y} \;\; \mapsto \;\; \hat{\sigma}(\mathbf{Y}),\ \hat{\mathrm{M}}(\mathbf{Y}),\ \hat{\boldsymbol{\beta}}_{\hat{\mathrm{M}}(\mathbf{Y})}(\mathbf{Y}) \;\; \mapsto \;\; \mathrm{CI}_{j\cdot\hat{\mathrm{M}}}(K)\ (j \in \hat{\mathrm{M}})$$

for the long run of independent realizations of $\mathbf{Y}$ (by the LLN), and not for any particular realizations $\mathbf{y}^{(1)}$, $\mathbf{y}^{(2)}$. A standard formulation used to navigate these complexities after a realization $\mathbf{y}$ of $\mathbf{Y}$ has been analyzed is the following: "For $j \in \hat{\mathrm{M}}$ we have $1-\alpha$ *confidence* that the interval $\mathrm{CI}_{j\cdot\hat{\mathrm{M}}(\mathbf{y})}(K)$ contains $\beta_{j\cdot\hat{\mathrm{M}}(\mathbf{y})}$."

- Marginal guarantees for individual predictors require some care because $\beta_{j\cdot\hat{\mathrm{M}}}$ does not exist for $j \notin \hat{\mathrm{M}}$. This makes $\beta_{j\cdot\hat{\mathrm{M}}} \in \mathrm{CI}_{j\cdot\hat{\mathrm{M}}}(K)$ an incoherent statement that does not define an event. Guarantees are possible if the condition $j \in \hat{\mathrm{M}}$ is added with a conjunction or is being conditioned on: the marginal and conditional probabilities

$$\mathbf{P}\left[\, j \in \hat{\mathrm{M}} \;\&\; \beta_{j\cdot\hat{\mathrm{M}}} \in \mathrm{CI}_{j\cdot\hat{\mathrm{M}}}(K_{j\cdot}) \right] \quad \text{and} \quad \mathbf{P}\left[\, \beta_{j\cdot\hat{\mathrm{M}}} \in \mathrm{CI}_{j\cdot\hat{\mathrm{M}}}(K_{j\cdot}) \;\middle|\; j \in \hat{\mathrm{M}} \right]$$

are both well-defined and can be the subject of coverage guarantees; see the online Appendix B.4.

Finally, we note that the smallest constant K that satisfies the guarantee (4.5) is specific to the procedure $\hat{\mathrm{M}}$. Thus different variable selection procedures would require different constants. Finding procedure-specific constants is a challenge that will be intentionally bypassed by the present proposals.

4.4. *Universal Validity for all Selection Procedures.* 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 4.9.

(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 realistic 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) Related to the previous point is the fact that today there is a realization that a considerable fraction of published empirical work is unreproducible or reports exaggerated effects (well-known in this regard is Ioannidis 2005). A factor contributing to this problem might well be liberal handling of variable selection and absent accounting for it in subsequent inference.

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 $\mathcal{M}_{\text{all}}$ (4.4) are always being searched. Rather, the search is often 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 of interest, 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 allowable submodels to be an (almost) arbitrary but pre-specified non-empty subset of $\mathcal{M}_{\text{all}}$; w.l.o.g. we can assume $\bigcup_{\mathrm{M}\in\mathcal{M}} \mathrm{M} = \{1, 2, ..., p\}$. 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 $d = \mathrm{rank}(\mathbf{X})$. — Selection procedures are now maps

$$\hat{\mathrm{M}} : \ \mathbf{Y} \mapsto \hat{\mathrm{M}}(\mathbf{Y}), \qquad \mathbb{R}^n \to \mathcal{M}. \tag{4.7}$$

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: $VIF_{j\cdot\mathrm{M}} = \|\mathbf{X}_j\|^2 / \|\mathbf{X}_{j\cdot\mathrm{M}}\|^2$ if $\mathbf{X}_j$ is centered, hence does not make use of $\mathbf{Y}$, and elimination according to $VIF_{j\cdot\mathrm{M}} > c$ 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* $\mathbf{Y} \in \mathbb{R}^n$*:*

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

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

The right hand max-$|t|$ bound of the lemma is sharp in the sense that there exists a variable selection procedure $\hat{\mathrm{M}}$ that attains the bound; see Section 4.9. — Next we introduce the $1-\alpha$ quantile of the right hand max-$|t|$ statistic of the lemma: Let K be the minimal value that satisfies

$$\mathbf{P}\left[\max_{\mathrm{M}\in\mathcal{M}}\max_{j\in\mathrm{M}}|t_{j\cdot\mathrm{M}}|\le K\right] \;\ge\; 1-\alpha. \tag{4.8}$$

This value will be called "the PoSI constant". It does not depend on any model selection procedures, but it does depend on the design matrix $\mathbf{X}$, the universe $\mathcal{M}$ of models subject to selection, the desired coverage $1-\alpha$, and the degrees of freedom r in $\hat{\sigma}$, hence $K = K(\mathbf{X}, \mathcal{M}, \alpha, r)$.

THEOREM 4.1. *For all model selection procedures* $\hat{\mathrm{M}} : \mathbb{R}^n \to \mathcal{M}$ *we have*

$$\mathbf{P}\left[\max_{j\in\hat{\mathrm{M}}}|t_{j\cdot\hat{\mathrm{M}}}|\le K\right] \;\ge\; 1-\alpha, \tag{4.9}$$

where $K = K(\mathbf{X}, \mathcal{M}, \alpha, r)$ *is the PoSI constant.*

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. The following is just a repackaging of Theorem 4.1:

COROLLARY 4.1. *"Simultaneous Post-Selection Confidence Guarantees" hold for any model selection procedure* $\hat{\mathrm{M}}\colon \mathbb{R}^n \to \mathcal{M}$*:*

$$\mathbf{P}\left[\,\beta_{j\cdot\hat{\mathrm{M}}} \in \mathrm{CI}_{j\cdot\hat{\mathrm{M}}}(K)\;\forall j \in \hat{\mathrm{M}}\,\right] \;\ge\; 1-\alpha, \tag{4.10}$$

where $K = K(\mathbf{X}, \mathcal{M}, \alpha, r)$ *is the PoSI constant.*

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

COROLLARY 4.2. *"Strong Post-Selection Error Control" holds for any model selection procedure* $\hat{\mathrm{M}}\colon \mathbb{R}^n \to \mathcal{M}$*:*

$$\mathbf{P}\left[\,\exists j \in \hat{\mathrm{M}} : \beta_{j\cdot\hat{\mathrm{M}}} \ne 0 \;\&\; |t^{(0)}_{j\cdot\hat{\mathrm{M}}}| > K\,\right] \;\le\; \alpha,$$

where $K = K(\mathbf{X}, \mathcal{M}, \alpha, r)$ *is the PoSI constant and* $t^{(0)}_{j\cdot\mathrm{M}}$ *is the t-statistic for the null hypothesis* $\beta_{j\cdot\mathrm{M}} = 0$.

The proof is standard (see online appendix B.3). The corollary states that, with probability $1-\alpha$, in a selected model *all* PoSI-significant rejections have detected true alternatives.

