Submitted to the Annals of Statistics

VALID POST-SELECTION INFERENCE ONLINE APPENDIX

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

The Wharton School, University of Pennsylvania

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

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

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

Inferential problems with the full model view of variable selection are pointed out by the "Vienna School" in the series of articles referenced in

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

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

Section 1. An insightful illustration is given by Leeb and Pötscher (2005) with a two-predictor situation where one predictor is protected from selection and only the covariate is subject to selection. They explicitly describe the sampling distribution of the coefficient estimate of the protected predictor, as the covariate is randomly selected/deselected according to tests that perform consistent or conservative model selection, respectively (ibid., p. 29). Their analysis shows (ibid., Figure 2) that the sampling distribution (1) depends critically on the unknown true coefficient of the covariate and the sample size, and (2) deviates from the fixed-model sampling distribution with features such as bi-modality or inflated variance. Because the true covariate slope is not known, it cannot be known either whether the sample size puts the sampling distribution in this realm of deviation from classical theory. — Generalizing to arbitrary linear models Pötscher, Leeb and Schneider prove that sampling distributions cannot be estimated after model selection and thresholding, not even asymptotically. They show that asymptotic normality is prone to non-uniformity of convergence especially near submodels, or should be considered as converging at a speed slower than root-N to a non-normal distribution. They prove that these effects are more pronounced under consistent than conservative model selection. — In the simplified context of what may be called "marginally thresholded" estimators, Pötscher and Schneider (2010) produce conservative confidence intervals, and they show that these intervals are wider than conventional ones as their widths need to account for the bias caused by thresholding.

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

B.2 "Omitted Variables Bias". By allowing each $\hat{\beta}_{j\cdot \mathrm{M}}$ to estimate its own submodel target $\beta_{j\cdot \mathrm{M}}$ rather than the full model parameter β_j, we sidestep the problem of "omitted variables bias" and with it a major driver of the problems analyzed by Leeb and Pötscher (Section B.1). In the present framework $\beta_j - \beta_{j\cdot \mathrm{M}}$ is not a bias as these are two different parameters that

answer two different questions. Just the same, we may consider the difference between β_j and $\beta_{j\cdot\mathrm{M}}$ in the classical case $d=p\le n$. Compare the following two definitions:

$$\boldsymbol{\beta}_{\mathrm{M}} \triangleq \mathbf{E}[\hat{\boldsymbol{\beta}}_{\mathrm{M}}] \quad \text{and} \quad \boldsymbol{\beta}^{\mathrm{M}} \triangleq (\beta_j)_{j\in\mathrm{M}}, \tag{0.1}$$

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

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

$$\mathbf{X}_{\mathrm{M}}^T \mathbf{X}_{\mathrm{M}^c} \boldsymbol{\beta}^{\mathrm{M}^c} = \mathbf{0}, \tag{0.2}$$

where M^c is the full model complement of M. Special cases of (0.2) include: (1) the column space of $\mathbf{X}_{\mathrm{M}}$ is orthogonal to that of $\mathbf{X}_{\mathrm{M}^c}$, and (2) $\boldsymbol{\beta}^{\mathrm{M}^c} = \mathbf{0}$, that is, the approximation to $\boldsymbol{\mu}$ in M_F is no better than in M.

B.3 Proof of Corollary 4.2. We start with the statement of strong family-wise error control by defining the true null hypotheses and true alternatives for the true $\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\mu}$, as well as the sets of insignificant and significant tests for the observed $\mathbf{Y}$:

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

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

LEMMA 0.1. *"Strong Family-Wise Error Control" holds for $K(\mathbf{X},\alpha)$:*

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

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

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

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

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

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

Proof: Define $\hat{\mathrm{M}}' \triangleq \{(j, \hat{\mathrm{M}}) \,|\, j \in \hat{\mathrm{M}}\}$. The event $H_1 \supset \hat{H}_1$ implies the event $H_1 \cap \hat{\mathrm{M}}' \supset \hat{H}_1 \cap \hat{\mathrm{M}}'$, hence, using Lemma 0.1:

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

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

Theorem 0.1. *For any selection procedure* $\hat{\mathrm{M}} : \mathbb{R}^n \to \mathcal{M}$ *it holds:*

$$\mathbf{P}\left[\, j \in \hat{\mathrm{M}} \;\&\; |t_{j\cdot\hat{\mathrm{M}}}| \leq K_{j\cdot}\right] \;\geq\; \mathbf{P}\left[\, j \in \hat{\mathrm{M}}\right] - \alpha, \tag{0.3}$$

