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 model selection and derive statistical inference from the selected model. Such inference is generally invalid. We propose to produce valid "post-selection inference" by reducing the problem to one of simultaneous inference. Simultaneity is required for all linear functions that arise as coefficient estimates in all submodels. By purchasing "simultaneity insurance" for all possible submodels, the resulting post-selection inference is rendered universally valid under all possible model selection procedures. This inference is therefore generally conservative for particular selection procedures, but it is always less conservative than full Scheffé protection. Importantly it does *not* depend on the truth of the selected submodel, and hence it produces valid inference even in wrong models. We describe the structure of the simultaneous inference problem and give some asymptotic results.

1. Introduction.

1.1. *The Problem.* In the classical theory of statistical inference data is assumed to be generated from a known model, and the properties of the parameters in the model are of interest. In applications, however, it is often the case that the model that generates the data is unknown, and as a consequence a model is often chosen based on the data. It is common practice to apply classical statistical inference, such as F- and t-tests and confidence intervals, to models that have been chosen with model selection methods. The practice is so pervasive that it appears in classical undergraduate textbooks on statistics, such as Moore and McCabe (2003).

Despite its prevalence, this practice is problematic because it ignores the fact that the inference is preceded by a selection process that produces a model that is itself stochastic. The stochastic nature of the selected model affects and distorts sampling distributions of the post-selection parameter estimates. The distortion of distribution occurs with all data-dependent (more

[*]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

precisely: response-dependent) model selection methods, including stepwise forward or backward search driven by F-to-enter or F-to-drop criteria, or all-subset searches driven by complexity penalties such as C_p, AIC, BIC, information-theoretic criteria, risk-inflation, the Lasso, LARS, or prediction criteria such as cross-validation, or recent proposals such as the Dantzig selector. For general descriptions of these selection rules, see Hastie, Tibshirani, and Friedman (2009).

The problem of post-selection inference has been recognized long ago by Buehler and Fedderson (1963), Brown (1967), Olshen (1973), Sen (1979), Sen and Saleh (1987), Dijkstra and Veldkamp (1988), Pötscher (1991), and discussed more recently by Leeb and Pötscher (2005; 2006; 2008). In particular, Leeb and Pötscher (2006; 2008) show from a natural perspective that it is impossible to find the distribution of post-selection estimates, not even asymptotically.

1.2. *A Basic Example.* As an illustration of the problem, consider a situation where the outcomes $\mathbf{Y}$ are generated by the Gaussian Linear Model

$$\mathrm{M}_1: \quad \mathbf{Y} = \beta_0 \mathbf{X}_0 + \beta_1 \mathbf{X}_1 + \boldsymbol{\epsilon}. \tag{1.1}$$

(One of the two predictors might be a constant intercept vector, but this is immaterial.) For simplicity we assume $n \gg p = 2$ and hence consider $\mathrm{sd}(\epsilon) = \sigma$ as known and without loss of generality $\sigma = 1$. Also, β_0 and β_1 might possibly be 0, but no other covariate is involved in generating $\mathbf{Y}$. Suppose the covariates $\mathbf{X}_j$ are vectors such that

$$\|\mathbf{X}_0\|^2 = \|\mathbf{X}_1\|^2 = 1, \qquad c = \langle \mathbf{X}_0, \mathbf{X}_1 \rangle.$$

Suppose also that we would like to obtain inference for β_0 only. However, we do not know if we should exclude β_1 from the model and retain instead

$$\mathrm{M}_0: \quad \mathbf{Y} = \beta_0 \mathbf{X}_0 + \boldsymbol{\epsilon}.$$

The following is a stylized routine that captures aspects of what might be taught in an undergraduate course on statistics and what is often practiced in statistical applications:

1. Fit the full model M_1 of (1.1) and test the null hypothesis $\beta_1 = 0$ in M_1 (at the usual 5% significance level, say). If it is rejected, use the full model M_1, else use the model M_0. Denoting the selected model by $\hat{\mathrm{M}}$, we have $\hat{\mathrm{M}} = \mathrm{M}_1$ or $\hat{\mathrm{M}} = \mathrm{M}_0$ depending on the outcome of the test.

2. If $\hat{\mathrm{M}} = \mathrm{M}_1$, use the least squares (LS) estimate of β_0 in M_1: $\hat{\beta}_{0\cdot\mathrm{M}_1} = \langle \mathbf{a}, \mathbf{Y} \rangle$ where $\mathbf{a} = (\mathbf{X}_0 - c\mathbf{X}_1)/(1-c^2)$. Its nominal standard error is $\sigma_{0\cdot\mathrm{M}_1} = \|\mathbf{a}\| = (1-c^2)^{-1/2}$ (recall we assume $\sigma = 1$ to be known). The routine $1-\alpha$ interval will be for β_0 relative to the model M_1:

$$\left[\hat{\beta}_{0\cdot\mathrm{M}_1} \pm \Phi^{-1}(1-\alpha/2)\,\sigma_{0\cdot\mathrm{M}_1}\right].$$

3. If, however, $\hat{\mathrm{M}} = \mathrm{M}_0$, use the LS estimate of β_0 in M_0: $\hat{\beta}_{0\cdot\mathrm{M}_0} = \langle \mathbf{X}_0, \mathbf{Y} \rangle$. Its nominal standard error is $\sigma_{0\cdot\mathrm{M}_0} = \|\mathbf{X}_0\| = 1$. The routine $1-\alpha$ interval will be for β_0 relative to the model M_0:

$$\left[\hat{\beta}_{0\cdot\mathrm{M}_0} \pm \Phi^{-1}(1-\alpha/2)\,\sigma_{0\cdot\mathrm{M}_0}\right].$$

The interval used in practice can therefore be summarized as follows:

$$\left[\hat{\beta}_{0\cdot\hat{\mathrm{M}}} \pm \Phi^{-1}(1-\alpha/2)\,\sigma_{0\cdot\hat{\mathrm{M}}}\right]. \tag{1.2}$$

It is assumed to have coverage probability $1-\alpha$, which would follow if the associated test statistic

$$z_{0\cdot\hat{\mathrm{M}}} = \frac{\hat{\beta}_{0\cdot\hat{\mathrm{M}}} - \beta_0}{\sigma_{0\cdot\hat{\mathrm{M}}}} \tag{1.3}$$

had a standard normal distribution. Written in this explicit manner, however, doubt arises immediately and justifiably so as simple simulations show. Consider, for example, the scenario where M_1 is true with $\beta_0 = 0$, $\beta_1 = 1.5$, $c = \langle \mathbf{X}_0, \mathbf{X}_1 \rangle = 1/\sqrt{2}$, and $\alpha = 0.05$ both for the two-sided testing of $\beta_1 = 0$ and for the intended nominal coverage of (1.2). For this case Figure 1 shows a comparison of the actual distribution of (1.3) with the nominal distribution $\mathcal{N}(0,1)$ based on one million simulations of the response vector $\mathbf{Y}$. We see that the actual distribution is leaning to the right compared to the nominal distribution. As a result, the coverage probability falls below the nominal coverage probability: $\mathrm{P}[|z_{0\cdot\hat{\mathrm{M}}}| \le 1.96] = 0.796 < .95$. Thus routine statistical inference after model selection is invalid. Key to this inferential breakdown is that 81.5% of the time the wrong model M_0 is selected due to substantial collinearity between the "wrong" predictor $\mathbf{X}_0$ and the "correct" predictor $\mathbf{X}_1$.

In this example we assumed that $\mathbf{X}_0$ is forced to be in the model. When the selection method is unconstrained, the predictor $\mathbf{X}_0$ may or may not be included in the model, hence statistical inference for β_0 is further complicated by the potential absence of $\mathbf{X}_0$ from the selected model. In general,

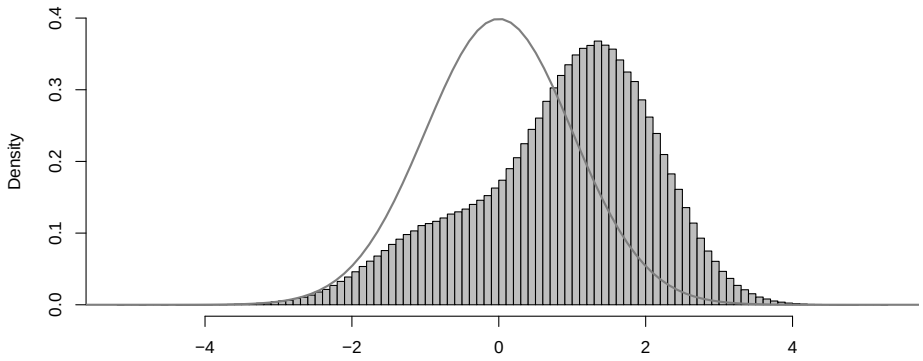

FIG 1. *Distribution of* $z_{0\cdot\hat{\mathrm{M}}}$ *when* $\boldsymbol{\beta} = (0, 1.5)^T$.

routine statistical inference for a parameter β_j after model selection neglects the fact that (1) the presence of β_j is conditional on $\mathbf{X}_j$ being selected into the model, (2) what $\hat{\beta}_j$ estimates depends on the other predictors in the model, (3) the presence of these other predictors is also conditional on being selected into the model, and (4) the selected model cannot be assumed to be true. — The pernicious effects of selection on inference for coefficients β_j is described in a companion paper by Berk, Brown and Zhao (2010).

1.3. *Valid Post-Selection Inference – "PoSI" – and Its Framework.* In this study we propose a method for statistical inference that is valid after model selection in linear models. The approach is essentially a reduction of the post-selection inference problem to one of simultaneous inference. The result is inference that is conservative in that it protects against the multiplicity of all models that are potential outcomes of model selection. In what follows we will refer to this approach to valid post-selection inference as "PoSI". A critical aspect of the methodology is that selected models are *not* assumed to be first-order 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. It is thought that some predictor columns are unnecessary or undesirable, and the goal of model selection is to choose a subset of the columns to generate a more parsimonious predictor matrix. For convenience, we only consider the case of full-rank $\mathbf{X}$, but the results easily generalize to the rank-deficient $\mathbf{X}$.

We will generally assume the full model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I}),$$

to be true, but the assumption of first-order correctness, $\mathrm{E}[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}$, is strictly speaking unnecessary. Its only purpose is to produce a valid independent estimate s_F of σ in terms of the mean squared error (MSE) of the full model: $s_F^2 = \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2/(n-p)$. However, other possibilities for producing an estimate s of σ exist:

1. Exact replications of the response obtained under identical conditions might be available in sufficient numbers. In this case an estimate s could be obtained as the MSE of the one-way ANOVA of the groups of replicates.
2. A larger linear model than the full model might be considered as true, in which case s could be the MSE from this larger model.
3. Another dataset, similar to the one currently being analyzed, might be available, and it might lend itself to produce an estimate s.

In any of these cases, even the full model need not be first-order correct. The technically indispensible assumptions are second-order correctness, that is, homoscedasticity, and distributional correctness, that is, normality: $\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbf{I})$. This is the framework in which we will be able to carry out PoSI.

The remainder of this article is organized as follows. In Section 2 we introduce methodology for reducing the PoSI problem to a problem of simultaneous inference and then a problem of seeking a constant K that depends on $\mathbf{X}$ only and controls the width of the PoSI simultaneous confidence intervals. We propose in Section 2.5 a novel selection method that makes the constant sharp on post-selection inference. After some structural results for the PoSI problem given in Section 3, we show in Section 4 that with increasing number of predictors p the constant K can range between the asymptotic rates $O(\sqrt{\log p})$ and $O(\sqrt{p})$. We give examples for both rates and, inspired by problems in sphere packing and covering, we give upper bounds for the limiting constant in the $O(\sqrt{p})$ case. In Section 5 we discuss computations that can be used to determine the constant for any $\mathbf{X}$ to satisfactory accuracy when $p \leq 20$, and we also provide computations for non-asymptotic upper bounds on these constants that work for larger p. We conclude with a discussion in Section 6. Some proofs are deferred to the appendix.

Finally, we note that although we describe the methodology for unconstrained selection, it is easily adapted to situations such as the one considered

in Section 1.2 above, where inference is sought for a specific coefficient whose predictor is forced to be in the model and only the other predictors are subject to selection (see also Sections 2.5 and 4.2). Similarly, the methodology can be adapted to cases where interest centers on small models, such as models of size ≤ 5 in the presence of $p = 50$ predictors.