4.7. *Computation of the POSI Constant.* Several portions of the following treatment are devoted to a better understanding of the structure and value of the POSI constant $K(\mathbf{X}, \mathcal{M}, \alpha, r)$. Except for very special choices it does not seem possible to provide closed form expressions for its value. However the structural geometry and other properties to be described later do enable a reasonably efficient computational algorithm. R-code for computing the POSI constant for small to moderate values of p is available on the authors' web pages. This code is accompanied by a manuscript that will be published elsewhere describing the computational algorithm and generalizations. For the basic setting involving $\mathcal{M}_{\text{all}}$ the algorithm will conveniently provide values of $K(\mathbf{X}, \mathcal{M}_{\text{all}}, \alpha, r)$ for matrices $\mathbf{X}$ of rank ≤ 20, or slightly larger depending on available computing speed and memory. It can also be adapted to compute K for some other families contained within $\mathcal{M}_{\text{all}}$, such as some discussed in Section 4.5.

4.8. *Scheffé Protection.* Realizing the idea that the LS estimators in different submodels are generally unbiased estimates of 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 (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 $\mathbf{x} \in \text{span}(\mathbf{X})$:

$$t_{\mathbf{x}} \;\triangleq\; \frac{(\mathbf{Y}-\boldsymbol{\mu})^T\mathbf{x}}{\hat{\sigma}\|\mathbf{x}\|}. \tag{4.11}$$

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

$$\mathbf{P}\left[\sup_{\mathbf{x}\in\text{span}(\mathbf{X})} |t_{\mathbf{x}}| \le K_{\text{Sch}}\right] \;=\; 1-\alpha, \tag{4.12}$$

where the Scheffé constant is

$$K_{\text{Sch}} \;=\; K_{\text{Sch}}(\alpha, d, r) \;=\; \sqrt{d\mathrm{F}_{d,r,1-\alpha}}. \tag{4.13}$$

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

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

Thus for $j \in \hat{\text{M}}$ a parameter estimate $\hat{\beta}_{j \cdot \hat{\text{M}}}$ whose t-ratio exceeds K_{Sch} in magnitude is universally safe from having the rejection of "$H_0 : \beta_{j \cdot \hat{\text{M}}} = 0$" invalidated by variable 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 worthwhile. An indication is given by the following comparison as $r \to \infty$:

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

(For orthogonal designs see Section 5.5.) Thus the PoSI constant K_{orth} is much smaller than K_{Sch}. The large gap between the two suggests that the Scheffé constant may be too conservative at least in some cases. We will study 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 an example 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-root-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.

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

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

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 submodel $\hat{\text{M}}_{SPAR}(\mathbf{Y})$ 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.

4.10. *One Primary Predictor and Controls — "PoSI1".* Sometimes a regression analysis is centered on a predictor of interest, $\mathbf{X}_j$, and on inference for its coefficient $\beta_{j\cdot\mathrm{M}}$. The other predictors in M act as controls, so their purpose is to adjust the primary predictor for confounding effects and possibly to boost the primary predictor's own "effect". This situation is characterized by two features:

- Variable selection is limited to models that contain the primary predictor. We therefore define for any model universe $\mathcal{M}$ a sub-universe $\mathcal{M}_{j\cdot}$ of models that contain the primary predictor $\mathbf{X}_j$:

$$\mathcal{M}_{j\cdot} \triangleq \{\mathrm{M} \,|\, j \in \mathrm{M} \in \mathcal{M}\},$$

  so that for $\mathrm{M} \in \mathcal{M}$ we have $j \in \mathrm{M}$ iff $\mathrm{M} \in \mathcal{M}_{j\cdot}$.
- Inference is sought for the primary predictor $\mathbf{X}_j$ only, hence the relevant test statistic is now $|t_{j\cdot\mathrm{M}}|$ and no longer $\max_{j\in\mathrm{M}} |t_{j\cdot\mathrm{M}}|$. The former statistic is coherent because it is assumed that $j \in \mathrm{M}$.

We call this the "PoSI1" situation in contrast to the unconstrained PoSI situation. Similar to PoSI, PoSI1 starts with a "significant triviality bound":

LEMMA 4.2. *("Primary Predictor's Significant Triviality Bound") For a fixed predictor* $\mathbf{X}_j$ *and model selection procedure* $\hat{\mathrm{M}} \colon \mathbb{R}^n \to \mathcal{M}_{j\cdot}$*, it holds:*

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

For a "proof", the only thing to note is $j \in \hat{\mathrm{M}}(\mathbf{Y})$ by the assumption $\hat{\mathrm{M}}(\mathbf{Y}) \in \mathcal{M}_{j\cdot}$. — We next define the "PoSI1" constant for the predictor $\mathbf{X}_j$ as the $1-\alpha$ quantile of the max-$|t|$ statistic on the right side of the lemma: Let $K_{j\cdot} = K_{j\cdot}(\mathbf{X}, \mathcal{M}, \alpha, r)$ be the minimal value that satisfies

$$\mathbf{P}\left[\max_{\mathrm{M}\in\mathcal{M}_{j\cdot}} |t_{j\cdot\mathrm{M}}| \le K_{j\cdot}\right] \;\ge\; 1-\alpha. \tag{4.14}$$

Importantly, this constant is dominated by the general PoSI constant:

$$K_{j\cdot}(\mathbf{X}, \mathcal{M}, \alpha, r) \;\le\; K(\mathbf{X}, \mathcal{M}, \alpha, r), \tag{4.15}$$

for the obvious reason that the present max-$|t|$ is smaller than the general PoSI max-$|t|$ due to $\mathcal{M}_{j\cdot} \subset \mathcal{M}$ and the restriction of inference to $\mathbf{X}_j$. The constant $K_{j\cdot}$ provides the following "PoSI1" guarantee shown as the analog of Theorems 4.1 and Corollary 4.1 folded into one:

THEOREM 4.2. *Let* $\hat{\mathrm{M}}: \mathbb{R}^n \to \mathcal{M}_{j\cdot}$ *be a selection procedure that always includes the predictor* $\mathbf{X}_j$ *in the model. Then we have*

$$\mathbf{P}\left[\,|t_{j\cdot\hat{\mathrm{M}}}| \le K_{j\cdot}\right] \;\ge\; 1-\alpha, \tag{4.16}$$

and accordingly we have the following post-selection confidence guarantee:

$$\mathbf{P}\left[\,\beta_{j\cdot\hat{\mathrm{M}}} \in \mathrm{CI}_{j\cdot\hat{\mathrm{M}}}(K_{j\cdot})\right] \;\ge\; 1-\alpha. \tag{4.17}$$

Inequality (4.16) is immediate from Lemma 4.2. The "triviality bound" of the lemma is attained by the following variable selection procedure which we name "SPAR1":

$$\hat{\mathrm{M}}_{j\cdot}(\mathbf{Y}) \;\triangleq\; \operatorname*{argmax}_{\mathrm{M}\in\mathcal{M}_{j\cdot}} |t_{j\cdot\mathrm{M}}(\mathbf{Y})|. \tag{4.18}$$

It is a potentially realistic description of some data analyses when a predictor of interest is determined a priori, and the goal is to optimize *this* predictor's "effect". This procedure requires the full protection of the PoSI1 constant $K_{j\cdot}$.

In addition to its methodological interest, the PoSI1 situation addressed by Theorem 4.2 is of theoretical interest: Even though the PoSI1 constant $K_{j\cdot}$ is dominated by the unrestricted PoSI constant K, we will construct in Section 6.2 an example of predictor matrices for which the PoSI1 constant increases at the Scheffé rate and is asymptotically more than 63% of the Scheffé constant K_{Sch}. It follows that near-Scheffé protection can be needed even for SPAR1 variable selection.