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

$$\mathbf{P}\left[\, j \in \hat{\mathrm{M}} \;\&\; \beta_{j\cdot\hat{\mathrm{M}}} \in \mathrm{CI}_{j\cdot\hat{\mathrm{M}}}(K_{j\cdot})\right] \;\geq\; \mathbf{P}\left[\, j \in \hat{\mathrm{M}}\right] - \alpha. \tag{0.4}$$

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

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

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

COROLLARY 0.1. *For any selection procedure* $\hat{\mathrm{M}} : \mathbb{R}^n \to \mathcal{M}$ *we have:*

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

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

B.5 PoSI P-Value Adjustment for Model Selection. Statistical inference for regression coefficients is more often carried out in terms of p-values than confidence intervals. The usual p-values are for null hypotheses $\beta_{j\cdot\mathrm{M}} = 0$, hence the test statistics are

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

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

$$F_{j\cdot\mathrm{M}}(t) = \mathbf{P}[\, |t^{(0)}_{j\cdot\mathrm{M}}| < t\,], \qquad F_{\max}(t) \;=\; \mathbf{P}[t^{(0)}_{\max} < t]. \tag{0.5}$$

The former measures marginal null coverage of a two-sided retention interval $[-t, +t]$, while the latter measures simultaneous coverage of a retention cube

$[-t,+t]^k$ where $k = |\{(j,\mathrm{M}) \,|\, j \in \mathrm{M} \in \mathcal{M}\}|$ is the number of tests performed, which can be as many as $p\,2^{p-1}$ in the classical case $d = p \le n$ for $\mathcal{M} = \mathcal{M}_{\text{all}}$. Denoting by $t^{obs}_{j\cdot\mathrm{M}}$ and $t^{obs}_{\max}$ the observed values of $t^{(0)}_{j\cdot\mathrm{M}}$ and $t^{(0)}_{\max}$, respectively, the following p-values can be defined:

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

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

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

Comments:

(1) The marginal p-value ignores the fact that k tests are being performed.
(2) The global adjusted p-value establishes whether at least the strongest "effect" is statistically significant, and it is therefore an overall test similar to, but more specific than, the overall F-test. Because the latter is derived from Scheffé protection, the global adjusted PoSI p-value is more powerful and still protects against any model selection in the model universe $\mathcal{M}$.
(3) The individual adjusted p-value adjusts each $|t_{j\cdot\mathrm{M}}|$ as if it were a max statistic, hence results in an over-adjustment for all but $t_{\max}$. A sharper method than this "one-step adjustment" would be a simulation-based "step-down" method. We have not examined this route though we suspect that the computational expense may be prohibitive and the gain in statistical efficiency may be small.

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

B.6 The PoSI Process. An alternative way of looking at the PoSI problem is in terms of a stochastic process indexed by (j,M) for $j \in \mathrm{M}$. We mention this view because it is the basis of some software implementations used to solve simultaneous inference and coverage problems, even though in this case it does not result in a practicable approach.

In the PoSI problem the obvious process is $\mathbf{W} = (t_{j\cdot\mathrm{M}})_{j\in\mathrm{M}\in\mathcal{M}}$, which is a t-process for finite degrees of freedom r in $\hat{\sigma}$ and a Gaussian process in the limit $r \to \infty$.

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

$$\mathbf{\Sigma} \;=\; (\Sigma_{j\cdot\mathrm{M};\,j'\cdot\mathrm{M}'}), \qquad \Sigma_{j\cdot\mathrm{M};\,j'\cdot\mathrm{M}'} \;\triangleq\; \bar{\boldsymbol{l}}^{T}_{j\cdot\mathrm{M}}\bar{\boldsymbol{l}}_{j'\cdot\mathrm{M}'}. \tag{0.6}$$

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

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

B.7 Figures.

