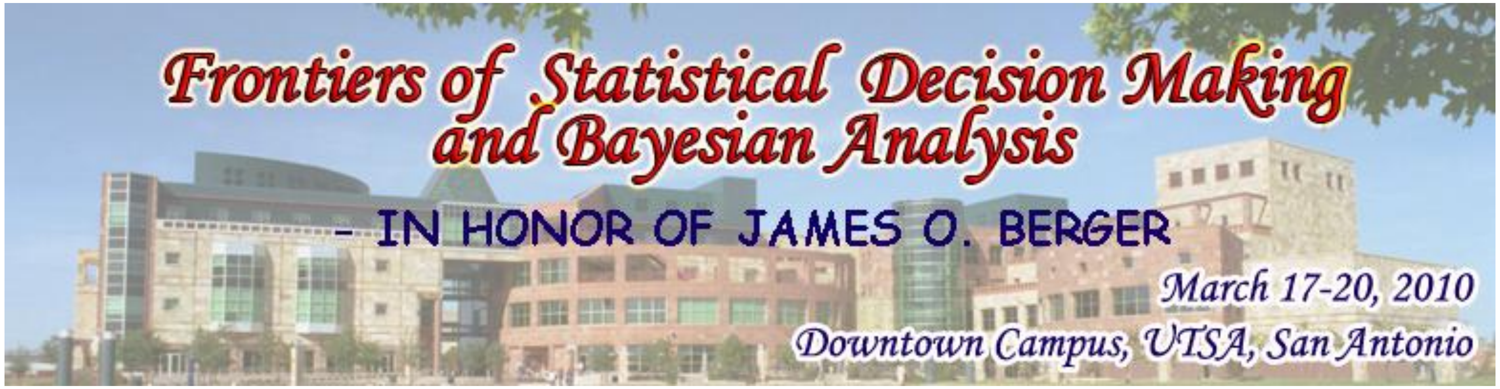

Frontiers of Statistical Decision Making and Bayesian Analysis
- IN HONOR OF JAMES O. BERGER
March 17-20, 2010
Downtown Campus, UTSA, San Antonio

Valid Statistical Inference After Model Selection

Lawrence D. Brown
Statistics Department, Wharton School
University of Pennsylvania

Joint work with: R. Berk, L. Zhao; & A. Buja, M. Freiman, K. Zhang.

Model Selection

- Common in Pedagogy
- Many instances involve Gaussian Linear Models and choice of $r \leq p$ effects based on n observations
- Common in Applications:

1. In social sciences

   p smallish & $p << n$

2. In observational studies

   larger p & $p < n$

3. In physical sciences, biology and genomics

   {very} large p & $p<n$ [or (often) $p>n$].

[Ground rule: We're interested in this case, but present talk is only about $p<n$.]

Inference after Model Selection

- Generally uses the selected model,
- And ignores the fact it was preceded by model selection
- Here are some examples:
  1. From a textbook.
  2. A prototypical applied analysis
  3. A “Toy” example to illustrate post model selection distributions
- Then I’ll describe the nature of our research and results
- The problem is OF COURSE not new. See Pötscher, Leeb, and collaborators for a recent series of papers and other references.

A Textbook Example

From Moore and McCabe, *Intro to Practice of Statistics* (Chap 11)

Sample of 1st year computer science majors.

Y = Cumulative GPA (after three semesters)

X_1 = HSMath, X_2 = HSSci, X_3 = HSEng, X_4 = SATM, X_5 = SATV

- Several models are tried as “predictors” of Y. For each, the resulting ANOVA table is examined:
  eg, $X_1 – X_3$, $X_4 – X_5$, $X_1 – X_5$, X_1 & X_3.
- X_1 & X_3 are chosen as the final set of predictors of Y.
- There are several appropriate cautionary clauses.
- I’ll focus on two

Two Important Cautions about Model Selection

(one included in Moore and MCCabe, and one omitted!)

“Individual regression coefficients, their standard errors and significance tests are meaningful only when interpreted in the context of the ….”