2. Methodology.

2.1. *Framing the Problem 1: Multiplicity of Regression Coefficients.* The main point of this section is to be serious about the fact that the meaning of a regression coefficient depends on what the other predictors in the model are. As a consequence, the p regression coefficients of the full model will be proliferating into a plethora of as many as $p\,2^{p-1}$ distinct regression coefficients depending on which submodel they appear in. We start with notation.

To denote submodels we use the index set $\mathrm{M} = \{j_1, j_2, ..., j_m\} \subset \mathrm{M}_F = \{1, \ldots, p\}$ of the predictors $\mathbf{X}_{j_i}$ in the submodel; the size of the submodel is $m = |\mathrm{M}|$ and that of the full model is $p = |\mathrm{M}_F|$. Let $\mathbf{X}_\mathrm{M} = (\mathbf{X}_{j_1}, ..., \mathbf{X}_{j_m})$ denote the $n \times m$ submatrix of $\mathbf{X}$ with columns indexed by M, and let $\hat{\boldsymbol{\beta}}_\mathrm{M}$ denote the corresponding least squares estimate:

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

Now that $\hat{\boldsymbol{\beta}}_\mathrm{M}$ is an estimate, what is it estimating? One answer is that it estimates the true slopes if the submodel M is true, but this gets us into theorizing under assumptions that might not be true – a view that underlies Leeb and Pötscher's (2005, 2006, 2008) criticism. A more fruitful view is *not* to assume that M is first-order correct but rather to define the target of $\hat{\boldsymbol{\beta}}_\mathrm{M}$ by the requirement that it be an unbiased estimate of its target:

$$\boldsymbol{\beta}_\mathrm{M} \triangleq \mathrm{E}[\hat{\boldsymbol{\beta}}_\mathrm{M}] = (\mathbf{X}_\mathrm{M}^T \mathbf{X}_\mathrm{M})^{-1} \mathbf{X}_\mathrm{M}^T \mathrm{E}[\mathbf{Y}]. \tag{2.1}$$

Thus a target for $\hat{\boldsymbol{\beta}}_\mathrm{M}$ can be defined by the right hand side for any $\mathbf{X}_\mathrm{M}$ without the assumption of its or any model's first-order correctness: $\boldsymbol{\mu} = \mathrm{E}[\mathbf{Y}]$ can be entirely arbitrary and unconstrained.

If a conventional interpretation of regression coefficients in a first-order incorrect model is desired, it can no longer be cast in terms of "average difference in the response for a unit difference in X_j, ceteris paribus." Instead, the phrase "in the response" should be replaced with the unwieldy but more correct phrase "in the response approximated in the submodel M". The reason is that the fit in the submodel M is $\hat{\mathbf{Y}}_\mathrm{M} = \mathbf{H}_\mathrm{M}\mathbf{Y}$ (where $\mathbf{H}_\mathrm{M} = \mathbf{X}_\mathrm{M}(\mathbf{X}_\mathrm{M}^T\mathbf{X}_\mathrm{M})^{-1}\mathbf{X}_\mathrm{M}^T$) whose target is $\boldsymbol{\mu}_\mathrm{M} = \mathrm{E}[\hat{\mathbf{Y}}_\mathrm{M}] = \mathbf{H}_\mathrm{M}\mathrm{E}[\mathbf{Y}] = \mathbf{H}_\mathrm{M}\boldsymbol{\mu}$. Thus

in the submodel M we estimate $\boldsymbol{\mu}_{\mathrm{M}}$ which is the Least Squares approximation to the "truth" $\boldsymbol{\mu}$ with regard to the design $\mathbf{X}_{\mathrm{M}}$. Strictly speaking these statements are true for all models, including the full model, since according to G.E.P. Box "all models are wrong, but some are useful." The premise of our study is that it is possible to provide *valid inference in first-order incorrect models*.

If the full model is not assumed to be first-order correct, one can define the target of $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ just as a special case of (2.1) with $\mathrm{M} = \mathrm{M}_F$:

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

This is the same as the "true" coefficient vector in the full model if the full model is first-order correct: $\mathrm{E}[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}$. Without this assumption, using only $\mathbf{H}_{\mathrm{M}} = \mathbf{H}_{\mathrm{M}}\mathbf{H}$ where $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, we obtain for any submodel M the following:

$$\boldsymbol{\beta}_{\mathrm{M}} = (\mathbf{X}_{\mathrm{M}}^T\mathbf{X}_{\mathrm{M}})^{-1}\mathbf{X}_{\mathrm{M}}^T\mathbf{X}\boldsymbol{\beta}. \tag{2.2}$$

Thus the target $\boldsymbol{\beta}_{\mathrm{M}}$ of $\hat{\boldsymbol{\beta}}_{\mathrm{M}}$ is an estimable linear function of the coefficients $\boldsymbol{\beta}$ in the full model even if the latter is not first-order correct.

Compare next the following two definitions:

$$\boldsymbol{\beta}_{\mathrm{M}} \triangleq \mathrm{E}[\hat{\boldsymbol{\beta}}_{\mathrm{M}}] \quad \text{and} \quad \boldsymbol{\beta}^{\mathrm{M}} \triangleq (\beta_{j\in\mathrm{M}})^T, \tag{2.3}$$

the latter being the coefficients from the full model M_F subsetted to the submodel M. While $\hat{\boldsymbol{\beta}}_{\mathrm{M}}$ estimates $\boldsymbol{\beta}_{\mathrm{M}}$, it does *not* generally estimate $\boldsymbol{\beta}^{\mathrm{M}}$. Instead, the definition of $\boldsymbol{\beta}_{\mathrm{M}}$ involves $\mathbf{X}$ and all coefficients in $\boldsymbol{\beta}$ through (2.2). Actually, a little algebra shows that $\boldsymbol{\beta}_{\mathrm{M}} = \boldsymbol{\beta}^{\mathrm{M}}$ if and only if

$$\mathbf{X}_{\mathrm{M}}^T\mathbf{X}_{\mathrm{M}^c}\boldsymbol{\beta}^{\mathrm{M}^c} = \mathbf{0}, \tag{2.4}$$

where M^c denotes the complement of M in M_F. Special cases of (2.4) include: (1) the column space of $\mathbf{X}_{\mathrm{M}}$ is orthogonal to that of $\mathbf{X}_{\mathrm{M}^c}$, and (2) $\boldsymbol{\beta}^{\mathrm{M}^c} = \mathbf{0}$, meaning that the submodel M is first-order correct. However, in general (2.4) does not hold. Ignoring the difference between $\boldsymbol{\beta}_{\mathrm{M}}$ and $\boldsymbol{\beta}^{\mathrm{M}}$ leads to invalid inference in the conventional post-selection inference routine, as seen in the basic example in Section 1.2.

In order to distinguish regression coefficients as a function of the model they appear in, we write $\beta_{j\cdot\mathrm{M}} = \mathrm{E}[\hat{\beta}_{j\cdot\mathrm{M}}]$ for the components of $\boldsymbol{\beta}_{\mathrm{M}} = \mathrm{E}[\hat{\boldsymbol{\beta}}_{\mathrm{M}}]$. An important convention we adopt throughout this article is that the index j of a coefficient refers to the coefficient's index in the original full model M_F: $\beta_{j\cdot\mathrm{M}}$ refers not to the j'th coordinate of $\boldsymbol{\beta}_{\mathrm{M}}$, but to the coordinate of

$\boldsymbol{\beta}_{\mathrm{M}}$ corresponding to the j'th predictor in the full design matrix $\mathbf{X}$. We refer to this convention as "*full model indexing*".

We conclude this section with some observations that may alleviate investigators apprehensions about their estimates $\hat{\beta}_{j\cdot\mathrm{M}}$ if they are not to assume the truth of their selected submodel M. The reality is that investigators never know whether a selected submodel is true. Yet, we just showed that the estimates they obtain can be given meaning, and we will show next that they can be given valid statistical inference — *not* assuming first-order correctness of the submodel. The following considerations may be helpful in making sense of estimates without burdening them with untenable assumptions:

- The interpretation of parameters $\beta_{j\cdot\mathrm{M}}$ and their estimates $\hat{\beta}_{j\cdot\mathrm{M}}$ can be usefully framed in the language of "adjustment": the j'th coefficient and its estimate are "adjusted" for the other predictors in the submodel M. If two investigators have selected different submodels, they have estimated differently adjusted parameters for the predictors shared by the two submodels. Statistical common sense dictates that caution should be used when comparing coefficients for a predictor obtained in two different submodels. We consider a difference in adjustment as defining different parameters: $\beta_{j\cdot\mathrm{M}}$ and $\beta_{j\cdot\mathrm{M}'}$ are not the same parameters if $\mathrm{M} \neq \mathrm{M}'$.
- A purpose of this study is to provide valid statistical inference in *any* submodel. One might, however, object that if *any* choice of submodel can be provided with valid statistical inference irrespective of how "wrong" it might be, then something must be wrong with the approach because not just any submodel should be used. This objection fails to allow that correctness of submodels is not always the primary concern — a greater priority might be well-reasoned adjustment for factors that subject matter experts consider relevant. Such experts may want to tread a balance between empirical evidence of model fit and subject matter knowledge about the relative importance of predictors. It may therefore be wise for the statistician to be prepared to provide inferential protections for any outcome of model selection.

2.2. *Framing the Problem 2: Simultaneous Inference for PoSI.* After defining $\boldsymbol{\beta}_{\mathrm{M}}$ as the target of the estimate $\hat{\boldsymbol{\beta}}_{\mathrm{M}}$, we consider inference for it. To this end we require a normal homoscedastic model for $\mathbf{Y}$, but we can leave its mean $\boldsymbol{\mu} = \mathrm{E}[\mathbf{Y}]$ unspecified, $\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbf{I}_n)$. We then have:

$$\hat{\boldsymbol{\beta}}_{\mathrm{M}} \sim \mathcal{N}(\boldsymbol{\beta}_{\mathrm{M}}, \sigma^2(\mathbf{X}_{\mathrm{M}}^T\mathbf{X}_{\mathrm{M}})^{-1}).$$

For inference we need a valid estimate of σ that is independent of all estimates $\hat{\beta}_{j\cdot\mathrm{M}}$. Since most frequently in practice the full model is assumed to be first-order correct so that its MSE provides a valid estimate of σ, we write $s = s_F$ and assume $(n-p)s_F^2/\sigma^2 \sim \chi^2_{n-p}$. However, other sources for an estimate of σ, discussed in Section 1.3, should be kept in mind as they allow us to drop the assumption that the full model is first-order correct.

Let $t_{j\cdot\mathrm{M}}$ denote a t-ratio for $\beta_{j\cdot\mathrm{M}}$ that uses this MSE value:

$$t_{j\cdot\mathrm{M}} \triangleq \frac{\hat{\beta}_{j\cdot\mathrm{M}} - \beta_{j\cdot\mathrm{M}}}{\left((\mathbf{X}_\mathrm{M}^T\mathbf{X}_\mathrm{M})^{-1}\right)_{jj}^{\frac{1}{2}}\ s_F}. \tag{2.5}$$

(Consistent with full model indexing, the notation $(...)_{jj}$ refers not to the j'th diagonal element but to the diagonal element corresponding to the predictor $\mathbf{X}_j$.) The quantity $t_{j\cdot\mathrm{M}}$ has a central t-distribution with $n-p$ degrees of freedom.

Important is that the standard error used in (2.5) does *not* involve the MSE s_M from the submodel M, for two reasons: (1) We do not assume that the submodel M is first-order correct; therefore each MSE s^2_M could have a distribution that is a multiple of a non-central χ^2 distribution with unknown non-centrality parameter. (2) More disconcertingly, the MSE is the result of selection, $s^2_{\hat{\mathrm{M}}}$, and hence a blend of MSEs, $s^2_{\hat{\mathrm{M}}} = \sum_\mathrm{M} s^2_\mathrm{M} I(\hat{\mathrm{M}} = \mathrm{M})$; nothing is known about its distribution. These problems are avoided by using a valid MSE s_F^2 that is independent of submodels.