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 $d = \mathrm{rank}(X) \le \min(n, p)$, 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.* We simplify the PoSI coverage problem (4.8) as follows: Due to pivotality of t-statistics, the problem is invariant under translation of $\boldsymbol{\mu}$ and rescaling of σ (see equation (4.1)). Hence it suffices to solve coverage problems for $\boldsymbol{\mu} = \mathbf{0}$ 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 use the more familiar notation $\mathbf{Z}$ instead of $\tilde{\mathbf{Y}}$. The random vector $\mathbf{Z}/\hat{\sigma}$ 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 (3.5) and their t-statistics (4.1) simplify to

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

which 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:

$$\boldsymbol{l}_{j\cdot\mathrm{M}} \triangleq \frac{\mathbf{X}_{j\cdot\mathrm{M}}}{\|\mathbf{X}_{j\cdot\mathrm{M}}\|^2}, \quad \bar{\boldsymbol{l}}_{j\cdot\mathrm{M}} \triangleq \frac{\boldsymbol{l}_{j\cdot\mathrm{M}}}{\|\boldsymbol{l}_{j\cdot\mathrm{M}}\|} = \frac{\mathbf{X}_{j\cdot\mathrm{M}}}{\|\mathbf{X}_{j\cdot\mathrm{M}}\|}. \tag{5.2}$$

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

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

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

PROPOSITION 5.2. *The PoSI problem* (4.8) *is equivalent to a d-dimensional coverage problem for linear functions of the multivariate t-vector $\mathbf{Z}/\hat{\sigma}$:*

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

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.1) 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* $j \ne j'$ *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 $\mathcal{M}=\mathcal{M}_{\text{all}}$: 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 $\mathcal{M}=\mathcal{M}_{\text{all}}$ 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 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

$$\mathbf{\Pi}_K \;=\; \mathbf{\Pi}_K(\mathbf{X},\mathcal{M}) \;\triangleq\; \{\mathbf{z}\in\mathbb{R}^d \mid |\bar{\boldsymbol{l}}^T\mathbf{z}| \le K,\ \forall\,\bar{\boldsymbol{l}}\in\mathcal{L}(\mathbf{X},\mathcal{M})\,\}, \tag{5.5}$$

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 $r\to\infty$:

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

Many properties of the polytopes $\mathbf{\Pi}_K$ are not specific to PoSI because they hold for polytopes (5.5) 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 $2K$:

$$\mathbf{\Pi}_K \;=\; \bigcap_{\bar{\boldsymbol{l}}\in\mathcal{L}} \{\mathbf{z}\in\mathbb{R}^d \mid |\mathbf{z}^T\bar{\boldsymbol{l}}\,| \;\le K\,\}.$$

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.

5.5. *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 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:

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*

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

By Proposition 5.4 the PoSI problem is bounded below by orthogonal designs, and by Proposition 4.1 it is loosely bounded above by the Scheffé ball (both for all α, d, and r). The question of how close to the Scheffé bound PoSI problems 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.

5.6. *A Duality Property of PoSI Vectors.* In the classical case $d = p$ and $\mathcal{M} = \mathcal{M}_{\text{all}}$ 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. 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 $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ (see (3.5) and (3.4)). In a change of perspective, we interpret the transpose matrix

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

as a predictor matrix, 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 $\mathcal{M} = \mathcal{M}_{\text{all}}$:

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.2. 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*

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

6. Illustrative Examples and Asymptotic Results. We consider examples in the classical case $d = p$ and $\mathcal{M} = \mathcal{M}_{\text{all}}$. Also, we work with the Gaussian limit $r \to \infty$, that is, σ^2 known, and w.l.o.g. $\sigma^2 = 1$.

6.1. *Example 1: Exchangeable Designs.* In exchangeable designs all pairs of predictor vectors enclose the same angle. In canonical coordinates a convenient parametrization of a family of symmetric exchangeable designs is

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

where $-1/p < a < \infty$, and $\mathbf{E}_{p\times p}$ is a matrix with all entries equal to 1. The range restriction on a assures that $\mathbf{X}^{(p)}$ is positive definite. We will write $\mathbf{X} = \mathbf{X}^{(p)} = \mathbf{X}(a) = \mathbf{X}^{(p)}(a)$ depending on which parameter matters in a given context. We will make use of the fact that

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

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 $\mathbf{1} = (1, 1, ..., 1)^T \in \mathbb{R}^p$.

As collinearity drives the fracturing of the regression coefficients into model-dependent quantities $\beta_{j\cdot\mathrm{M}}$, it is of interest to analyze $K(\mathbf{X}(a))$ as $\mathbf{X}(a)$ 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 5.6): for exchangeable 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.

We next address the asymptotic behavior of $K = K(\mathbf{X}^{(p)}, \alpha)$ for increasing p. As noted in Section 4.8, 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* $\mathbf{X}^{(p)}(a)$ *(defined in* (6.1) *above) have the following limiting behavior:*

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

The proof can be found in Appendix A.3. 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.

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 $K(\mathbf{X}^{(p)}, \alpha)$ is $O(\sqrt{p})$, hence equal to the rate of the Scheffé constant $K_{\mathrm{Sch}}(\alpha, p)$. Perhaps surprisingly, it is sufficient to consider PoSI1 (Section 4.10) whose constant is dominated by that of full PoSI. Let the PoSI1 predictor of interest be $\mathbf{X}_p^{(p)}$, so the search is over all models $\mathrm{M} \ni p$, but inference is sought only for $\beta_{p\cdot\mathrm{M}}$.

The task is to construct a design for which simultaneous inference for all adjusted coefficients $\beta_{p\cdot\mathrm{M}}$ requires the PoSI1 constant $K_{p\cdot}(\mathbf{X})$ of Section 4.10 to be in the order of $\sqrt{p}$. To this end consider the following upper triangular $p \times p$ design matrix in canonical coordinates:

$$\mathbf{X}^{(p)}(c) \;=\; (\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_{p-1}, \mathbf{X}_p(c)), \tag{6.2}$$

where $\mathbf{X}_p(c) = (c, c, ..., c, \sqrt{1-(p-1)c^2})^T \in \mathbb{R}^T$ is the primary predictor and the canonical basis vectors $\mathbf{e}_1, ..., \mathbf{e}_{p-1} \in \mathbb{R}^p$ are the controls. The vector $\mathbf{X}_p(c)$ has unit length, hence the parameter c is the correlation between the primary predictor and the controls. It is constrained to $c^2 < 1/(p-1)$, so $\mathbf{X}^{(p)}(c)$ has full rank. For $c^2 = 1/(p-1)$ the primary predictor $\mathbf{X}_p(c)$ becomes fully collinear with the controls, and it is on the approach to this boundary where the rate of the following theorem is attained:

THEOREM 6.2. *For σ^2 known, the designs (6.2) have PoSI1 constants $K_{p\cdot}(\mathbf{X}^{(p)}(c), \alpha)$ with the following asymptotic rate:*

$$\lim_{p\to\infty} \sup_{c^2<1/(p-1)} \frac{K_{p\cdot}(\mathbf{X}^{(p)}(c), \alpha)}{\sqrt{p}} \;=\; 0.6363....$$

The proof is in Appendix A.4. As mentioned, $K(\mathbf{X}, \alpha) \geq K_{j\cdot}(\mathbf{X}, \alpha)$, hence the theorem provides a lower bound on the rate of the full PoSI constant. The value 0.6363... is not maximal, and we currently have indications that the supremum over all designs may exceed 0.78. Together with the upper bound of Corollary 6.1 this would provide a narrow asymptotic range for worst-case PoSI. — Most importantly, the example shows that for some designs PoSI constants can be much larger than the $O(1)$ $|t|$-quantiles used in common practice.