A. … “other explanatory variables in the model”

B. … model selection procedures that have been explicitly or implicitly applied to the same data to select the model.

[*There are no quotation marks on* B *because it isn’t in the text*. **But it should be**.]

A Prototypical Application: Length of Criminal Sentence

- L = Length of criminal sentence (months)
- $Y = \text{Log}(L + ½)$
- X_i = 13 possible “predictors” of sentence length; $i = 1,..,13$.

eg, Drug possession, burglary, assault or gun-related as the primary offense (4 variables), # of juvenile arrests, # of prior adult arrests, age, sex, etc.

- $n = 250$
- Model selection (all subsets BIC) chose 6 variables for the model.
- Re-analysis of a new, validation data set of 250 using the model chosen on the test set gave quite different results. (And, only 4 of the 6 chosen variables are significant in the re-analysis.)

“Toy” Example (the simplest of several)

Gaussian linear model with p = 2 & $n \approx \infty$; so let

- $Y \sim N\left(X\beta, \sigma^2 I\right)$ [Set $\sigma^2 = 1$, known, *wlog*.]
- with $\left(X'X\right)^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & +.55\cdot & -.44\cdot \\ 0 & -.44\cdot & +.55\cdot \end{pmatrix}$.
- All models contain β_1, the intercept coefficient AND ALSO β_2.
- Model selection chooses $\hat{M}$ with $\hat{M}$ = full model if $\left|\hat{\beta}_3\right| > \sqrt{5/9}$; otherwise $\hat{M}$ chooses only β_1, β_2. [This is roughly AIC.]

Confidence Intervals

- If $M = \{1,2,3\}$, a fixed choice, then $\operatorname{var}\left(\hat{\beta}_{2\cdot M}\right) = 5/9 = \operatorname{var}\left(\hat{\beta}_{3\cdot M}\right)$.
- Hence the routine intervals are

$$\beta_{2\cdot\hat{M}} \in \hat{\beta}_{2\cdot\hat{M}} \pm \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\sqrt{\frac{5}{9}} \quad \& \quad \beta_{3\cdot\hat{M}} \in \hat{\beta}_{3\cdot\hat{M}} \pm \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\sqrt{\frac{5}{9}}$$

- If $\hat{M} = \{1,2\}$ then the routine interval (for $\beta_{2\cdot\hat{M}}$) is

$$\beta_{2\cdot\hat{M}} \in \hat{\beta}_{2\cdot\hat{M}} \pm \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\sqrt{\frac{1}{5}} \quad (\text{since } \operatorname{var}\left(\beta_{2\cdot\{1,2\}}\right) = 1/5).$$

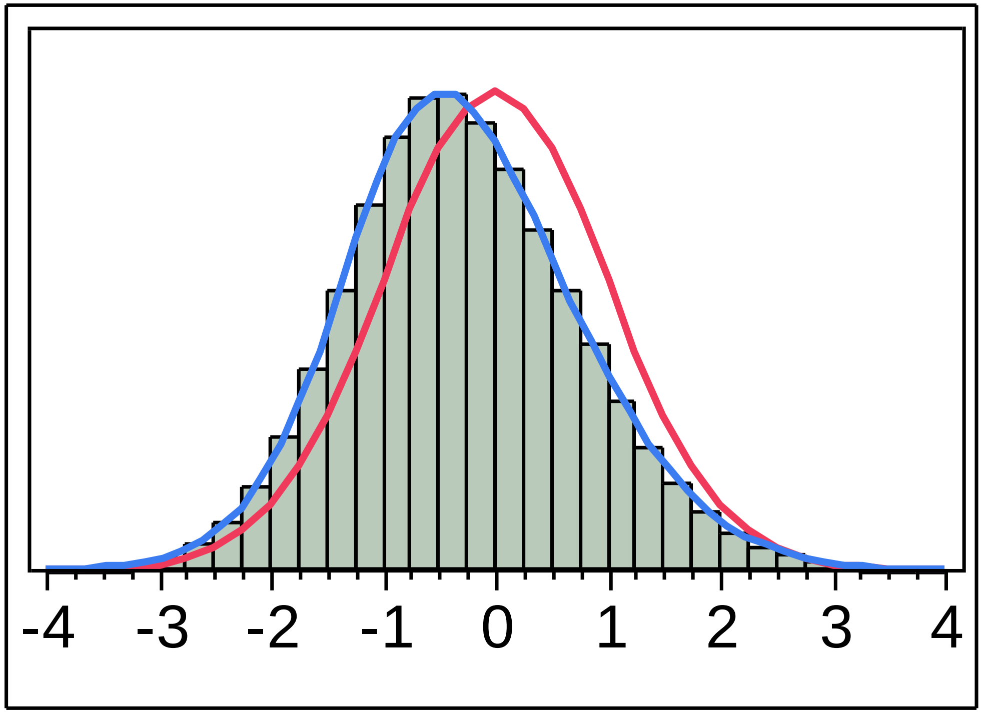
- Proper coverage of these intervals assumes