With the above notations, a routine $1-\alpha$ confidence interval for $\beta_{j\cdot\mathrm{M}}$ is formed by

$$\mathrm{CI}(j,\mathrm{M};K) \ = \ \left[\hat{\beta}_{j\cdot\mathrm{M}} \pm K\left((\mathbf{X}_\mathrm{M}^T\mathbf{X}_\mathrm{M})^{-1}\right)_{jj}^{\frac{1}{2}} s_F\right], \tag{2.6}$$

where $K = t_{n-p,1-\alpha/2}$ is the $1-\alpha/2$ quantile of a t-distribution with $n-p$ degrees of freedom. This construction of $\mathrm{CI}(j,\mathrm{M};K)$ is valid under the assumption that the submodel M is chosen independently of the response $\mathbf{Y}$. In practice, data analysts usually proceed in a different fashion (see the references in Section 1): They may first examine and process the data in order to produce a selected model; call this model $\hat{\mathrm{M}}$ to emphasize that it depends in some fashion on $\mathbf{Y}$: $\hat{\mathrm{M}} = \hat{\mathrm{M}}(\mathbf{Y})$. They may then use the estimates $\hat{\beta}_{j\cdot\hat{\mathrm{M}}}$ $(j \in \hat{\mathrm{M}})$, and for statistical inference they may rely on confidence intervals $\mathrm{CI}(j,\hat{\mathrm{M}};K)$ and tests, but using the MSE $s^2_{\hat{\mathrm{M}}}$ from the selected model $\hat{\mathrm{M}}$.

It is often asserted or assumed in practice that for fixed j it holds true that $\mathrm{P}[\beta_j \in \mathrm{CI}(j,\hat{\mathrm{M}};K)] \geq 1-\alpha$. This statement, however, poses a conundrum:

To make sense there must be an understanding that the event inside the probability means $\{j \in \hat{\mathrm{M}}\} \cap \{\beta_j \in \mathrm{CI}(j, \hat{\mathrm{M}}; K)\}$, but the probability of this event cannot be controlled because $\mathrm{P}[j \in \hat{\mathrm{M}}]$ alone is unpredictable and may be "small" (for example, 0.7) even for a strong predictor. The obvious solution would seem to be conditioning on the event $[j \in \hat{\mathrm{M}}]$, but the statement $\mathrm{P}[\beta_j \in \mathrm{CI}(j, \hat{\mathrm{M}}; K) \,|\, j \in \hat{\mathrm{M}}] \geq 1-\alpha$ is generally not valid either (Leeb and Pötscher (2005; 2006; 2008), Berk, Brown and Zhao (2010)).

In this study we propose to construct valid PoSI inference by replacing the constant $K = t_{n-p,1-\alpha/2}$ in (2.6) with a larger value such that a family-wise error rate is controlled:

$$\mathrm{P}\left[\max_{\mathrm{M}} \max_{j \in \mathrm{M}} |t_{j\cdot\mathrm{M}}| \leq K\right] \;=\; 1-\alpha, \tag{2.7}$$

where the maxima are taken over all submodels M and all indexes $j \in \mathrm{M}$. That is, we require K to be sufficiently large that we obtain simultaneously valid inference for all parameters $\beta_{j\cdot\mathrm{M}}$. This tends to be a large simultaneous inference problem as the number of parameters can be as large as $p\,2^{p-1}$ (each predictor j is contained in 2^{p-1} submodels).

If K satisfies (2.7), then we will show that the confidence intervals (2.6) satisfy

$$\mathrm{P}\left[\beta_{j\cdot\hat{\mathrm{M}}} \in \mathrm{CI}(j, \hat{\mathrm{M}}; K)\ \forall j \in \hat{\mathrm{M}}\right] \;\geq\; 1-\alpha \tag{2.8}$$

for *any* model selection procedure $\hat{\mathrm{M}} = \hat{\mathrm{M}}(\mathbf{Y})$. Universal validity irrespective of the model selection procedure is a strong property that raises some questions but also has benefits, both of which we now discuss:

- Most fundamentally, the inference guarantee (2.8) solves the conundrum mentioned above: the event $\{\beta_{j\cdot\hat{\mathrm{M}}} \in \mathrm{CI}(j, \hat{\mathrm{M}}; K)\}$ is not coherent because it fails to explicitly require $\{j \in \hat{\mathrm{M}}\}$. We solve this problem by inserting the "for-all" quantifier ($\forall$) in (2.8). The event $\{\beta_{j\cdot\hat{\mathrm{M}}} \in \mathrm{CI}(j, \hat{\mathrm{M}}; K)\ \forall j \in \hat{\mathrm{M}}\}$ is fully coherent.
- Since (2.8) states validity of inference for any model selection procedure, it entails that choosing K according to (2.7) produces a rather conservative inference procedure. A natural question is whether there exists a model selection procedure that actually requires the family-wise bound K of (2.7). The answer is affirmative, and the requisite selection procedure is discussed in Section 2.5. Thus, in a certain respect our approach to PoSI cannot be improved.
- Universal validity of statistical inference with respect to any model selection procedure seems desirable or even essential for applications

in which the model selection routine is not specified in advance or for which the routine involves some ad hoc elements that cannot be accurately pre-specified. 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.

- If one insists on inference for only the parameters in the selected model, one has to face the fact that an estimate $\hat{\beta}_{j\cdot\mathrm{M}}$ is only seen for those $\mathbf{Y}$ for which $\hat{\mathrm{M}}(\mathbf{Y}) = \mathrm{M}$. We bypass this problem by proposing that all estimates $\hat{\beta}_{j\cdot\mathrm{M}}$ are always at least potentially seen, even though few are ever actually looked at. This view allows data analysts to change their minds, to improvise and consider models other than those produced by the selection procedure, and to experiment with multiple selection procedures. The PoSI method proposed here covers all such eventualities.

We conclude with a caveat: The treatment in this article does not address prediction of future observations. We are only concerned here with honest inference for regression coefficients.

2.3. *Simultaneous Inference for the PoSI Problem.* We state and prove the theorem that derives the PoSI guarantee (2.8) from the simultaneous inference condition (2.7):

THEOREM 2.1. *Suppose* $\hat{\mathrm{M}} = \hat{\mathrm{M}}(\mathbf{Y})$ *is a model selection procedure. Let* K *be such that*

$$\mathrm{P}\left[\max_{\mathrm{M}} \max_{j\in\mathrm{M}} |t_{j\cdot\mathrm{M}}| \le K\right] \ge 1-\alpha. \tag{2.9}$$

Then

$$\mathrm{P}\left[\max_{j\in\hat{\mathrm{M}}} |t_{j\cdot\hat{\mathrm{M}}}| \le K\right] \ge 1-\alpha. \tag{2.10}$$

Preparatory note: We have used repeatedly the notion that a selection procedure $\hat{\mathrm{M}} = \hat{\mathrm{M}}(\mathbf{Y})$ is a mapping from the response vectors $\mathbf{Y}$ to the submodels M. To fix ideas, we note that there are 2^p submodels (the empty model being a zero fit with no parameters). Using set theory notation $2^{\{1,2,\ldots,p\}} = \{\mathrm{M} : \mathrm{M} \subset \mathrm{M}_F\}$ for the set of all submodels, a model selection procedure $\hat{\mathrm{M}}$ is a mapping

$$\hat{\mathrm{M}} : \mathbb{R}^N \to 2^{\{1,2,\ldots,p\}}, \qquad \mathbf{Y} \mapsto \hat{\mathrm{M}}(\mathbf{Y}) = \{j_1, \ldots, j_q\},$$

where $1 \le j_1 < ... < j_q \le p$. Such a mapping partitions the space $\mathbb{R}^N$ of response vectors $\mathbf{Y}$ into at most 2^p regions within each of which the selected submodel $\hat{\mathrm{M}}(\mathbf{Y})$ is shared. The discussion surrounding the definitions (2.3) describes that, properly viewed, each submodel M has its own targets of estimation, the submodel-specific regression coefficients $\beta_{j\cdot \mathrm{M}}$, whose number is generally $p\,2^{p-1}$. Thus the requirement (2.9) asks for considerable simultaneity protection.

PROOF: The link between the PoSI guarantee (2.10) and the simultaneous inference problem (2.9) is easily established by upper-bounding the expression

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

with a universal term that no longer depends on the selected $\hat{\mathrm{M}}(\mathbf{Y})$ as follows:

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

This inequality holds for all $\mathbf{Y} \in \mathbb{R}^n$. We have then for any selection procedure $\hat{\mathrm{M}}$ the universal inequality

$$P\left[\max_{j \in \hat{\mathrm{M}}(\mathbf{Y})} |t_{j\cdot \hat{\mathrm{M}}(\mathbf{Y})}(\mathbf{Y})| \le K\right] \;\ge\; P\left[\max_{\mathrm{M}} \max_{j \in \mathrm{M}} |t_{j\cdot \mathrm{M}}(\mathbf{Y})| \le K\right],$$

where the right hand probability no longer depends on the selection procedure $\hat{\mathrm{M}}$. This inequality implies the assertion of the theorem. □

As the constant K depends on $\mathbf{X}$ only and not on the selection procedure $\hat{\mathrm{M}}$, we write $K = K(\mathbf{X})$. Depending on the context we may list more arguments, as in $K = K(\mathbf{X}, \alpha, p, n)$. The constant satisfies the PoSI guarantee (2.8), which is a reformulation of (2.10). We call the interval

$$\left[\hat{\beta}_{j\cdot \hat{\mathrm{M}}} \pm K(\mathbf{X}) \left((\mathbf{X}_{\hat{\mathrm{M}}}^T \mathbf{X}_{\hat{\mathrm{M}}})^{-1}\right)_{jj}^{1/2} s_F\right]$$

the "PoSI simultaneous confidence interval", and we call $K(\mathbf{X})$ the "PoSI constant."

2.4. *Scheffé Protection.* By being serious about the fact that the LS estimators in different submodels in general estimate different parameters, we generated a simultaneous inference problem involving up to $p\,2^{p-1}$ linear combinations $\beta_{j\cdot \mathrm{M}}$ of the p regression coefficients $\beta_1, ..., \beta_p$ from the full model. In view of the enormous number of linear combinations or "contrasts"

for which simultaneous inference is sought, one should wonder whether the problem is not best solved by Scheffé's method (1953; 1959) which provides simultaneous inference for *all* linear combinations or contrasts:

$$\mathrm{P}\left[\sup_{\mathbf{a}} \frac{(\mathbf{a}^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}))^2}{\mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a}\, s_F^2} \le K_{Sch}^2\right] \;=\; 1-\alpha, \tag{2.11}$$

where the Scheffé constant K_{Sch} is found to be $K_{Sch} = \sqrt{pF_{p,n-p,1-\alpha}}$ (assuming the full model is true). It has the pleasing property that it provides an upper bound for *all* PoSI constants:

$$K(\mathbf{X}) \;\le\; K_{Sch} \quad \forall\, \mathbf{X}.$$

Thus we find that parameter estimates $\hat{\beta}_{j\cdot\mathrm{M}}$ whose t-ratio exceeds K_{Sch} in magnitude are universally safe from invalidation due to model selection in *any* regression context.

The universality of the Scheffé constant is, however, 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 a worthwhile endeavor. An indication toward this end is provided by the following comparison as $n - p \to \infty$ and $p \to \infty$:

- For the Scheffé constant it holds $K_{Sch} \sim \sqrt{p}$.
- For the special case of an orthogonal design $\mathbf{X}$, however, it holds $K_{orth} \sim \sqrt{2\log p}$ (see Section 3.5 below).

Thus the PoSI constant $K_{orth} \sim \sqrt{2\log p}$ for orthogonal designs is much smaller than the Scheffé constant $K_{Sch} \sim \sqrt{p}$. The big gap between the two suggests that the Scheffé constant may be too conservative at least in some cases. We will study the case of certain non-orthogonal designs for which the PoSI constant is $O(\sqrt{\log(p)})$ in Section 4.1. On the other hand, the PoSI constant can approach the order $O(\sqrt{p})$ of the Scheffé constant K_{Sch} as well, and we will study one such case in Section 4.2.

