This lecture

Information splitting for more powerful conditional inference:

- The Gaussian case: Garcia Rasines, Young (2022). "Splitting approaches for post-selection inference", *Biometrika*.

- Extension: Leiner, Duan, Wasserman, Ramdas (2023). "Data fission: splitting a single data point", *arXiv:2112.11079*.

Issues with the conditional approach

The conditional approach provides error guarantees which are specific to a parameter.

As such, in many situations conditional methods enjoy a sounder theoretical justification than unconditional ones.

However, stronger requirements lead to a **loss of robustness against misspecification of the model and/or the selection rule**.

In this lecture we will explore ways for providing conditional inference that bypass these issues.

Sensitivity to misspecification

Most conditional methods rely on a **normality assumption to get rid of nuisance parameters** by conditioning on the direction of interest.

In non-selective regimes, a Gaussian model can sometimes be justified on asymptotic grounds after some sort of dimensionality reduction.

However, asymptotic approximations tend to perform poorly in low-probability regions.

In high dimensions, **selection events** are almost always **low-probability regions**, which makes conditional methods very sensitive to deviations from Gaussianity.

Sensitivity to misspecification

Suppose we have n IID samples $Y_1, \ldots, Y_n$ from an unknown distribution.

Under mild conditions, we can use the CLT approximation

$$\bar{Y} \dot{\sim} N(\mu, \sigma^2/n), \tag{1}$$

to provide inference for $\mu = \mathbb{E}(Y_1)$, where $\sigma^2 = \text{Var}(Y_1)$ can be estimated with the data.

Now, suppose we are in a selective regime, where μ is only analysed if $\bar{Y} > 0$, say. The corresponding approximation would be

$$\bar{Y} \mid \bar{Y} > 0 \dot{\sim} N(\mu, \sigma^2/n) \mid [0, \infty). \tag{2}$$

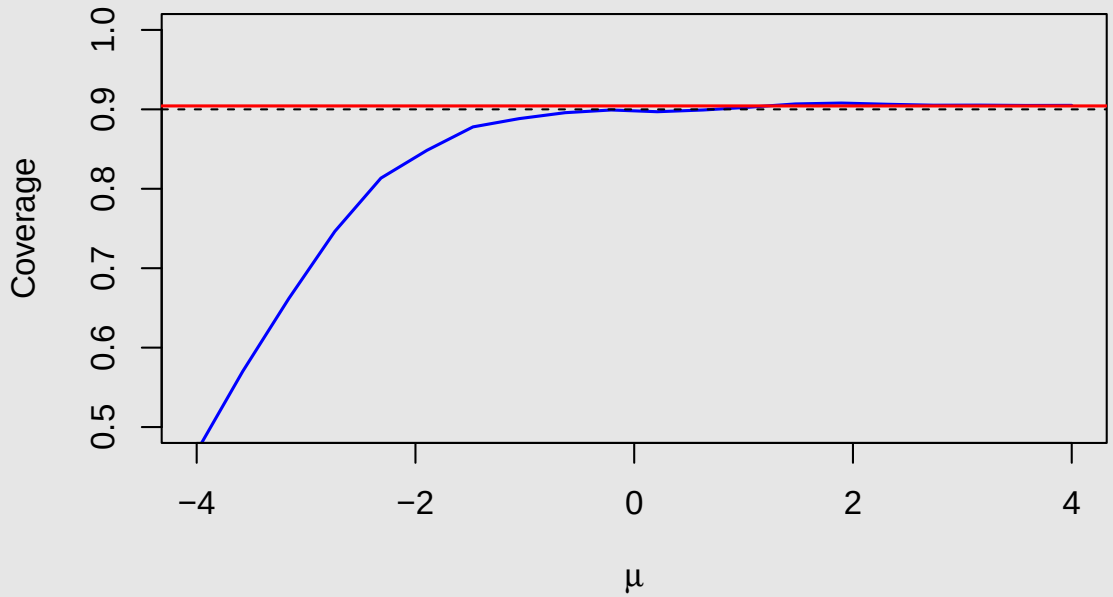
$\rightarrow$ If $\mu << 0$, this approximation will likely be inaccurate.

$\rightarrow$ Furthermore, estimating σ^2 is difficult.

