Recap

A typical data analysis follows this structure:

1. Obtain data $D = [Y, X]$, from a potentially very complex distribution, with an objective in mind.

2. Statistician performs an exploratory analysis on D to select a working model M to fit and a set of inferential objectives.

3. Model-fitting and inference are carried out using the same data set.

The key idea is that the selection step injects additional sampling variability that needs to be accounted for to ensure repeated-sampling validity of the inferences in the last step.

Recap

For a quantitative response Y, with $\mathbb{E}[Y] = \mu$, we introduced the projection-parameter as the solution of the misspecified least-squares problem,

$$\beta_M = X_M^{\dagger} \mu, \tag{1}$$

which can be thought of as the best approximation of μ in the linear model $\{\mu = X_M \beta^M : \beta^M \in \mathbb{R}^p\}$.

PoSI framework (and extensions) provides valid inference for a selected $\beta_{\hat{M}}$ **universally over all model selection procedures** $\hat{M}$ by expanding the classical confidence intervals via the PoSI constant K, for which

$$P\left(\left|\frac{\hat{\beta}_{j\hat{M}} - \beta_{j\hat{M}}}{\hat{\sigma}[(X_{\hat{M}}^T X_{\hat{M}})^{-1}_{jj}]^{1/2}}\right| \leq K \ \forall\, j = 1, \ldots, |\hat{M}|\right) \geq 1 - \alpha. \tag{2}$$

This lecture

More powerful selective inference:

- Zrnic, Jordan (2022). "Post-selection inference via algorithmic stability", *arXiv:2011.09462*.

Selective inference with a random design:

- Kuchibhotla, Brown, Buja, Cai, George, Zhao (2020). "Valid post-selection inference in model-free linear regression", *The Annals of Statistics*.
- Rinaldo, Wasserman, G'Sell (2019). "Bootstrapping and sample splitting for high-dimensional, assumption-free inference", *The Annals of Statistics*.

More powerful selective inference

PoSI is necessarily conservative because it is required to protect against *any* selection rule, even those which make little sense.

To remedy this, in addition to restricting the model space, one can decrease the size of the PoSI confidence regions by limiting the set of allowed selection rules.

Motivated by ideas from differential privacy, a powerful restriction comes via the notion of **algorithmic stability**.

Loosely speaking, a selection rule is stable if it is **not very sensitive to the data**.

Broadly, the idea is that, for a stable selection rule, the degree of dependence between model selection and inference is bounded, and therefore the selection bias can be controlled.

More powerful selective inference

Stability is a property of **randomised selection rules**.

The output of a randomised selection rule $\hat{M}$, $\hat{M}(y)$, is a *random variable* on the power set of $\{1, \ldots, p\}$.

In general, starting from a deterministic rule $\hat{M}\colon \mathbb{R}^n \to \mathcal{M}$, $\mathcal{M} \subseteq 2^{\{1,\ldots,p\}}$, a randomised version can be obtained by applying it to a *noisy* version of the data, such as $\tilde{Y} = Y + W$, where W is artificial noise from a known distribution generated by the statistician.

Conditionally on $Y = y$, $\hat{M}(y + W)$ is a random variable whose output is more or less similar to $\hat{M}(y)$ depending on the distribution of W.

In general, stability is a property of randomised algorithms (not necessarily model selection rules), $\mathcal{A}\colon \mathbb{R}^n \to \mathcal{S}$, where $\mathcal{S}$ is a set of random variables.

More powerful selective inference

Closeness between the output of two randomised algorithms is formalised via the concept of *indistinguishability* (aka *max-divergence*).

Definition

A random variable A is (η, τ)-**indistinguishable** from a random variable B, denoted $A \approx_{\eta,\tau} B$, if for all measurable sets E,

$$P(A \in E) \leq e^{\eta}\mathbb{P}(B \in E) + \tau. \tag{3}$$

The parameters are $\eta > 0$ and $\tau \in [0, 1]$.

$\rightarrow$ In the context of differential privacy, η is a user-specified parameter that controls the trade-off between security and accuracy, and τ is the probability of having a security breach.

$\rightarrow$ In selective inference, we think of η as **balancing the information** allocated to selection and inference respectively, and of τ as being proportional to the **miscoverage probability** of a confidence set.

More powerful selective inference

With this notion in mind, we define a stable randomised algorithm as one for which, for most samples y, we can "guess" the distribution of $\mathcal{A}(y)$ (given y) knowing only the underlying distribution P but not the data y.

Definition

Let $\mathcal{A}$ be a randomised algorithm. We say it is (η, τ, ν)-**stable** with respect to a distribution P on $\mathbb{R}^n$ if there exists a random variable A_0, possibly depending on P, such that

$$P\left(y \in \mathbb{R}^n \colon \mathcal{A}(y) \approx_{\eta,\tau} A_0\right) \geq 1 - \nu. \tag{4}$$

When the algorithm is a model-selection rule, we write $\mathcal{A}$ and A_0 as $\hat{M}$ and M_0.

More powerful selective inference

Crucially, stability entails the existence of a random variable M_0, which can be thought of as an **oracle selection rule**, which is **independent of the data**, and therefore does not generate selection bias.

So, if we had access to M_0, we could use it instead and report classical confidence intervals for the selected parameter.

The following result makes this intuition rigorous.