$$z_{j,\hat{M}} = \left(\hat{\beta}_{j\cdot\hat{M}} - \beta_j\right) \Big/ \sqrt{\operatorname{var}\left(\hat{\beta}_{j\cdot\hat{M}}\right)} \sim N(0,1)$$

- **Is this true???**

Distribution of $z_{2,\hat{M}} = \left(\hat{\beta}_{2\square\hat{M}} - \beta_1\right) \Big/ \sqrt{\mathrm{var}\left(\hat{\beta}_{2\square\hat{M}}\right)}$

Suppose (as a special case) $\beta_2 = \beta_3 = \sqrt{5/9}$.



___ = Actual distribution ____= Ideal N(0,1) dist.

Consequences

- The actual distribution is mis-centered and doesn't have $\sigma = 1$

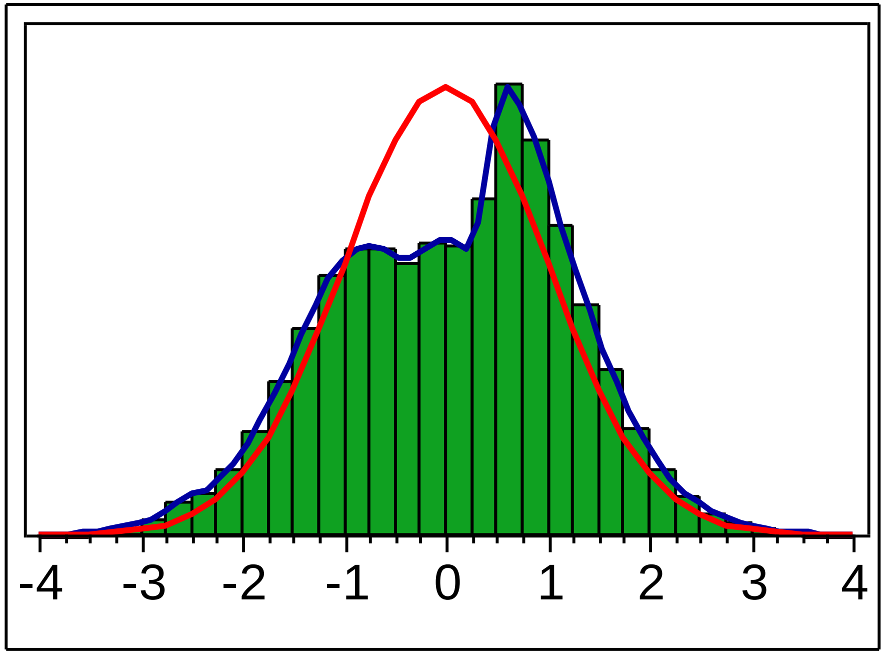
So

- Confidence intervals formed under the standard assumption that $z_{2,\hat{M}} = \left(\hat{\beta}_{2\cdot\hat{M}} - \beta_{2\cdot\hat{M}}\right) \Big/ \sqrt{\operatorname{var}\left(\hat{\beta}_{2\cdot\hat{M}}\right)} \sim N(0,1)$ will not have the nominal coverage.
- In this "toy", coverage isn't deficient by a lot –

$$P\left(\beta_{2\cdot\hat{M}} \in \hat{\beta}_{2\cdot\hat{M}} \pm 1.96\sqrt{5/9}\right) \approx 0.938 \neq 0.95.$$

- Suppose you look at both $\beta_{2\cdot\hat{M}}$ and $\beta_{3\cdot\hat{M}}$ (when $\hat{M} = \{1,2,3\}$). And then look at the distribution of $z_{2,\hat{M}}$ if $\hat{M} = \{1,2\}$ or of the more extreme of $z_{j,\hat{M}}$, $j = 2,3$ if $\hat{M} = \{1,2,3\}$ (for simultaneous CI^{s}).

