This lecture

The conditional approach to selective inference:

- Motivation and main ideas: Fithian, Sun, Taylor (2017). "Optimal inference after model selection", *arXiv:1410.2597*.

- Conditional inference with affine selection rules: Lee, Sun, Sun, Taylor (2016). "Exact post-selection inference, with application to the lasso", *The Annals of Statistics*.

- Selective inference for clustering: Gao, Bien, Witten (2022). "Selective inference for hierarchical clustering", *Journal of the American Statistical Association*.

The conditional approach

So far we have explored construction of confidence regions

$$\{R_M(Y)\colon M \in \mathcal{M}\} \tag{1}$$

satisfying (at least approximately)

$$P\left(\beta_{\hat{M}} \in R_{\hat{M}}(Y)\right) \geq 1-\alpha, \quad \hat{M}\colon \mathbb{R}^n \to \mathcal{M}. \tag{2}$$

Alternatively, one might require the confidence regions to be **valid conditionally on the selected model**, that is, to satisfy

$$P\left(\beta_{\hat{M}} \in R_{\hat{M}}(Y) \mid \hat{M} = M\right) \tag{3}$$

$$= P\left(\beta_M \in R_M(Y) \mid \hat{M} = M\right) \geq 1-\alpha, \tag{4}$$

where the **selection event** $\{\hat{M} = M\} = \{y \in \mathbb{R}^n\colon \hat{M}(y) = M\}$ is formed by all the data points that would have led to selection of the same model as the data actually observed.

The conditional approach

We shall refer to this approach to selective inference as the **conditional approach**.

For each $M \in \mathcal{M}$, denote by $E_M = \{\hat{M} = M\}$ the selection event.

Clearly, inferential procedures with conditional guarantees are also valid unconditionally, as

$$P\left(\beta_{\hat{M}} \in R_{\hat{M}}(Y)\right) = \sum_{M \in \mathcal{M}} P(E_M) P\left(\beta_M \in R_M(Y) \mid E_M\right) \geq 1 - \alpha, \quad (5)$$

but the reverse is not true in general. We can see this very clearly in the following example.

The conditional approach

Example

(*Selected mean problem*) Let $Y_i \sim N(\mu_i, 1)$, $i = 1, \ldots, n$, independently, and define the data-dependent parameter of interest $\psi = \mu_I$, where $I = \arg\max\{Y_i\}$, i.e. inference is provided for the mean of the maximum observation.

The standard PoSI intervals $R_I = [Y_I \pm K]$, where K satisfies

$$\mathbb{P}\left(\max_{i=1,\ldots,n} |Y_i - \mu_i| \leq K\right) = 1 - \alpha, \tag{6}$$

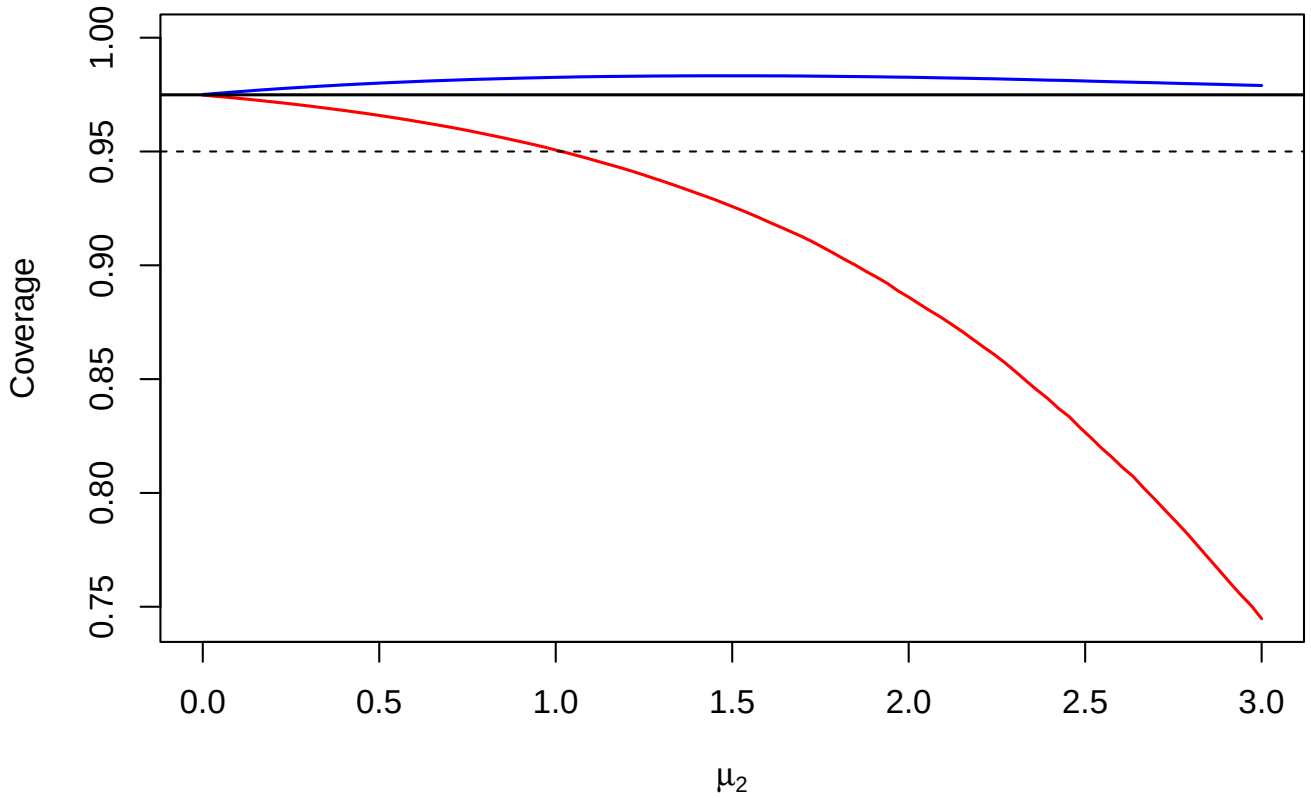
are unconditionally valid $1 - \alpha$ selective CIs.

However, conditionally on having selected one of the means, the coverage can be well below the nominal one.

The conditional approach

Example

Figure: Coverages of 95% unconditional selective CIs with $n = 2$ samples, for $\mu_1 = 0$ and $\mu_2 \in [0, 3]$. In black, unconditional coverage; in red, coverage conditionally on selecting μ_1; in blue, coverage conditionally on selecting μ_2.



The conditional approach

The conditional approach asserts that the answer to an inferential question selected with the data must be valid **given that the question was asked**.

This ensures that, for example, $(1-\alpha)\%$ of the confidence intervals reported for a specific parameter β_M contain the true value β_M^0.

Abstractly, we can think about the conditional approach as a form of sample splitting: the component of the data used for selection is the random variable $Z = \mathbf{1}(Y \in E_M)$, and the data used for inference is $Y \mid Z = 1$.

$\longrightarrow$ Conditioning on selection effectively **discards all the information contained in the data that has been used for selection**, making the inferential analysis independent of the selection step.

The conditional approach

Example

"Publication bias" refers to the influence that the results of a statistical analysis have on the probability of reporting/publishing the findings.

As an idealised example, suppose n research groups take independent measurements of a quantity μ, obtaining respective samples $Y_i = \mu + \varepsilon_i$, but that they only report the analysis if the data indicates that μ is significant, e.g. if $|Y_i| > 2$.

Then, the collection of published analyses on this particular effect μ will, on average, overstate the size of the true underlying effect.

According to the conditional approach, the published analysis ought to be carried out conditionally on the event that the mean appeared significant, i.e. on $\{|Y_i| > 2\}$.

Affine selection rules

Most analytic methodology for conditional inference is restricted to the class of **affine selection rules**, for which all the selection events can be written in the form

$$E_M = \{y \in \mathbb{R}^n : Ay \leq b\} \tag{7}$$

for some $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$ independent of the data (but dependent on M).

This class is general enough to accommodate some important, non-trivial types of selection rules.

The events of the **selected mean problem** can also be written in this way. For example, $E_1 = \{y \in \mathbb{R}^n : y_1 = \max_{i=1,\dots,n}\{y_i\}\}$ has $b = \mathbf{0}$ and

$$A = \begin{pmatrix} -1 & 1 & 0 & 0 & \cdots & 0 \\ -1 & 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ -1 & 0 & 0 & 0 & \dots & 1 \end{pmatrix}. \tag{8}$$

Affine selection rules

A popular tool for high-dimensional regression is the lasso estimator and its variants.