Lemma

Let $\hat{M}$ be (η, τ, ν)-stable selection rule and M_0 be the corresponding oracle rule. Then

$$[Y, \hat{M}(Y)] \approx_{\eta, \tau+\nu} [Y, M_0], \tag{5}$$

with Y and M_0 independent.

More powerful selective inference

Consider a selective inference problem with data $Y \in \mathbb{R}^n$ and model-dependent targets of inference $\{\beta_M : M \in \mathcal{M}\}$.

Suppose that we can find a collection $\{\hat{R}^\alpha_M(\cdot) : M \in \mathcal{M}, \alpha \in (0,1)\}$ such that

$$P\left(\beta_M \in \hat{R}^\alpha_M(Y)\right) \geq 1 - \alpha \text{ for all } M \in \mathcal{M}, \alpha \in (0,1). \tag{6}$$

Theorem

Fix $\delta \in (0,1)$ and let $\hat{M}$ be a (η, τ, ν)-stable selection algorithm. Then

$$P\left\{\beta_{\hat{M}} \in \hat{R}^{\delta e^{-\eta}}_{\hat{M}}(Y)\right\} \geq 1 - (\delta + \tau + \nu). \tag{7}$$

$\rightarrow$ Under stability, we can effectively ignore selection and report classical intervals with a "selection-adjusted" nominal coverage.

For a given a (δ, τ, ν) and a desired coverage $1 - \alpha$, the adjusted nominal coverage needs to be $\tilde{\alpha} = (\alpha - \tau - \nu)e^{-\eta}$.

More powerful selective inference

Proof.

By the lemma,

$$\begin{aligned} P\left\{\beta_{\hat{M}} \notin \hat{R}_{\hat{M}}^{\delta e^{-\eta}}(Y)\right\} &\leq e^{\eta} P\left\{\beta_{M_0} \notin \hat{R}_{M_0}^{\delta e^{-\eta}}(Y)\right\} + \tau + \nu \\ &= e^{\eta}\mathbb{E}\left[P\left\{\beta_{M_0} \notin \hat{R}_{M_0}^{\delta e^{-\eta}}(Y) \mid M_0\right\}\right] + \tau + \nu. \end{aligned}$$

By construction,

$$P\left\{\beta_{M_0} \notin \hat{R}_{M_0}^{\delta e^{-\eta}}(Y) \mid M_0\right\} \leq \delta e^{-\eta}, \tag{8}$$

and therefore

$$P\left\{\beta_{\hat{M}} \notin \hat{R}_{\hat{M}}^{\delta e^{-\eta}}(Y)\right\} \leq e^{\eta} e^{-\eta}\delta + \tau + \nu = \delta + \tau + \nu. \tag{9}$$

□

More powerful selective inference

In the PoSI framework, where $Y \sim N(\mu, \sigma^2 I_n)$ and $\beta_M = X_M^\dagger \mu$, these confidence regions specialise to the joint confidence intervals

$$\left[\hat{\beta}_{j\hat{M}} \pm K_{\hat{M}}(\delta e^{-\tau})[(X_{\hat{M}}^T X_{\hat{M}})^{-1}]_{jj}^{1/2}\hat{\sigma}\right], \quad j = 1, \ldots, |\hat{M}| \tag{10}$$

where $\hat{\sigma}$ is a suitable estimator of σ, and $K_M(\alpha)$ is a model-dependent PoSI constant: the minimum value of K satisfying

$$P\left(\max_{1\leq j\leq |M|}\left|\frac{\hat{\beta}_{jM} - \beta_{jM}}{[(X_M^T X_M)^{-1}]_{jj}^{1/2}\hat{\sigma}}\right| \leq K\right) \geq 1 - \alpha. \tag{11}$$

More powerful selective inference

In many cases, stable selection rules can be derived from existing ones via addition of Laplace noise.

Let $Y \sim N(\mu, \sigma^2 I_n)$ for a known $\sigma > 0$, and suppose that the selection rule is a function of the randomised algorithm $\mathcal{A}(Y) = a^T Y + W$ for some fixed $a \in \mathbb{R}^n$, where

$$W \sim \text{Laplace}\left(\frac{z_{1-\nu/2}\sigma\|a\|}{\eta}\right), \tag{12}$$

This is, $\hat{M} = M \iff \mathcal{A}(Y) \in E_M$ for some $E_M \subseteq \mathbb{R}$.

We have that $\hat{M}$ is $(\eta, 0, \nu)$-stable.

More powerful selective inference

To see this, let $\mathcal{Y}_\nu = \{x \in \mathbb{R}^n : |a^T x - a^T \mu| \leq z_{1-\nu/2}\sigma\|a\|\}$, so that

$$P(\mathcal{Y}_\nu) = P(|N(0,1)| \leq z_{1-\nu/2}) = 1 - \nu. \tag{13}$$

Consider the oracle randomised algorithm $\mathcal{A}(\mu) = a^T\mu + W$, which uses the true unknown mean.

Since the density ratio of a Laplace(b) and a u+Laplace(b) distribution is upper bounded by $e^{|u|/b}$, we have, for all $x \in \mathcal{Y}_\nu$ and all measurable E,

$$\frac{P\{\mathcal{A}(x) \in E\}}{P\{\mathcal{A}(\mu) \in E\}} \leq e^{\eta}. \tag{14}$$

Therefore, $\mathcal{A}$ and any function of it are $(\eta, 0, \nu)$-stable.