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

2.5. *PoSI-Sharp Model Selection — "SPAR" and "SPAR1".* We describe the model selection procedure that requires the full protection of the simultaneous inference procedure (2.7). It is found by using the criterion of the simultaneous inference procedure itself for selection:

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

In words: the selected submodel is found by looking for the most significant adjusted predictor across all submodels. In this submodel, the less significant predictors matter only in so far as they boost the significance of the winning predictor by adjusting it accordingly. For this reason we call this selection procedure "*Single Predictor Adjusted Regression* or "SPAR". This procedure has nothing to do with the quality of the fit to $\mathbf{Y}$ provided by the model. While our present interest is only in pointing out the existence of a selection procedure that requires full PoSI protection, SPAR could be of practical interest when the analysis is centered on strength of effects, not quality of model fit.

Practically of greater interest might be a restricted version of SPAR whereby a predictor of interest is determined a priori and the search is for adjustment that optimizes this predictor's effect. We name the resulting procedure "SPAR1". If the predictor of interest is $\mathbf{X}_p$, say, the model selection is

$$\hat{\mathrm{M}}(\mathbf{Y}) \triangleq \mathrm{argmax}_{\mathrm{M}:\, p \in \mathrm{M}} |t_{p \cdot \mathrm{M}}(\mathbf{Y})|.$$

The associated "PoSI1" guarantee that we seek is

$$\mathrm{P}\left[\max_{\mathrm{M}: p \in \mathrm{M}} |t_{p \cdot \mathrm{M}}| \le K^{(p)}\right] = 1 - \alpha.$$

Clearly, the unrestricted PoSI constant K dominates the PoSI1 constant: $K \ge K^{(p)}$. Even so, we will construct in Section 4.2 an example where the PoSI1 constant increases at the Scheffé rate, and indeed is asymptotically more than 63% of the Scheffé constant. This is our main reason for introducing SPAR1 and PoSI1 at this point.

3. The Structure of the PoSI Problem.

3.1. *Canonical Coordinates.* We introduce canonical coordinates to reduce the dimensionality of the design matrix from $n \times p$ to $p \times p$. This reduction is important both geometrically and computationally because the coverage problem really takes place in the column space of $\mathbf{X}$.

DEFINITION: *For any orthonormal basis* $\mathbf{Q} = (\mathbf{q}_1, ..., \mathbf{q}_p)$ *of the column space of* $\mathbf{X}$*, we call* $\tilde{\mathbf{X}} = \mathbf{Q}^T\mathbf{X}$ *and* $\tilde{\mathbf{Y}} = \mathbf{Q}^T\mathbf{Y}$ *canonical coordinates of* $\mathbf{X}$ *and* $\hat{\mathbf{Y}}$*, where* $\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \mathbf{Q}\mathbf{Q}^T\mathbf{Y}$*. Completing the basis to all of* $\mathbb{R}^n$ *by adding* $\mathbf{Q}_\perp = (\mathbf{q}_{p+1}, ..., \mathbf{q}_n)$*, canonical coordinates of the residuals* $\mathbf{r}$ *are given by* $\tilde{\mathbf{r}} = \mathbf{Q}_\perp^T\mathbf{r}$.

The sizes in canonical coordinates are $\tilde{\mathbf{X}}_{p\times p}$, $\tilde{\mathbf{Y}}_{p\times 1}$, and $\tilde{\mathbf{r}}_{(n-p)\times 1}$. In what follows we extend the notation $\mathbf{X}_M$ for extraction of subsets of columns to canonical coordinates $\tilde{\mathbf{X}}_M$. Accordingly slopes obtained from canonical coordinates will be denoted by $\hat{\boldsymbol{\beta}}_M(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}) = (\tilde{\mathbf{X}}_M^T\tilde{\mathbf{X}}_M)^{-1}\tilde{\mathbf{X}}_M^T\tilde{\mathbf{Y}}$ to distinguish them from the slopes obtained from the original data $\hat{\boldsymbol{\beta}}_M(\mathbf{X}, \mathbf{Y}) = (\mathbf{X}_M^T\mathbf{X}_M)^{-1}\mathbf{X}_M^T\mathbf{Y}$, if only to state in the following proposition that they are identical. The same holds for the MSEs $s_F^2(\tilde{\mathbf{r}}) = \|\tilde{\mathbf{r}}\|^2/(n-p)$ and $s_F^2(\mathbf{r}) = \|\mathbf{r}\|^2/(n-p)$, and the ensuing t-statistics.

PROPOSITION 3.1. *Properties of canonical coordinates:*

1. $\tilde{\mathbf{Y}} = \mathbf{Q}^T\mathbf{Y}$ *and* $\tilde{\mathbf{r}} = \mathbf{Q}_\perp^T\mathbf{Y}$,
2. $\tilde{\mathbf{X}}_M^T\tilde{\mathbf{X}}_M = \mathbf{X}_M^T\mathbf{X}_M$ *and* $\tilde{\mathbf{X}}_M^T\tilde{\mathbf{Y}} = \mathbf{X}_M^T\mathbf{Y}$,
3. $\hat{\boldsymbol{\beta}}_M(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}) = \hat{\boldsymbol{\beta}}_M(\mathbf{X}, \mathbf{Y})$ *for all submodels* M,
4. $s_F(\tilde{\mathbf{r}}) = s_F(\mathbf{r})$,
5. $\tilde{\mathbf{Y}} \sim \mathcal{N}(\tilde{\mathbf{X}}\boldsymbol{\beta}, \sigma^2\mathbf{I}_p), \quad \tilde{\mathbf{r}} \sim \mathcal{N}(\mathbf{0}_{(n-p)\times 1}, \sigma^2\mathbf{I}_{n-p})$.
6. $t_{j\cdot M} = \frac{\hat{\beta}_{j\cdot M}(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}) - \beta_{j\cdot M}}{\left((\tilde{\mathbf{X}}_M^T\tilde{\mathbf{X}}_M)^{-1}\right)_{jj}^{1/2} s_F(\tilde{\mathbf{r}})}$
7. *The canonical coordinates* $\tilde{\mathbf{X}}$ *of the design matrix can be chosen to form an upper triangular matrix or a symmetric matrix.*

The proofs of *1.-6.* are elementary. As for *7.*, an upper triangular version $\tilde{\mathbf{X}}$ can be obtained from a QR-decomposition based on, for example, a (modified) Gram-Schmidt procedure: $\mathbf{X} = \mathbf{QR}$, $\tilde{\mathbf{X}} = \mathbf{R}$; a symmetric version of $\tilde{\mathbf{X}}$ 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 $\mathbb{R}^p$. In what follows we will freely assume when convenient that all objects are rendered in canonical coordinates, implying that the design matrix is of size $p \times p$ and the response (or its component that matters for estimation of slopes) is of size $p \times 1$.

3.2. *PoSI as a Gaussian Coverage Problem.* We show that after some simplifications the PoSI coverage problem (2.7) can be reduced to a Gaussian coverage problem for a specific polytope in $\mathbb{R}^p$ whose shape is determined by the design matrix $\mathbf{X}$. We start with the simplifications: Due to pivotality

of t-statistics, the PoSI problem is invariant under translation of $\boldsymbol{\beta}$ and rescaling of σ. Hence it suffices to verify coverage statements for $\boldsymbol{\beta} = \mathbf{0}$ and $\sigma = 1$. Furthermore, since we consider the classical case $n \gg p$, the limit $n \to \infty$ is equivalent to assuming σ^2 known and hence $s_F^2 = \sigma^2 = 1$. The t-statistics become z-statistics, which we write as

$$z_{j\cdot \mathrm{M}} \;=\; \frac{\hat{\beta}_{j\cdot \mathrm{M}}}{\big((\mathbf{X}_{\mathrm{M}}^T \mathbf{X}_{\mathrm{M}})^{-1}\big)_{jj}^{1/2}}. \tag{3.1}$$

We will concentrate on this special case because it simplifies the analytics without loss of structurally important features.

Simultaneous inference for linear functions in linear models has a simple well-known geometry which we now specialize to the case of the PoSI problem. The linear functions for which we seek simultaneous inference are the slope estimates in all submodels. As linear functions of the response they and their associated z-statistics have the form

$$\hat{\beta}_{j\cdot \mathrm{M}} \;=\; \boldsymbol{l}_{j\cdot \mathrm{M}}^T \mathbf{Y}, \qquad z_{j\cdot \mathrm{M}} \;=\; \bar{\boldsymbol{l}}_{j\cdot \mathrm{M}}^T \mathbf{Y} \tag{3.2}$$

where the "PoSI coefficient vectors" $\boldsymbol{l}_{j\cdot \mathrm{M}}$ and $\bar{\boldsymbol{l}}_{j\cdot \mathrm{M}}$ are (up to a scale factor) the predictor vector $\mathbf{X}_j$ adjusted for (orthogonalized w.r.t.) the other predictors in the submodel M. Notation to mathematically express this fact is unavoidably cumbersome:

$$\boldsymbol{l}_{j\cdot \mathrm{M}} \;=\; \frac{(\mathbf{I} - \mathbf{P}_{\mathrm{M}\setminus j})\mathbf{X}_j}{\|(\mathbf{I} - \mathbf{P}_{\mathrm{M}\setminus j})\mathbf{X}_j\|^2}, \qquad \bar{\boldsymbol{l}}_{j\cdot \mathrm{M}} \;=\; \frac{\boldsymbol{l}_{j\cdot \mathrm{M}}}{\|\boldsymbol{l}_{j\cdot \mathrm{M}}\|}, \tag{3.3}$$

where $\mathbf{P}_{\mathrm{M}\setminus j} = \mathbf{X}_{\mathrm{M}\setminus j}(\mathbf{X}_{\mathrm{M}\setminus j}^T \mathbf{X}_{\mathrm{M}\setminus j})^{-1} \mathbf{X}_{\mathrm{M}\setminus j}^T$ is the projection matrix in the submodel M but leaving out the predictor j (we write $\mathrm{M} \setminus j$ instead of the more correct $\mathrm{M} \setminus \{j\}$). The connection between (3.1) and (3.2) follows from

$$\big((\mathbf{X}_{\mathrm{M}}^T \mathbf{X}_{\mathrm{M}})^{-1}\big)_{jj} \;=\; (\mathbf{X}_j^T (\mathbf{I} - \mathbf{P}_{\mathrm{M}\setminus j})\mathbf{X}_j)^{-1} \;=\; \|\boldsymbol{l}_{j\cdot \mathrm{M}}\|^2.$$

To complete the structural description of the PoSI problem we let

$$\mathcal{L}(\mathbf{X}) \;=\; \{\bar{\boldsymbol{l}}_{j\cdot \mathrm{M}} : j, \mathrm{M},\ j \in \mathrm{M}\} \tag{3.4}$$

and note that in canonical coordinates $\bar{\boldsymbol{l}}_{j\cdot \mathrm{M}} \in \mathbb{R}^p$. Using the simplifications $\boldsymbol{\beta} = \mathbf{0}_p$ and $\sigma = 1$, we further have $\tilde{\mathbf{Y}} \sim \mathcal{N}(\mathbf{0}_p, \mathbf{I}_p)$, and therefore:

PROPOSITION 3.2. *In the limit $(n - p) \to \infty$ the PoSI problem (2.7) is equivalent to a p-dimensional Gaussian coverage problem: Find $K(\mathbf{X})$ such that*

$$\mathrm{P}\left[\max_{\mathrm{M}} \max_{j \in \mathrm{M}} |z_{j\cdot \mathrm{M}}| \le K\right] \;=\; \mathrm{P}\left[\max_{\bar{\boldsymbol{l}} \in \mathcal{L}(\mathbf{X})} |\bar{\boldsymbol{l}}^T \mathbf{Z}| \le K(\mathbf{X})\right] \;=\; 1 - \alpha,$$

where $\mathbf{Z} = (Z_1, ..., Z_p)^T \sim \mathcal{N}(\mathbf{0}_p, \mathbf{I}_p)$.

