LTCC: Selective Inference

Daniel García Rasines

Imperial College London

daniel.garcia-rasines16@imperial.ac.uk

February 2024

Structure of the course

1. Introduction.
2. Unconditional inference.
    - Fixed design.
    - Random design.
3. Conditional inference.
    - Most powerful conditional inference.
    - Information-splitting methods.
4. Bayesian approaches.
5. Other topics.

Key references

Review papers:

- Zhang, Khalili, Asgharian (2022). "Post-model-selection inference in linear regression models: an integrated review", *Statistics Surveys*.
- Kuchibhotla, Kolassa, Kuffner (2022). "Post-selection inference", *Annual Review of Statistics and Its Application*.

Week 1:

- Berk, Brown, Buja, Zhang, Zhao (2013). "Valid post-selection inference", *Annals of Statistics*.
- Bachoc, Preinerstorfer, Steinberg (2020). "Uniformly valid confidence intervals post-model-selection", *Annals of Statistics*.

Motivation

The **classical approach** to statistical inference assumes that all the **models to fit** and all the **inferential objectives** are **fixed prior to the data analysis**.

For example, a common regression problem assumes observation of a vector $Y \sim N(X\beta, \sigma^2 I)$ and seeks inference for β.

However, this is **not how statistics operates in practice**.

Motivation

Typically, the practitioner **interacts with the data** in order to select a suitable model to fit and/or a set of relevant inferential questions to address—the **selection stage**.

Such data exploration allows the practitioner to **focus only on the most relevant aspects** of the data-generating process.

BUT it invalidates the assumptions of a fixed model or inferential objectives, leading to a (possible) loss of the inferential guarantees indicated by classical theory.

Motivation

The further **sampling variability** introduced in the pre-analysis stage often leads to:

- Overstatement of statistical significance (exaggerated p-values).
- Confidence intervals with low coverage.
- Overestimation of effect sizes.
- Underestimation of variances.

Motivation

Selection effects are **regularly overlooked in statistical practice**, and are often cited as one of the main causes of the replicability crisis in science.

Famously, Breiman (1992) referred to this issue as a "quiet scandal in the statistical community".

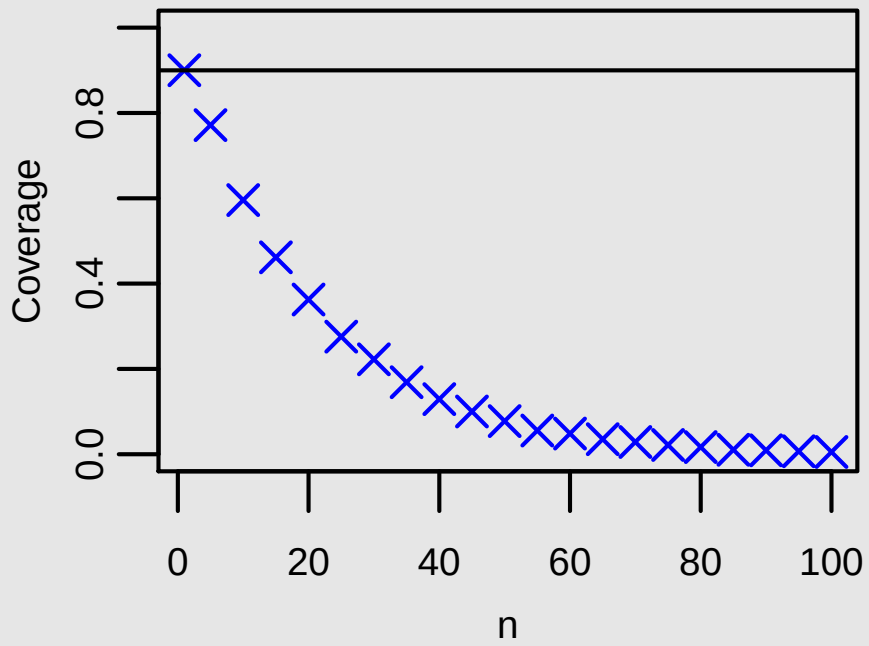
$\rightarrow$ The goal of **selective inference** is to **restore validity of inference** after selection.

Example: inference on winners

Let $Y_i \sim N(\theta_i, 1)$ independently for $i = 1, \ldots, n$, and suppose a confidence interval is required for the mean of the largest observation, $\theta_I(Y)$, where