More powerful selective inference

From a basic stable rule such as the previous one, more sophisticated stable rules can be derived using the following two properties.

Postprocessing:

If $\mathcal{A}$ is (η, τ, ν)-stable, then any composition $\mathcal{B} \circ \mathcal{A}$ is also (η, τ, ν)-stable.

For example, if we have a stable version of the lasso algorithm, with maximiser $\hat{\beta}$, the selection rule $\hat{M} = \{1 \leq j \leq p : \hat{\beta}_j \neq 0\}$ is also stable.

Composition:

Consider a sequence of algorithms $\mathcal{A}_t : \mathcal{S}_1 \times \cdots \times \mathcal{S}_{t-1} \times \mathbb{R}^n \to \mathcal{S}_t$ for $t = 1, \ldots, k$. If each algorithm is stable, their composition is also stable, and the stability parameters can be written as a function of the stability parameters of the $\mathcal{A}_t$'s.

In its most simple form, if each algorithm is $(\eta, 0, 0)$-stable, their composition is $(k\eta, 0, 0)$-stable.

For example, this can be applied to stepwise algorithms if each step is known to be stable.

More powerful selective inference

As an example, consider a stable version of the lasso algorithm.

For data $[Y, X] \in \mathbb{R}^n \times \mathbb{R}^{n \times p}$, the (standard) lasso estimator is the solution of

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{2} \| Y - X\beta \|_2^2 \text{ s.t. } \|\beta\|_1 \leq C, \tag{15}$$

for a chosen L_1 restriction penalty $C > 0$.

The stable lasso is obtained by running a numerical optimisation routine for (15) adding Laplacian noise at each step.

More powerful selective inference

The tuning parameters of the algorithm are the *penalty* C, the *number of steps* k, and the *stability-inducing parameters* δ and η. Furthermore, let $\{e_i\}_{i=1}^p$ be the canonical basis of $\mathbb{R}^p$. For theoretical guarantees, we require $Y \sim N(\mu, \sigma^2 I_n)$ and assume we have an estimator $\hat{\sigma}$ of σ.

Stable lasso: Set $\beta^{(1)} = 0$. For $t = 1, \ldots, k$:

1. For all $\phi \in C \cdot \{\pm e_i\}_{i=1}^p$, sample

$$W_{t,\phi} \sim \text{Laplace}\left(\frac{4t_{r,1-\delta/(2d)} C \|X\|_{2,\infty}}{\eta n}\right), \|X\|_{2,\infty} = \max_i \|X_i\|_2. \tag{16}$$

2. For all $\phi \in C \cdot \{\pm e_i\}_{i=1}^p$, let $\alpha_\phi = -\frac{2}{n\hat{\sigma}} \phi^T X^T (Y - X\beta_t) + W_{t,\phi}$.
3. Set $\phi_t = \arg\min \alpha_\phi$.
4. Set $\beta^{(t+1)} = (1 - \Delta_t)\beta^{(t)} + \Delta_t \phi_t$, where $\Delta_t = 2/(t+1)$.

Report $\hat{\beta} = \beta^{(k+1)}$.

More powerful selective inference

This is essentially a randomised version of the classical Frank-Wolfe algorithm for the standard lasso which uses the stability of the additive Laplace noise as a building block. We have the following result.

Theorem

Assume that $Y \sim N(\mu, \sigma^2 I_n)$ *and* $\hat{\sigma}^2 \sim \sigma^2 \chi^2_r / r$. *The randomised lasso algorithm is both*

- $(k\eta^2/2, \sqrt{2k \log(1/\delta)}, \delta)$-*stable.*
- $(k\eta, 0, \delta)$-*stable.*

By the post-processing property, any model-selection procedure based on the output of this algorithm, such as

$$\hat{M} = \{j \in \{1, \ldots, p\} : \hat{\beta}_j \neq 0\}, \tag{17}$$

is stable.

$\rightarrow$ Similar results hold for stable versions of other well-established procedures such as marginal screening.

Fixed vs. random design

So far we have assumed that the design X is **fixed**.

This makes modelling and computation easier.

Conceptually, it can be justified in one of the following cases:

- The covariate values are fixed by the experimenter.
- Ancillarity argument: X is random but its distribution is independent of any parameter of interest. Then, by the *Conditionality Principle*, inference ought to be provided conditionally on the observed X.

The latter assumption is often sensible when the working model is well-specified, but is problematic under misspecification.