Sensitivity to misspecification

Suppose we construct selective CIs for μ according to the model $Y_i = \mu + N(0,1)$, $i = 1, \ldots, 100$, and selection event $\bar{Y} > 0$, but the true error distribution is $t_3/\sqrt{2}$ (which has unit variance). The following plot shows how the coverage deteriorates as μ decreases. In the analogous situation without selection, the coverage is independent of μ.

Figure: In blue, coverage of 90% selective CIs under misspecification; in red, coverage of 90% standard CIs under misspecification.



Low probability events

This is especially **problematic in high dimensions**, where it is very common for all the selection events to have a very low probability.

This is very simple to illustrate in the variable-selection setting, where there are $2^p - 1$ (non-empty) models to choose from.

Suppose, for example, that the lasso is applied to data from a linear model, and that it wrongly selects a few non-significant predictors.

What is the probability that, on repeated sampling from the same distribution, the exact same false discoveries are made if $p = 200$, say?

Low probability events

Consider a case with $n = 80$ and $10 \leq p \leq 40$. For each p, we sampled 10^4 data points from $Y \sim N(X\beta, I_n)$, with $\beta = (2, 2, 2, 2, 2, 0, \ldots, 0)^T$, and each time we run the lasso with cross-validation and recorded the selected model.

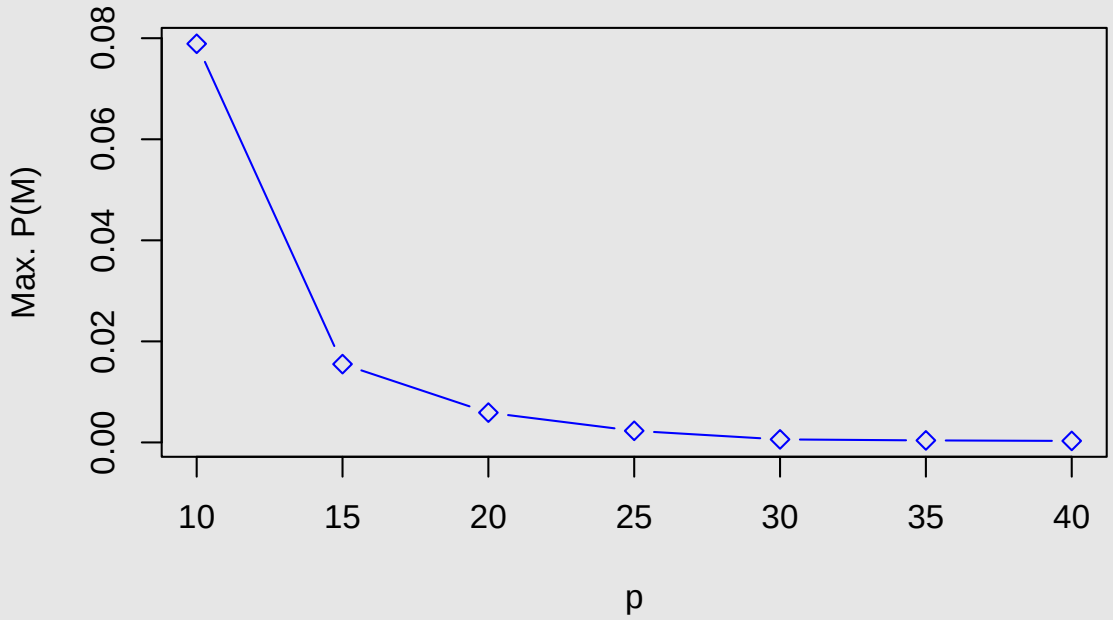


Figure: Proportion of samples in which the most selected model was picked.

Uniform validity

Another natural criticism of the conditional approach is that it **requires knowledge of the selection event**.

This precludes the use of very complex selection rules, and more importantly of data-based decisions during the selection process.

This is not even possible when we can partially restrict the collection of selection rules.

For example, suppose it is suspected an analysis for an effect μ is only reported if $Y = \mu + \varepsilon \geq t$ for some unknown t. Upon observing y, we cannot discard any $t \leq y$. In particular, for the boundary case $t = y$ we obtain a semi-infinite CI.

In general, if we want to protect against a collection of selection events $\{E_i : i \in \mathcal{I}\}$, we need to have conditional validity for the smaller event

$$E = \bigcap_{i\,:\, y \in E_i} E_i. \tag{3}$$

Lack of power

Intervals can be arbitrarily wide whenever the truncation region is upper/lower bounded, as the resulting truncation models are not identifiable in the respective limits.