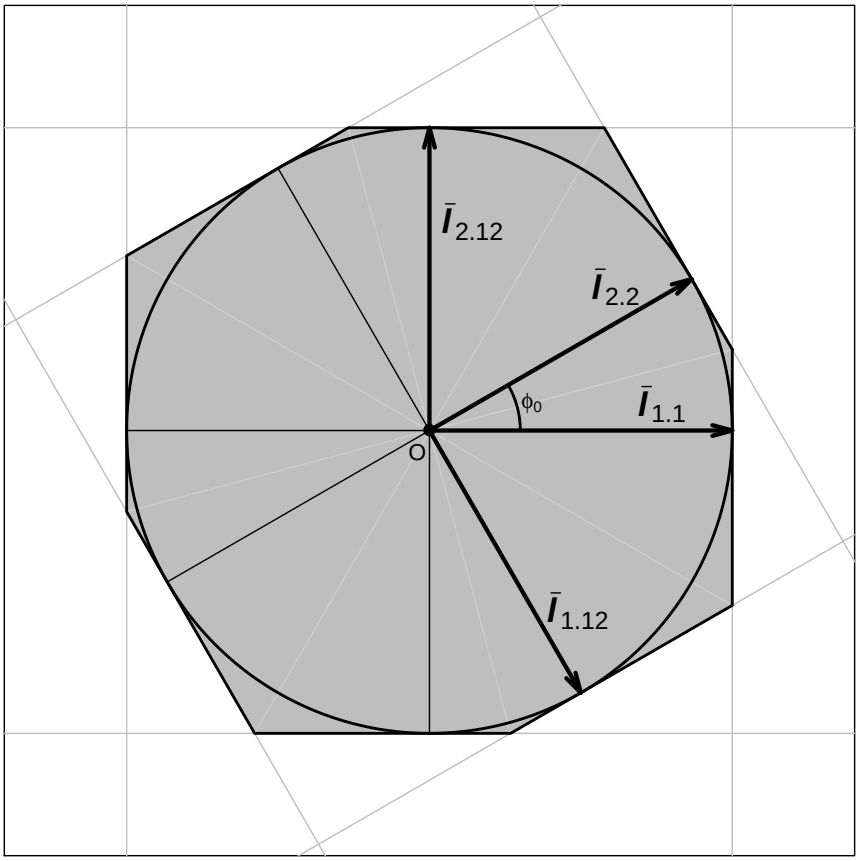

Fig 1: *The PoSI polytope* $\mathbf{\Pi}_{K=1}$ *(Section 5.4) is tangent to the Scheffé disk (2-D ball)* $\mathbf{B}_{K=1}$ *for* $d = p = 2$*. The normalized raw predictor vectors are* $\bar{\boldsymbol{l}}_{1\cdot\{1\}} \sim \mathbf{X}_1$ *and* $\bar{\boldsymbol{l}}_{2\cdot\{2\}} \sim \mathbf{X}_2$*, and the normalized adjusted versions are* $\bar{\boldsymbol{l}}_{1\cdot\{1,2\}}$ *and* $\bar{\boldsymbol{l}}_{2\cdot\{1,2\}}$*. Shown in gray outline are the two squares (2-D cubes) generated by the o.n. bases* $(\bar{\boldsymbol{l}}_{1\cdot\{1\}}, \bar{\boldsymbol{l}}_{2\cdot\{1,2\}})$ *and* $(\bar{\boldsymbol{l}}_{2\cdot\{2\}}, \bar{\boldsymbol{l}}_{1\cdot\{1,2\}})$*, respectively. The PoSI polytope is the intersection of the two squares.*