Distribution of the more extreme $z_{j,\hat{M}}$



___ = Actual distribution **____= Ideal N(0,1) dist.**

Here, $P\left(\max\left(\left|z_{j,\hat{M}}\right|\right) < 1.96\right) = \mathbf{0.92}$ and $P\left(\max\left(\left|z_{j,\hat{M}}\right|\right) < \mathbf{2.18}\right) = 0.95$.

Formulation

- Pre model-selection observations are

$$Y \,\square\, N_n\left(X\beta, \sigma^2 I\right),\ X \text{ is } p \times p \text{ and full rank}.$$

Convention: Usually the "intercept" is not of model selection interest. If so, then we assume the columns of *X* have been centered.

[Generalizations such as non-normal models are of interest, but not treated here.]

- **Also assume,** $\boldsymbol{n > p}$.
- Then $\hat{\sigma}^2 = MSE_{\text{fullmodel}}$

is a valid estimate of σ^2, free of any model selection effects.

- A (sub)Model, *M*, is a subset of $\{1, .., p\}$, and leads to

$$X_M = \left[X_{(j)} : j \in M \right]$$

$X_{(j)}$ denotes the *j*-th column of *X*. X_M is an $n \times (\# M)$ matrix..

For a given model, *M*

Denote the corresponding LS estimate by

$$\hat{\beta}_{\square M} = \left(X_M' X_M \right)^{-1} X_M' Y \ \square \ B_M Y \, .$$

Coordinates of $\hat{\beta}_{\square M}$ are $\hat{\beta}_{k\square M}$, $k \in M$.

Let $P_M = XB_M$ = Projection matrix on *M*.

Let

$$x_{k\square M} = \left(I - P_{M-\{k\}} \right) x_{(k)} .$$

Then

$$\hat{\beta}_{k\square M} = x_{k\square M}' Y \Big/ \left\| x_{k\square M} \right\|^2 \ \square \ \ell_{k\square M}' Y$$

Note $\ell_{k\square M} \propto x_{k\square M}$, the residual vector of $x_{(k)}$ from $\mathrm{ColSp}\left(X_{M-\{k\}} \right)$.

Meaning of Correlation Coefficients within *M*

- For given *M* define $\beta_{k\cdot M}$ by

$$\beta_{k\cdot M} = E\left(\hat{\beta}_{k\cdot M}\right) = \ell'_{k\cdot M} X\beta .$$

So, the centering and interpretation of each coefficient, $\beta_{k\cdot M}$, also depends on the identity of the other coefficients in *M*.

- **Conventional test** of $H_0 : \beta_{k\cdot M} = 0$ is based on

$$t_{k\cdot M} = \frac{\ell'_{k\cdot M} Y}{\left\|\ell_{k\cdot M}\right\| \hat{\sigma}}$$

with Student’s-t null distribution & *n*−*p* df.

- **Conventional CI** is $\mathrm{CI}_{k\cdot M} \equiv \hat{\beta}_{k\cdot M} \pm \mathrm{t}_{n-p;1-\alpha/2}\, \hat{\sigma} \big/ \left\|x_{k\cdot M}\right\|$.

Note: *n*−*p* df, not *n* − #*M*. If σ^2 is known then replace $\hat{\sigma}$ by σ and **t** by Z (a standard normal).

Model Selection

The data is examined and a “model” $\hat{M} = M\left(Y\right)$ is chosen.

[Model selection is (only) about choice of predictor variables, not about – eg – transformation of *Y*.]