6.3. *Bounding Away from Scheffé.* We provide a rough asymptotic upper bound on all PoSI constants $K(\mathbf{X}, \mathcal{M}, \alpha)$. It has the Scheffé rate but with a multiplier that is strictly less than Scheffé's. The bound is loose because it ignores the rich structure of the sets $\mathcal{L}(\mathbf{X}, \mathcal{M})$ (Section 5.3) and only uses 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 $|\mathcal{L}_d| \le a_d$ where $a_d^{1/d} \to a$ (> 1). Denote by $K(\mathcal{L}_d, \alpha)$ the $(1-\alpha)$-quantile of $\sup_{\bar{l} \in \mathcal{L}_d} |\bar{l}^T \mathbf{Z}|$. Then the following describes an asymptotic worst-case bound for $K(\mathcal{L}_d, \alpha)$ and its attainment:*

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

The proof of Theorem 6.3 (see the Appendix A.5) is an adaptation of Wyner's (1967) techniques for sphere packing and sphere covering. The worst-case bound ($\le$) is based on a surprisingly crude Bonferroni-style inequality for caps on spheres. Attainment of the bound ($\ge$) makes use of the artifice of picking the vectors $\bar{l} \in \mathcal{L}$ 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 $K(\mathbf{X}_{n\times p}, \mathcal{M}_{\text{all}}, \alpha)$ is*

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

The corollary shows that the asymptotic rate of the PoSI constant, if it reaches the Scheffé rate, will always have a multiplier that is strictly below that of the Scheffé constant. We do not know whether there exist designs for which the bound of the corollary is attained, but the theorem says the bound is sharp for unstructured sets $\mathcal{L}$.

7. 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, $d=\mathrm{rank}(\mathbf{X})<p$, 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 principle is that submodels have their own regression coefficients; put differently, $\beta_{j\cdot\mathrm{M}}$ 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 $K(\mathbf{X}_{n\times p},\mathcal{M},\alpha,r)$ 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 $K(\mathbf{X},\mathcal{M},\alpha,r)$ with $d=\mathrm{rank}(\mathbf{X})$ 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 $d=p$. When the size of models $|\mathrm{M}|$ is limited as a function of p ("sparse models"), better rates can be achieved, and we will report these results elsewhere.
- 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.
- Situations to which the PoSI framework should be extended include generalized linear models, mixed effects models, models with random predictors, as well as prediction problems. Results for the last two situations will be reported elsewhere.
- It would be desirable to devise post-selection inference for specific selection procedures for cases in which a strict model selection protocol is being adhered to.

R code for computing the PoSI constant for up to $p=20$ can be obtained

from the authors' web pages (a manuscript describing the computations is available from the authors).

Acknowledgments. We thank E. Candes, L. Dicker, M. Freiman, E. George, A. Krieger, M. Low, Z. Ma, E. Pitkin, L. Shepp, N. Sloane, P. Shaman and M. Traskin for very helpful discussions. The acronym "SPAR" is due to M. Freiman. We are indebted to an anonymous reviewer for extensive and constructive criticism that deeply influenced the positioning of this article.

APPENDIX A: PROOFS

A.1. Proof of Proposition 5.3.

1. The matrix $\mathbf{X}^*_{\mathrm{M}} = \mathbf{X}_{\mathrm{M}}(\mathbf{X}_{\mathrm{M}}^T\mathbf{X}_{\mathrm{M}})^{-1}$ 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 $\mathbf{X}_{\mathrm{M}}^T\mathbf{X}^*_{\mathrm{M}} = \mathbf{I}_p$. 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.2. 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 $\boldsymbol{l}^T_{j\cdot \mathrm{M}}\, \boldsymbol{l}^*_{j\cdot \mathrm{M}^*}$. 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

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

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

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

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.3. Proof of Theorem 6.1. The parameter a in equation (6.1) 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 $j \in \mathrm{M}$.

Consider first the case $|\mathrm{M}| = 1$, hence $\mathrm{M} = \{j\}$: We have $\boldsymbol{l}_{j\cdot\mathrm{M}} = \mathbf{X}_j$, the j-th column of $\mathbf{X}$, and $\bar{\boldsymbol{l}}_{j\cdot\mathrm{M}} = \boldsymbol{l}_{j\cdot\mathrm{M}}/\sqrt{pa^2+2a+1}$. For any $\mathbf{Z} \in \mathbb{R}^p$ it follows

$$|\bar{\boldsymbol{l}}_{j\cdot\mathrm{M}}^T \mathbf{Z}| \;\le\; |Z_j| + \left|\frac{1}{\sqrt{p}}\sum_k Z_k\right| \;\le\; \|\mathbf{Z}\|_\infty + \left|\frac{1}{\sqrt{p}}\sum_k Z_k\right| . \tag{A.8}$$

Consider next the case $|\mathrm{M}| > 1$, and for notational convenience let $j = 1$ and $\mathrm{M} = \{1, \ldots, m\}$ where $1 < m \le p$. The following results can then be applied to arbitrary M and $j \in \mathrm{M}$ by permuting coordinates. The projection of $\mathbf{X}_1$ on the space spanned by $\mathbf{X}_2, \ldots, \mathbf{X}_m$ must be of the form

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

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

$$(\boldsymbol{l}_{1\cdot\mathrm{M}})_k \;=\; \begin{cases} 1 + da & (k=1) \\ -\frac{1-d}{m-1} + da & (2 \le k \le m) \\ da & (k \ge m+1) \end{cases} . \tag{A.9}$$

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 da is non-negative, maximal wrt m for $m = 2$, and thereafter maximal wrt a for $a = 1/\sqrt{p}$, whence $\max_{a\ge 0, m\ge 2} da = 1/(2(\sqrt{p}+1))$ and finally

$$0 \le da < \frac{1}{2\sqrt{p}} . \tag{A.10}$$

This fact will make the term da in (A.9) asymptotically irrelevant. Using $\|\boldsymbol{l}_{1\cdot\mathrm{M}}\| \ge 1$ and $\bar{\boldsymbol{l}}_{1\cdot\mathrm{M}} = \boldsymbol{l}_{1\cdot\mathrm{M}}/\|\boldsymbol{l}_{1\cdot\mathrm{M}}\|$ as well as (A.9) and (A.10) we obtain

$$\begin{aligned} |\bar{\boldsymbol{l}}_{1\cdot\mathrm{M}}^T \mathbf{Z}| \;&\le\; |Z_1| + \frac{1}{m-1}\sum_{j=2}^{m}|Z_j| + \left|\frac{1}{2\sqrt{p}}\sum_{j=1}^{p} Z_j\right| \\ &\le\; \|\mathbf{Z}\|_\infty + \|\mathbf{Z}\|_\infty + \left|\frac{1}{2\sqrt{p}}\sum_{j=1}^{p} Z_j\right| . \end{aligned} \tag{A.11}$$

Combining (A.8) and (A.11) we obtain for $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$ the following:

$$\begin{aligned}\sup_{a\geq 0; j, \mathrm{M}: j\in \mathrm{M}} |\bar{\boldsymbol{l}}_{j\cdot \mathrm{M}}^T \mathbf{Z}| &\leq 2\|\mathbf{Z}\|_\infty + \left|\frac{1}{\sqrt{p}} \sum_{j=1}^p Z_j\right| \\ &\leq 2\sqrt{2\log p}(1+o_p(1)) + O_p(1)\,.\end{aligned}$$

This verifies that

$$\limsup_{p\to\infty} \frac{\sup_{a\in(-1/p,\infty)} K(\mathbf{X}(a))}{\sqrt{2\log p}} \leq 2 \quad \text{in probability.} \tag{A.12}$$

It remains to prove that equality holds in (A.12). To this end let $Z_{(1)} < Z_{(2)} < \ldots < Z_{(p)}$ denote the order statistics of Z_1, Z_2, ..., Z_p. Fix m. We have in probability

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

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 a given $\mathbf{Z}$ we choose $\boldsymbol{l}_{j^*\cdot \mathrm{M}^*}$ such that $j^* = j^*(\mathbf{Z})$ is the index of $Z_{(1)}$ and $\mathrm{M}^* = \mathrm{M}^*(\mathbf{Z})$ includes j^* as well as the set of indices of $Z_{(k)}$ for $p-m+2 \leq k \leq p$. From (A.9) we then obtain 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.12) yields the desired conclusion.