Consider the model $Y \sim N(\mu, 1) \mid [0, \infty)$. Upon observing $y = 0.0001$, say, one cannot discard $\mu_0 = -10$ as the true generating mean, even if this value is nonsensical for the problem.

Furthermore, for such extreme values of the parameter, any distributional assumption is probably violated, and this has an important effect on the inferential conclusions.

This can be partially solved with prior distributions, but lack of information remains problematic, as the posterior is almost identical to the prior in the boundary cases. Furthermore, inferences turn out to be too sensitive to the prior assumptions (more on this next lecture).

Solution

Many of these problems can be bypassed by limiting the information available in the selection stage.

There are, essentially, two ways of doing this:

- **Data splitting**: if the data can be written as $Y = [Y^{(1)}, Y^{(2)}]$, with $Y^{(1)}$ and $Y^{(2)}$ independent, use only $Y^{(1)}$, say, for selection.

- **Randomisation**: base selection on a **noisy version of the data**, denoted by U.

  In general, write $U = u(Y, W)$, where Y is the original data and W is artificial noise generated by the statistician.

  For example, $U = Y + W$, $W \sim N(0, I_n)$.

Data splitting

If selection is based on a subsample $Y^{(1)}$, we have two possibilities:

- Base inference on remaining data $Y^{(2)}$: this ignores the specificities of the selection step, and the problem becomes a non-selective one, to which we can apply standard machinery.
  - This ensures that the analysis is valid conditionally and **universally** for any selection rule.
  - In particular, the validity of the conclusions is less affected by model misspecification.
  - BUT, inference is **not admissible**, as it discards the information about the selected parameter provided by $Y^{(1)}$.
- **Data carving**: base inference on the conditional distribution $Y \mid \{Y^{(1)} \in E_1\}$, where E_1 is the selection event in terms of $Y^{(1)}$.
  - It is more powerful, but requires knowledge (and tractability) of the selection event.

Randomisation

The same ideas apply to randomisation.

Given a randomisation scheme U, there exist two possibilities:

- Base inference on $Y \mid U$, thereby discarding all the information provided by U.
- Base inference on $Y \mid \{U \in E_U\}$, where E_U is the selection event in terms of U.

The main difference with respect to data splitting is that for randomisation, even the first option might be difficult to apply if the conditional distribution $Y \mid U$ is complicated.