This yields a post-selection design –

$$X_{\hat{M}} = \text{the columns of } X \text{ with indices in } \hat{M}.$$

Conventional inference following Model Selection is Invalid:

Typically $P\left(\beta_{k\in\hat{M}} \in \mathrm{CI}_{k\in\hat{M}}\right) < 1-\alpha$ [instead of desired $\geq 1-\alpha$].

We propose to construct **Po**st **S**election **I**nference with valid tests and multiple confidence statements (Family Wise Error).

PoSI Criterion (for CIs)

Define constant **K** so that CI of the form

$$\mathrm{CI}^{*}_{k\cdot M} \equiv \hat{\beta}_{k\cdot\hat{M}} \pm \mathrm{K}\hat{\sigma}\big/\left\|x_{k\cdot\hat{M}}\right\|$$

satisfies

(*) $$P_{\beta}\left(\beta_{k\cdot\hat{M}} \in \mathrm{CI}^{*}_{k\cdot\hat{M}}\right) \geq 1-\alpha \text{ for all } k \in \hat{M}\,.$$

K depends on α, p, $n-p$ and X.

But K does not depend on the rule leading to $\hat{M}$, or on $\hat{M}$ itself, and (*) is true for all β.

A restricted version of (*) – call it (*k) is also of interest. This is (*) only for the fixed k, under the restriction $\hat{M} \supset k$. See later for some results for (*k).

Because of normality of $\hat{\beta}_{k\cdot M}$ (*) is equivalent to

$$1-\alpha \le P_{\beta}\left[\max_{k\in\hat{M}}\left|t_{k\cdot\hat{M}}\right| \le \mathrm{K}\right]$$

Then, because of linearity and centered-ness of $\hat{\beta}_{k\cdot M}$ this is implied by

(**) $$1-\alpha \le P_{0}\left[\max_{M\,;\,k\in M}\left|t_{k\cdot M}\right| \le \mathrm{K}\right].$$

Canonical Form

Rotation of Y reduces problem to canonical form with new $X, Y, \hat{\sigma}^2$ without affecting meaning of β (*eg*, *TSH*, Chapter 7), where now

(CF) $X \sim p \times p, Y \sim N_p\left(X\beta, \sigma^2 I_p\right), \left(n-p\right)\hat{\sigma}^2 \sim \sigma^2 \chi^2_{n-p}.$

A further rotation {of $M \to \left(X'X\right)^{-1/2} X'M$ } transforms to a form in which *wlog* ***X*** **is symmetric** in (CF) – *ie* $X = \left(X'X\right)^{1/2}$.

Changing the measurement scale of individual β_k also does not affect the nature of the problem. Hence *wlog*, **can assume** $X = \left(X'X\right)^{1/2}$ is a correlation matrix (or that $X'X$ is one).

PoSI is Possible

Review: PoSI needs K such that

(**) $$1-\alpha \le P_0\left[\max_{M;\,k\in M}\left|t_{k\cdot M}\right| \le \mathrm{K}\right].$$

Theorem: Scheffe’s constant $\mathrm{K_S} = \sqrt{pF_{p,n-p;1-\alpha}}$ satisfies (**).

Proof: By its construction

(1) $$1-\alpha = P_0\left[\max_{\mathbf{c}} \frac{\mathbf{c}'Y}{\|\mathbf{c}\|\hat{\sigma}} \le \mathrm{K}\right] < P_0\left[\max_{M;\,k\in M} \frac{\ell'_{k\cdot M}Y}{\|\ell_{k\cdot M}\|\hat{\sigma}} \le K\right]. \|$$

- $\mathrm{K_S}$ does not depend on *X*.
- $\mathrm{K_S}$ may give very conservative CIs. (Inequality in (1) can be Big.)
- **It’s possible to do better**. Here’s our plan:

Proposal for Improved PoSI

- For fixed α, *p*, *n – p* and *X* computationally find K=K(*X*) such that (**) holds – ie, $1-\alpha = P_0\left[\max_{M\,;\,k\in M}\left|t_{k\cdot M}\right| \le \mathrm{K}\right]$.
- For modest *p* we can always do so by simulating the null distribution of Y, and using Monte-Carlo.