For data $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, the lasso estimate is defined as an L_1-penalised version of the least squares problem,

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1, \tag{9}$$

where $\lambda > 0$ is a user-specified parameter that controls the amount of regularisation (we shall assume uniqueness of the solution, which is guaranteed under mild assumptions on the design).

Since the nature of the L_1 norm forces some of the coefficients of $\hat{\beta}$ to be zero, a natural model selection rule is given by

$$\hat{M} = \{j = 1, \ldots, p : \hat{\beta}_j \neq 0\}. \tag{10}$$

Affine selection rules

It turns out that the selection events for this model selector are given by the union of $2^{|M|}$ affine regions, where each region corresponds to a combination of signs of the active lasso coefficients $\hat{\beta}_j \neq 0$.

For every $M \in 2^{\{1,\dots,p\}} \setminus \emptyset$ and $s \in \{-1,1\}^{|M|}$, define

$$A_M(s) = \begin{pmatrix} \lambda^{-1} X_{-M}^T (I - P_M) \\ -\lambda^{-1} X_{-M}^T (I - P_M) \\ -\text{diag}(s) X_M^\dagger \end{pmatrix}, b_M(s) = \begin{pmatrix} \mathbf{1} - X_{-M}^T (X_M^T)^\dagger s \\ \mathbf{1} + X_{-M}^T (X_M^T)^\dagger s \\ -\lambda \text{diag}(s)(X_M^T X_M)^{-1} s \end{pmatrix},$$

where X_{-M} contains the columns that are not in M and P_M is the projection onto $\text{span}(X_M)$.

For a lasso solution $\hat{\beta}$ such that $\hat{M} = M$, let $\hat{s} \in \{-1,1\}^{|M|}$ be the vector of signs of the active lasso coefficients, e.g. if $\hat{\beta} = (2, 0, -3, 0)^T$, $\hat{s} = (1, -1)^T$.

Affine selection rules

Theorem

For the lasso model selector $\hat{M}$ given in (10)*, it holds that*

$$\{\hat{M} = M, \hat{s} = s\} = \{A_M(s)y \leq b_M(s)\}. \tag{11}$$

The full selection event can be then recovered as

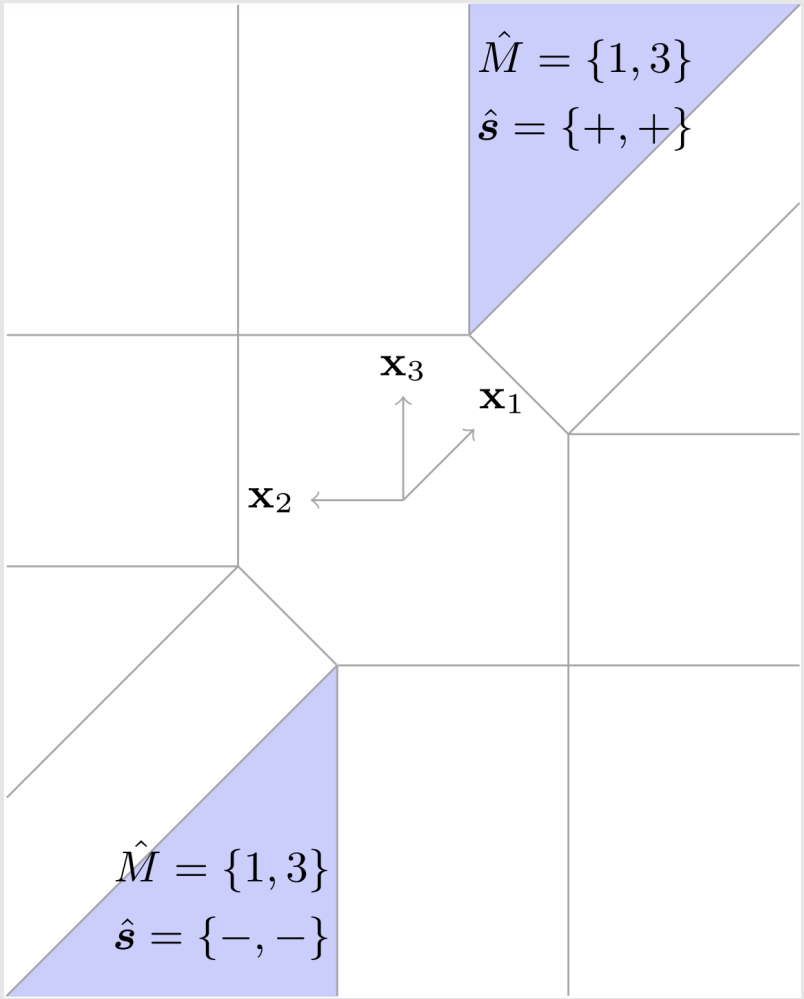
$$E_M = \bigcup_{s \in \{-1,1\}^{|M|}} \{A_M(s)y \leq b_M(s)\}. \tag{12}$$

However, we will often work with the finer events $E_M(s) = E_M \cap \{\hat{s} = s\}$, as they are easier to handle analytically and computationally, and the extra conditioning does not affect the validity of the inferences.

In some cases there are stronger reasons for conditioning on the signs conceptually; e.g. if an effect is only included in the model if it is positive.

Affine selection rules

Figure: Plot of a lasso selection region for $n = 2$, $p = 3$ (Lee et al., 2020).



Affine selection rules

Proof.

If the lasso solution is unique, for almost every y, a vector of coefficients $\hat{\beta} \in \mathbb{R}^p$ and a full vector of signs $\hat{s} \in \mathbb{R}^p$ are the solution of the lasso problem if and only if they satisfy the KKT conditions:

$$X^T(X\hat{\beta} - y) + \lambda\hat{s} = 0; \tag{13}$$

$$\hat{s}_j = \text{sign}(\hat{\beta}_j) \text{ if } \hat{\beta}_j \neq 0; \tag{14}$$

$$\hat{s}_j \in (-1, 1) \text{ if } \hat{\beta}_j = 0. \tag{15}$$

Partition the equations into the components relative to M and $-M$:

$$X_M^T(X_M\hat{\beta}_M - y) + \lambda\hat{s}_M = 0; \tag{16}$$

$$X_{-M}^T(X_M\hat{\beta}_M - y) + \lambda\hat{s}_{-M} = 0; \tag{17}$$

$$\text{sign}(\hat{\beta}_M) = \hat{s}_M; \tag{18}$$

$$\|\hat{s}_M\|_\infty < 1. \tag{19}$$

Affine selection rules

Proof.

The KKT conditions are necessary and sufficient. Therefore, we have that $\{(\hat{M}, \hat{s}) = (M, s)\}$ if and only if there exists vectors w and u such that

$$X_M^T(X_M w - y) + \lambda s = 0; \tag{20}$$

$$X_{-M}^T(X_M w - y) + \lambda u = 0; \tag{21}$$

$$\text{sign}(w) = s; \tag{22}$$

$$\|u\|_\infty < 1. \tag{23}$$

Solve the first two equations for w and u:

$$w = (X_M^T X_M)^{-1}(X_M^T y - \lambda s); \tag{24}$$

$$u = X_{-M}^T (X_M^T)^\dagger s + \frac{1}{\lambda} X_{-M}^T (I_n - P_M) y; \tag{25}$$

and combine them with the condition $\text{sign}(w) = s$ and $\|u\|_\infty < 1$. □

Affine selection rules

Proof.

Writing out these two conditions explicitly gives the desired result:

$$\{\text{sign}(w) = s\} = \{\text{diag}(s)w > 0\} \tag{26}$$

$$= \{\text{diag}(s)(X_M^T X_M)^{-1}(X_M^T y - \lambda s) > 0\}; \tag{27}$$

$$\{\|u\|_\infty < 1\} = \left\{-\mathbf{1} < X_{-M}^T (X_M^T)^\dagger s + \frac{1}{\lambda} X_{-M}^T (I_n - P_M) y < \mathbf{1}\right\}. \tag{28}$$

□

Affine selection rules

Other affine selection rules:

- **Elastic net**, based on the $L_1 + L_2$-penalised estimator

$$\hat{\beta} = \arg\min_{\beta} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1 + \gamma\|\beta\|_2, \quad \lambda, \gamma > 0. \tag{29}$$

- **Marginal screening**, which selects the predictors with largest correlations $|X_j^T Y|$, where X has been standardised to have unit norm columns.

- **Stepwise regression** with a fixed number of steps, whereby predictors are sequentially added/extracted from $\hat{M}$ to increase the RSS of a linear fit.