$$I(Y) = \underset{i=1,\ldots,n}{\arg\max}\, Y_i.$$

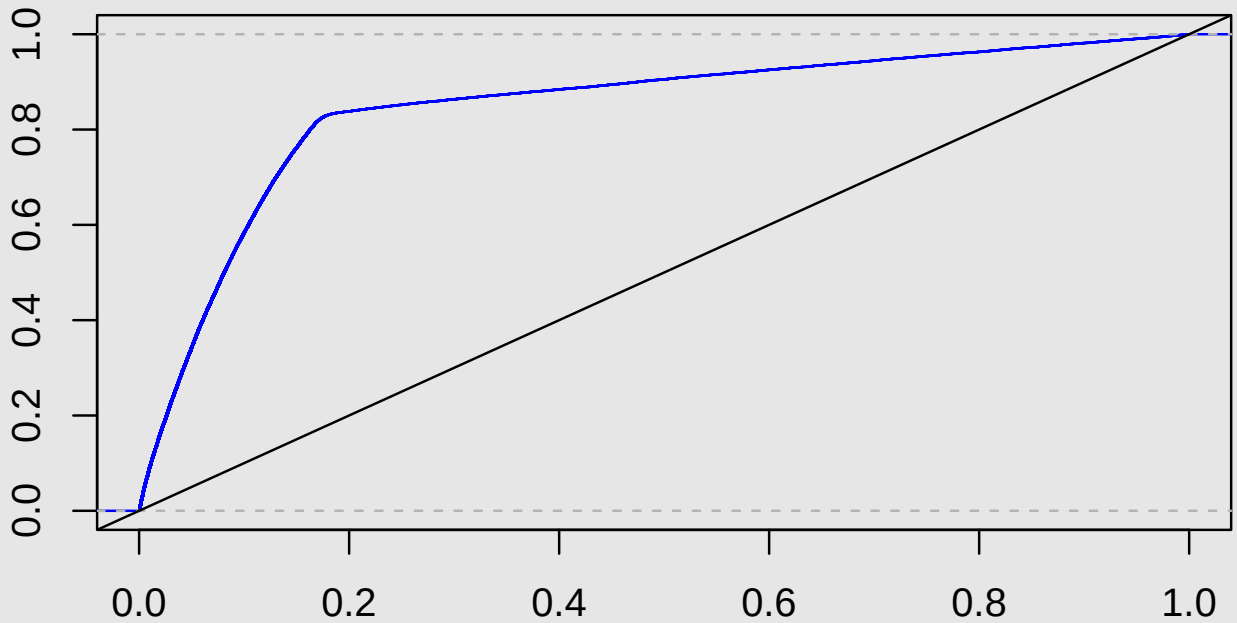
Plot shows the coverage of conventional 90% confidence intervals $[y_{I(Y)} - 1.64, y_{I(Y)} + 1.64]$ for the case $\theta_1 = \ldots = \theta_n$ as $n \to \infty$.



Example: inference after model selection

$Y \sim N(\mathbf{0}, I_{80})$ and $X \in \mathbb{R}^{80\times 20}$ with $N(0,1)$ entries independent of Y.

A forward stepwise algorithm minimising the AIC is used to select a linear submodel, and a marginal p-value is computed to test significance for each variable in the selected model.



$\sim$ **35% of selected coefficients deemed significant** at a 0.05 level, even though none of them are!

Approaches

Selective inference has a long and rich history.

Many approaches have been advocated to estimate/control selection effects, such as

- Simultaneous inference (Bonferroni, BH, ...).
- Bayesian methods: model averaging, empirical Bayes.
- Bootstrap.
- Differential privacy.

The goal of this course is to give an overview of recent proposals and describe how they fit within the general framework.

Framework

Data matrix denoted by $[Y, X]$, where

- $Y = (Y_1, \ldots, Y_n)^T \in \mathbb{R}^n$ is a quantitative response vector.
- $X = [X_1, \ldots, X_p] \in \mathbb{R}^{n \times p}$ is a $n \times p$ design matrix, containing observations of p covariates. For now assume **design is fixed**.

In general (provided finite expectation), one can write

$$Y = \mu + \varepsilon, \text{ where } \mu = E[Y] \text{ and } \varepsilon = Y - \mu. \tag{1}$$

Often μ is assumed to belong to a parametric class, such as the linear model $\{X\beta \colon \beta \in \mathbb{R}^p\}$, and inference is sought for the corresponding parameter.

Framework

If $p \sim n$ or $p > n$ it is common to seek a smaller model due to identifiability issues and lack of interpretability of the larger model.

An index set $M = \{i_1, \ldots, i_m\} \subseteq \{1, \ldots, p\}$ will denote a **linear submodel** containing only the covariates in M, where $m = |M|$.

If $X_M = [X_{i_1}, \ldots, X_{i_m}]$ denotes the submatrix of X that contains only the covariates in M, the linear submodel corresponding to M posits

$$Y = X_M \beta^M + \varepsilon, \quad \beta^M \in \mathbb{R}^m. \tag{2}$$

The projection parameter

In selective inference one normally treats **models as approximations** of a (potentially very complex) true underlying distribution.

This aligns more naturally with the model-selection framework and is considerably more realistic.

Even if $\mu \notin \text{span}(X_M) = \{X_M b \colon b \in \mathbb{R}^m\}$ for some M, one can still define a meaningful model-dependent parameter, the **projection parameter**:

$$\beta_M = \underset{b \in \mathbb{R}^m}{\arg\min}\, E\|Y - X_M b\|^2 = (X_M^T X_M)^{-1} X_M^T \mu = X_M^\dagger \mu \in \mathbb{R}^m, \qquad (3)$$

for all M with $m < n$ and $\text{rank}(X_M) = m$, where $A^\dagger := (A^T A)^{-1} A^T$.

$\rightarrow$ It is the **best linear predictor** of μ in M with respect to the squared error loss.

The projection parameter

The entries of the projection parameter, denoted by β_{jM}, $j \in M$, have a different interpretation to the conventional regression coefficients.