A.4. Proof of Theorem 6.2. Recall from (6.2) the designs

$$\mathbf{X} = (\mathbf{e}_1, \mathbf{e}_3, ..., \mathbf{e}_{p-1}, \mathbf{X}_p(c)),$$

where $\mathbf{X}_p(c) = (c, c, ..., c, \sqrt{1-(p-1)c^2})^T$ is the primary predictor. The matrix $\mathbf{X}$ will be treated according to PoSI1 (Section 4.10), hence we will examine the distribution of $\max_{\mathrm{M}: p\in \mathrm{M}} |\bar{\boldsymbol{l}}_{p\cdot \mathrm{M}}^T \mathbf{Z}|$ (assuming $\sigma^2 = 1$ known). We determine $\bar{\boldsymbol{l}}_{p\cdot \mathrm{M}}$ for a fixed model M ($\ni p$) with $|\mathrm{M}| = m$:

$$\bar{\boldsymbol{l}}_{p\cdot \mathrm{M}, j} = \begin{cases} \sqrt{\frac{1-(p-1)c^2}{1-(m-1)c^2}} & j = p \\ 0 & j \in \mathrm{M}\backslash\{p\} \\ \frac{c}{\sqrt{1-(m-1)c^2}} & j \in \mathrm{M}^c \end{cases}$$

Therefore,

$$z_{p\cdot \mathrm{M}} \;=\; \vec{l}_{p\cdot \mathrm{M}}^{\,T}\mathbf{Z} \;=\; \sqrt{\frac{1-(p-1)c^2}{1-(m-1)c^2}}\,Z_1 + \frac{c}{\sqrt{1-(m-1)c^2}}\sum_{j\in \mathrm{M}^c} Z_j. \tag{A.13}$$

For fixed m we can explicitly maximize the sum on the right hand side:

$$\max_{\mathrm{M:}\ |\mathrm{M}|=m}\left|\sum_{j\in \mathrm{M}^c} Z_j\right| \;=\; \max\left(\sum_{j=1}^{p-m} Z_{(p-j)},\ -\sum_{j=1}^{p-m} Z_{(j)}\right),$$

where $Z_{(j)}$ is the j-th order statistic of Z_1, Z_2, ..., Z_{p-1}, omitting Z_p. We can also explicitly maximize the factor $c/\sqrt{1-(m-1)c^2}$ in (A.13):

$$\sup_{c^2<1/(p-1)} \frac{c}{\sqrt{1-(m-1)c^2}} \;=\; \frac{1}{\sqrt{p-m}},$$

and equality is attained as $c^2 \uparrow 1/(p-1)$. Therefore, for fixed m, we can continue from (A.13) as follows:

$$\begin{aligned}\sup_{c^2<1/(p-1)}\ \max_{|\mathrm{M}|=m} \frac{|z_{p\cdot \mathrm{M}}|}{\sqrt{p}} \;&=\; O_p\left(\sqrt{\frac{1}{p}}\right)\\ &\quad+\; \sqrt{\frac{p}{p-m}}\max\left(\sum_{j=1}^{p-m} Z_{(p-j)}\frac{1}{p},\ -\sum_{j=1}^{p-m} Z_{(j)}\frac{1}{p}\right).\end{aligned}$$

The reason for writing the two sums in this manner is that we will interpret them as approximations to Riemann sums. To this end we borrow from Bahadur (1966) the following approximations for $j=1,...,p-1$:

$$Z_{(j)} \;=\; \Phi^{-1}\left(\frac{j}{p}\right) \;+\; O_p(p^{-1/2}).$$

Reparametrizing $m=rp$, the anticipated Riemann approximation is

$$\int_r^1 \Phi^{-1}(x)dx \;=\; \sum_{j=1}^{p-m} \Phi^{-1}\left(\frac{p-j}{p}\right)\frac{1}{p} \;+\; O(p^{-2}).$$

Therefore,

$$\sum_{j=1}^{p-m} Z_{(p-j)}\,\frac{1}{p} \;=\; \int_r^1 \Phi^{-1}(x)dx \;+\; O_p(p^{-1/2}),$$

and similarly

$$-\sum_{j=1}^{p-m} Z_{(j)}\,\frac{1}{p} \;\;=\;\; \int_r^1 \Phi^{-1}(x)dx \;+\; O_p(p^{-1/2}).$$

Summarizing,

$$\begin{aligned}
&\sup_c \max_{|\mathrm{M}|=m} \left|\frac{z_{p\cdot\mathrm{M}}}{\sqrt{p}}\right| \\
&= \frac{1}{\sqrt{p-m}}\,\max\left(\sum_{j=1}^{p-m} Z_{(p-j)}\frac{1}{p}, -\sum_{j=1}^{p-m} Z_{(j)}\frac{1}{p}\right) \;+\; O_p(\sqrt{1/p}) \\
&= \frac{1}{\sqrt{1-r}}\int_r^1 \Phi^{-1}(x)dx \;+\; O_p(p^{-1/2}) \;+\; O_p(\sqrt{1/p}) \\
&= \frac{1}{\sqrt{1-r}}\,\phi(\Phi^{-1}(r)) \;+\; O_p(\sqrt{1/p}).
\end{aligned}$$

The function $f(r) = \frac{1}{\sqrt{1-r}}\phi(\Phi^{-1}(r))$ is maximized at $r^* \approx 0.72972$ with $f(r^*) \approx 0.6363277$. Therefore,

$$\limsup_{p\to\infty} \sup_c \frac{1}{\sqrt{p}} \max_{\mathrm{M}} |z_{p\cdot\mathrm{M}}| \;=\; 0.636\ldots. \tag{A.14}$$

The bound is sharp because it is attained by the models that include the first or last $m^* = r^*p$ order statistics of $\mathbf{Z}$ when $p \to \infty$ and $c^2 \uparrow \frac{1}{p-1}$. From (A.14) we conclude that $K_{1\cdot}(\mathbf{X}) \sim 0.6363\sqrt{p}$.