- **Least angle regression** with a fixed number of steps, a "stepwise version" of the lasso.

The polyhedral lemma

A key methodological result for conditional selective inference is the **polyhedral lemma**, which provides exact and analytically tractable confidence intervals (or p-values) when the data is **Gaussian** and the selection rule is **affine**.

Suppose that $Y \sim N(\mu, \Sigma)$, with $\Sigma > 0$ known, and that the parameters of interest for a selected model M can be written as linear combinations of the mean: $\psi_M = \eta(M)^T \mu$ for some $\eta(M) \in \mathbb{R}^n$.

Note that the coefficients of the projection parameter $\beta_M = X_M^\dagger \mu$ can be written in this form, with $\eta_j(M) = (X_M^\dagger)^T e_j$ for $j = 1, \ldots, |M|$.

Importantly, this method does not provide joint inference for the whole vector β_M, only marginal inferences for the coefficients.

The polyhedral lemma

For a generic direction of interest $\eta \in \mathbb{R}^n$, decompose Y into two independent components,

$$\eta^T Y \quad \text{and} \quad Z = (I_n - c\eta^T)Y, \tag{30}$$

where $c = (\eta^T \Sigma^{-1} \eta)^{-1} \Sigma \eta$.

Lemma

(Polyhedral lemma) For any A and b,

$$\{Ay \leq b\} = \{\mathcal{V}^-(z) \leq \eta^T y \leq \mathcal{V}^+(z), \mathcal{V}^0(z) \geq 0\}, \tag{31}$$

where

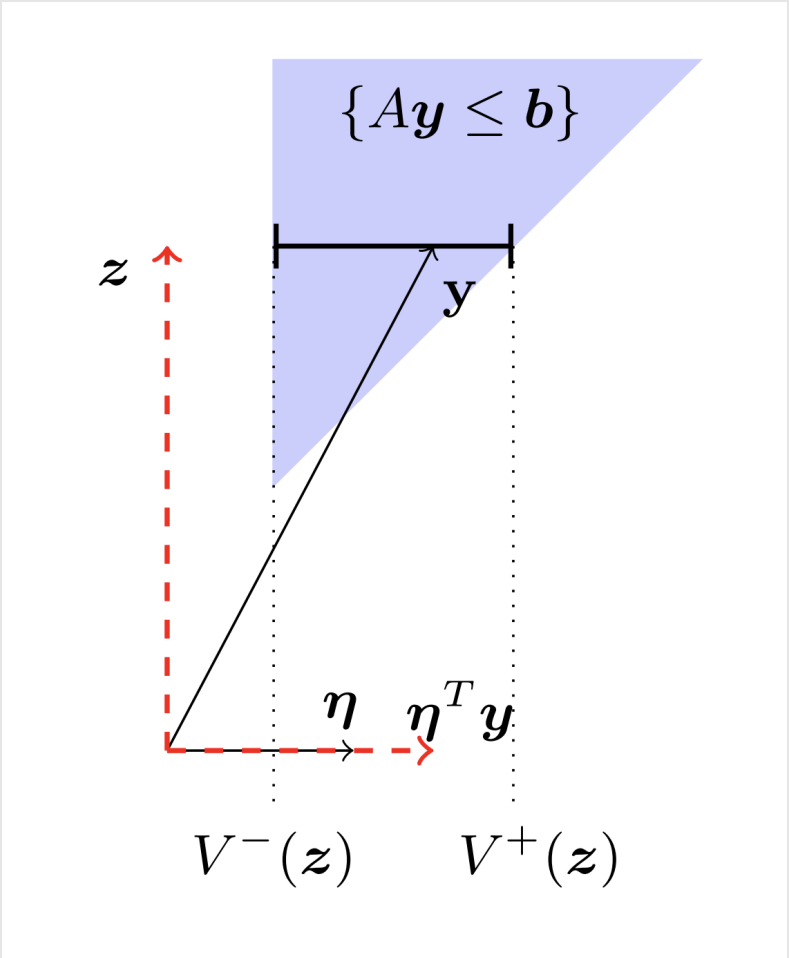
$$\mathcal{V}^-(z) = \max_{j:\,(Ac)_j<0} \frac{b_j - (Az)_j}{(Ac)_j}; \tag{32}$$

$$\mathcal{V}^+(z) = \min_{j:\,(Ac)_j>0} \frac{b_j - (Az)_j}{(Ac)_j}; \tag{33}$$

$$\mathcal{V}^0(z) = \min_{j:\,(Ac)_j=0} b_j - (Az)_j. \tag{34}$$

The polyhedral lemma

Figure: Geometric interpretation of the polyhedral lemma (Lee et al., 2016).



The polyhedral lemma

Proof.

Write $y = c(\eta^T y) + z$ and rewrite the polyhedron as

$$\begin{aligned}
\{Ay \leq b\} &= \left\{A(c(\eta^T y) + z) \leq b\right\} \\
&= \left\{Ac(\eta^T y) \leq b - Az\right\} \\
&= \left\{(Ac)_j(\eta^T y) \leq b_j - (Az)_j \text{ for all } j\right\} \\
&= \left\{\begin{array}{ll} \eta^T y \leq \frac{b_j - (Az)_j}{(Ac)_j} & \text{for } j\colon (Ac)_j > 0, \\ \eta^T y \geq \frac{b_j - (Az)_j}{(Ac)_j} & \text{for } j\colon (Ac)_j < 0, \\ 0 \leq b_j - (Az)_j & \text{for } j\colon (Ac)_j = 0 \end{array}\right\}.
\end{aligned}$$

Since $\eta^T y$ is the same quantity for all j, it must be at least the maximum of the lower bounds, which is $\mathcal{V}^-(z)$, and no more than the minimum of the upper bounds, which is $\mathcal{V}^+(z)$. □

The polyhedral lemma

By the polyhedral lemma, we have that

$$\eta^T Y \mid \{AY \leq b\} \stackrel{d}{=} \eta^T Y \mid \{\mathcal{V}^-(Z) \leq \eta^T Y \leq \mathcal{V}^+(Z), \mathcal{V}^0(Z) \geq 0\}, \quad (35)$$

with Z independent of $\eta^T Y$. Thus, by further conditioning on $Z = z$, where z is such that $Ay \leq b$, we have that

$$\eta^T Y \mid \{AY \leq b, Z = z\} \stackrel{d}{=} \eta^T Y \mid \{\mathcal{V}^-(z) \leq \eta^T Y \leq \mathcal{V}^+(z)\}, \quad (36)$$

which is simply a truncated univariate Gaussian distribution, for which inference is available in closed form.

Specifically, we require inference for $\psi = \eta^T \mu$ in the model

$$\hat{\psi} \sim N(\psi, \gamma^2) \mid [a, b], \quad \gamma = \eta^T \Sigma \eta, a = \mathcal{V}^-(z), b = \mathcal{V}^+(z). \quad (37)$$

We can have $a = -\infty$ and/or $b = \infty$, in which case the corresponding direction is not truncated.

The polyhedral lemma

Let $Y \sim N(\mu, \sigma^2)$ and consider a generic truncation event $E \subseteq \mathbb{R}$. The CDF of $Y \mid E$ is

$$F_\mu^E(y) = P_\mu(Y \leq y \mid Y \in E) = \int_{-\infty}^{y} \frac{\sigma^{-1}\phi\{\sigma^{-1}(\mu - y')\}}{\varphi(\mu)} \mathrm{d}y', \quad (38)$$

where $\varphi(\mu) = P_\mu(Y \in E)$ is the selection probability.