An alternative way of looking at the PoSI problem is in terms of a Gaussian process. We mention this view because it is the basis of some software implementations used to solve Gaussian simultaneous inference and coverage problems, even though in this case it does not result in a practicable approach. In the PoSI problem the obvious Gaussian process is $\mathbf{W} = (z_{j\cdot \mathrm{M}})$. The covariance structure of $\mathbf{W}$ is $\mathbf{\Sigma} = (\Sigma_{j\cdot \mathrm{M};\, j'\cdot \mathrm{M}'})$ where

$$\Sigma_{j\cdot \mathrm{M};\, j'\cdot \mathrm{M}'} \;=\; \bar{\boldsymbol{l}}_{j\cdot \mathrm{M}}^T \bar{\boldsymbol{l}}_{j'\cdot \mathrm{M}'}\,. \tag{3.5}$$

The coverage problem can be written as $\mathrm{P}[\|\mathbf{W}\|_\infty \le K] = 1 - \alpha$. Software that computes such coverage allows users to specify a covariance structure $\mathbf{\Sigma}$ and intervals such as $[-K, +K]$ for the components. In our experiments this approach worked up to $p = 7$, the obvious 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 Section 5 worked for up to $p = 20$.

3.3. *Geometry of the PoSI Coefficient Vectors.* The set $\mathcal{L}(\mathbf{X})$ of coefficient 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 p-dimensional.

PROPOSITION 3.3. *Orthogonalities in* $\mathcal{L}(\mathbf{X})$*:*

1. *Successive orthogonalization:*
   $\bar{\boldsymbol{l}}_{j\cdot\{j\}} = \mathbf{X}_j / \|\mathbf{X}_j\|$,
   $\bar{\boldsymbol{l}}_{j\cdot \mathrm{M}} \in \mathrm{span}\{\, \mathbf{X}_j : j \in \mathrm{M}\}$ *and* $\bar{\boldsymbol{l}}_{j\cdot \mathrm{M}} \perp \mathbf{X}_{j'}$ *for* $j \ne j'$ *both* $\in \mathrm{M}$.
2. *The following forms an o.n. basis of* $\mathbb{R}^p$*:*

$$\{\bar{\boldsymbol{l}}_{1\cdot\{1\}},\ \bar{\boldsymbol{l}}_{2\cdot\{1,2\}},\ \bar{\boldsymbol{l}}_{3\cdot\{1,2,3\}},\ \ldots,\ \bar{\boldsymbol{l}}_{p\cdot\{1,2,\ldots,p\}}\}.$$

   Other o.n. bases are obtained by permuting the order of $\{1, 2, ..., p\}$ *(not all of which may result in distinct bases).*
3. *Two 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. *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).*

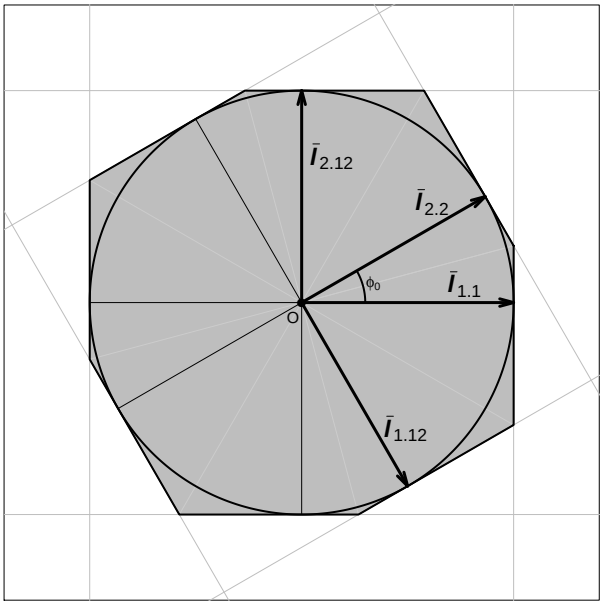

FIG 2. *The PoSI polytope* $\mathbf{\Pi}_{K=1}$ *for* $p = 2$*: The normalized raw predictor vectors are* $\bar{\boldsymbol{l}}_{1\cdot\{1\}} \sim \mathbf{X}_1$ *and* $\bar{\boldsymbol{l}}_{2\cdot\{2\}} \sim \mathbf{X}_2$*, and the normalized adjusted versions are* $\bar{\boldsymbol{l}}_{1\cdot\{1,2\}}$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

A note on cardinalities: If the predictor vectors $\mathbf{X}_j$ have no orthogonal pairs among them, then $|\mathcal{L}(\mathbf{X})| = p\,2^{p-1}$. If, however, there exist orthogonal pairs, then $|\mathcal{L}(\mathbf{X})|$ is less. For example, if there exists exactly one orthogonal pair, then $|\mathcal{L}(\mathbf{X})| = (p-1)\,2^{p-1}$. In the extreme, when $\mathbf{X}$ is a fully orthogonal design, then $|\mathcal{L}(\mathbf{X})| = p$.

The proposition implies that there exist certain necessary orthogonalities in $\mathcal{L}(\mathbf{X})$. In terms of the covariance structure $\mathbf{\Sigma}$ (3.5), orthogonalities in $\mathcal{L}(\mathbf{X})$ correspond to zero correlations in $\mathbf{\Sigma}$. Part *4.* of the proposition states that 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 certainly rich in zeros. (It can be much sparser in the presence of orthogonalities among the predictors.)

3.4. *The PoSI Polytope.* Coverage problems can be framed geometrically in terms of probability coverage of polytopes in $\mathbb{R}^p$. For the PoSI problem the polytope is defined by

$$\mathbf{\Pi}_K \;=\; \{\mathbf{u} \in \mathbb{R}^p : \, |\bar{\boldsymbol{l}}^T \mathbf{u}| \le K, \; \forall \bar{\boldsymbol{l}} \in \mathcal{L}(\mathbf{X}) \},$$

henceforth called the "PoSI polytope". The PoSI coverage problem is to calibrate K such that

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

In this notation the Scheffé ellipsoid (2.11) turns into the "Scheffé ball" that has a root-χ^2_p coverage probability in the limit $(n-p) \to \infty$:

$$\mathbf{B}_K \;=\; \{\mathbf{u} \in \mathbb{R}^p : \, \|\mathbf{u}\| \le K \}, \qquad \mathrm{P}[\mathbf{Z} \in \mathbf{B}_K] = F_{\chi^2,p}(K^2).$$

Many properties of the polytopes $\mathbf{\Pi}_K$ are not specific to PoSI because they hold for polytopes derived from simultaneous inference problems for linear functions of $\mathbf{Y}$ with coefficient vectors forming arbitrary sets $\mathcal{L}$ of unit vectors:

1. The polytopes form a scale family of geometrically similar bodies: $\mathbf{\Pi}_K = K\mathbf{\Pi}_1$.
2. They are point symmetric about the origin: $\mathbf{\Pi}_K = -\mathbf{\Pi}_K$.
3. They contain the Scheffé ball: $\mathbf{B}_K \subset \mathbf{\Pi}_K$.
4. They are intersections of "slabs" of width $2K$: $\mathbf{\Pi}_K = \bigcap_{\bar{\boldsymbol{l}} \in \mathcal{L}} \{\mathbf{u} \in \mathbb{R}^p : |\mathbf{u}^T \bar{\boldsymbol{l}}| \le K \}$.
5. They 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 many orthogonalities in $\mathcal{L}(\mathbf{X})$ described in Proposition 3.3. Of particular interest are the potentially many o.n. bases in $\mathcal{L}(\mathbf{X})$ because each o.n. basis generates a hypercube as its polytope. Therefore:

PROPOSITION 3.4. *The PoSI polytope is the intersection of up to $p!$ congruent hypercubes.*

The simplest case of PoSI polytopes, those for $p = 2$, is illustrated in Figure 2.

3.5. *An Optimality Property of Orthogonal Designs.* In orthogonal designs, adjustment has no effect: $\bar{\boldsymbol{l}}_{j\cdot \mathrm{M}} = \bar{\boldsymbol{l}}_{j\cdot\{j\}} = \mathbf{X}_j/\|\mathbf{X}_j\|$. This fact together with the properties of $\mathbf{\Pi}_K$ from the previous subsection imply $\mathcal{L}(\mathbf{X}) = \{\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_p\}$, and hence $\mathbf{\Pi}_K$ is a hypercube. This latter fact implies an

obvious optimality property of orthogonal designs: In general, $\mathbf{\Pi}_K$ is an intersection of hypercubes of width $2K$, hence if $\mathbf{\Pi}_K$ *is* a hypercube of width $2K$, it is of maximal extent (keeping in mind point symmetry at the origin of the polytopes and radial symmetry about the origin of the distribution of $\mathbf{Z}$). Therefore:

PROPOSITION 3.5. *Among p-dimensional (non-singular) designs, orthogonal designs* $\mathbf{X}$ *yield*

- *the maximal coverage probability* $\mathrm{P}[\mathbf{Z} \in \mathbf{\Pi}_K]$ *for fixed* K*, and*
- *the minimal PoSI constant* $K = K(\mathbf{X}, \alpha, p)$ *satisfying* $\mathrm{P}[\mathbf{Z} \in \mathbf{\Pi}_K] = 1 - \alpha$ *for fixed* α*.*

Optimality of orthogonal designs translates to optimal asymptotic behavior of their constant $K(\mathbf{X})$ for large p:

PROPOSITION 3.6. *The asymptotic lower bound for* $K(\mathbf{X}, p, \alpha)$*, attained for orthogonal designs* $\mathbf{X}$*, is*

$$\inf_{\mathbf{X}} K(\mathbf{X}, p, \alpha) \;=\; \sqrt{2 \log p} + o(p),$$

where $p \to \infty$ *and* $n - p \to \infty$*.*

The above facts show that the PoSI problem is bounded on one side by orthogonal designs: $\inf_{\mathbf{X}} K(\mathbf{X}, \alpha, p, n) = K_{orth}(\alpha, p, n)$, for all α, n and p. On the other side, the Scheffé ball yields a loose upper bound: $\sup_{\mathbf{X}} K(\mathbf{X}, \alpha, p, n) < K_{Sch}(\alpha, p, n)$. The question of how close to the Scheffé bound $\sup_{\mathbf{X}} K(\mathbf{X}, \alpha, p)$ can get will occupy us in Section 4.2. Unlike the infimum problem, the supremum problem does not appear to have a unique optimizing design $\mathbf{X}$ uniformly in α, p and n.

3.6. *A Duality Property of PoSI Vectors.* There exists a duality in terms of PoSI vectors $\mathcal{L}(\mathbf{X})$ which we will use in Section 4.1 below but which is also of independent interest. We require some preliminaries: Letting $F = \{1, 2, ..., p\}$ be the full model, we observe that the transposes of the (unnormalized) PoSI vectors $\boldsymbol{l}_{j \cdot \mathrm{M}_F}$ form the rows of the matrix $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, for the simple reason that $\hat{\boldsymbol{\beta}}_F = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$. In a change of perspective, we interpret the transpose matrix

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

as a design matrix as well, to be called the "dual design" of $\mathbf{X}$. It is also of size $n \times p$ ($p \times p$ 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:

THEOREM 3.1. $\mathcal{L}(\mathbf{X}^*) = \mathcal{L}(\mathbf{X}),\ \ \mathbf{\Pi}_K(\mathbf{X}^*) = \mathbf{\Pi}_K(\mathbf{X}),\ \ K(\mathbf{X}^*) = K(\mathbf{X}).$

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:

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

where as always $\mathrm{M}_F = \{1, 2, ..., p\}$ denotes the full model. 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 3.1. *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 design matrix (in canonical coordinates) is chosen to be symmetric according to Proposition 3.1 (*7.*): if $\mathbf{X}^T = \mathbf{X}$, then $\mathbf{X}^* = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} = \mathbf{X}^{-1}$, and hence:

COROLLARY 3.1. *If* $\mathbf{X}$ *is symmetric in canonical coordinates, then*

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

4. Illustrative Examples and Asymptotic Results.

4.1. *An Example: Exchangeable Designs.* In this section we illustrate the PoSI problem with the example of exchangeable designs $\mathbf{X}$ in which all pairs of predictor vectors enclose the same angle. In canonical coordinates a convenient parametrization of a family of symmetric exchangeable design is

$$\mathbf{X} \ = \ \mathbf{I}_p + a\mathbf{E}_{p\times p}, \tag{4.1}$$

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

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

is also an exchangeable design. The function $c_p(a) = -a/(1+pa)$ maps the interval $(-1/p, \infty)$ onto itself, and it holds $c_p(0) = 0$, $c_p(a) \downarrow -1/p$ as $a \uparrow +\infty$, and vice versa. Exchangeable designs have the nice feature that they are analytically tractable, and at the same time they are sufficiently rich to include orthogonal designs ($a = 0$) as well as to extend to two types of collinearities: the collapse of the predictor vectors to a single dimension span($\mathbf{e}$)for $a \uparrow \infty$ on the one hand, and to a subspace span$(\mathbf{e})^{\perp}$ of dimension $(p-1)$ for $a \downarrow -1/p$ on the other hand ($\mathbf{e} = (1, 1, ..., 1)^T$).