A.5. Proof of Theorem 6.3. We show that if $a_p^{1/p} \to a$ (>1), then

- we have a uniform asymptotic worst-case bound,

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

- which is attained when $|\mathcal{L}_p| = a_p$ and $\bar{l} \in \mathcal{L}_p$ are i.i.d. Unif(S^{p-1}) independent of $\mathbf{Z}$:

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

These facts imply the assertions about $(1-\alpha)$-quantiles $K(\mathcal{L}_p)$ of $\max_{\bar{l}\in\mathcal{L}_p} |\bar{l}^T\mathbf{Z}|$ 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 $R/\sqrt{p} \overset{\mathbf{P}}{\to} 1$ it is sufficient to show the following:

- uniform asymptotic worst-case bound:

$$\lim_{p\to\infty} \sup_{|\mathcal{L}_p|\le a_p} \max_{\bar{l}\in\mathcal{L}_p} |\bar{l}^T \mathbf{U}| \overset{\mathbf{P}}{\le} \sqrt{1-1/a^2}\,; \tag{A.15}$$

- attainment of the bound when $|\mathcal{L}_p| = a_p$ and $\bar{l} \in \mathcal{L}_p$ are i.i.d. Unif(S^{p-1}) independent of $\mathbf{U}$:

$$\lim_{p\to\infty} \max_{\bar{l}\in\mathcal{L}_p} |\bar{l}^T \mathbf{U}| \overset{\mathbf{P}}{\ge} \sqrt{1-1/a^2}\,. \tag{A.16}$$

To show (A.15), 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):

$$\begin{aligned} \mathbf{P}[\max_{\bar{l}\in\mathcal{L}} |\bar{l}^T \mathbf{U}| > K'] &= \mathbf{P} \bigcup_{\bar{l}\in\mathcal{L}} [\,|\bar{l}^T \mathbf{U}| > K'] \\ &\le \sum_{\bar{l}\in\mathcal{L}} \mathbf{P}[\,|\bar{l}^T \mathbf{U}| > K'] \\ &= |\mathcal{L}_p|\, \mathbf{P}[\,|U| > K'], \end{aligned} \tag{A.17}$$

where U is any coordinate of $\mathbf{U}$ or projection of $\mathbf{U}$ onto a unit vector. We will show that the bound (A.17) converges to zero. We use the fact that $U^2 \sim \text{Beta}(1/2, (p-1)/2)$, hence

$$\mathbf{P}[\,|U| > K'] = \frac{1}{\mathrm{B}(1/2,(p-1)/2)} \int_{K'^2}^1 x^{-1/2}(1-x)^{(p-3)/2} dx \tag{A.18}$$

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 $\Gamma(1/2) = \sqrt{\pi}$.

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

where we used $x^{-1/2} \le 1/K'$ on the integration interval. Continuing with the chain of bounds from (A.17) we have:

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

Since $|\mathcal{L}_p|^{1/(p-1)} \to a\ (>0)$, 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.15).

To show (A.16), 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):

$$\begin{aligned} \mathbf{P}[\max_{\bar{\boldsymbol{l}}\in\mathcal{L}_p} |\bar{\boldsymbol{l}}^T\mathbf{U}| \le K'] &= \prod_{\bar{\boldsymbol{l}}\in\mathcal{L}_p} \mathbf{P}[\,|\bar{\boldsymbol{l}}^T\mathbf{U}| \le K'] = \mathbf{P}[\,|U| \le K']^{|\mathcal{L}_p|} \\ &= \left(1-\mathbf{P}[\,|U|>K']\right)^{|\mathcal{L}_p|} \\ &\le \exp\left(-|\mathcal{L}_p|\,\mathbf{P}[\,|U|>K']\right). \end{aligned} \tag{A.19}$$

We will lower-bound the probability $\mathbf{P}[\,|U|>K']$ recalling (A.18) 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).

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

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

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

Since $|\mathcal{L}_p|^{1/(p-1)} \to a\ (>0)$, the r.h.s. converges to $+\infty$ at nearly geometric speed if $a\sqrt{1-K'^2} > 1$, that is, $K' < \sqrt{1-1/a^2}$. This proves (A.16).

APPENDIX B: SUPPLEMENTARY MATERIALS

B.1. The Full Model Interpretation of Parameters. In the full model interpretation, coefficients always have the fixed meaning as full model parameters. Variable selection then means setting some coefficient estimates to zero, and these estimates always exist for all predictors, irrespective of whether they are selected or deselected.

The full model interpretation of parameters is appropriate, for example, if the full model is viewed as "data generating" and the predictors are hence causal for the response, or if the full model describes a physical system where the full set of predictors is needed to capture the system fully. In such situations it is natural to consider the coefficients in the full model as targets of estimation, even though "nature" may choose to set some of them to zero. This view is meaningful for example in tomography applications where the variables constitute voxels and their coefficients are rates of absorption, hence variable selection amounts to selection of voxels with high absorption. The use of the selected voxels is for display and medical diagnosis, and there is no meaning in interpreting these voxels as constituting a submodel.

If full model parameters are estimated by forcing some of them to zero and estimating the remainder via least squares, then the result is a type of shrinkage estimator for the full model parameters. Such estimators are often referred to as "preliminary test estimators"; see Saleh (2006) for a comprehensive treatment and many references. These estimators are also closely related to more recently studied "hard threshold" and "soft threshold" estimators; for a taste of the extensive literature on these and related estimators see Tsybakov (2009) and references therein. The "submodel" corresponding to non-zero parameter estimates is viewed as a computational compression and a parsimonious statistical summary of the data, but it is viewed neither as a model in its own right nor as an object of future scientific research.

Inferential problems with the full model view of variable selection are pointed out by the "Vienna School" in the series of articles referenced in Section 1. An insightful illustration is given by Leeb and Pötscher (2005) with a two-predictor situation where one predictor is protected from selection and only the covariate is subject to selection. They explicitly describe the sampling distribution of the coefficient estimate of the protected predictor, as the covariate is randomly selected/deselected according to tests that perform consistent or conservative model selection, respectively (ibid., p. 29). Their analysis shows (ibid., Figure 2) that the sampling distribution (1) depends critically on the unknown true coefficient of the covariate and the sample size, and (2) deviates from the fixed-model sampling distribution with features such as bi-modality or inflated variance. Because the

true covariate slope is not known, it cannot be known either whether the sample size puts the sampling distribution in this realm of deviation from classical theory. — Generalizing to arbitrary linear models Pötscher, Leeb and Schneider prove that sampling distributions cannot be estimated after model selection and thresholding, not even asymptotically. They show that asymptotic normality is prone to non-uniformity of convergence especially near submodels, or should be considered as converging at a speed slower than root-N to a non-normal distribution. They prove that these effects are more pronounced under consistent than conservative model selection. — In the simplified context of what may be called "marginally thresholded" estimators, Pötscher and Schneider (2010) produce conservative confidence intervals, and they show that these intervals are wider than conventional ones as their widths need to account for the bias caused by thresholding.

The criticisms of the "Vienna School" are important and may have far reaching implications. They indicate that in the framework of full model parameters the biases incurred by model selection pose problems for statistical inference. Some of these can be traced to the so-called "omitted variables bias" (see, for example, Angrist and Pischke 2009), that is, the fact that in the presence of partial collinearity the omission of covariates creates biases in the estimates of parameters of primary interest. The term "bias" is of course justified only in the framework of full model parameters. If we interpret submodel estimates as estimates of submodel parameters rather than full model parameters (as we will do next), then there is no bias problem and this source of defects in sampling distributions disappears.

B.2. "Omitted Variables Bias". By allowing each $\hat{\beta}_{j\cdot \mathrm{M}}$ to estimate its own submodel target $\beta_{j\cdot \mathrm{M}}$ rather than the full model parameter β_j, we sidestep the problem of "omitted variables bias" and with it a major driver of the problems analyzed by Leeb and Pötscher (Section B.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, we 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:

$$\boldsymbol{\beta}_{\mathrm{M}} \triangleq \mathbf{E}[\hat{\boldsymbol{\beta}}_{\mathrm{M}}] \quad \text{and} \quad \boldsymbol{\beta}^{\mathrm{M}} \triangleq (\beta_j)_{j\in \mathrm{M}}, \tag{B.1}$$

the latter being the coefficients β_j from the full model M_F subsetted to the submodel M. While $\hat{\boldsymbol{\beta}}_{\mathrm{M}}$ is an unbiased estimate for $\boldsymbol{\beta}_{\mathrm{M}}$, it is *not* generally for $\boldsymbol{\beta}^{\mathrm{M}}$. 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

$$\mathbf{X}_{\mathrm{M}}^{T}\mathbf{X}_{\mathrm{M}^c}\boldsymbol{\beta}^{\mathrm{M}^c} = \mathbf{0}, \tag{B.2}$$

where M^c is the full model complement of M. Special cases of (B.2) include: (1) the column space of $\mathbf{X}_{\mathrm{M}}$ is orthogonal to that of $\mathbf{X}_{\mathrm{M}^c}$, and (2) $\boldsymbol{\beta}^{\mathrm{M}^c} = \mathbf{0}$, that is, the approximation to $\boldsymbol{\mu}$ in M_F is no better than in M.