A $1-\alpha$ CI for μ valid conditionally on $\{Y \in E\}$ is given by $[L(Y), U[Y]]$, where

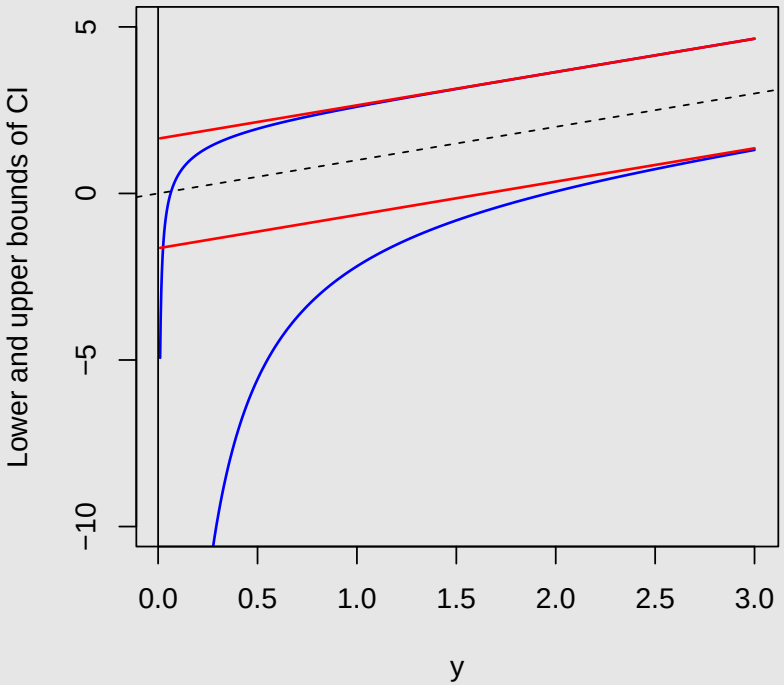
$$F^E_{L(Y)}(Y) = q_1; \quad (39)$$

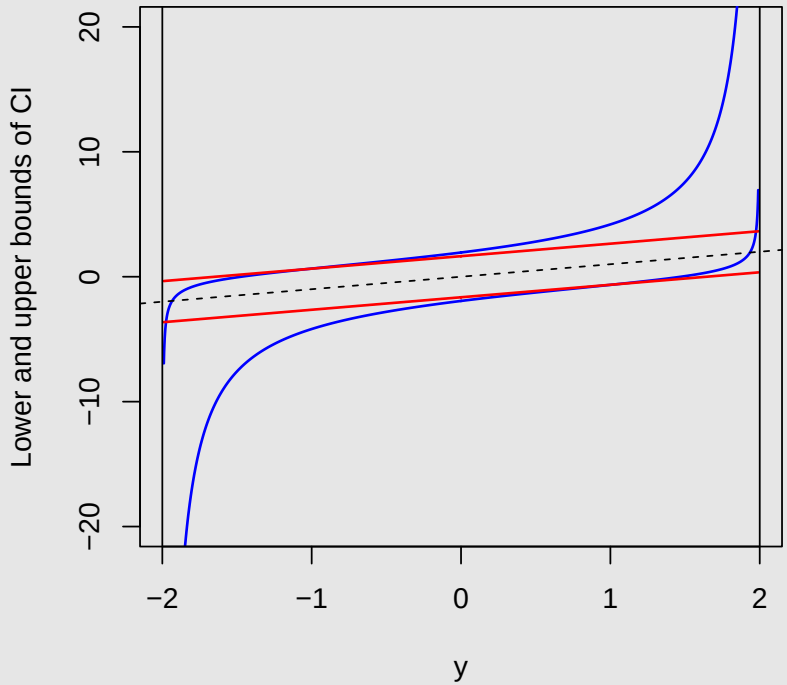
$$F^E_{U(Y)}(Y) = q_2; \quad (40)$$

with $q_1 - q_2 = 1 - \alpha$. For example, we can take $q_2 = 1 - q_1 = \alpha/2$.

The polyhedral lemma

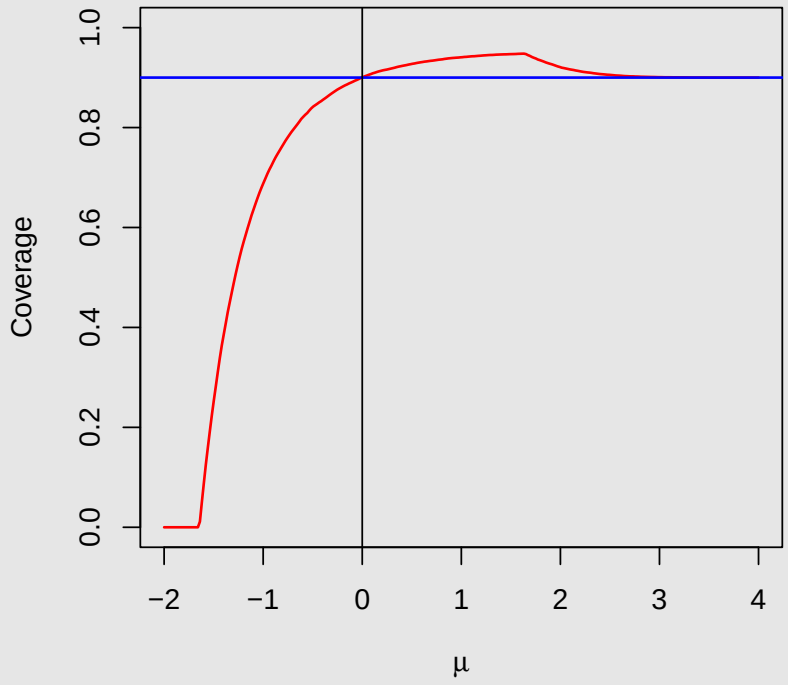
Figure: CIs for the truncated Gaussian model with $1 - \alpha = 0.9$ and $\sigma^2 = 1$, as a function of the data y. On the left, $E = [0, \infty)$; on the right, $E = [-2, 2]$. In blue, conditional intervals; in red, unadjusted (classical) intervals.

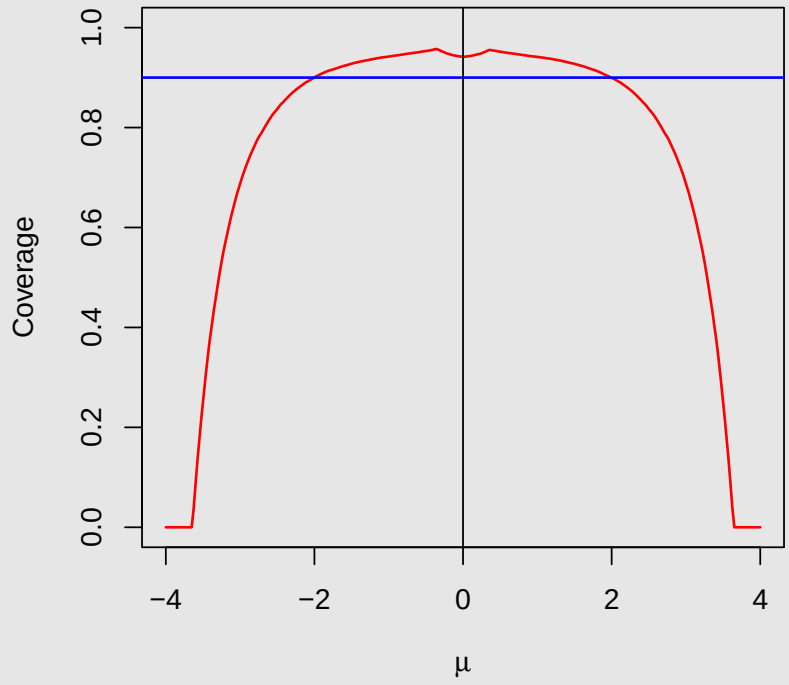




The polyhedral lemma

Figure: Coverage of the confidence intervals in the previous setting, as a function of μ.





The polyhedral lemma

This result can be used to provide conditional selective inference for the lasso and other important selection rules by conditioning on the signs of the selected coefficients.

While the extra conditioning makes the analysis computationally simple, it unavoidably results in a loss of power: the more conditioning, the less powerful is the inference.