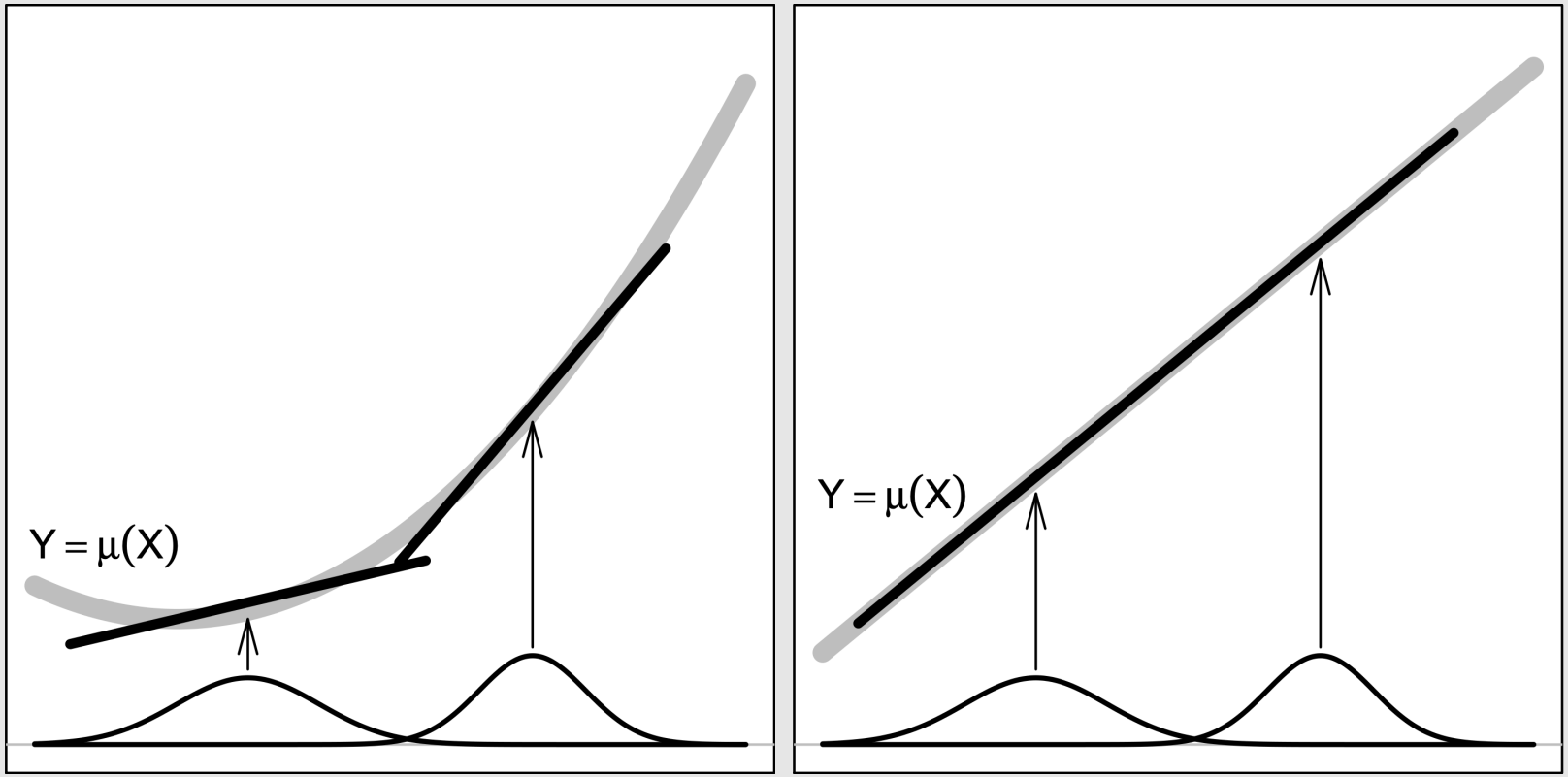


Figure: Buja et al. (2020). "Models as approximations I: consequences illustrated with linear regression".

$\rightarrow$ The best linear approximation of the true regression function $\mu(X) = \mathbb{E}[Y \mid X]$ **depends on the distribution of X**.

In view of this, some authors have approached the selective inference problem from a random-design perspective.

Two main proposals:

- UPoSI: reformulation of the original PoSI framework in the random-X setting.

- Data splitting: basing model selection and inference on independent subsamples of the data. Inferential step is thereby unaffected by selection, allowing usage of standard statistical techniques.

The inferential target

Let $(Y_i, X_i^T)^T \in \mathbb{R} \times \mathbb{R}^p$, $i = 1, \ldots, n$ be n samples of observations, **independent but not necessarily identically distributed**, and set $Y = (Y_1, \ldots, Y_n)^T$ and $X = [X_1, \ldots, X_p]^T$.

For a model $M \subseteq \{1, \ldots, p\}$ define the population and empirical risks of the squared loss:

$$R_n(\theta; M) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\{Y_i - X_i(M)^T \theta\}^2\right] \tag{18}$$

$$\hat{R}_n(\theta; M) = \frac{1}{n} \sum_{i=1}^{n} \{Y_i - X_i(M)^T \theta\}^2, \tag{19}$$

where $X_i(M) = (X_{ij})_{j \in M} \in \mathbb{R}^{|M|}$, and their respective minimisers

$$\beta_{n,M} = \arg \min_{\theta \in \mathbb{R}^{|M|}} R_n(\theta; M), \quad \hat{\beta}_{n,M} = \arg \min_{\theta \in \mathbb{R}^{|M|}} \hat{R}_n(\theta; M). \tag{20}$$

The inferential target

Define the full-model second order matrices and vectors

$$\Sigma_n = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[X_i X_i^T\right], \qquad \hat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^T \tag{21}$$

$$\Gamma_n = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[X_i Y_i^T\right], \qquad \hat{\Gamma}_n = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i^T \tag{22}$$

The analogous quantities for a given model M are defined with $X(M)$ instead of X. They are simply submatrices and subvectors of the full-model versions.