B.3. Proof of Corollary 4.2. 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 $\mathbf{Y}$:

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

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

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

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 $t^{(0)}_{j\cdot\mathrm{M}} = t_{j\cdot\mathrm{M}} \Leftrightarrow (j,\mathrm{M})\in H_0$,

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

by the definition of the PoSI constant $K = K(\mathbf{X},\mathcal{M},\alpha,r)$ (4.8). □

Corollary 4.2. *"Strong Post-Selection Error Control" holds for any model selection procedure* $\hat{\mathrm{M}} : \mathbb{R}^n \to \mathcal{M}$*:*

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

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 B.1:

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

B.4. Alternative PoSI Guarantees. PoSI and PoSI1 provide inferential guarantees for two distinct situations: In PoSI all predictors are subjected to selection, and all that are selected are the subject of inference; in PoSI1 one predictor of interest is forced into all models, and only the coefficients (plural!) of this predictor are the subject of inference. Invariably, however, there arises the question of an intermediate situation: Can any guarantee be given when there is a predictor of special interest, but it is subjected to selection and inference is sought only when it is selected? In what follows we give guarantees for this situation. Even though it differs from PoSI1, we can re-use the PoSI1 constant $K_{j\cdot}$. For the rest of this section let j be the index of *a fixed and a priori chosen predictor*.

THEOREM B.1. *For any selection procedure* $\hat{\mathrm{M}} : \mathbb{R}^n \to \mathcal{M}$ *it holds:*

$$\mathbf{P}\left[j \in \hat{\mathrm{M}} \;\&\; |t_{j\cdot\hat{\mathrm{M}}}| \le K_{j\cdot} \right] \;\ge\; \mathbf{P}\left[j \in \hat{\mathrm{M}} \right] - \alpha, \tag{B.3}$$

and accordingly we have the following post-selection confidence guarantee:

$$\mathbf{P}\left[j \in \hat{\mathrm{M}} \;\&\; \beta_{j\cdot\hat{\mathrm{M}}} \in \mathrm{CI}_{j\cdot\hat{\mathrm{M}}}(K_{j\cdot}) \right] \;\ge\; \mathbf{P}\left[j \in \hat{\mathrm{M}} \right] - \alpha. \tag{B.4}$$

Proof: In what follows we will use Lemma 4.2 and the definition of $K_{j\cdot}$ (4.14):

$$\begin{aligned}
&\mathbf{P}\left[j \in \hat{\mathrm{M}}(\mathbf{Y}) \right] - \mathbf{P}\left[j \in \hat{\mathrm{M}}(\mathbf{Y}) \;\&\; |t_{j\cdot\hat{\mathrm{M}}(\mathbf{Y})}(\mathbf{Y})| \le K_{j\cdot} \right] \\
&= \mathbf{P}\left[j \in \hat{\mathrm{M}}(\mathbf{Y}) \;\&\; |t_{j\cdot\hat{\mathrm{M}}(\mathbf{Y})}(\mathbf{Y})| > K_{j\cdot} \right] \\
&\le \mathbf{P}\left[j \in \hat{\mathrm{M}}(\mathbf{Y}) \;\&\; \max_{\mathrm{M}\in\mathcal{M}_{j\cdot}} |t_{j\cdot\mathrm{M}}(\mathbf{Y})| > K_{j\cdot} \right] \\
&\le \mathbf{P}\left[\max_{\mathrm{M}\in\mathcal{M}_{j\cdot}} |t_{j\cdot\mathrm{M}}(\mathbf{Y})| > K_{j\cdot} \right] \\
&\le \alpha \qquad \square
\end{aligned}$$

A deficiency of these inference guarantees is that they are vacuous for selection probabilities below α, and they have "bite" only if $\mathbf{X}_j$ is a "strong" predictor in the sense that its selection probability $\mathbf{P}[\, j \in \hat{\mathrm{M}}\,]$ is large. If one chooses $\alpha = 0.01$, for example, the guarantee says that the probability of *selection and coverage* never falls more than 0.01 below the probability of *selection*. The theorem can be rewritten in terms of guarantees conditional on $\mathbf{X}_j$ being selected:

COROLLARY B.1. *For any selection procedure* $\hat{\mathrm{M}} : \mathbb{R}^n \to \mathcal{M}$ *we have:*

$$\mathbf{P}\left[\ \beta_{j\cdot\hat{\mathrm{M}}} \in \mathrm{CI}_{j\cdot\hat{\mathrm{M}}}(K_{j\cdot})\ \middle|\ j \in \hat{\mathrm{M}}\right] \ \geq\ 1 - \frac{\alpha}{\mathbf{P}[\,j \in \hat{\mathrm{M}}]}.$$

Again, there is "bite" only when the selection probability $\mathbf{P}[j \in \hat{\mathrm{M}}]$ is large.

B.5. 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:

$$F_{j\cdot\mathrm{M}}(t) = \mathbf{P}[\,|t^{(0)}_{j\cdot\mathrm{M}}| < t\,], \qquad F_{\max}(t) \ =\ \mathbf{P}[t^{(0)}_{\max} < t]. \tag{B.5}$$

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 $\mathcal{M} = \mathcal{M}_{\mathrm{all}}$. 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: $\quad \mathrm{pval}_{j\cdot\mathrm{M}} \ = 1 - F_{j\cdot\mathrm{M}}(\,|t^{obs}_{j\cdot\mathrm{M}}|\,)$
(2) Global adjusted: $\quad \mathrm{pval}^{PoSI}_{j\cdot\mathrm{M}} = 1 - F_{\max}(t^{obs}_{\max})$
(3) Individual adjusted: $\quad \mathrm{pval}^{PoSI}_{j\cdot\mathrm{M}} = 1 - F_{\max}(|t^{obs}_{j\cdot\mathrm{M}}|)$

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. We have not examined this route though we suspect that 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.]

B.6. 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

$$\mathbf{\Sigma} \;=\; (\Sigma_{j\cdot\mathrm{M};\,j'\cdot\mathrm{M}'}), \qquad \Sigma_{j\cdot\mathrm{M};\,j'\cdot\mathrm{M}'} \;\triangleq\; \bar{\boldsymbol{l}}_{j\cdot\mathrm{M}}^{T}\bar{\boldsymbol{l}}_{j'\cdot\mathrm{M}'}. \tag{B.6}$$

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 $\mathbf{\Sigma}$, intervals such as $[-K, +K]$ for the components, and degrees of freedom r. In our experiments this approach worked in the classical case $d=p$ and $\mathcal{M}=\mathcal{M}_{\mathrm{all}}$ for $p\le 7$, the limiting factor being the space requirement $p\,2^{p-1}\times p\,2^{p-1}$ for the matrix $\mathbf{\Sigma}$. By comparison the approach described in Buja et al. (2012) 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 $\mathbf{\Sigma}$, orthogonalities in $\mathcal{L}(\mathbf{X}, \mathcal{M})$ correspond to zero correlations in $\mathbf{\Sigma}$. Part *4.* of the proposition states that in the classical case and $\mathcal{M}=\mathcal{M}_{\mathrm{all}}$ each "row" of $\mathbf{\Sigma}$ 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 $\mathbf{\Sigma}$ approaches half for increasing p. Thus $\mathbf{\Sigma}$, though not sparse, is rich in zeros. It can be much sparser in the presence of exact orthogonalities among the predictors.