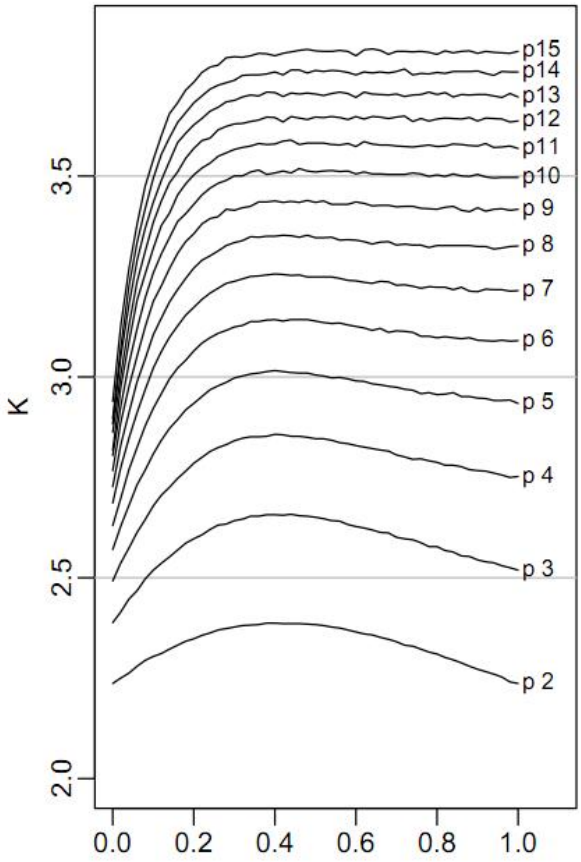

Fig 2: *The PoSI constant* $K(\mathbf{X}^{(p)}(a), \alpha = 0.05)$ *for exchangeable designs* $\mathbf{X}^{(p)} = \mathbf{I}_p + a\mathbf{E}_{p\times p}$ *for* $a \in [0, \infty)$ *(Section 6.1). The horizontal axis shows* $a/(1+a)$*, hence the locations 0, 0.5 and 1.0 represent* $a = 0,\ 1,\ \infty$*, respectively. Surprisingly, the largest* $K(\mathbf{X}^{(p)}(a))$ *is not attained at* $a = \infty$*, the point of perfect collinearity, at least not for dimensions up to* $p = 10$*. The graph is based on 10,000 random samples in* p *dimensions for* $p = 2, ..., 15$*.*

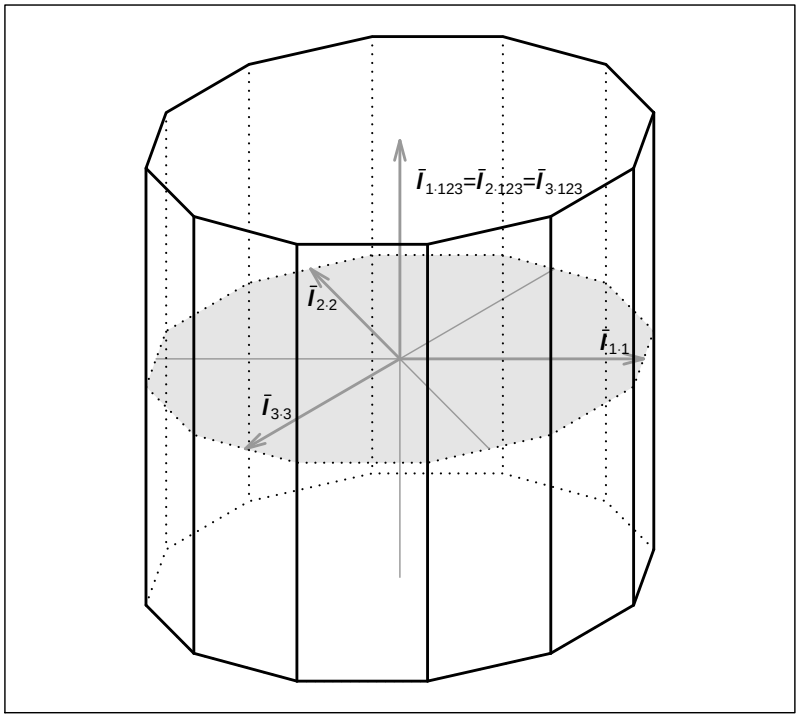

Fig 3: *Exchangeable Designs (Section 6.1): The geometry of the limiting PoSI polytope* $\mathbf{\Pi}_K$ *for* $p = 3$ *as the design* $\mathbf{X}^{(p)}(a)$ *of equation (6.1) approaches either of the two collinearities. For* $a \uparrow \infty$*, the predictor vectors fall into the 1-D subspace* $\mathrm{span}(\mathbf{1})$*, and for* $a \downarrow -1/p$ *they fall into* $\mathrm{span}(\mathbf{1})^\perp$*. With duality in mind and considering the permutation symmetry of exchangeable designs, it follows that the limiting polytope is a prismatic polytope with a* p*-simplex as its base in* $\mathrm{span}(\mathbf{1})^\perp$*. The figure shows this prism for* $p = 3$*. The unit vectors* $\bar{\boldsymbol{l}}_{1\cdot\{1\}} \sim \mathbf{X}_1$*,* $\bar{\boldsymbol{l}}_{2\cdot\{2\}} \sim \mathbf{X}_2$ *and* $\bar{\boldsymbol{l}}_{3\cdot\{3\}} \sim \mathbf{X}_3$ *form an equilateral triangle. The plane* $\mathrm{span}(\mathbf{1})^\perp$ *also contains the six once-adjusted vectors* $\bar{\boldsymbol{l}}_{j\cdot\{j,j'\}}$ *(*$j' \neq j$*), while the three fully adjusted vectors* $\bar{\boldsymbol{l}}_{j\cdot\{1,2,3\}}$ *collapse to* $\mathbf{1}/\sqrt{p}$*, turning the polytope into a prism.*

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