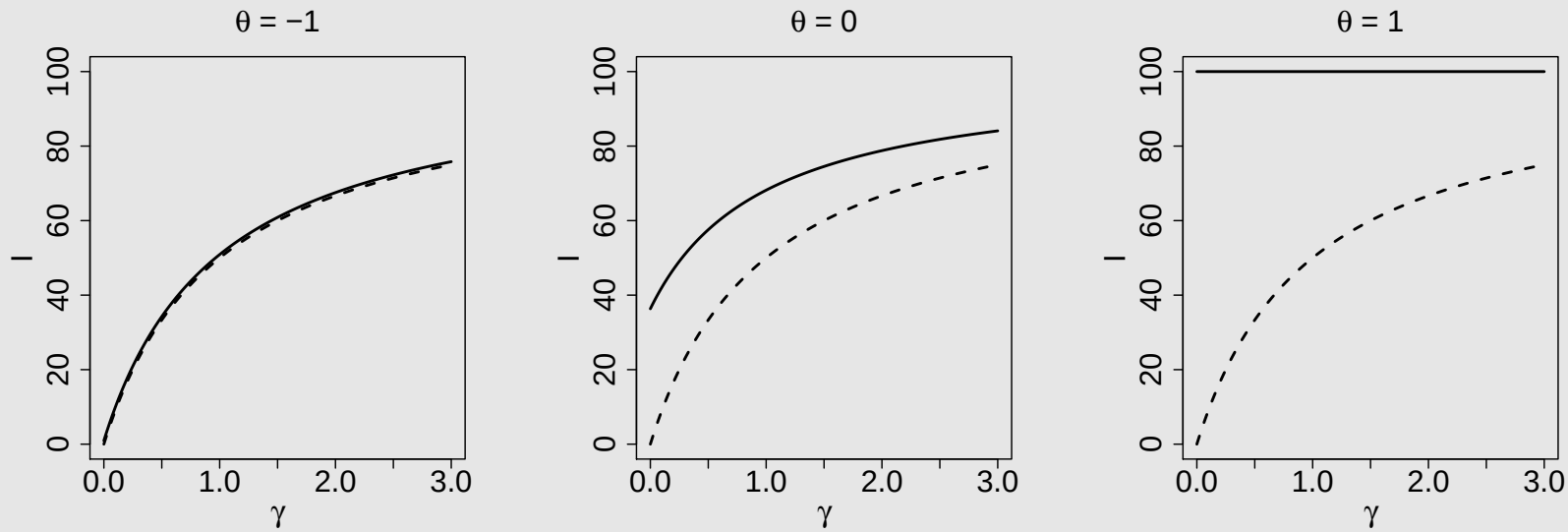
Information trade-off

The loss of power incurred by basing inference on a split vs. doing data carving depends on the parameter.

Consider the model $Y \sim N(\theta, 1/100)$, and suppose the selection event is $U = Y + W > 0$, where $W \sim N(0, \gamma/100)$.

$\rightarrow$ In solid black, Fisher information for θ in $Y \mid \{U > 0\}$.

$\rightarrow$ In dashed black, Fisher information for θ in $Y \mid U$.



Randomisation of Gaussian data

One case where the distribution of $Y \mid U$ admits a very simple form is when the data is Gaussian.

Let $Y \sim N(\mu, \Sigma)$, with $\mu \in \mathbb{R}^n$ unknown and $\Sigma \in \mathbb{R}^{n \times n}$ known and positive definite.

A convenient randomisation strategy is given by $U = Y + W$, where $W \sim N(0, \Sigma_W)$ for some positive definite covariance $\Sigma_W \in \mathbb{R}^{n \times n}$. We then have that

$$U \sim N(\mu, \Sigma + \Sigma_W). \tag{4}$$

Randomisation of Gaussian data

Furthermore, if we define

$$V = Y - \Sigma_W^{-1}\Sigma W, \tag{5}$$

we have that $[U, V]$ is jointly sufficient for Y and that $U \perp V$, so basing inference on $Y \mid U$ amounts to basing it on the marginal distribution of V, which is Gaussian. We call this the "(U, V) decomposition".

Note that this scheme can accommodate different covariance structures.

In regression we are often interested in $\Sigma = \sigma^2 I_n$, in which case a natural option is to take $W \sim N(0, \gamma^2\sigma^2 I_n)$ for some $\gamma > 0$ which **controls information split between selection and inference**, so that, **independently**

$$\begin{aligned} U &\sim N(\mu, \sigma^2\{1+\gamma^2\} I_n); \\ V &\sim N(\mu, \sigma^2\{1+\gamma^{-2}\} I_n). \end{aligned}$$

Randomisation of Gaussian data

Two main differences with data splitting:

- It is **applicable in more settings**, as it doesn't require independent observations.

  $\rightarrow$ For example, it can be used to split time series data.

- Unlike data splitting, it **requires knowledge of the variance**.

  $\rightarrow$ In practice, a plug-in approach turns out to work quite well, provided the variance can be estimated with reasonable accuracy.

Randomisation as information averaging

An appealing feature of randomisation is that it provides a way of **averaging information over multiple data splits using a single noise sample**.

This supports the intuition that it **provides a more balanced information split than data splitting**, and offers a possible a way of selecting the randomisation variance.

This also points to a formal **advantage in terms of inferential power** over the data splits it averages over.

Randomisation as information averaging

Let $Y \sim \mathcal{F}(\beta; X) \in \mathbb{R}^n$ be a random vector whose distribution depends on the design X and on a parameter $\beta \in \mathbb{R}^p$.

For a given subset $r \subseteq \{1, \ldots, n\}$, denote the corresponding data split by $Y^r = (Y_i)_{i \in r}$.

Moreover, denote the Fisher information matrix of Y by $\mathcal{I}_Y(\beta)$.

A data split r **distributes the total information** between the selection and inferential tasks as

$$\mathcal{I}_Y(\beta) = \mathcal{I}_{Y^r}(\beta) + \mathbb{E}_{Y^r}[\mathcal{I}_{Y^{r^c}|Y^r=y^r}(\beta)] \equiv \mathcal{I}_r(\beta) + \mathcal{I}_{r^c|r}(\beta), \tag{6}$$

while a randomisation rule $U = u(Y, W)$ divides the information as

$$\mathcal{I}_Y(\beta) = \mathcal{I}_U(\beta) + \mathbb{E}_U[\mathcal{I}_{Y|U=u}(\beta)] \equiv \mathcal{I}_U(\beta) + \mathcal{I}_{Y|U}(\beta). \tag{7}$$

Randomisation as information averaging

Consider an ordered collection of data splits $\mathcal{R} = (r_1, \ldots, r_m)$ and a collection of positive weights $\mathcal{P} = (p_1, \ldots, p_m)$ adding up to one.