We expect that if non-orthogonality/collinearity drives the fracturing of the regression coefficients into model-dependent quantities $\beta_{j\cdot \mathrm{M}}$ with the ensuing problem of inferential multiplicity, it should be insightful to analyze the PoSI constant $K(\mathbf{X})$ as the design matrix $\mathbf{X} = \mathbf{X}(a)$ moves from orthogonality 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 3.6): 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 3.1.
- A more basic intuition would suggest that $K(\mathbf{X})$ increases as $\mathbf{X}$ moves away from orthogonality and approaches collinearity. Even this intuition is not fully born out: In Figure 3 we depict numerical approximations to $K(\mathbf{X}(a), \alpha = 0.05)$ for $a \in [0, \infty)$ ($a \in (-1/p, 0]$ being redundant due to duality). As the traces show, $K(\mathbf{X}(a))$ increases as $\mathbf{X}(a)$ moves away from orthogonality, up to a point, whereafter it descends as it approaches collinarity, at least for dimensions $p \leq 10$.

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

We next address the asymptotic behavior of $K = K(\mathbf{X}, \alpha, p)$ for increasing p. As noted in Section 2.4, there is a wide gap between orthogonal designs with $K_{orth} \sim \sqrt{2 \log p}$ and the full Scheffé protection with $K_{Sch} \sim \sqrt{p}$. The following theorem shows how exchangeable designs fall into this gap:

THEOREM 4.1. *PoSI constants of exchangeable design matrices* $\mathbf{X}(a)$

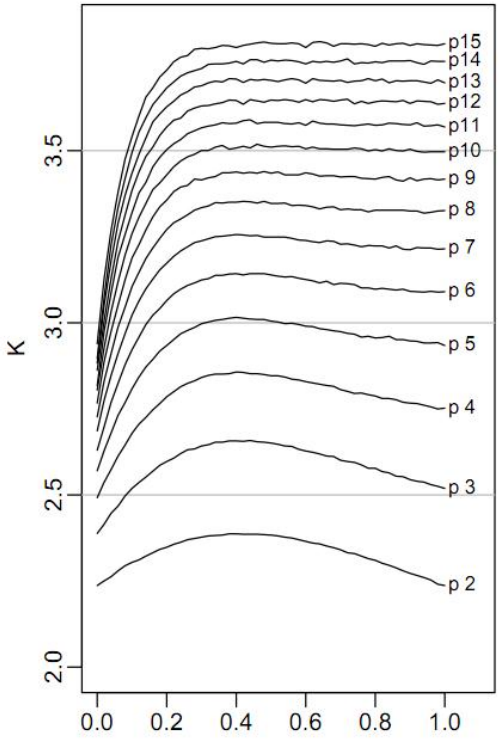

FIG 3. *The PoSI constant* $K(\mathbf{X}, \alpha = 0.05)$ *for exchangeable designs* $\mathbf{X} = \mathbf{I} + a\mathbf{E}$ *for* $a \in [0, \infty)$. *The horizontal axis shows* $a/(1+a)$, *hence the locations 0, 0.5 and 1.0 represent* $a = 0,\ 1,\ \infty$, *respectively. Surprisingly, the largest* $K(\mathbf{X})$ *is not attained at* $a = \infty$, *the point of perfect collinearity, at least not for dimensions up to* $p = 10$. *The graph is based on 10,000 random samples in* p *dimensions for* $p = 2, ..., 15$.

have the following limiting behavior:

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

The proof can be found in Appendix A.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.

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

We finish with a geometric depiction of the limiting polytope $\mathbf{\Pi}_K$ as $\mathbf{X}(a)$ approaches either collinearity: For $a\uparrow\infty$, the predictor vectors fall into the 1-D subspace $\mathrm{span}(\mathbf{e})$, and for $a\downarrow -1/p$ they fall into $\mathrm{span}(\mathbf{e})^\perp$. 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 $\mathrm{span}(\mathbf{e})^\perp$. In Figure 4 we show this prism for $p=3$. The unit vectors $\bar{\boldsymbol{l}}_{1\cdot\{1\}} \sim \mathbf{X}_1$, $\bar{\boldsymbol{l}}_{2\cdot\{2\}} \sim \mathbf{X}_2$ and $\bar{\boldsymbol{l}}_{3\cdot\{3\}} \sim \mathbf{X}_3$ form an equilateral triangle. The plane $\mathrm{span}(\mathbf{e})^\perp$ 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 $\mathbf{e}/\sqrt{p}$, turning the polytope into a prism.

4.2. *An Example where $K(\mathbf{X})$ is close to the Scheffé Bound.* In this section, we describe an example where the asymptotic upper bound for $K(\mathbf{X},\alpha,p)$ is $O(\sqrt{p})$, hence close to the Scheffé constant K_{Sch} in terms of the asymptotic rate. In this example we consider SPAR1 (Section 2.5) whereby a predictor of interest has been chosen, $\mathbf{X}_p$, say. The goal of model selection with SPAR1 is to "boost the effect" of $\mathbf{X}_p$ by adjusting it for optimally chosen predictors $\mathbf{X}_j$ $(j<p)$. The search is over the 2^{p-1} models that contain $\mathbf{X}_p$, but inference is sought only for the adjusted coefficient $\beta_{p\cdot\mathrm{M}}$.

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

$$\mathbf{X} = (\mathbf{e}_1,\ldots,\mathbf{e}_{p-1},\mathbf{1}_p)\,, \tag{4.2}$$

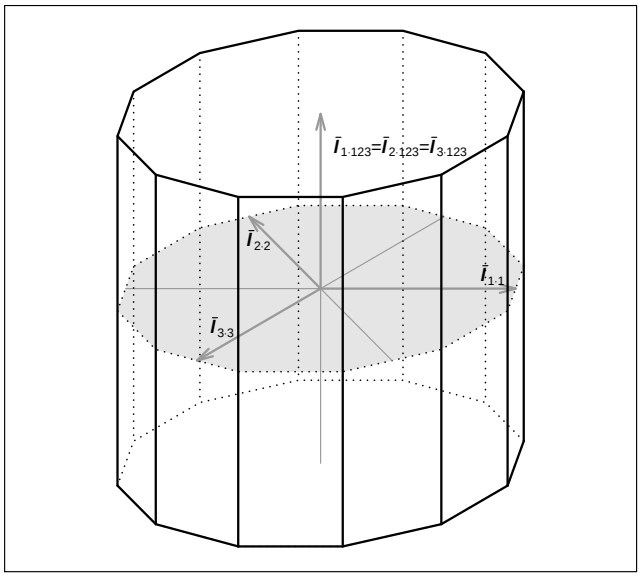

FIG 4. *Exchangeable Designs: The geometry of the limiting PoSI polytope for* $p = 3$ *as* $a \downarrow -1/p$ *or* $a \uparrow +\infty$ *in* (4.1).

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

THEOREM 4.2. *The designs* (4.2) *have PoSI1 simultaneous* $1 - \alpha$ *confidence intervals for* $\mathbf{X}_p$ *of the form* $\left[\hat{\beta}_p \pm K^{(p)}(\mathbf{X})\sqrt{(\mathbf{X}^T\mathbf{X})^{-1}_{pp}}\right]$ *where*

$$\lim_{p\to\infty} \frac{K^{(p)}(\mathbf{X})}{\sqrt{p}} \;=\; 0.6363....$$

A (partial) proof is in Appendix A.4. (We will only show the $\geq$ part). As always, we consider the case of "large $n - p$," that is, σ known; for small $n - p$ the constant is larger. The theorem shows that even if we restrict consideration to a single predictor $\mathbf{X}_p$ and its adjustments, the constant $K^{(p)}$ to reach valid simultaneous inference against all submodels contain-

ing that coefficient can be much greater than the $O(1)$ t-quantiles used in common practice. Also, since for the unrestricted PoSI constant $K(\mathbf{X})$ we have $K(\mathbf{X}) \geq K^{(p)}(\mathbf{X})$, the theorem shows that there exist design matrices for which the PoSI constants are of the asymptotic order of the Scheffé constants.

4.3. *Bounding Away from Scheffé.* We provide a rough asymptotic upper bound on all PoSI constants $K(\mathbf{X}, \alpha, p)$. It is strictly smaller than the Scheffé constant but not by much. The bound, however, is loose because it is based on letting go of the rich structure of the sets $\mathcal{L}(\mathbf{X})$ (Section 3.3) and only using their cardinality $|\mathcal{L}| = p\,2^{p-1}$. We state the bound for more general cardinalities $|\mathcal{L}|$ than required for the PoSI problem:

THEOREM 4.3. *Denote by $\mathcal{L}_p$ arbitrary finite sets of p-dimensional unit vectors, $\mathcal{L}_p \subset S^{p-1}$, such that $|\mathcal{L}_p| \leq a_p$ where $a_p^{1/p} \to a$ (> 0). Denote by $K(\mathcal{L}_p)$ the $(1-\alpha)$-quantile of $\sup_{\bar{l} \in \mathcal{L}_p} |\bar{l}^T \mathbf{Z}|$. Then the following describes an asymptotic worst-case bound and its attainment:*

$$\lim_{p\to\infty} \sup_{|\mathcal{L}_p| \leq a_p} \frac{K(\mathcal{L}_p)}{\sqrt{p}} \;=\; \left(1 - \frac{1}{a^2}\right)^{1/2}.$$

The proof of Theorem 4.3, to be found in Appendix A.5, is an adaptation of Wyner's (1967) techniques for sphere packing and sphere covering. The worst-case bound $(\leq)$ is based on a surprisingly crude Bonferroni-style inequality for caps on spheres. Attainment of the bound $(\geq)$ makes use of the artifice of picking the vectors $\bar{l} \in \mathcal{L}$ randomly and independently. — Applying the theorem to PoSI sets $\mathcal{L} = \mathcal{L}(\mathbf{X})$, we have $|\mathcal{L}| = p\,2^{p-1} = a_p$, hence $a_p^{1/p} \to 1/2$, and therefore the theorem applies with $a = 1/2$:

COROLLARY 4.1. *A universal asymptotic upper bound on PoSI constants is given by*

$$\lim_{p\to\infty} \sup_{\mathbf{X}} \frac{K(\mathbf{X})}{\sqrt{p}} \;\leq\; \frac{\sqrt{3}}{2} = 0.866\ldots.$$

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

5. Computations. The results in this section are for both finite n and finite p. The computations of the design-specific constant $K = K(\mathbf{X}, \alpha, p, n)$ are MC-based, while those of the universal upper bound $K = K_{univ}(\alpha, p, n)$ derive from an analytical formula for a lower bound on the coverage probability inspired by Theorem 4.3.