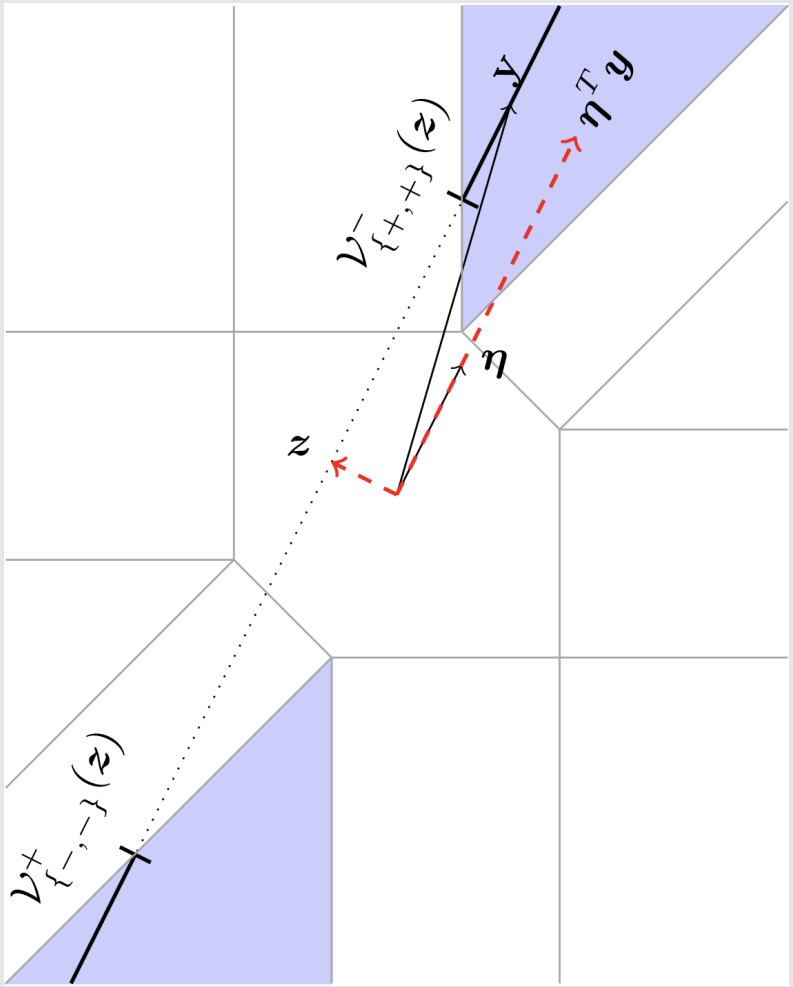
Assume the conditioning event is $E = \bigcup_s \{A(s)Y \leq b(s)\}$ and define the quantities $\mathcal{V}_s^-(z)$ and $\mathcal{V}_s^+(z)$ analogously. By a similar argument, we have

$$\eta^T Y \mid \{E, Z = z\} \stackrel{d}{=} \eta^T Y \mid \bigcup_s \{\mathcal{V}_s^-(z) \leq \eta^T Y \leq \mathcal{V}_s^+(z)\}, \qquad (41)$$

i.e. the problem is reduced to inference on the mean of a univariate Gaussian distribution truncated to a union of intervals (note that this requires computation of $2^{|M|+1}$ truncation limits, as opposed to only 2).

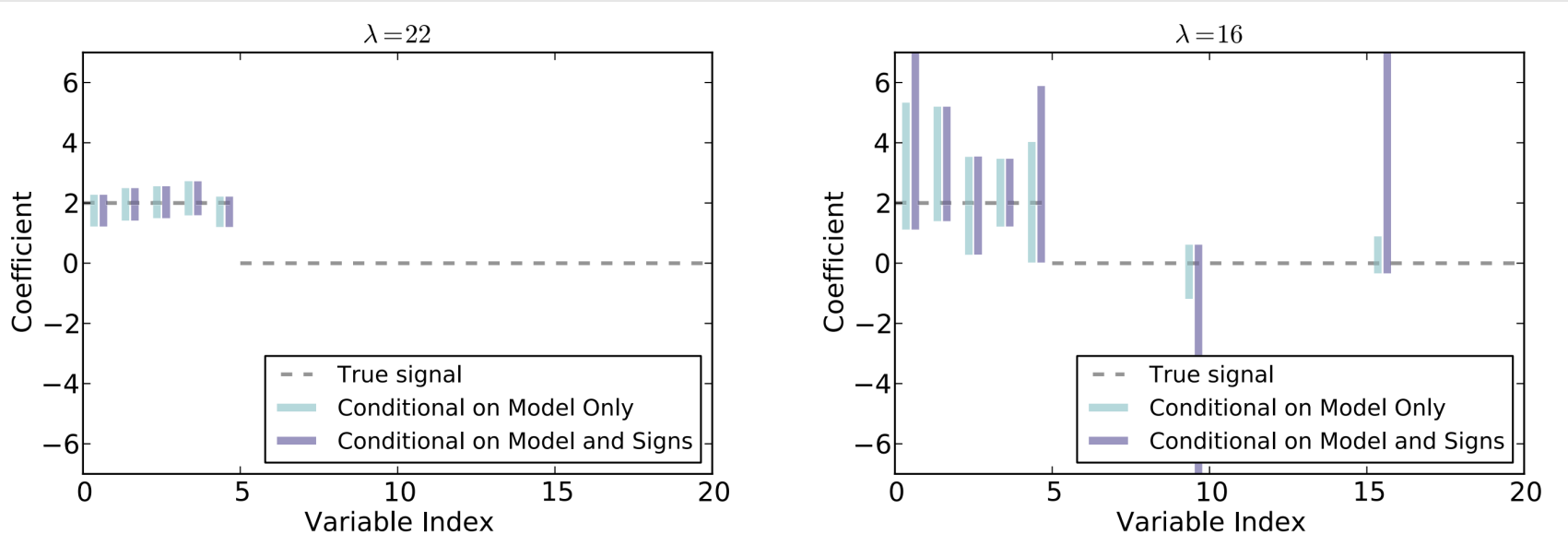
The polyhedral lemma

Figure: Conditioning on a union of polyhedra (Lee et al., 2016).



The polyhedral lemma

How much power is lost by conditioning on the coefficient signs depends on how strong the signal is, as demonstrated in the following figure (Lee et al, 2016). Data was simulated with $n = 25$ and $p = 50$, but only the first 20 coefficients are shown.



The polyhedral lemma

To be able to use these methods we need to know the variance or at least have a good estimate of it to plug in.

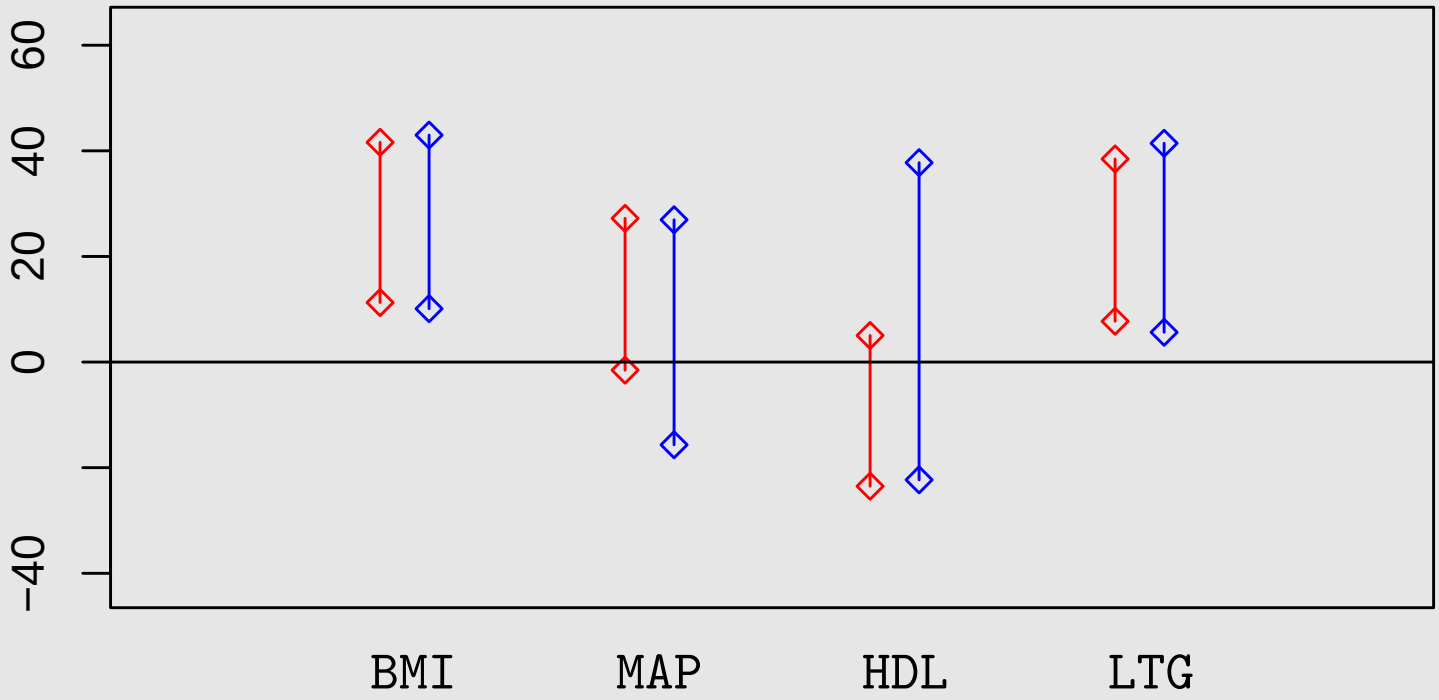
If p/n is small, estimating σ^2 via the residual sum of squares of the full model $\mu = X\beta$ tends to give reasonable results.