The least squares parameter and its estimator satisfy

$$\hat{\Sigma}_n(M)\hat{\beta}_{n,M} = \hat{\Gamma}_n(M) \text{ and } \Sigma_n(M)\beta_{n,M} = \Gamma_n(M). \tag{23}$$

The inferential target

Some remarks:

- $\beta_{n,M}$ is the best linear predictor of Y using $X(M)$ relative to the squared loss.
- The normal equations are written implicitly because the framework allows for non-unique minimisers. In such cases the corresponding confidence regions will contain a subspace of $\mathbb{R}^p$.
- For a fixed M,

$$\sqrt{n}\{\hat{\beta}_{n,M} - \beta_{n,M}\} \xrightarrow{d} N(0, V_M) \tag{24}$$

  for some matrix $V_M > 0$ under mild conditions.
- In general, for $M \subseteq M'$ it does not hold that $\beta_{n,M}$ is a subvector of $\beta_{n,M'}$ due to dependence among the columns of X.

The inferential target

As usual focus on inference via confidence regions.

Asymptotic normality lends itself for the construction of regions $\hat{R}_{n,M} \subseteq \mathbb{R}^{|M|}$ such that

$$\liminf_{n\to\infty} P\left(\beta_{n,M} \in \hat{R}_{n,M}\right) \geq 1-\alpha. \tag{25}$$

UPoSI confidence regions satisfy instead

$$\liminf_{n\to\infty} P\left(\beta_{n,\hat{M}} \in \hat{R}_{n,\hat{M}}\right) \geq 1-\alpha \tag{26}$$

for an arbitrary measurable $\hat{M}\colon \mathbb{R}^n \times \mathbb{R}^{n\times p} \to \mathcal{M}$, where $\mathcal{M}$ is the power set of $\{1,\ldots,p\}$.

Selective inference as simultaneous inference

Theorem

For any set of confidence regions $\{\hat{R}_{n,M} : M \in \mathcal{M}\}$ *and* $0 < \alpha < 1$,

$$P\left(\beta_{n,\hat{M}} \in \hat{R}_{n,\hat{M}}\right) \geq 1 - \alpha \text{ for all } \hat{M}: P(\hat{M} \in \mathcal{M}) = 1 \tag{27}$$

if and only if

$$P\left(\bigcap_{M \in \mathcal{M}} \left\{\beta_{n,M} \in \hat{R}_{n,M}\right\}\right) \geq 1 - \alpha. \tag{28}$$

$\rightarrow$ We can solve the selective inference problem (27) by solving the simultaneous inference problem (28).

For asymptotic validity of the confidence regions, require

$$\liminf_{n \to \infty} P\left(\bigcap_{M \in \mathcal{M}} \left\{\beta_{n,M} \in \hat{R}_{n,M}\right\}\right) \geq 1 - \alpha. \tag{29}$$

Selective inference as simultaneous inference

Proof.

For $M \in \mathcal{M}$, define the event $A_M = \left\{\beta_{n,M} \in \hat{R}_{n,M}\right\}$.

(28) $\Rightarrow$ (27): It is clear that

$$\bigcap_{M \in \mathcal{M}} A_M \subseteq A_{\hat{M}} \text{ for all } \hat{M}.$$

(27) $\Rightarrow$ (28): Let $\hat{M}$ be such that $\hat{M} \in \arg\min_{M \in \mathcal{M}} \mathbf{1}\{A_M\}$, so that $\mathbf{1}\{A_{\hat{M}}\} = \min_{M \in \mathcal{M}} \mathbf{1}\{A_M\}$. It follows that

$$\bigcap_{M \in \mathcal{M}} A_M = A_{\hat{M}}.$$

Selective inference as simultaneous inference

Worst possible selection rule:

The second part of the proof shows existence of an "adversarial selection rule" that picks a model if the corresponding parameter is not covered by the confidence region.

As in the fixed-X case, this selection rule is not actionable, as it depends on the population parameter.

Inherent high-dimensionality:

In view of the theorem, selective inference requires inference for a very large number of parameters $\beta_{n,M}$.

For instance, if $p < n$, there are $p2^{p-1}$ parameters in the various submodels.

UPoSI confidence regions

Recall the second-order arrays and their estimators:

$$\Sigma_n = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[X_i X_i^T\right], \qquad \hat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^T \tag{30}$$

$$\Gamma_n = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[X_i Y_i^T\right], \qquad \hat{\Gamma}_n = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i^T \tag{31}$$

and their submodel versions $\Sigma_n(M), \hat{\Sigma}_n(M), \Gamma_n(M), \hat{\Gamma}_n(M)$, and introduce the estimation errors

$$\mathcal{D}_n^{\Sigma} = \|\hat{\Sigma}_n - \Sigma_n\|_\infty = \max_{M\in\mathcal{M}_2} \|\hat{\Sigma}_n(M) - \Sigma_n(M)\|_\infty \tag{32}$$

$$\mathcal{D}_n^{\Gamma} = \|\hat{\Gamma}_n - \Gamma_n\|_\infty = \max_{M\in\mathcal{M}_1} \|\hat{\Gamma}_n(M) - \Gamma_n(M)\|_\infty, \tag{33}$$

where $\mathcal{M}_k = \{M\colon |M| \le k\}$.