$\rightarrow$ If $Y = X_M\beta^M + \varepsilon$, β_j^M is the **average difference in the response for a unit difference in X_{i_j}, ceteris paribus in the model M.**

$\rightarrow$ In the non-linear case, β_{jM} is the **average difference in the response approximated in the submodel M.**

Naturally, if the selected model is in fact correct, then $\beta_M = \beta^M$ and the usual interpretation holds.

Interpretation of the projection parameter

For $j \in M$, let X_{jM} be the corresponding column of X_M, and define

$$r_{jM} = \left\{ I_n - X_{M\setminus\{j\}} X^{\dagger}_{M\setminus\{j\}} \right\} X_{jM}, \tag{4}$$

the residual vector of the regression of X_{jM} on the other predictors in M.

We can rewrite the j-th coefficient of β_M as

$$\beta_{jM} = \frac{1}{\|r_{jM}\|^2} r_{jM}^T \mu, \tag{5}$$

so β_{jM} is a **measure the relevance** of the j-th covariate once we have adjusted for the other covariates in M.

Interpretation of the projection parameter

More specifically, denote by $P_M = X_M X_M^\dagger$ the projection matrix onto $\text{span}(X_M)$.

We can decompose it as

$$P_M = P_{M\setminus\{j\}} + (r_{jM}^T X_{jM}) r_{jM} r_{jM}^T. \tag{6}$$

$\rightarrow$ The **null hypothesis** $H_0 : \beta_{jM} = 0$ is equivalent to

$$P_M \mu = P_{M\setminus\{j\}} \mu. \tag{7}$$

PoSI

PoSI[1] (**Po***st* **S***election* **I***nference*) is a framework for **selective inference for projection parameters** with

- Finite sample guarantees for any dimension (n, p).
- Universal validity over (virtually) all model selection procedures.

Seminal work; introduced ideas which constitute the basis of much of the contemporary work on selective inference.

Crucially, it requires very **strong distributional assumptions**.

Extensions with less restrictive assumptions will be considered later.

[1] Berk, Brown, Buja, Zhang, Zhao (2013). "Valid post-selection inference", *Ann. Stat.*

PoSI

Assume a Gaussian response $Y \sim N(\mu, \sigma^2 I_n)$, with $\mu \in \mathbb{R}^n$ and $\sigma^2 > 0$.

PoSI requires an estimator $\hat{\sigma}$ which is **independent** of all estimates $\hat{\beta}_{jM}$, and such that

$$\hat{\sigma}^2 \sim \sigma^2 \frac{\chi^2_r}{r}. \tag{8}$$

We write $r = \infty$ if σ^2 is known.

PoSI

Distributional requirements constitute a major practical limitation of the approach.

However, they are only needed for finite-sample validity; asymptotic guarantees can be derived under much weaker conditions, as we shall see.

A valid variance estimator will be available if $p < n$ and the full model $Y \sim N(X\beta, \sigma^2 I_n)$ is (at least approximately) valid.

PoSI

For a **fixed** submodel M, a natural estimator of β_M is

$$\hat{\beta}_M = X_M^{\dagger} Y \sim N(\beta_M, \sigma^2 (X_M^T X_M)^{-1}). \tag{9}$$

If a suitable estimate of σ^2 is available, confidence sets and testing procedures for β_M are provided by classical normal theory.

Introduce the t-values relative to submodel M:

$$t_{jM} = \frac{\beta_{jM} - \hat{\beta}_{jM}}{[(X_M^T X_M)^{-1}]_{jj}^{1/2} \hat{\sigma}} \sim t_r, \quad j \in M. \tag{10}$$

A valid $1 - \alpha$ valid confidence interval for β_{jM} is given by

$$CI_{jM}(K) = [\hat{\beta}_{jM} \pm K[(X_M^T X_M)^{-1}]_{jj}^{1/2} \hat{\sigma}], \quad K = t_{r,1-\alpha/2}. \tag{11}$$

PoSI

In the selective context, however, the submodel M is allowed to depend on the data, i.e. it is **random**, which invalidates the previous CI.

Write a random submodel as $\hat{M} \equiv \hat{M}(Y) \subseteq \{1, \ldots, p\}$.

Think of $\hat{M}$ as the result of some variable-selection procedure such as LASSO, stepwise regression, visual diagnostics, etc.

In principle, $\hat{M} \colon \mathbb{R}^n \to \mathcal{M}_{\text{all}}$ can be any *measurable* map, where

$$\mathcal{M}_{\text{all}} = \{M \subseteq \{1, \ldots, p\} \colon \text{rank}(X_M) = |M|\}, \tag{12}$$

i.e. the projection parameter needs to be identifiable.

PoSI

Associated with a random model $\hat{M}$ there is a random projection parameter $\beta_{\hat{M}} = X_{\hat{M}}^{\dagger}\mu$, which constitutes the **moving target of inference**.

Note that

- $\beta_{\hat{M}}$ has random dimension $|\hat{M}|$.
- For a fixed j, it might not be the case that $j \in \hat{M}$.
- Conditionally on $j \in \hat{M}$, the parameter $\beta_{\hat{M}}$ is random.

Furthermore, the corresponding estimator

$$\hat{\beta}_{\hat{M}} = X_{\hat{M}}^{\dagger} Y \tag{13}$$

is **not normally distributed** due to the extra variability from the selection step, so classical normal theory does not apply.

PoSI

Given the stochastic nature of the parameter, what constitutes a valid inferential procedure this setting?