[The limitation is that the max step involves looking at $p2^{p-1}$ possibilities. So, “naive” version of this requires (approx) $p \le 16$.]

- Alternatives to naive computation are being explored!
- Also of interest: fix α, *p*, *n – p* and [using theory] find $\inf_X \mathrm{K}\left(X\right)$ and $\sup_X \mathrm{K}\left(X\right)$.
- Then let $p \to \infty$ for limit asymptotics.

- In **special cases** linear model theory and geometry yield useful results and perspectives.
- Two tantalizing (and somewhat useful) facts

    1. Gram-Schmidt orthogonalities:

    $$k \in M, l \notin M \Rightarrow x_{k \cdot M} \perp x_{l \cdot (M \cup \{l\})}$$

    2. Duality: There are several duality facts. The easiest to state (though perhaps not the most intuitive) is

    $$K(X) = K\left(X^{-1}\right)$$

(And, there is a corresponding match-up of partial correlation vectors from X and X^{-1}.)

Bounds (1)

- Lower Bound:

$$\mathrm{K}\left(X\right) \geq \mathrm{K}\left(I\right) = \Phi^{-1}\left(\left(1+\left(1-\alpha\right)^{1/p}\right)\Big/2\right)$$

.

$$\approx \sqrt{2\log p} \text{ as } p \to \infty, n-p \to \infty, \ \alpha \text{ fixed}$$

- Upper Bound: From an example (discussed later)

$$\sqrt{p}\Big/\sqrt{\pi} + o\left(\sqrt{p}\right) \leq \sup\nolimits_X \mathrm{K}\left(X\right) < K_S = \sqrt{p} + o\left(\sqrt{p}\right).$$

- Moral from comparison of these bounds:

Calculation of $\mathrm{K}\left(X\right)$ matters since the value can turn out to be anywhere from about $\sqrt{2\log p}$ to about $\sqrt{p}\Big/\sqrt{\pi}$.

Bounds (2)

The logic in the PoSI proposal ignores the variable selection method. This is reflected in the computation of K via (**) which involves an upper bound over all possible model selections.

Justification(s) for this perspective –

(a) Many model selections are informal, and the methodology is not clearly specified in advance

(b) (**) provides a relatively straightforward possibility for creating valid (conservative) inference; this can be useful even if more complicated proposals could better take account of the model selection algorithm

(c) What we can so far idealize as possible alternatives seem to require computations of possible model-selection outcomes at the (unknown) value of β and/or split-sample or other re-sampling options having uncertain fixed-sample properties. See Leeb and Pötscher for some ideas (and many negative properties).

(d) (**) can easily be modified for certain restricted model-selection plans.

“SPAR”

- A model-selection routine for which (**) is sharp.
- This is “**S**ingle **P**arameter **A**djusted **R**egression”; formally

$$\hat{M}_{\text{spar}} = \left\{ \hat{M} = \hat{M}\left(Y\right) : \exists \hat{k} \in \hat{M} \ni \left| t_{\hat{k}\square\hat{M}} \right| = \max{}_{M,k\in M} \left| t_{k\square M} \right| \right\}.$$

- Though artificial, this approximates what a naive scientist might do - one who combs a large data set looking for the most “publishable” result.
- A modification of this is somewhat more plausible in a setting in which one first settles on a co-variate of principle interest, and looks for the set of control variates that make this have the largest apparent effect after including those controls.

Examples (1):