UPoSI confidence regions

The UPoSI confidence regions are

$$\hat{R}_{n,M} = \{\theta \in \mathbb{R}^{|M|} \colon \|\hat{\Sigma}_n\{\hat{\beta}_{n,M} - \theta\}\|_\infty \leq C_n^\Gamma(\alpha) + C_n^\Sigma(\alpha)\|\theta\|_1\} \tag{34}$$

$$\hat{R}^\dagger_{n,M} = \{\theta \in \mathbb{R}^{|M|} \colon \|\hat{\Sigma}_n\{\hat{\beta}_{n,M} - \theta\}\|_\infty \leq C_n^\Gamma(\alpha) + C_n^\Sigma(\alpha)\|\hat{\beta}_{n,M}\|_1\} \tag{35}$$

where $C_n^\Gamma(\alpha)$ and $C_n^\Sigma(\alpha)$ are bivariate joint upper α quantiles of $\mathcal{D}_n^\Gamma$ and $\mathcal{D}_n^\Sigma$:

$$P\left(\mathcal{D}_n^\Gamma \leq C_n^\Gamma(\alpha) \text{ and } \mathcal{D}_n^\Sigma \leq C_n^\Sigma(\alpha)\right) \geq 1 - \alpha. \tag{36}$$

Note:

- $\hat{R}_{n,M}$ has finite-sample guarantees and requires less assumptions, but it is more difficult to interpret and analyse.
- By contrast, $\hat{R}^\dagger_{n,M}$ provides asymptotic guarantees and requires additional assumptions, but they are more transparent.

UPoSI confidence regions

Some considerations:

- Bivariate quantiles, and therefore the confidence regions, are not unique. One may marginally increase one and decrease the other suitably, maintaining the bivariate coverage probability.
- The quantiles need to be estimated from the data. A bootstrap approach can be used to this end.
- Under mild conditions, $\max\{C_n^{\Gamma}(\alpha), C_n^{\Sigma}(\alpha)\} \to 0$ as $n \to \infty$.
- As opposed to fixed-X PoSI, there is no gain here from using sparse model spaces. This is because the estimation errors depend only on models of sizes 1 and 2.

UPoSI confidence regions

Theorem

The confidence regions $\{\hat{R}_{n,M} : M \in \mathcal{M}\}$, *where*

$$\hat{R}_{n,M} = \{\theta \in \mathbb{R}^{|M|} : \|\hat{\Sigma}_n\{\hat{\beta}_{n,M} - \theta\}\|_\infty \leq C_n^{\Gamma}(\alpha) + C_n^{\Sigma}(\alpha)\|\theta\|_1\}, \quad (37)$$

satisfy

$$P\left(\bigcap_{M\in\mathcal{M}} \left\{\beta_{n,M} \in \hat{R}_{n,M}\right\}\right) \geq 1 - \alpha. \quad (38)$$

An analogous result holds for the alternative regions $\hat{R}^{\dagger}_{n,M}$ under extra regularity conditions. Note:

- If $|M| > n$, $\hat{\Sigma}_n$ is singular, and $\hat{R}_{n,M}$ contains a nontrivial affine subspace of $\mathbb{R}^p$.
- Independence is not required (though it is needed for quantile estimation).

UPoSI confidence regions

Proof.

The proof is based on deterministic inequalities. Start from

$$\hat{\Sigma}_n(M)(\hat{\beta}_{n,M} - \beta_{n,M}) + (\hat{\Sigma}_n(M) - \Sigma_n(M))\beta_{n,M} = \hat{\Gamma}_n(M) - \Gamma_n(M).$$

The triangle inequality gives

$$\begin{aligned}
\|\hat{\Sigma}_n(M)(\hat{\beta}_{n,M} - \beta_{n,M})\|_\infty & \\
&\leq \|\hat{\Gamma}_n(M) - \Gamma_n(M)\|_\infty + \|(\hat{\Sigma}_n(M) - \Sigma_n(M))\beta_{n,M}\|_\infty \\
&\leq \|\hat{\Gamma}_n(M) - \Gamma_n(M)\|_\infty + \|\hat{\Sigma}_n(M) - \Sigma_n(M)\|_\infty \|\beta_{n,M}\|_1 \\
&\leq \|\hat{\Gamma}_n - \Gamma_n\|_\infty + \|\hat{\Sigma}_n - \Sigma_n\|_\infty \|\beta_{n,M}\|_1.
\end{aligned}$$

Therefore

$$\mathbb{P}\left(\bigcap_{M \in \mathcal{M}} \left\{ \|\hat{\Sigma}_n(M)(\hat{\beta}_{n,M} - \beta_{n,M})\|_\infty \leq \mathcal{D}_n^\Gamma + \mathcal{D}_n^\Sigma \|\beta_{n,M}\|_1 \right\} \right) = 1. \quad (39)$$

Substituting the errors by the quantiles gives the result. □

UPoSI confidence regions

Some important remarks:

- A bootstrap algorithm can be used to estimate the error quantiles such that the confidence sets are asymptotically valid under the quasi-exponential asymptotic regime $\log(p)^7 = o(n)$.
- The UPoSI confidence sets are not hyperrectangles, and therefore do not immediately imply marginal confidence intervals for each component, but there exists algorithms for constructing the smallest covering hyperrectangle for the sets.
- This method can be applied also in the fixed-X setting and can in fact produce smaller, though less interpretable, regions.

UPoSI confidence regions

Simulation study with fixed-X:[1]

Comparison of PoSI, UPoSI and UPoSI smallest hyperrectangle.