In moderate and high-dimensional settings, an effective option is to construct $\hat{\sigma}^2$ from the residual sum of squares using a lasso fit tuned by cross validation.

Alternatively, conservative but safer confidence intervals can be obtained by using an overestimate of σ^2, such as the unconditional sample variance of Y.

The polyhedral lemma

To see the polyhedral lemma in practice we return to the diabetes data example ($n = 442$, $p = 10$). In blue, conditional 90% selective intervals for the coefficients of the projection parameter after lasso selection with $\lambda = 7$ (conditioning on coefficient signs). In red, unadjusted (classical) intervals. The variance was estimated in the classical way.



The conditional approach

The conditional approach has two important drawbacks:

- It relies on the selection rule being fixed prior to the analysis: universal guarantees are unattainable conditionally, as the procedure would need to be valid given *any* event containing the observed y, whose intersection is $E = \{y\}$, leaving no information for inference.
- Even though inferences are specific to a selection rule, they can be very conservative, as conditioning discards too much information about the parameter of interest. For example, if $Y \sim N(\mu, 1)$ and the selection event E has $\inf(E) \neq -\infty$ or $\sup(E) \neq \infty$, the conditional CIs have infinite expected length:

$$\mathbb{E}[U(Y) - L(Y) \mid Y \in E] = \infty. \tag{42}$$

  This result extends to the multidimensional setting when we condition on the direction of interest.

The conditional approach

Outside the realm of affine selection rules, there are few rules which admit nice and tractable selection events.

Loftus (2015)[1] characterised the selection events of the lasso tuned with cross validation (rather than with a prespecified penalty) as intersections of quadratic sets in y, of the form $y^T Ay + b^T y + c > 0$.

Alternatively, if the shape of the selection events is unknown but the selection rule can be reapplied to simulated data, Monte Carlo procedures might be used for estimating the conditional distribution, but these tend to be very costly.

[1] "Selective inference after cross-validation", *arxiv:511.08866*

Selective inference for clustering

Reduction of conditional selective problems to univariate Gaussian models have been exploited in other areas outside regression.

An important framework of application is **inference after clustering**.

A common inferential objective in clustering is to **test differences in the means** of the different clusters.

Classical testing procedures are useful in situations when the groups are known or have been selected **independently of the data**.

However, when the groups are **selected using the data**, classical tests yield an extremely inflated false positive rate.

Selective inference for clustering

Consider the following matrix Gaussian model for n observations of p features:

$$X \sim N_{n \times p}(\mu, I_n, \sigma^2 I_p), \tag{43}$$

where $\mu \in \mathbb{R}^{n \times p}$, with rows μ_i, is unknown, and $\sigma^2 > 0$ is known.

In other words, the rows of X are independent $N(\mu_i, \sigma^2 I_n)$ random variables.

For $\mathcal{G} \subseteq \{1, \ldots, n\}$, define

$$\bar{\mu}_{\mathcal{G}} = \frac{1}{|\mathcal{G}|} \sum_{i \in \mathcal{G}} \mu_i, \quad \bar{X}_{\mathcal{G}} = \frac{1}{|\mathcal{G}|} \sum_{i \in \mathcal{G}} X_i, \tag{44}$$

which will be referred to as the mean of $\mathcal{G}$ and the empirical mean of $\mathcal{G}$ in X.

Selective inference for clustering

Given a realisation x of X, consider the two-step adaptive procedure:

1. A clustering algorithm $\mathcal{C}$ is used to obtain a partition $\mathcal{C}(x)$ of $\{1, \ldots, n\}$.

2. For two clusters $\mathcal{C}_1, \mathcal{C}_2$, the same data x is used to test

$$H_0 : \bar{\mu}_{\mathcal{C}_1} = \bar{\mu}_{\mathcal{C}_2} \quad \text{versus} \quad H_1 : \bar{\mu}_{\mathcal{C}_1} \neq \bar{\mu}_{\mathcal{C}_2}. \tag{45}$$

The classical (non-selective) Wald test computes the p-value as

$$P_{H_0}\left(\|\bar{X}_{\mathcal{C}_1} - \bar{X}_{\mathcal{C}_2}\|_2 \geq \|\bar{x}_{\mathcal{C}_1} - \bar{x}_{\mathcal{C}_2}\|_2 \right), \tag{46}$$

where, under H_0, $\|\bar{X}_{\mathcal{C}_1} - \bar{X}_{\mathcal{C}_2}\|_2^2 \sim \sigma^2(|\mathcal{C}_1|^{-1} + |\mathcal{C}_2|^{-1})\chi^2_p$.

Selective inference for clustering

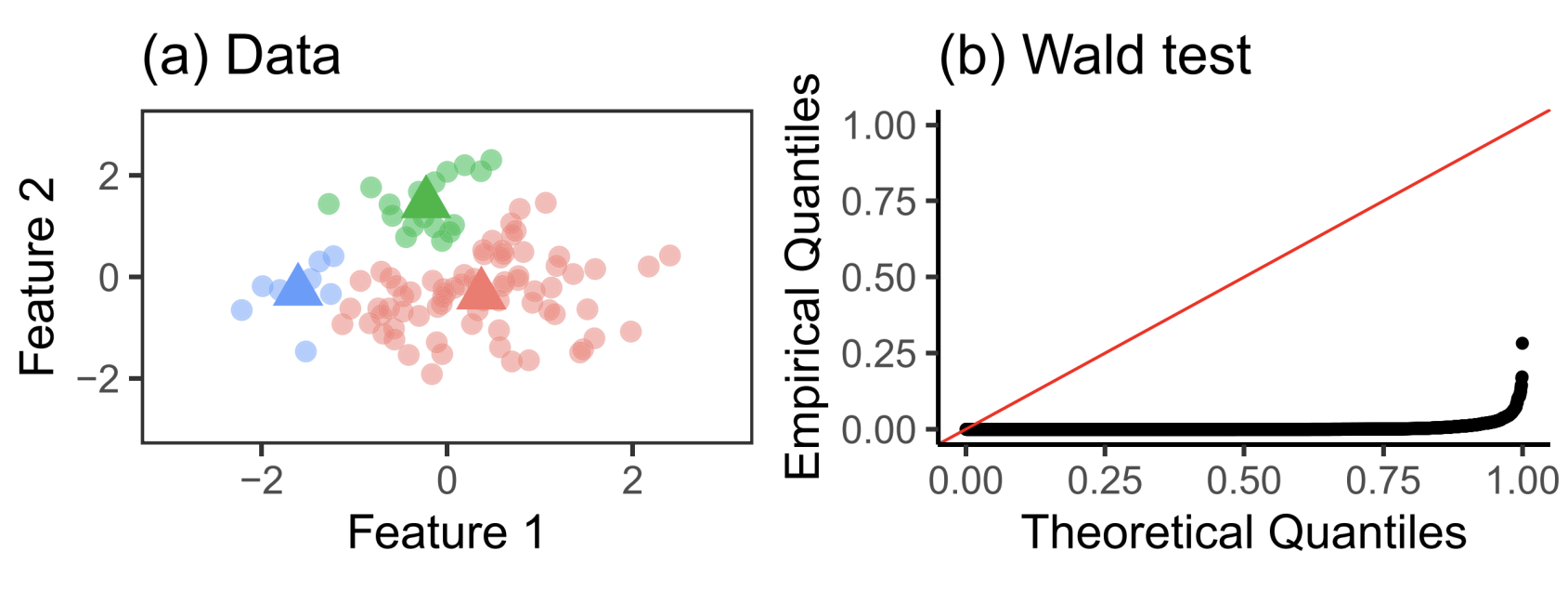
However, this test does not account for the double-use of the data and is therefore affected by selection bias.

Since the clusters are selected using the data, it is natural that we will observed significant differences between the group means, even if they are equal (just as selected effects in a regression model tend to be overestimated by their face-value estimators).

We can see this empirically in the following figure.