1. Exchangeable Designs

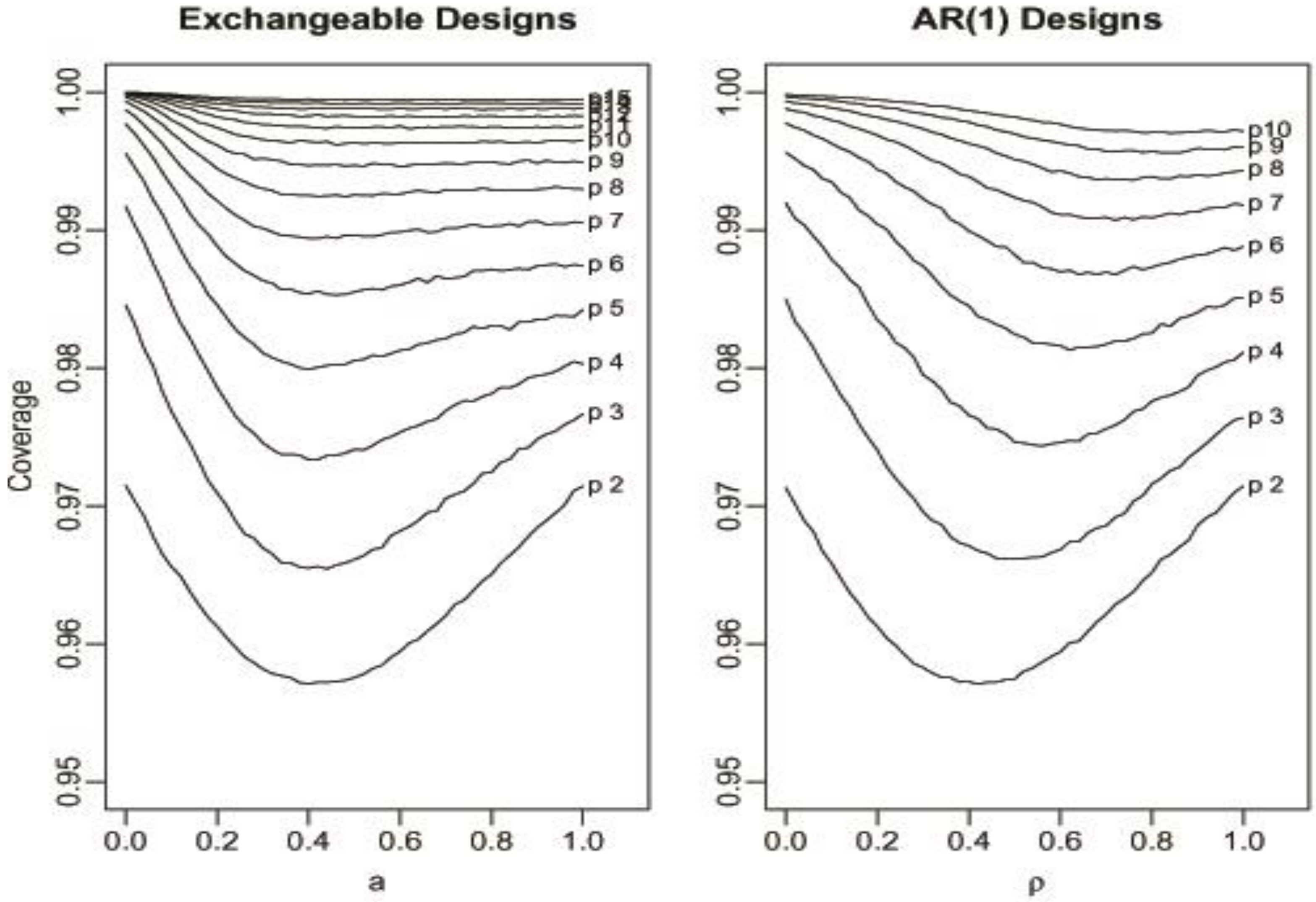
$$X_{ii} = 1; \; X_{ij} = r, \; -1/(p-1) < r < 1$$