The data-generating model is $Y_i = X_i^T \beta + \varepsilon_i$, with $\beta = (0, \ldots, 0)^T$ and $\varepsilon \sim N(0, 1)$ independently, $n = 200$, $p = 15$.

Three covariate settings considered:

A. *Orthogonal design*: $\hat{\Sigma}_n = I_p$.

B. *Exchangeable design*: $\hat{\Sigma}_n = I_p - (p+2)^{-1} \mathbf{1}_p \mathbf{1}_p^T$.

C. *Worst-case design*:

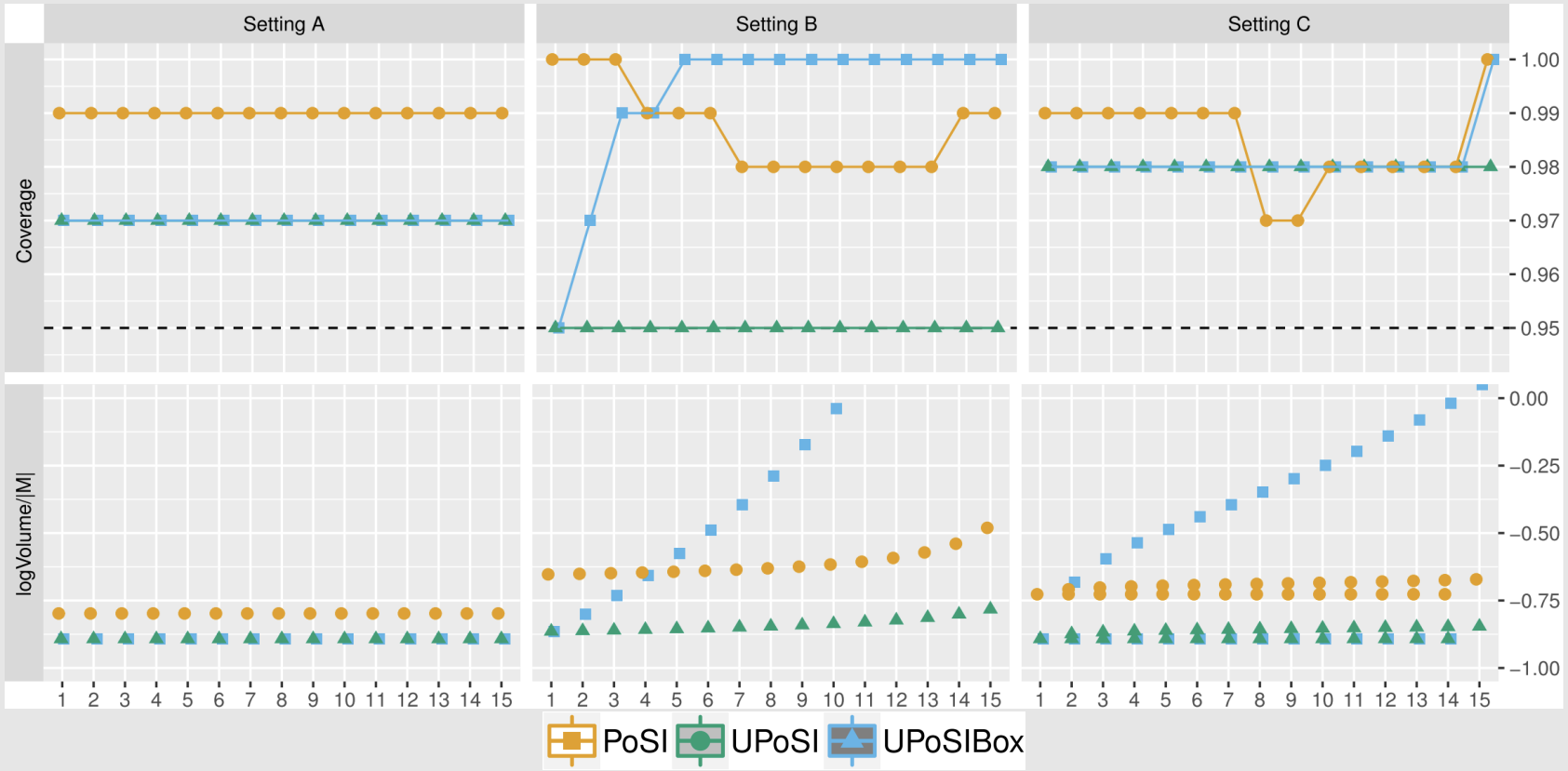
$$\hat{\Sigma}_n = \begin{bmatrix} I_{p-1} & c\mathbf{1}_{p-1} \\ \mathbf{0}_{p-1}^T & \sqrt{1/2} \end{bmatrix}. \tag{40}$$

Settings A and B are theoretically optimal for the original PoSI approach, while setting C leads to the largest PoSI constant.

[1] From Kuchibhotla, Brown, Buja, Cai, George, Zhao (2020). "Valid post-selection inference in model-free linear regression", *The Annals of Statistics*.

UPoSI confidence regions

Figure: Simulation results for 100 replications. Nominal coverage set at 95%. Plot shows simultaneous coverage and size of the confidence sets for different model sizes $1 \leq |M| \leq 15$. In setting C, models containing the last covariate produce significantly larger regions than those which do not.



Data splitting

An alternative, "obvious" approach to selective inference with a random design is sample splitting, whereby only a subset of the observations is used for model selection and the rest are used for inference.