Selective inference for clustering

Figure: Gao el al. (2022)

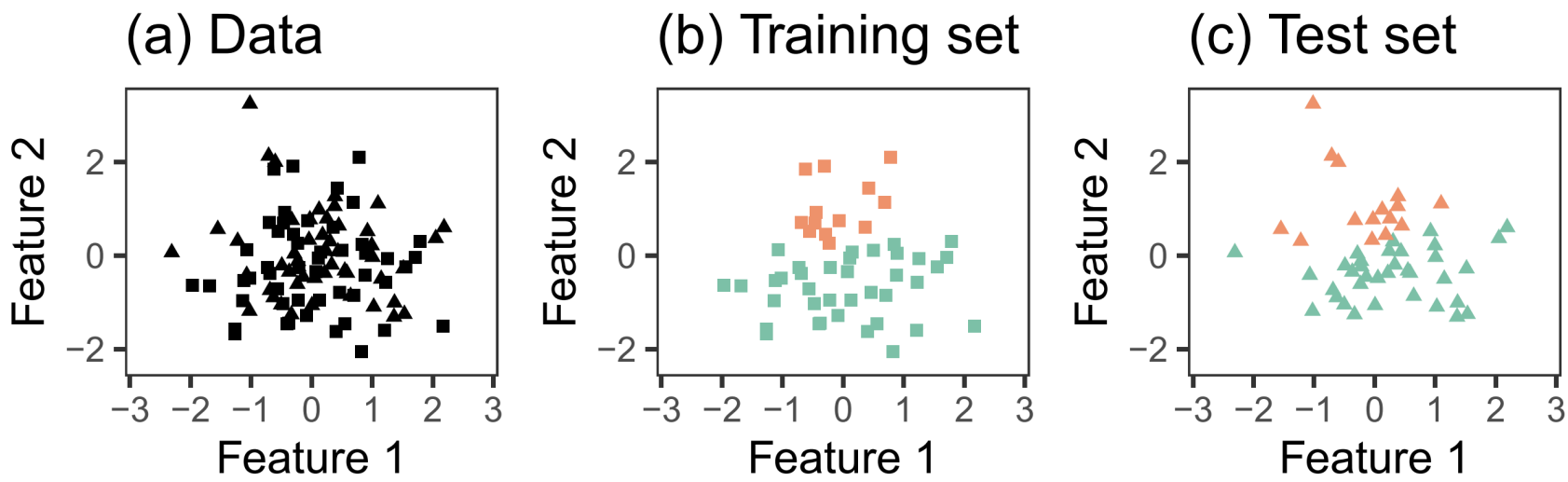


Selective inference for clustering

Perhaps surprisingly, **data splitting does not work** for this problem.

$\rightarrow$ In the testing stage we need to assign the observations to the clusters using the training set, thereby breaking independence.

Figure: Gao el al. (2022)



Selective inference for clustering

In the spirit of the conditional approach, a testing procedure is deemed valid if the errors guarantees hold conditionally on the groups having been selected by the clustering algorithm, i.e. if the p-value

$$P_{H_0}\left(\|\bar{X}_{\mathcal{C}_1} - \bar{X}_{\mathcal{C}_2}\|_2 \geq \|\bar{x}_{\mathcal{C}_1} - \bar{x}_{\mathcal{C}_2}\|_2 \mid \mathcal{C}_1, \mathcal{C}_2 \in \mathcal{C}(X)\right) \tag{47}$$

is less than a prespecified level α.

Interpretation: Among all realisations of X for which $\mathcal{C}_1$ and $\mathcal{C}_2$ are chosen, what proportion have a difference in empirical cluster means at least as large as the one observed in the dataset, when, in fact, $\bar{\mu}_{\mathcal{C}_1} = \bar{\mu}_{\mathcal{C}_2}$?

Selective inference for clustering

The unconditional distribution of $\bar{X}_{\mathcal{C}_1} - \bar{X}_{\mathcal{C}_2}$ under H_0 is independent of the unknown μ, but conditioning on selection breaks the independence, so we need to get rid of the nuisance parameters.

Write $\bar{X}_{\mathcal{C}_1} - \bar{X}_{\mathcal{C}_2} = X^T \nu$, where

$$[\nu]_i = \frac{\mathbf{1}(i \in \mathcal{C}_1)}{|\mathcal{C}_1|} - \frac{\mathbf{1}(i \in \mathcal{C}_2)}{|\mathcal{C}_2|}. \tag{48}$$

Just as in the polyhedral lemma, we decompose the data into two orthogonal components, the projection in the direction of interest,

$$D = \frac{(\bar{X}_{\mathcal{C}_1} - \bar{X}_{\mathcal{C}_2})}{\|\bar{X}_{\mathcal{C}_1} - \bar{X}_{\mathcal{C}_2}\|_2}, \tag{49}$$

and its complement,

$$Z = (I_n - P_\nu)X. \tag{50}$$

Selective inference for clustering

We can then define the p-value conditionally on Z and D:

$$p(x) = P_{H_0}\left(\|\bar{X}_{\mathcal{C}_1} - \bar{X}_{\mathcal{C}_2}\|_2 \geq \|\bar{x}_{\mathcal{C}_1} - \bar{x}_{\mathcal{C}_2}\|_2 \mid \mathcal{C}_1, \mathcal{C}_2 \in \mathcal{C}(X), Z = z, D = d\right).$$

In the direction of interest, the conditioning event can be rewritten as

$$\begin{aligned}\mathcal{S}(x; \{\mathcal{C}_1, \mathcal{C}_2\}) &= \left\{\phi > 0 \colon \mathcal{C}_1, \mathcal{C}_2 \in \mathcal{C}\left(z + \left(\frac{\phi}{\frac{1}{|\mathcal{C}_1|} + \frac{1}{|\mathcal{C}_2|}}\right)\nu(\mathcal{C}_1, \mathcal{C}_2)d^T\right)\right\} \\ &\equiv \left\{\phi > 0 \colon \mathcal{C}_1, \mathcal{C}_2 \in \mathcal{C}\left(x'(\phi)\right)\right\}.\end{aligned}$$

Theorem

For any non-overlapping $\mathcal{C}_1$ and $\mathcal{C}_2$,

$$p(x) = 1 - \mathbb{F}\left(\|\bar{x}_{\mathcal{C}_1} - \bar{x}_{\mathcal{C}_2}\|_2; \sigma\sqrt{\frac{1}{|\mathcal{C}_1|} + \frac{1}{|\mathcal{C}_2|}}; \mathcal{S}(x; \{\mathcal{C}_1, \mathcal{C}_2\})\right), \quad (51)$$

where $\mathbb{F}(t; c, \mathcal{S})$ is the CDF of a $c\chi^2_p$ distribution truncated to $\mathcal{S}$.

Selective inference for clustering

To understand this better let's analyse the function $x'(\phi)$.

Since $x^T \nu = \bar{x}_{\mathcal{C}_1} - \bar{x}_{\mathcal{C}_2}$, the i-th row of $x'(\phi)$ is

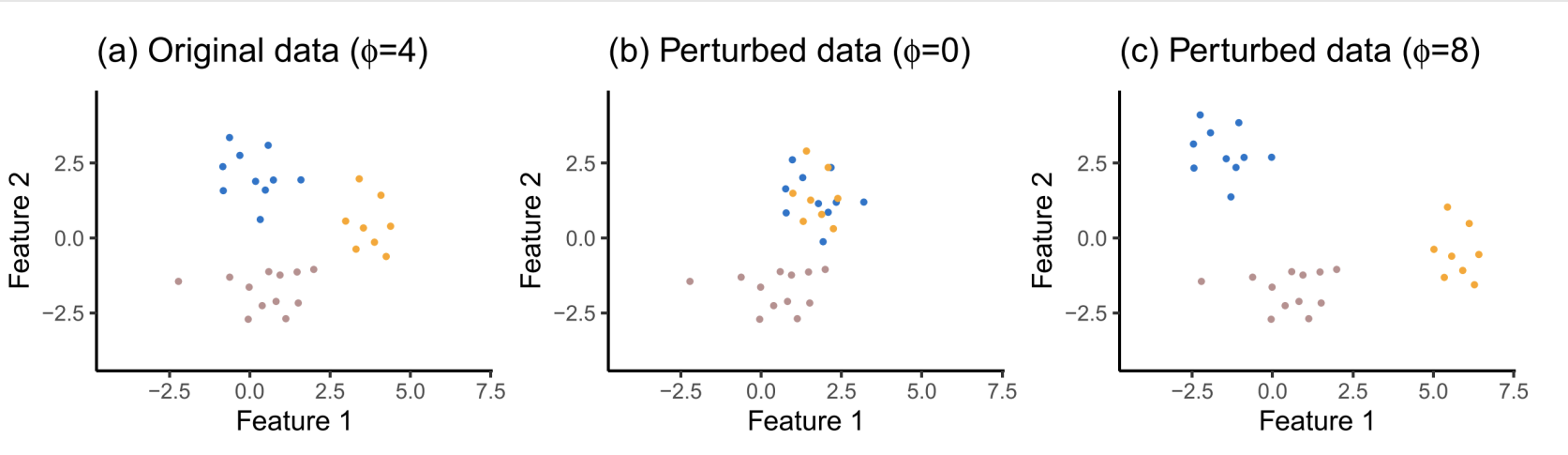
$$[x'(\phi)]_i = \begin{cases} x_i + \left(\frac{|\mathcal{C}_2|}{|\mathcal{C}_1|+|\mathcal{C}_2|}\right)(\phi - \|\bar{x}_{\mathcal{C}_1} - \bar{x}_{\mathcal{C}_2}\|_2)\, d, & \text{if } i \in \mathcal{C}_1; \\ x_i - \left(\frac{|\mathcal{C}_1|}{|\mathcal{C}_1|+|\mathcal{C}_2|}\right)(\phi - \|\bar{x}_{\mathcal{C}_1} - \bar{x}_{\mathcal{C}_2}\|_2)\, d, & \text{if } i \in \mathcal{C}_2; \\ x_i, & \text{otherwise.} \end{cases} \tag{52}$$