In the fixed-M setting, the $1-\alpha$ confidence interval $CI_{jM}(K)$ for β_{jM} satisfies

$$P\left(\beta_{jM} \in CI_{jM}(K)\right) = 1-\alpha. \tag{14}$$

In the random-model setting, the **PoSI** (Post-Selection Inference) framework seeks a value of K such that

$$P\left(\beta_{j\hat{M}} \in CI_{j\hat{M}}(K)\ \forall\ j \in \hat{M}\right) \geq 1-\alpha, \tag{15}$$

for **any** selection procedure $\hat{M}$. K is called the **PoSI constant**.

PoSI

Some key aspects of this approach:

- **Universality**: CIs are valid **regardless of the selection procedure**, even if this involves subjective and informal decisions; the practitioner is even allowed to change their mind and report a different model post-hoc.
- Intervals tend to be very **conservative** as a result: the actual coverage can be well above the nominal one for some selection rules.
- However, **there exists a selection procedure that requires full protection**: unless there is a strong reason for discarding certain ill-behaved selection rules, PoSI is optimal.
- It provides only **unconditional** guarantees (more on this later).
- Implemented via simulation (computationally demanding if $p \approx 20$).

PoSI: restricted model space

The conservative nature of PoSI can be partially alleviated under the assumption that not all models in $\mathcal{M}_{\text{all}}$ are being searched.

In many applications, there is **a priori knowledge about the set of plausible selected models**, e.g.

- A subset of the covariates is forced into the model (e.g. an intercept).
- There is a size restriction on the model: $|M| \leq k$ (*sparsity*).
- Hierarchical restrictions: polynomial regression, interactions, etc.

PoSI: restricted model space

If such assumption can be made, we denote by $\mathcal{M} \subseteq \mathcal{M}_{\text{all}}$ the *pre-specified* set of allowed models.

With sufficiently strong restrictions on $\mathcal{M}$ (particularly the sparsity one), the PoSI approach becomes **computationally manageable for large** p.

Further reduction can be achieved by discarding variables ignoring the response, e.g. if there is collinearity.

PoSI: the selection-adjusted constant

The PoSI constant K is formally defined as

$$K(X, \mathcal{M}, \alpha, r) = \min \left\{ K > 0 \colon \mathbb{P}\left(\max_{M \in \mathcal{M}} \max_{j \in M} |t_{jM}| \leq K \right) \geq 1 - \alpha \right\}, \tag{16}$$

where, recall,

$$t_{jM} = \frac{\beta_{jM} - \hat{\beta}_{jM}}{[(X_M^T X_M)^{-1}]_{jj}^{1/2} \hat{\sigma}} = \frac{e_j^T X_M^{\dagger}(Y - \mu)}{[(X_M^T X_M)^{-1}]_{jj}^{1/2} \hat{\sigma}}. \tag{17}$$

$T = \max_{M \in \mathcal{M}} \max_{j \in M} |t_{jM}|$ is **distribution constant**, so K is computable.

PoSI: proof of coverage control

For any measurable model-selection procedure $\hat{M} : \mathbb{R}^n \to \mathcal{M}$, we have the trivial bound

$$\max_{j \in \hat{M}} |t_{j\hat{M}}| \leq \max_{M \in \mathcal{M}} \max_{j \in M} |t_{jM}|. \tag{18}$$

Thus,

$$\mathbb{P}\left(\max_{j \in \hat{M}} |t_{j\hat{M}}| \leq K\right) \geq \mathbb{P}\left(\max_{M \in \mathcal{M}} \max_{j \in M} |t_{jM}| \leq K\right) \geq 1 - \alpha, \tag{19}$$

where $K = K(X, \mathcal{M}, \alpha, r)$.

PoSI: achievement of nominal coverage

Among all possible model-selection procedures, there is one for which the nominal covarage is achieved, i.e. for which the PoSI constant is **sharp**.

It is the **significance-hunting procedure**, which seeks the model with the most significant observed effect:

$$\hat{M}^*(Y) = \arg\max_{M \in \mathcal{M}} \max_{j \in M} |t_{jM}|. \tag{20}$$

Although it is generally not advisable to select a model via $\hat{M}^*$, protection against it is a **guarantee against bad practice**.

PoSI: computation of the PoSI constant

Closed-form expressions for K are **not available**.

Brute force **Monte Carlo** used to approximate the $1-\alpha$ quantile of

$$T = \max\{|t_{jM}| : M \in \mathcal{M}, j \in M\}. \tag{21}$$

$\rightarrow$ If $\text{rank}(X) = p$ and $\mathcal{M} = \mathcal{M}_{\text{all}}$, need to evaluate $p2^{p-1}$ t-values.

$\rightarrow$ Computations are specific to the design X.

$\rightarrow$ Universal bounds needed in high-dimensional problems.

PoSI: one primary predictor

Sometimes the analysis is centred on a predictor of interest, X_j, while the other predictors in M act as controls, so that

- The submodel space is $\mathcal{M}_j = \{M \in \mathcal{M}_{\text{all}} : j \in M\}$.
- Only the t-statistic associated with X_j is relevant.