Assume we have two subsets of IID observations

$$D_1 = \{(Y_i, X_i^T)^T : i = 1, \ldots, n\}, D_2 = \{(Y_i, X_i^T)^T : i = n+1, \ldots, 2n\}.$$

Since the data is IID, $\beta_{n,M}$ is independent of n, so we shall write β_M.

Furthermore, we restrict $\mathcal{M}$ to the set of M's such that the least squares problem has a unique solution, so that

$$\beta_M = \Sigma(M)^{-1}\Gamma(M),$$

where $\Sigma(M) = \mathbb{E}[X(M)X(M)^T]$ and $\Gamma(M) = \mathbb{E}[X(M)Y^T]$, and similarly for $\hat{\beta}_M$.

Data splitting

Clearly, if $\{\hat{R}_M : M \in \mathcal{M}\}$ are valid confidence sets for β_M for fixed M, they are also valid under selection for any selection rule $\hat{M}$, as

$$\mathbb{P}\left(\beta_{\hat{M}} \in \hat{R}_{\hat{M}}\right) = \sum_{M \in \mathcal{M}} \mathbb{P}\left(\beta_M \in \hat{R}_M \mid \hat{M} = M\right) \mathbb{P}\left(\hat{M} = M\right) \tag{41}$$

$$= \sum_{M \in \mathcal{M}} \mathbb{P}\left(\beta_M \in \hat{R}_M\right) \mathbb{P}\left(\hat{M} = M\right) \geq 1 - \alpha. \tag{42}$$

Trivially, they are also valid conditionally on the selected model, $\hat{M} = M$.

For a fixed M, standard inferential procedures can be used to construct asymptotic non-parametric confidence sets.

Data splitting: confidence sets based on asymptotic normality

Under suitable conditions, $\sqrt{n}\left\{\beta_M - \hat{\beta}_M\right\} \xrightarrow{d} N(0, V_M)$ for a positive definite matrix V_M.

The confidence sets based on a normal approximation are

$$\hat{R}_M = \left\{\beta \in \mathbb{R}^{|M|} : \|\beta - \hat{\beta}_M\|_\infty \leq \frac{t_\alpha}{\sqrt{n}}\right\}, \tag{43}$$

where t_α is such that

$$\mathbb{P}\left(\|\hat{V}_M^{1/2} Z\|_\infty \leq t_\alpha\right) = \alpha, \tag{44}$$

Z follows an $|M|$-dimensional standard Gaussian distribution independent of the data, and $\hat{V}_M$ is a plug-in estimator of V_M.

Under certain distributional assumptions and bounded model spaces $\mathcal{M}_n$ with $\max_{M \in \mathcal{M}_n} |M| = o(n^{1/5})$, we have, for all $\hat{M}$,

$$\liminf_{n \to \infty} \mathbb{P}(\beta_{\hat{M}} \in \hat{R}_M) \geq 1 - \alpha. \tag{45}$$

Data splitting: bootstrap confidence sets

Evaluation of $\hat{V}_M$ and computation of the quantile t_α turns out to be computationally demanding in some settings.

Alternatively, let t^*_α be the smallest positive number such that

$$\mathbb{P}\left(\sqrt{n}\|\hat{\beta}^*_M - \hat{\beta}_M\|_\infty \leq t^*_\alpha \mid D_2\right) \geq 1-\alpha, \tag{46}$$

where $\hat{\beta}^*_M$ is a (standard) bootstrap copy of $\hat{\beta}_M$ using D_2.

The bootstrap confidence sets are defined as

$$\hat{R}^*_M = \left\{\beta \in \mathbb{R}^{|M|} : \|\beta - \hat{\beta}_M\|_\infty \leq \frac{t^*_\alpha}{\sqrt{n}}\right\}. \tag{47}$$

These confidence sets are asymptotically valid under similar conditions as the previous ones.

Data splitting

We have presented these constructions for the LS parameter β_M, but the framework can be applied to other type of inferential targets:

- The **Leave out covariate inference (LOCO)** parameters, which measure the importance of the selected covariates:

$$\gamma_j(M) = \mathbb{E}\left[|Y - \hat{\beta}_{M_j}^T X(M_j)| - |Y - \hat{\beta}_M^T X(M)|\right], \tag{48}$$

  where $\hat{\beta}_M$ is any estimator of β_M, and M_j and $\hat{\beta}_{M_j}$ are obtained by re-running model selection and estimation after removing the j-th covariate from the data.

- The **prediction parameter**, which measures how well the selected model will predict future observations:

$$\rho_M = \mathbb{E}\left[|Y - \hat{\beta}_M^T X(M)|\right]. \tag{49}$$

Data splitting

Some important remarks about data splitting:

- The main drawback is that it relies on the practitioner committing themselves to look only at a subset of the data during the selection stage. Thus, it does not protect against the bad practice of selecting the most convenient split and not acknowledging it.
- Ignoring some observations in the selection stage might even by inadvisable; e.g. if selection is the main goal or if one wants to study the stability of the selection procedure with respect to the split.
- It can sometimes have very little power in both stages (more powerful splitting strategies will be discussed later).
- The arbitrariness of the data split can be unsettling: two different statisticians using the same dataset but different splits can end up providing inference for different parameters.
- It requires identically distributed observations, while UPoSI does not.