5.1. *Computation of PoSI Constants* $K = K(\mathbf{X}, \alpha, p, n)$. We derive a simple algorithm for computing $K(\mathbf{X}, \alpha, p, n)$ for up to $p = 20$ predictors. For finite n we revert from z- to t-statistics with the assumption that all $t_{j\cdot\mathrm{M}}$ share the estimate s_F of σ from the full model in the denominator. For calibration of coverage probabilities we rely on the pivotality of t-statistics in $\boldsymbol{\beta}$ and σ, which allows us to perform simulations conveniently for $\boldsymbol{\beta} = \mathbf{0}$ and $\sigma = 1$. We also work in canonical coordinates (Section 3.1) which are fully constructive and reduce design vectors $\mathbf{X}_j$ and their adjustments $\bar{\boldsymbol{l}}_{j\cdot\mathrm{M}}$ from n to p dimensions. Thus we write the t-statistics as $t_{j\cdot\mathrm{M}} = \bar{\boldsymbol{l}}_{j\cdot\mathrm{M}}\mathbf{Z}/s_F$ where $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}_p, \mathbf{I}_p)$, $\bar{\boldsymbol{l}}_{j\cdot\mathrm{M}} \in S^{p-1}$, and $(n-p)s_F^2 \sim \chi^2_{n-p}$ is independent of $\mathbf{Z}$. The criterion is the maximal magnitude of t's, which, using the usual notation $\mathcal{L} = \{\bar{\boldsymbol{l}}_{j\cdot\mathrm{M}} : j \in \mathrm{M}\}$ from (3.4), we write as

$$\max_{j,\mathrm{M}:j\in\mathrm{M}} |t_{j\cdot\mathrm{M}}| \;=\; \max_{\bar{\boldsymbol{l}}\in\mathcal{L}} |\bar{\boldsymbol{l}}^T\mathbf{Z}|/s_F\,.$$

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

$$\mathrm{P}[\max_{\bar{\boldsymbol{l}}\in\mathcal{L}} |\bar{\boldsymbol{l}}^T\mathbf{Z}|/s_F \le K] \;\approx\; \frac{1}{I}|\{i : \max_{\bar{\boldsymbol{l}}\in\mathcal{L}} |\bar{\boldsymbol{l}}^T\mathbf{Z}^{(i)}|/s_F^{(i)} \le K\}| \;=\; 1-\alpha.$$

However, it is possible to increase precision by removing some of the variance from this MC simulation by integrating out the radial component of $\mathbf{Z}/s_F$. As the distribution of $\mathbf{Z}/s_F$ is spherically symmetric about $\mathbf{0}_p$, it can be decomposed into a radial and an angular component (see also the proof of Theorem 4.3, Appendix A.5):

$$\mathbf{Z}/s_F = R\mathbf{U}, \quad R^2/p \sim F_{p,n-p}\,, \quad \mathbf{U} \sim \mathrm{Unif}(S^{p-1}), \quad R, \mathbf{U} \text{ independent}.$$

The maximal t-statistic becomes

$$\max_{j,\mathrm{M}:j\in\mathrm{M}} |t_{j\cdot\mathrm{M}}| \;=\; \max_{\bar{\boldsymbol{l}}\in\mathcal{L}} |\bar{\boldsymbol{l}}^T(R\mathbf{U})| \;=\; R\max_{\bar{\boldsymbol{l}}\in\mathcal{L}} |\bar{\boldsymbol{l}}^T\mathbf{U}| \;=\; R\,C,$$

where $C = \max_{\bar{\boldsymbol{l}}\in\mathcal{L}} |\bar{\boldsymbol{l}}^T\mathbf{U}|$ is the random variable that measures the maximal magnitude of the cosine between $\mathbf{U}$ and the vectors in the set $\mathcal{L}$. We integrate

the radial component:

$$\begin{aligned} \mathrm{P}[\max_{\mathrm{M},j:j\in\mathrm{M}} |t_{j\cdot\mathrm{M}}| \le K] &= \mathrm{P}[RC \le K] \\ &= \mathrm{E}[\mathrm{P}[\, R^2/p \le (K/C)^2/p \,|\, C\,]\,] \\ &= \mathrm{E}[F_{p,n-p}((K/C)^2/p)]\,, \end{aligned}$$

where $F_{p,n-p}(...)$ denotes the c.d.f. of the F-distribution (not its quantiles).

PROPOSITION 5.1. *Let* $\mathbf{U} \sim \mathrm{Unif}(S^{p-1})$ *and* $C = \max_{\bar{\boldsymbol{l}}\in\mathcal{L}} |\bar{\boldsymbol{l}}^T\mathbf{U}|$*; then the PoSI coverage probability for* K *is*

$$\mathrm{P}[\max_{\mathrm{M},j:j\in\mathrm{M}} |t_{j\cdot\mathrm{M}}| \le K] \;=\; \mathrm{E}[F_{p,n-p}((K/C)^2/p)] \tag{5.1}$$

The expectation on the right hand side refers to the random variable C. The connection of this formula to Scheffé protection is as follows: The Scheffé case arises for $\mathcal{L} = S^{p-1}$ and hence $C \equiv 1$, in which case calibration of the right hand side of (5.1) requires $F_{p,n-p}(K^2/p) = 1-\alpha$, which reproduces the Scheffé constant $K_{Sch} = \sqrt{pF^{-1}_{p,n-p}(1-\alpha)}$. To gain on Scheffé, one needs $C < 1$ in distribution, which is the case for any $|\mathcal{L}| < \infty$.

The PoSI constant $K = K(\mathbf{X}, \alpha, p, n)$ can be approximated by calibrating an MC estimate of (5.1). To this end, we sample i.i.d. unit vectors $\mathbf{U}^{(1)}, ..., \mathbf{U}^{(I)} \sim \mathrm{Unif}(S^{p-1})$, calculate their maximal absolute cosines $C^{(i)} = \max_{\bar{\boldsymbol{l}}\in\mathcal{L}} |\bar{\boldsymbol{l}}^T\mathbf{U}^{(i)}|$ and calibrate:

$$\mathrm{P}[\max_{\mathrm{M},j:j\in\mathrm{M}} |t_{j\cdot\mathrm{M}}| \le K] \;\approx\; \frac{1}{I}\sum_{i=1,...,I} F_{p,n-p}((K/C^{(i)})^2/p) \;=\; 1-\alpha\,. \tag{5.2}$$

The real computational expense is in calculating the values $C^{(i)}$, which involves maximizing the magnitude of the inner product of $\mathbf{U}^{(i)}$ with the $p\,2^{p-1}$ vectors $\bar{\boldsymbol{l}} \in \mathcal{L}$. Currently we completely enumerate the set $\mathcal{L}$, and this is why the computations are limited to about $p \le 20$. For $p = 20$ ($p\,2^{p-1} = 10,485,760$), elapsed time on 2010 computing equipment is about one hour. Once the values $C^{(i)}$ are calculated (for $I = 1,000$, say), they are re-used in (5.2) for all values of K that are tried in the bisection search, which therefore takes up negligible time. — An implementation of this algorithm in *R*-code (`http://www.r-project.org/`) can be obtained from the authors' web pages.

5.2. *Computation of Universal Upper Bounds $K_{univ}(\alpha, p, n)$.* The same technique that produced the universal asymptotic upper bound in Section 4.3 (see the proof in Appendix A.5) can be used to compute a universal finite-n/finite-p upper bound: $K_{univ}(\alpha, p, n) \geq K(\mathbf{X}, \alpha, p, n)$ $(\forall \mathbf{X})$, yet strictly less than the Scheffé bound, $K_{univ}(\alpha, p, n) < K_{Sch}(\alpha, p, n)$. This exercise has two purposes: (a) providing a computational method, and (b) giving the message that the asymptotic bound 0.866 of Theorem 4.3 should be taken with a grain of salt; for $p \to \infty$ it is approached from above, hence it is not conservative for any finite p in providing a bound for unstructured $\mathcal{L}$ of size $|\mathcal{L}| = p\, 2^{p-1}$.

As in Section 4.3 we ignore all structure in $\mathcal{L}$ except for its cardinality, hence the technique really works for arbitrary coverage problems involving many linear estimable functions. We start from the expression (5.1) for the coverage probability and observe that for any random variable C' that dominates C stochastically, $C' \overset{\mathcal{D}}{\geq} C$, we have

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

because $c \mapsto F_{p,n-p}((K/c)^2/p)$ is monotone decreasing for $c > 0$. Thus calibrating the left hand side results in a more conservative constant $K' \geq K$. We can create C' following the lead of Theorem 4.3 by using its proof's Bonferroni-style bound (A.15) and relying on the fact that every $(\vec{l}^T\mathbf{U})^2$ has a Beta$(1/2, (p-1)/2)$-distribution (A.16):

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

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

PROPOSITION 5.2. *Let C' be a random variable with the following c.d.f.:*

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

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

$$\mathrm{E}[F_{p,n-p}((K'/C')^2/p)] \;\leq\; \mathrm{P}[\max_{j,\mathrm{M}:j\in\mathrm{M}} |t_{j\cdot\mathrm{M}}| \leq K'] . \tag{5.3}$$

Thus, if the l.h.s. in (5.3) is calibrated with a search over K' so that

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

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

$$K(\mathbf{X}, \alpha, p, n) \;\le\; K_{univ}(\alpha, p, n) \;<\; K_{Sch}(\alpha, p, n).$$

For the last inequality, note that the Scheffé constant is for $C \equiv 1$, whereas $\mathrm{P}[C' < 1] = 1$.

Good approximations of the l.h.s. in (5.3) for arbitrary K' are obtained by calculating a grid of equi-probability quantiles for the distribution of C' once for all, and re-use it in the bisection search for K'. For a grid of length I, the grid points can be obtained by solving $F_{C'}(c_i) = i/(I+1)$ or, equivalently but more conveniently, $1 - F_{C'}(c_i) = i/(I+1)$:

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

Numerically this works for up to about $p = 40$; for larger p the "Bonferroni" denominator $|\mathcal{L}| = p\,2^{p-1}$ creates quantiles that are too extreme for conventional numerical routines of Beta quantiles. If it works, the approximation to the l.h.s. of (5.3) is

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

which can be used for calibration.

A comparison of the distribution of the variable C' with the values $C \equiv 1$ (Scheffé) and $C = \sqrt{3}/2 = 0.866...$ (asymptotic bound) is of interest because it shows (a) to what degree simultaneous inference problems involving $|\mathcal{L}|$ estimable functions are necessarily less stringent than Scheffé, and (b) to what degree the asymptotic bound is approximated by C'. Such comparisons are given in Figure 5: all distributions are strictly below 1, beating Scheffé as they obviously should, but the asymptotic bound 0.866 is approached from above, which means that this value should be enlarged somewhat. In view of Figure 5, a good rough and ready rule for practice might be using a fraction 0.9 of the Scheffé constant as an approximate universal PoSI constant.

5.3. *P-Value Adjustment for Simultaneity.* Statistical inference for regression coefficients is probably more often carried out in terms of p-values than confidence intervals. There exists an immediate translation between the two modes of inference: Informally, the two-sided p-value is the complement of the coverage probability of a confidence interval that exactly touches the null hypothesis $\beta_{j\cdot\mathrm{M}} = 0$. This statement translates to

$$\mathrm{pval}_{j\cdot\mathrm{M}} = 1 - F_{j\cdot\mathrm{M}}(|t^{obs}_{j\cdot\mathrm{M}}|) \quad \text{where} \quad F_{j\cdot\mathrm{M}}(t) = \mathrm{P}[|t_{j\cdot\mathrm{M}}| < t]$$

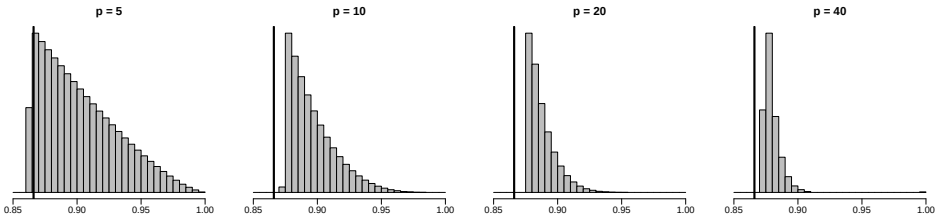

FIG 5. *Distribution of the variable C' of Section 5.2 used for calculating the universal upper bound on the PoSI constants. The vertical bold line on the left shows the asymptotic bound $\sqrt{3}/2$. The distributions approach the asymptotic value from above, thereby showing that universal upper bounds for finite p tend to be somewhat larger than the asymptotic limit, yet less than 1 and hence less than the Scheffé constant.*

is the c.d.f. of the $|t_{n-p}|$ distribution and $t^{obs}_{j\cdot \mathrm{M}}$ is the observed value of the t-statistic for $\beta_{j\cdot \mathrm{M}}$. This is the familiar definition of the *marginal p-value*: the probability of a test statistic more extreme than the observed one. The marginal p-value suffers from ignoring multiplicity/simultaneity of performing as many as $|\mathcal{L}|$ tests.