In this context, the PoSI constant is defined differently:

$$K_j(X, \mathcal{M}, \alpha, r) = \min \left\{ K > 0 : \mathbb{P}\left(\max_{M \in \mathcal{M}_j} |t_{jM}| \leq K \right) \geq 1 - \alpha \right\}. \quad (22)$$

As in the unrestricted case, exact coverage with this constant is achieved by the significance-hunting procedure

$$\hat{M}_j^*(Y) = \arg\max_{M \in \mathcal{M}_j} |t_{jM}|. \quad (23)$$

PoSI: Scheffé bound

PoSI provides simultaneous inference for up to $p2^{p-1}$ linear contrasts β_{jM}.

Scheffé's method provides **simultaneous protection for all linear combinations** without all the computational burden.

Write

$$t_x = \frac{(Y - \mu)^T x}{\hat{\sigma}\|x\|}, \quad x \in \text{span}(X) \setminus \{\mathbf{0}\}. \tag{24}$$

Recall that for $x \propto r_{jM} \in \text{span}(X)$, $t_x = t_{jM}$, so simultaneous inference for all the directions in the column space of X is an overkill for our problem.

PoSI: Scheffé bound

Scheffé's constant, explicitly given by $K_S(\alpha, d, r) = \sqrt{dF_{d,r,1-\alpha}}$, with $d = \dim\{\text{span}(X)\}$, satisfies

$$\mathbb{P}\left(\sup_{x \in \text{span}(X)} |t_x| \leq K_S\right) = 1 - \alpha, \tag{25}$$

and thus it provides **valid selective intervals regardless of the design**.

Naturally $K \leq K_S$ for any X, and often the **difference is substantial**, but provides a valid solution when simulation of PoSI constant is too costly.

PoSI: size of the PoSI constant

Asymptotic bounds for $n \geq p$, $p \to \infty$ and $\mathcal{M} = \mathcal{M}_{\text{all}}$:

- **Lower bound**: $K = \Omega(\sqrt{\log p})$, achieved by orthogonal designs.
- **Upper bound**: $K = O(\sqrt{p})$, achieved by equicorrelated designs.

$\to$ For orthogonal X, $\beta_{jM} \equiv \beta_j$ for all (j, M), so only p directions need to be covered.

$\to$ Large range $(\sqrt{\log p}, \sqrt{p})$ suggests strong dependence on X.

$\to$ Scheffé constant has $K_S \sim \sqrt{p}$, so its optimality is case-dependent.

For **sparse model spaces**, $\mathcal{M}_s = \{M \subseteq \{1, \ldots, p\} : |M| \leq s\}$,

$$K = O\left(\sqrt{s \log(p/s)}\right). \tag{26}$$

PoSI: other universal bounds

A general upper bound for K is given by

$$Q_T\{g(\mathcal{M}, X), r, 1-\alpha/2\} \leq \frac{g(\mathcal{M}, X) + A}{1 - Br^{-1/2}} \tag{27}$$

for some known constants $A, B > 0$, where $Q_T(x, r, \alpha)$ is the α quantile of a non-central t distribution with r degrees of freedom and non-centrality parameter x, and

$$g(\mathcal{M}, X) = \mathbb{E}\left[\max_{M\in\mathcal{M}} \max_{j\in M} |w_{jM}^T Z|\right], Z \sim N(0, I_n), w_{jM}^T = e_{jM}^T X_M^\dagger / \|e_{jM}^T X_M^\dagger\|. \tag{28}$$

Two important cases:

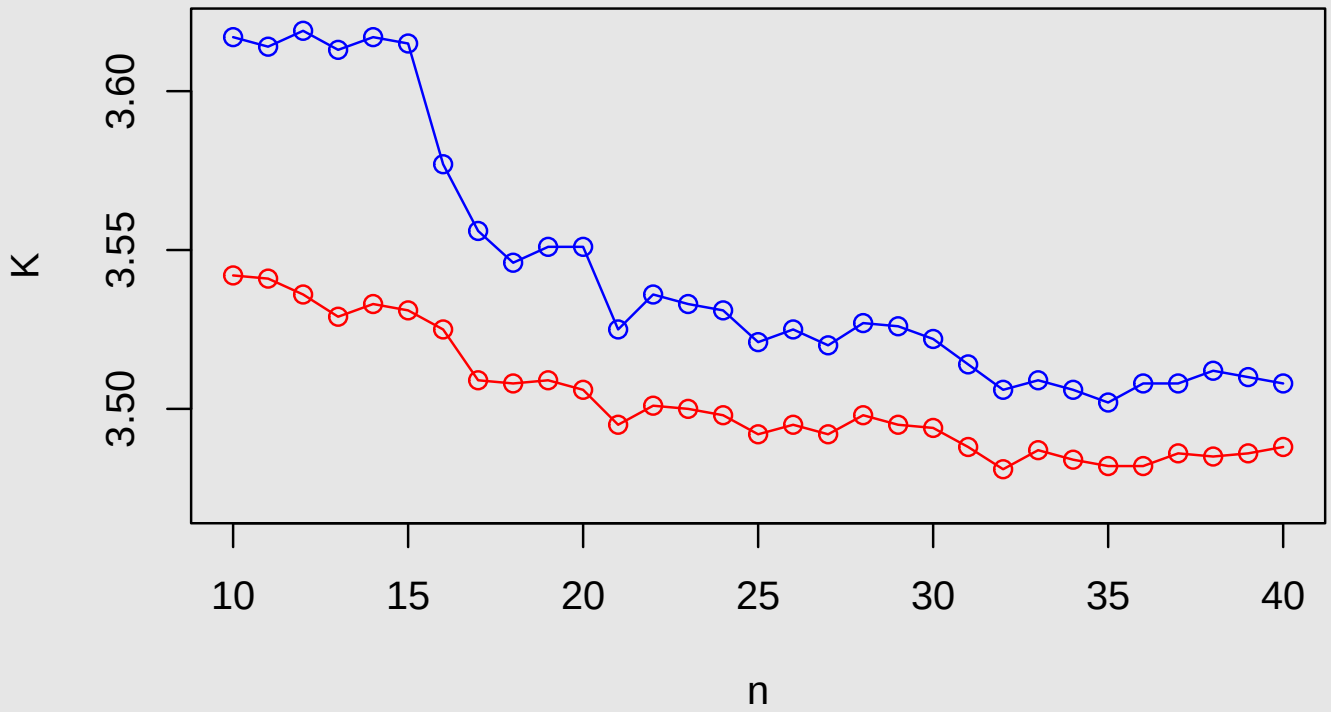
- Orthogonal designs: $g(\mathcal{M}, X) = \sqrt{2\log(2p)}$.
- Sparse models, $\mathcal{M}_s = \{M\colon |M| \leq s\}$: $g(\mathcal{M}, X) = \sqrt{2s\log(6p/s)}$.