B.7. Figures.

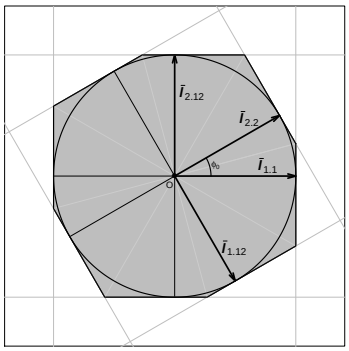

Fig 1: *The PoSI polytope* $\mathbf{\Pi}_{K=1}$ *tangent to the Scheffé disk (2-D ball)* $\mathbf{B}_{K=1}$ *for* $d = p = 2$*: The normalized raw predictor vectors are* $\overline{\boldsymbol{l}}_{1\cdot\{1\}} \sim \mathbf{X}_1$ *and* $\overline{\boldsymbol{l}}_{2\cdot\{2\}} \sim \mathbf{X}_2$*, and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

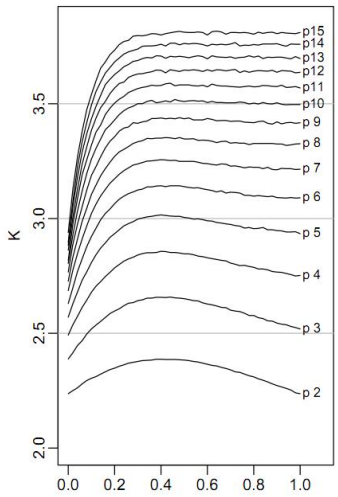

Fig 2: *The PoSI constant* $K(\mathbf{X}^{(p)}(a), \alpha = 0.05)$ *for exchangeable designs* $\mathbf{X}^{(p)} = \mathbf{I}_p + a\mathbf{E}_{p\times p}$ *for* $a \in [0, \infty)$*. The horizontal axis shows* $a/(1+a)$*, hence the locations 0, 0.5 and 1.0 represent* $a = 0,\ 1,\ \infty$*, respectively. Surprisingly, the largest* $K(\mathbf{X}^{(p)}(a))$ *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* $p = 2, ..., 15$*.*

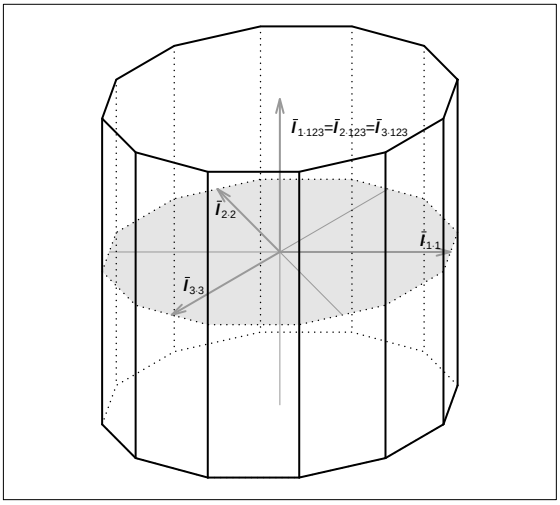

Fig 3: *Exchangeable Designs (Section 6.2): The geometry of the limiting PoSI polytope* $\mathbf{\Pi}_K$ *for* $p = 3$ *as the design* $\mathbf{X}^{(p)}(a)$ *in* (6.1) *approaches either of the two collinearities. For* $a \uparrow \infty$*, the predictor vectors fall into the 1-D subspace* $\mathrm{span}(\mathbf{1})$*, 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* $\mathrm{span}(\mathbf{1})^\perp$*. The figure shows 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.*

REFERENCES

[1] Angrist, J. D. and Pischke, J.-S. (2009). *Mostly Harmless Econometrics*, Princeton: Princeton University Press.

[2] Bahadur, R. (1966). Note on Quantiles in Large Samples, *Annals of Mathematical Statistics* **37**, 577–580.

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

[4] 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.

[5] Buja, A., Zhang, K., Berk, R., Brown, L. D. and Zhao, L. (2012). Computational Methods for Post-Selection Inference. (Forthcoming; partly contained in an older version of the present article and available at `http://stat.wharton.upenn.edu/ buja/PAPERS/PoSI.pdf`).

[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] Hall, P. and Carroll, R. (1989). Variance Function Estimation in Regression: The Effect of Estimating the Mean. *Journal of the Royal Statistical Society* B **51**, 3–14.

[10] Ioannidis, J.P.A. (2005). Why Most Published Research Findings Are False. *PLoS Med* 2(8): e124. doi:10.1371/journal.pmed.0020124.

[11] Kabaila, P. (1998). Valid Confidence Intervals in Regression after Variable Selection. *Econometric Theory* **14**, 463–482.

[12] 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.

[13] Kabaila, P. (2009). The Coverage Properties of Confidence Regions After Model Selection. *International Statistical Review* **77** (3), 405–414.

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

[15] 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.

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

[17] 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.

[18] 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.

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

[20] Leeb, H. and Pötscher, B. M. (2008b). Can One Estimate the Unconditional Distribution of Post-Model-Selection Estimators? *Econometric Theory* **24**, 338–376.
[21] 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.
[22] Moore, D. S. and McCabe, G. P. (2003). *Introduction to the Practice of Statistics*, 4th ed., New York: W. H. Freeman and Company.
[23] Olshen, R. A. (1973). The Conditional Level of the F-Test. *Journal of the American Statistical Association* **68**, 692–698.
[24] Pötscher, B. M. (1991). Effects of Model Selection on Inference. *Econometric Theory* **7**, 163–185.
[25] Pötscher, B. M. (2006). The Distribution of Model Averaging Estimators and an Impossibility Result Regarding its Estimation. *IMS Lecture Notes - Monograph Series* **52**, 113-129.
[26] 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.
[27] Pötscher, B. M. and Schneider, U. (2009). On the Distribution of the Adaptive LASSO Estimator. *Journal of Statistical Planning and Inference* **139**, 2775-2790.
[28] Pötscher, B. M. and Schneider, U. (2010). Confidence Sets Based on Penalized Maximum Likelihood Estimators in Gaussian Regression. *Electronic Journal of Statistics* **4**, 334-360.
[29] Pötscher, B. M. and Schneider, U. (2011). Distributional Results for Thresholding Estimators in High-Dimensional Gaussian Regression Models. *Electronic Journal of Statistics* **5**, 1876-1934.
[30] Saleh, A. K. Md. Ehsanes (2006). *Theory Of Preliminary Test And Stein-Type Estimation With Applications*, Hoboken, NJ: Wiley.
[31] Scheffé, H. (1959). *The Analysis of Variance*, New York: John Wiley & Sons.
[32] Sen, P. K. (1979). Asymptotic Properties of Maximum Likelihood Estimators Based on Conditional Specification. *The Annals of Statistics*, **7**, 742–755.
[33] 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.
[34] Tsybakov, A. (2009). *Introduction to Nonparametric Estimation*, Springer Series in Statistics. New York: Springer.
[35] 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