The globally simultaneous or *PoSI p-value* is defined as

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

where

$$F_{\max}(t) = \mathrm{P}[\max_{j,\mathrm{M}:\, j\in \mathrm{M}} |t_{j\cdot \mathrm{M}}| < t] \tag{5.4}$$

is the c.d.f. of the max test statistic under the global null hypothesis. The calculation of this c.d.f. is the subject of Section 5.1, in particular Equation 5.1 in Proposition 5.1. The role of the global PoSI p-value is that of a competitor of the p-value of the overall F-test, which in turn derives from Scheffé protection. Thus the PoSI p-value may provide sharper overall inference than the overall F-test, even though it was designed to adjust for the multiplicity of arbitrary model selection.

To correct the deficiencies of the marginal p-values $\mathrm{pval}_{j\cdot \mathrm{M}}$, it is useful to introduce adjusted versions $\mathrm{pval}^{PoSI}_{j\cdot \mathrm{M}}$ that do account for multiplicity. The informal definition of an adjusted p-value starts again with a confidence interval whose width is such that it exactly touches the null hypothesis; the difference to the marginal p-value is in the assigned coverage, which can be conservatively calculated as the simultaneous coverage of all intervals of this width across all M and all $j \in \mathrm{M}$. As an equivalent informal characterization,

we can define the adjusted p-value to be the probability of observing a t-statistic greater in magnitude than the observed t statistic *anywhere* among all submodels M and all coefficients $j \in \mathrm{M}$:

$$\mathrm{pval}_{j\cdot\mathrm{M}}^{PoSI} = 1 - F_{\max}(|t_{j\cdot\mathrm{M}}^{obs}|), \tag{5.5}$$

where $F_{\max}(t)$ is as in (5.4) above. This "one-step" adjustment is practicable once the basic computations outlined in Section 5.1 are completed. The adjustment is an over-adjustment for all but the maximal $|t_{j\cdot\mathrm{M}}|$. A sharper form of adjustment could be achieved with a "step-down" method such as the algorithm described by Pacifico et al. (2004, Section 2), but the computational expense may be prohibitive and the gain in statistical efficiency may be minimal. The adjustment (5.5) provides strong family-wise error control and has the simultaneity guarantee given in Pacifico et al. (2004): The set $\{\,(j,\mathrm{M}) \,|\, \mathrm{pval}_{j\cdot\mathrm{M}}^{PoSI} > \alpha\,\}$ characterizes a superset of the true null hypotheses $\beta_{j\cdot\mathrm{M}} = 0$ with probability $\geq 1-\alpha$. Conversely: $\{\,(j,\mathrm{M}) \,|\, \mathrm{pval}_{j\cdot\mathrm{M}}^{PoSI} \leq \alpha\,\}$ characterizes a subset of the true alternatives $\beta_{j\cdot\mathrm{M}} \neq 0$ with probability $\geq 1-\alpha$.

(Final note: "Adjustment" of p-values for multiplicity and "adjustment" of predictors for other predictors are two concepts that share nothing except the partial homonym.)

6. Summary and Discussion. We investigated the Post-Selection Inference or "PoSI" problem in the classical Gaussian linear model under the classical assumption $n \gg p$. We showed that, while finding the distribution of post-selection estimates is hard, valid post-selection inference is possible via simultaneous inference. The principal idea is to consider the regression coefficient of a given predictor as distinct when it occurs in different submodels: $\beta_{j\cdot\mathrm{M}}$ and $\beta_{j\cdot\mathrm{M}'}$ are different parameters if $\mathrm{M} \neq \mathrm{M}'$. We show that simultaneity protection for all parameters $\beta_{j\cdot\mathrm{M}}$ provides valid post-selection inference. In practice this means enlarging the constant $t_{1-\alpha/2,n-p}$ used in conventional inference to a constant $K(\mathbf{X},\alpha,p,n)$ that provides simultaneity protection for up to $p\,2^{p-1}$ parameters $\beta_{j\cdot\mathrm{M}}$. We showed (for known σ or "$n=\infty$") that the constant depends strongly on the design matrix $\mathbf{X}$ as the asymptotic bound for $K(\mathbf{X},\alpha,p)$ ranges between the minimum of $\sqrt{2\log p}$ achieved for orthogonal designs on the one hand, and a large fraction of the Scheffé bound $\sqrt{p}$ on the other hand. This wide asymptotic range of constants $K(\mathbf{X},\alpha,p)$ suggests that computation is critical for large p. Our current computational methods are limited to $p \leq 20$.

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

- Among open problems is the quest for the largest fraction of the asymptotic Scheffé rate $\sqrt{p}$ attained by PoSI constants. So far we know this fraction to be at least 0.6363 but no more than 0.8660...
- Computations for $p > 20$ are a challenge. Straight enumeration of the set of up to $p\,2^{p-1}$ linear contrasts should be replaced with heuristic shortcuts that yield practically useful upper bounds on $K(\mathbf{X}, \alpha, p, n)$ that are specific to $\mathbf{X}$, unlike the 0.8660 fraction of the Scheffé bound which is universal.
- The methodology is easily adapted to practically useful variations such as these: (1) Data analysts might be interested only in small submodels, $|\mathrm{M}| \le 5$, say, when p is large. (2) We introduced SPAR ("Single Predictor Adjusted Regression", Section 2.5) defined as the search among all predictors for strong adjusted "effects". Practitioners might be more interested in strong adjusted effects in one predetermined predictor of special interest, as in SPAR1. — Any limitation to a lesser number of submodels or regression coefficients to be searched increases the computationally accessible size of p.
- Among models to which the PoSI framework should be extended next are generalized linear models and mixed effects models.
- Alternative PoSI frameworks with quite different interpretations could possibly be based on bootstrap resampling.

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

Acknowledgments. We thank M. Freiman, E. George, E. Pitkin, L. Shepp, N. Sloane and M. Traskin for very helpful discussions.

APPENDIX A: PROOFS

A.1. Proof of Proposition 3.3.

1. The matrix $\mathbf{X}^*_M = \mathbf{X}_M(\mathbf{X}_M^T\mathbf{X}_M)^{-1}$ 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 $j' \ne j$ follows from $\mathbf{X}_M^T\mathbf{X}^*_M = \mathbf{I}_p$. 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 3.1 and Theorem3.1. The proof relies on a careful analysis of orthogonalities as described in Proposition 3.3, 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 3.1. 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 (3.3), 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 (3.3) 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 (3.3), we obtain

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

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

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

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

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

This proves the lemma and the theorem. □

A.3. Proof of Theorem 4.1. The parameter a can range from $-1/p$ to ∞, but because of duality there is no loss of generality in considering only the case in which $a \geq 0$, and we do so in the following. Let $\mathrm{M} \subset \{1,\ldots,p\}$ and $j \in \mathrm{M}$. If $\mathrm{M} = \{j\}$ then $\boldsymbol{l}_{j\cdot\mathrm{M}} = \mathbf{X}_j$, the j-th column of $\mathbf{X}$, and $\bar{\boldsymbol{l}}_{j\cdot\mathrm{M}} = \boldsymbol{l}_{j\cdot\mathrm{M}}/\sqrt{pa^2+2a+1}$. It follows that for $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}_p, \mathbf{I}_p)$,

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

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

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

When $m \geq 2$ the projection of $\mathbf{X}_1$ on the space spanned by $\mathbf{X}_2,\ldots,\mathbf{X}_m$ 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; no calculation of projection matrices is needed to verify this. Let $d = 1-c$. Then

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

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

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

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

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

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

This verifies that

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

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

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

Note that

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

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

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

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

A.4. Partial Proof of Theorem 4.2. While the theorem is correct as stated, of interest is only the inequality

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

the point being that a non-zero fraction of the Scheffé rate $\sqrt{p}$ can be attained by PoSI1 constants. As the proof of the reverse inequality is lengthy,

we provide here only the more straightforward proof of inequality (A.10) and indicate below the missing part which can be obtained from the authors on request. The following preparations are required for both inequalities.

To find $\bar{\boldsymbol{l}}_{p\cdot\mathrm{M}}$, we need to adjust the predictor of interest $\mathbf{X}_p = \mathbf{1}_p$ for other predictors $\mathbf{X}_j = \mathbf{e}_j$ $(j < p)$ in the model M. In this case adjusting means zeroing out the components of $\mathbf{X}_p$ for $j \in \mathrm{M}$ with the exception of $j = p$, hence the z-statistic (3.1) for the predictor of interest are

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

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

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

where

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

Next we observe that for a fixed model size $|\mathrm{M}| = \sum_{j\in\mathrm{M}} 1 = m$ (> 1) and a given $\mathbf{Z}$ the maximizing model has to include the predictors j for which Z_j is smallest, hence is of the form $\mathrm{M}_B = \{\, j :\, Z_j < B \,\} \cup \{p\}$, where B is chosen such that $|\mathrm{M}| = m$. It is therefore sufficient to consider only models of the form M_B. Furthermore, we can limit the search to $B \geq 0$ because adding j with $Z_j < 0$ to the model increases the numerator of the above ratio and makes it positive, and also it decreases the numerator, thereby increasing the ratio again:

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

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

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

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

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

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

The pointwise assertion of the lemma follows. □

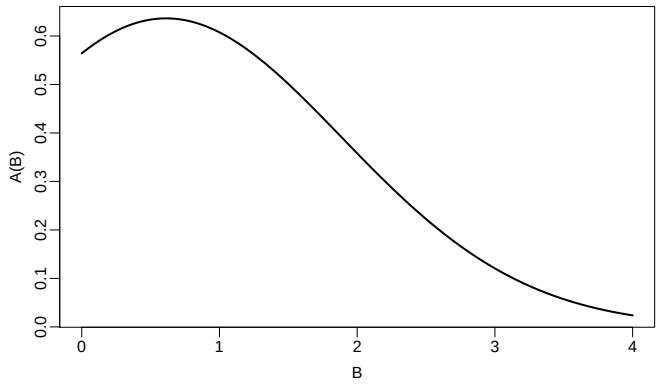

FIG 6. *The function* $A(B) = \frac{\phi(B)}{\sqrt{1-\Phi(B)}}$ *from the proof of Theorem 4.2 in Appendix A.4.*

Continuing with the proof of the theorem, we look for $B \geq 0$ that makes $z_{p\cdot \mathrm{M}_B}$ asymptotically the largest. Following the lead of the lemma we obtain the maximum of $A(B)$ as $A_{\max} = \max_{B\geq 0} \phi(B)/\sqrt{1-\Phi(B)} = 0.6363...$, attained at $B_{max} \approx 0.6121$. (The graph of $A(B)$ is shown in Figure 6.) We have therefore $z_{p\cdot \mathrm{M}_{B_{max}}} = (0.6363... + o_P(1))\sqrt{p}$. Since $\max_B z_{p\cdot \mathrm{M}_B} \geq z_{p\cdot \mathrm{M}_{B_{max}}}$, the PoSI1 constant is lower-bounded by $K^{(p)} \geq 0.6363...\sqrt{p}\,(1 + o_p(1))$. □

A.5. Proof of Theorem 4.3. We will show that if $a_p^{1/p} \to a\ (> 0)$, we have

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

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 4.3. We decompose $\mathbf{Z} = R\mathbf{U}$ where $R^2 = \|\mathbf{Z}\|^2 \sim \chi^2_p$ and $\mathbf{U} = \mathbf{Z}/\|\mathbf{Z}\| \sim \mathrm{Unif}(S^{p-1})$ are independent. Due to $R/\sqrt{p} \xrightarrow{\mathrm{P}} 1$ it is sufficient to show the following:

- uniform asymptotic worst-case bound:

$$\lim_{p\to\infty} \sup_{|\mathcal{L}_p|\le a_p} \max_{\bar{l}\in\mathcal{L}_p} |\bar{l}^T\mathbf{U}| \overset{\mathrm{P}}{\le} \sqrt{1-1/a^2}\,; \tag{A.11}$$

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

$$\lim_{p\to\infty} \max_{\bar{l}\in\mathcal{L}_p} |\bar{l}^T\mathbf{U}| \overset{\mathrm{P}}{\ge} \sqrt{1-1/a^2}\,. \tag{A.12}$$

To show (A.11), we upper-bound the non-coverage probability and show that it converges to zero for $K' > \sqrt{1-1/a^2}$. To this end we start with a Bonferroni-style bound, as in Wyner (1967):

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

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

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

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

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

We bound the Beta function and the integral separately:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

REFERENCES

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

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

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

[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] 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.

[6] 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.

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

[8] 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.

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

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

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

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

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

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

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

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

[17] 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.

[18] 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