Further refinements are available in the literature combining these two cases under the **Restricted Isometry Property**.

PoSI: size of the PoSI constant (empirical)

Set $p = 10$, $n \in \{10, \ldots, 40\}$, $r = \infty$ and $\alpha = 0.05$; predictors generated as Gaussian vectors with covariance $\Sigma_{ii} = 1$, $\Sigma_{ij} = 0.5$; Nsim $= 5 \times 10^4$.

$\mathcal{M} = \mathcal{M}_{\text{all}}$; $\mathcal{M} = \{M\colon |M| \leq 5\}$.



$\rightarrow$ The Scheffé constant for this problem is **4.28**.

PoSI: coverage

Empirical coverages of 95% CIs over 5000 samples; setting as before.

Data generated as $Y = X\beta + N(\mathbf{0}, I_n)$ under two parameters:

- $\beta^{(1)} = \mathbf{0}$.
- $\beta^{(2)} = (1, 1, -1, -1, 0, \ldots, 0)^T$.

Variable-selection rules:

- Lasso with cross-validation.
- Screening: a predictor is selected iff its significance p-value in the full linear model $\mu = X\beta$ is below 0.05.

$\mathbf{p = 10, n = 30}$

		Lasso	Screen.
$\beta^{(1)}$	PoSI	97.6	97.2
	Scheffé	99.8	99.8
	Unadj.	63.6	62.5
$\beta^{(2)}$	PoSI	99.5	99.5
	Scheffé	99.9	100
	Unadj.	93.4	92.8

$\mathbf{p = 10, n = 1000}$

	Lasso	Screen.
PoSI	95.7	95.4
Scheffé	99.7	99.8
Unadj.	50.5	50.4
PoSI	99.2	99.1
Scheffé	100	99.9
Unadj.	92.5	88.7

PoSI: extensions

The original PoSI framework has some practical limitations, most notably the restrictive distributional assumptions

$$Y \sim N(\mu, \sigma^2 I_n), \quad \hat{\sigma}^2 \sim \sigma^2 \frac{\chi^2_r}{r}. \tag{29}$$

Bachoc et al.[2] develop a more general framework that

- Provides **asymptotically valid** (*but* fixed-p) confidence intervals **without parametric assumptions** on the errors ε.
- Does **not require a consistent estimator** of σ.
- Is **applicable with other types of data (e.g. binary)**.

[2] Bachoc, Preinerstorfer, Steinberg (2020). "Uniformly valid confidence intervals post-model-selection". *Ann. Stat.*

Generalised PoSI: framework

- Data $Y = (Y_1, \ldots, Y_n)^T \sim \mathbb{P}_n$ has **independent but not necessarily identically distributed** components.
- $\mathbb{P}_n \in \mathbf{P}_n$, where $\mathbf{P}_n$ is a large non-parametric family of distributions.
- Statistician has a set of models $\mathcal{M}_n = \{M_{1,n}, \ldots, M_{d,n}\}$, possibly misspecified, where each $M_{j,n}$ is a set of distributions over $\mathcal{B}(\mathbb{R}^n)$.
- For each model $M \in \mathcal{M}_n$ there is a **prespecified parameter of interest**, $\theta_{M,n}(\mathbb{P}_n) \equiv \theta_{M,n}$, of dimension $m(M)$, and a corresponding estimator $\hat{\theta}_{M,n}$. Typically $\theta_{M,n}$ is a projection of P_n onto $\mathbf{P}_n$.

Note: We can phrase the original PoSI problem as a special case of this framework with $\mathbb{P}_n = N(\mu_n, \sigma^2 I_n)$, $M = \{N(X_M \beta^M, \sigma^2 I_n) \colon \beta^M, \sigma^2\}$, and $\theta_{M,n} = X_M^\dagger \mu_n$.