Thus, the function $x'(\phi)$ is a **perturbation of the observed data** in the direction of $\bar{x}_{\mathcal{C}_1} - \bar{x}_{\mathcal{C}_2}$: it pulls apart $\mathcal{C}_1$ and $\mathcal{C}_2$ if $\phi > \|\bar{x}_{\mathcal{C}_1} - \bar{x}_{\mathcal{C}_2}\|_2$, and it pushes together $\mathcal{C}_1$ and $\mathcal{C}_2$ otherwise.

$\rightarrow$ The goal is to find the range of perturbations such that the clustering algorithm selects the groups of interest $\mathcal{C}_1$ and $\mathcal{C}_2$.

Selective inference for clustering

Figure: Gao el al. (2022)



Selective inference for clustering

To compute $p(x)$ it suffices to characterise the truncation set $\mathcal{S}$, which constitutes the main challenge in the application of the method.

For a restricted class of algorithms, $\mathcal{S}$ can be computed exactly, as we shall see.

In general, the p-value needs to be approximated via Monte Carlo. If $\|\bar{x}_{\mathcal{C}_1} - \bar{x}_{\mathcal{C}_2}\|_2$ is large, vanilla Monte Carlo is very slow and importance sampling should be used instead:

1. Sample $\omega_1, \ldots, \omega_N \sim N\left(\|\bar{x}_{\mathcal{C}_1} - \bar{x}_{\mathcal{C}_2}\|_2, \sigma^2(|\mathcal{C}_1|^{-1} + |\mathcal{C}_2|^{-1})\right)$.
2. Set $\pi_i = f_1(\omega_i)/f_2(\omega_i)$, where f_1 is the density of the scaled chi-square distribution and f_2 is the density of the Gaussian proposal, and

$$\hat{p}(x) = \frac{\sum_{i=1}^{N} \pi_i \mathbf{1}\left\{\omega_i > \|\bar{x}_{\mathcal{C}_1} - \bar{x}_{\mathcal{C}_2}\|_2, \mathcal{C}_1, \mathcal{C}_2 \in \mathcal{C}(x'(\omega_i))\right\}}{\sum_{i=1}^{N} \pi_i \mathbf{1}\left\{\mathcal{C}_1, \mathcal{C}_2 \in \mathcal{C}(x'(\omega_i))\right\}}. \quad (53)$$

Selective inference for clustering

A set of algorithms for which the truncation set has been characterised analytically is the family of hierarchical clustering algorithms, defined as follows.

Let $\gamma(\mathcal{G}, \mathcal{G}'; x)$ be a function that quantifies the dissimilarity between two groups of observations

An **agglomerative hierarchical clustering algorithm** proceeds as follows:

Let $\mathcal{C}^{(1)}(x) = \{\{1\}, \ldots, \{n\}\}$. For $t = 1, \ldots, n-1$:

1. $\{\mathcal{W}_1^{(t)}(x), \mathcal{W}_2^{(t)}(x)\} = \arg\min\{\gamma(\mathcal{G}, \mathcal{G}'; x) : \mathcal{G}, \mathcal{G}' \in \mathcal{C}^{(t)}(x), \mathcal{G} \neq \mathcal{G}'\}$.

2. $\mathcal{C}^{(t+1)}(x) = \mathcal{C}^{(t)}(x) \cup \{\mathcal{W}_1^{(t)}(x) \cup \mathcal{W}_2^{(t)}(x)\} \setminus \{\mathcal{W}_1^{(t)}(x), \mathcal{W}_2^{(t)}(x)\}$.

Selective inference for clustering

For hierarchical algorithms, the set $\mathcal{S}$ can be written as a function of the distance metric between groups, γ.

First, we have the following lemma, which states that $x'(\phi)$ leads to the same final cluster as x if and only if all the intermediate merges coincide.

Lemma

Let $\mathcal{C} = \mathcal{C}^{n-K+1}$. We have that $\mathcal{C}_1, \mathcal{C}_2 \in \mathcal{C}(x'(\phi))$ if and only if $\mathcal{C}^{(t)}(x'(\phi)) = \mathcal{C}^{(t)}(x)$ for all $t = 1, \ldots, n-K+1$.

To characterise all the intermediate merges, define the set of "losing pairs" as all coexisting pairs of clusters that are never the "winning pairs" after $n - K$ steps:

$$\mathcal{L}(x) = \bigcup_{t=1}^{n-K} \left\{ \{\mathcal{G}, \mathcal{G}'\} : \mathcal{G}, \mathcal{G}' \in \mathcal{C}^{(t)}(x), \mathcal{G} \neq \mathcal{G}', \{\mathcal{G}, \mathcal{G}'\} \neq \{\mathcal{W}_1^{(t)}(x), \mathcal{W}_2^{(t)}(x)\} \right\}.$$

Selective inference for clustering

Each pair $\{\mathcal{G}, \mathcal{G}'\} \in \mathcal{L}(x)$ has a "lifetime", $[l_{\mathcal{G},\mathcal{G}'}(x), u_{\mathcal{G},\mathcal{G}'}(x)] \equiv [l, u]$, defined as the lowest and highest steps in the algorithm in which they coexist.

This leads to the following representation.

Theorem

Let $\mathcal{C} = \mathcal{C}^{n-K+1}$. Then

$$\mathcal{S} = \bigcap_{\mathcal{L}(x)} \left\{ \phi > 0 \colon \gamma(\mathcal{G}, \mathcal{G}'; x'(\phi)) > \max_{l \leq t \leq u} \gamma(\mathcal{W}_1^{(t)}, \mathcal{W}_2^{(t)}; x) \right\}. \tag{54}$$

The remaining step is to compute the function

$$\gamma(\mathcal{G}, \mathcal{G}'; x'(\phi)) \tag{55}$$

for a given distance γ.

Selective inference for clustering

This can be done, among others, for a class of squared Euclidean distances, defined by

$$\gamma(\{i\}, \{i'\}; x) = \|x_i - x_{i'}\|_2^2 \tag{56}$$

for all singletons $\{i\}, \{i'\}$, and then recursively via the linkage function

$$\gamma(\mathcal{G}_1 \cup \mathcal{G}_2, \mathcal{G}_3; x) = \alpha_1 \gamma(\mathcal{G}_1, \mathcal{G}_3; x) + \alpha_2 \gamma(\mathcal{G}_2, \mathcal{G}_3; x) + \beta \gamma(\mathcal{G}_1, \mathcal{G}_2; x), \tag{57}$$

for some coefficients $\alpha_1, \alpha_2, \beta$.

Proposition.

For these distances, $\gamma(\mathcal{G}_1 \cup \mathcal{G}_2, \mathcal{G}_3; x'(\phi))$ can be written as a quadratic function in ϕ whose coefficients can be recursively derived from those of $\gamma(\mathcal{G}_1, \mathcal{G}_2; x'(\phi))$, $\gamma(\mathcal{G}_1, \mathcal{G}_3; x'(\phi))$ and $\gamma(\mathcal{G}_2, \mathcal{G}_3; x'(\phi))$ in $O(1)$ time.

The conditional approach: an overview

Some important remarks about the conditional approach:

- Conditional methods require **the selection rule to be known** and fixed prior to the data analysis. In particular, they cannot account for informal model-checking, unlike PoSI and related approaches.
- Due to the need for working with truncated distributions, in high-dimensional settings they can only be applied analytically with (approximate) Gaussian data, as this allows dimension reduction via conditioning.
- Conditional methods are not robust against model misspecification when the selection probability is low.
- They tend to produce very wide confidence intervals, sometimes with infinite expected length; very long intervals are less likely to be reported in practice, which defeats the goal of enforcing error guarantees.