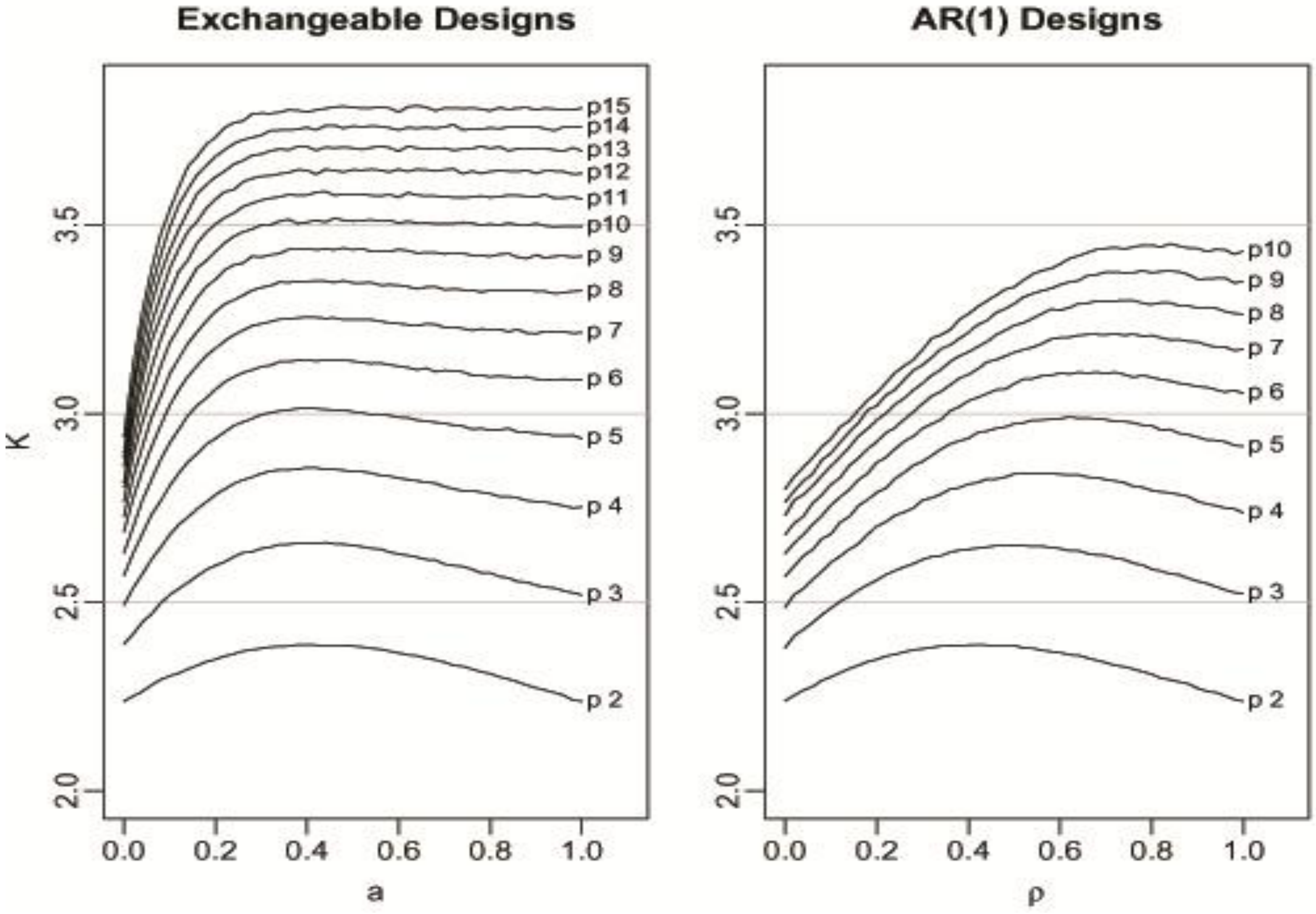
Note: The limiting cases $r = 1$ or $-1/(p-1)$ also make sense under an appropriate interpretation.

2. AR(1) Designs

$$X_{ij} = r^{|i-j|}$$

3.


Exchangeable Designs
AR(1) Designs
Coverage
1.00
0.99
0.98
0.97
0.96
0.95
0.0
0.2
0.4
0.6
0.8
1.0
a
ρ
p15
p10
p 9
p 8
p 7
p 6
p 5
p 4
p 3
p 2
p10
p 9
p 8
p 7
p 6
p 5
p 4
p 3
p 2


Exchangeable Designs
AR(1) Designs
K
a
ρ
2.0
2.5
3.0
3.5
0.0
0.2
0.4
0.6
0.8
1.0
p15
p14
p13
p12
p11
p10
p 9
p 8
p 7
p 6
p 5
p 4
p 3
p 2