Generalised PoSI: objective

For a fixed $\alpha \in (0,1)$, define a family of intervals for $\theta^{(j)}_{M,n}$,

$$\{CI^{(j)}_{1-\alpha,M} \colon M \in \mathcal{M}_n, 1 \leq j \leq m(M)\}, \tag{30}$$

satisfying

$$\liminf_{n\to\infty} \mathbb{P}_n \left(\theta^{(j)}_{M,n} \in CI^{(j)}_{1-\alpha,M} \text{ for all } 1 \leq j \leq m(M), M \in \mathcal{M}_n \right) \geq 1-\alpha. \tag{31}$$

It then follows that

$$\liminf_{n\to\infty} \mathbb{P}_n \left(\theta^{(j)}_{\hat{M}_n,n} \in CI^{(j)}_{1-\alpha,\hat{M}_n} \text{ for all } 1 \leq j \leq m(\hat{M}_n) \right) \geq 1-\alpha \tag{32}$$

for *any* model-selection procedure $\hat{M}_n \colon \mathbb{R}^n \to \mathcal{M}_n$.

Generalised PoSI: notation

Let $\theta_n = (\theta^T_{M_1,n}, \ldots, \theta^T_{M_d,n})^T$ and $\hat{\theta}_n = (\hat{\theta}^T_{M_1,n}, \ldots, \hat{\theta}^T_{M_d,n})^T$.

Assume regularity conditions such that

$$d\left\{\text{diag}(V_n)^{\dagger/2}(\hat{\theta}_n - \theta_n), N(0, \text{corr}(V_n))\right\} \to 0 \quad \text{as } n \to \infty. \tag{33}$$

for a sequence of covariances V_n, where $d\{\cdot\}$ is any distance metrising convergence in distribution, and

- $\text{diag}(A)$ is the diagonal matrix sharing the diagonal with A.
- $A^{1/2}$ is the square root of a SPD matrix A.
- $A^{\dagger/2} = (A^{\dagger})^{1/2}$.
- $\text{corr}(A) = \text{diag}(A)^{\dagger/2} A \, \text{diag}(A)^{\dagger/2}$.

Generalised PoSI: notation

Define $K_{1-\alpha}(\Sigma)$ as the $1-\alpha$ quantile of $\|Z\|_\infty$, where $Z \sim N(0, \Sigma)$.

This new "PoSI constant" is based on a Gaussian distribution because the guarantees of this method are asymptotic for fixed p.

The original PoSI constant in the known-σ^2 case ($r = \infty$) can be written as $K_{1-\alpha}(\text{corr}(\Gamma_X))$, where the $|M_i| \times |M_j|$ block of Γ_X is given by

$$X^\dagger_{M_i}(X^\dagger_{M_j})^T, \tag{34}$$

for $M_i, M_j \in \mathcal{M}$.

Generalised PoSI: consistent variance estimation

Theorem

*(**Bachoc et al., Theorem 2.3**) Let $\hat{V}_n$ be a consistent estimator of V_n. Under certain asymptotic conditions on $\hat{\theta}_n$ and $\hat{V}_n$, the intervals*

$$CI^{(j)}_{1-\alpha,M} = \hat{\theta}^{(j)}_{M,n} \pm \sqrt{[\hat{V}_n]_{\rho(M)+j}} K_{1-\alpha}(corr(\hat{V}_n)) \tag{35}$$

are asymptotically valid post-selection $1-\alpha$ confidence intervals for θ_{jM}, where for $M = M_{j,n}$, $\rho(M) = \sum_{l=1}^{j-1} m(M_{l,n})$.

$\rightarrow$ Since guarantees are only aymptotic, the PoSI constant is computed from a multivariate Gaussian instead of a multivariate t distribution.

$\rightarrow$ In a fully non-parametric setting consistent estimation of the variance is rarely possible, so this construction is still of limited use.

Generalised PoSI: variance overestimation

Theorem

*(**Bachoc et al., Theorem 2.5**) Suppose there exists estimators $\hat{v}^2_{j,n}$ of $[V_n]_j$, and an estimator $\hat{K}_n$ of $K_{1-\alpha}(corr(V_n))$ such that*

$$\mathbb{P}\left(\frac{K_{1-\alpha}(corr(V_n))}{\hat{K}_n}\max_j\sqrt{\frac{[V_n]_j}{\hat{v}^2_{j,n}}} > 1+\varepsilon\right) \to 0 \text{ for all } \varepsilon > 0. \tag{36}$$

Then, under the same conditions as before, the intervals

$$CI^{(j)}_{1-\alpha,M} = \hat{\theta}^{(j)}_{M,n} \pm \sqrt{\hat{v}^2_{\rho(M)+j,n}}\hat{K}_n \tag{37}$$

are asymptotically valid post-selection $1-\alpha$ confidence intervals for θ_{jM}.

$\rightarrow$ Estimators $\hat{v}^2_{j,n}$ are possible to construct in a variety of settings.

$\rightarrow$ Widely applicable upper bounds for $K_{1-\alpha}(\text{corr}(V_n))$ are available.

Generalised PoSI: homoskedastic quantitative data

Consider a misspecified version of original PoSI, where

- The true distribution of Y is such that the entries are independent, have identical variance σ_n^2, and

$$\max_{i=1,\ldots,n} \left[\mathbb{E}\left(|Y_i - \mathbb{E}(Y_i)|^{2+\delta}\right)\right]^{\frac{2}{2+\delta}} \leq \tau\sigma_n^2 \tag{38}$$

  for some $\delta > 0$, $\tau \geq 1$.
- The set of working models $\mathcal{M}_n$ for Y contains homoskedastic linear models $\mathbb{E}(Y) = X_M\beta_M$, $M \subseteq \{1, \ldots, p\}$, with p fixed.

Asymptotically valid selective CIs for projection parameters are given by

$$CI_{1-\alpha,M}^{(j)} = \hat{\beta}_{M,n}^{(j)} \pm \sqrt{\hat{\sigma}_{M,n}^2[(X_M^T X_M)^{-1}]_{jj}}K_{1-\alpha}(\text{corr}(\Gamma_X)); \tag{39}$$

$$\hat{\sigma}_{M,n}^2 = \frac{\|(I_n - P_M)Y\|^2}{n - |M|}. \tag{40}$$

Generalised PoSI: homoskedastic quantitative data

The resulting intervals are very similar to original PoSI intervals, but there are two key differences:

- The variance is estimated using the **residual sum of squares from the fit of the selected model**. This will in general overestimate the true variance, but adapts to misspecification.
- Since validity is guaranteed only asymptotically, the PoSI constant is computed from a multivariate Gaussian (rather than t) distribution.

This framework can be applied to heteroskedastic and binary data.

Data example

Diabetes dataset of Hastie and Efron (2012).

442 patients, 10 covariates.

Response: quantitative measure of disease progression.

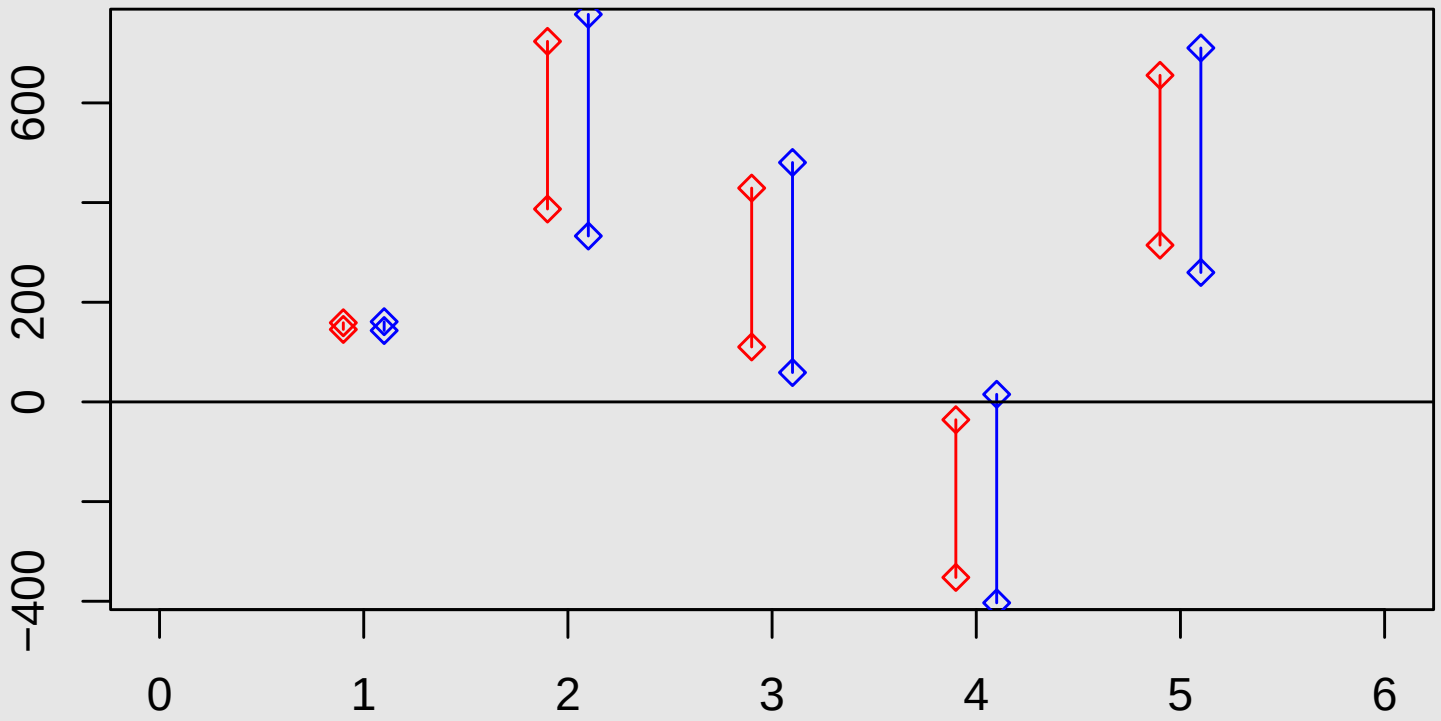
Covariates: age, sex, BMI, blood pressure, and 6 blood serum measurements.

We run the LASSO with penalty selected by cross-validation to identify the most significant predictors, obtaining:

```
BMI MAP HDL LTG
```

Data example

Comparison of unadjusted intervals with PoSI intervals at 95% level.



`HDL` is deemed significant (within the selected model) if the selection step is ignored, but when the appropriate adjustment is made there is no ground for rejection.