We say that a randomisation strategy $U = u(Y, W)$ **averages the information** over the splits in $\mathcal{R}$ with respect to $\mathcal{P}$ if

$$\mathcal{I}_U(\beta) = \sum_{i=1}^{m} p_i \mathcal{I}_{r_i}(\beta). \tag{8}$$

Note that this also implies that

$$\mathcal{I}_{Y|U}(\beta) = \sum_{i=1}^{m} p_i \mathcal{I}_{r_i^c|r_i}(\beta). \tag{9}$$

$\rightarrow$ The idea is that, if such a randomisation strategy exists, we can "borrow strength" from several splits with a single noise sample.

Randomisation as information averaging

For linear Gaussian models with known covariance, information averaging can be achieved through additive Gaussian randomisation.

Lemma

Let $Y \sim N(X\beta, \Sigma)$, with Σ invertible. For a given $(\mathcal{R}, \mathcal{P})$ such that $\cup_{i=1}^{m} r_i = \{1, \ldots, n\}$, a randomisation scheme that averages the information over $\mathcal{R}$ with respect to $\mathcal{P}$ is given by $U = Y + W$, where $W \sim N(0_n, \Sigma\Sigma_W)$,

$$\Sigma_W = \left\{ \sum_{i=1}^{m} p_i A_{r_i} \Sigma \right\}^{-1} - I_n, \tag{10}$$

$A_{r_i} = E_{r_i}^T (E_{r_i} \Sigma E_{r_i}^T)^{-1} E_{r_i}$ and E_{r_i} is the $0/1$ matrix such that $Y^{r_i} = E_{r_i} Y$.

Randomisation as information averaging

For the case $\Sigma = \sigma^2 I_n$, the appropriate noise distribution is $W \sim N_n(0, \sigma^2 \Sigma_W)$, where Σ_W is diagonal with elements $w_i^{-1} - 1$, with $w_i = \sum_{r: i \in r} p_r$.

In particular, if $\mathcal{R}$ contains all subsets of $\{1, \ldots, n\}$ of size n_1 and all weights are equal, we have $w_i = n_1/n \equiv f$.

Then, the appropriate randomisation has $\Sigma_W = (1 - f) f^{-1} I_n$.

Randomisation as information averaging

The optimality of the Fisher information is commonly measured through summary statistics of its inverse.

The following result shows that any randomisation rule which averages over a random data splitting strategy provides a **more efficient division of the information**.

Proposition.

Let R be a random data splitting rule induced by $(\mathcal{R}, \mathcal{P})$ and φ be a real-valued function defined on the set of $p \times p$ positive definite matrices which is convex and strictly increasing.

Let $U = u(Y, W)$ be randomisation scheme that averages over $\mathcal{R}$ with respect to $\mathcal{P}$, and assume that $\mathcal{I}_r(\beta)$ and $\mathcal{I}_{r^c|r}(\beta)$ are invertible for all $r \in \mathcal{R}$, and that $\mathcal{I}_{r_1}(\beta) \neq \mathcal{I}_{r_2}(\beta)$ for some $r_1, r_2 \in \mathcal{R}$. Then,

$$\varphi\left\{\mathcal{I}_U(\beta)^{-1}\right\} < \mathbb{E}\left[\varphi\left\{\mathcal{I}_R(\beta)^{-1}\right\}\right]; \tag{11}$$

$$\varphi\left\{\mathcal{I}_{Y|U}(\beta)^{-1}\right\} < \mathbb{E}\left[\varphi\left\{\mathcal{I}_{R^c|R}(\beta)^{-1}\right\}\right]. \tag{12}$$

Randomisation as information averaging

Common examples of φ include:

- $\varphi(A) = \text{tr}(A)$: average variance of the estimates of the regression coefficients.
- $\varphi(A) = v^T A v$ for $v \in \mathbb{R}^p$: estimation variance of $v^T \beta$.
- $\varphi(A) = \max\{\text{diag}(A)\}$: estimation variance of the regression coefficients.
- $\varphi(A) = \lambda_{\max}(A)$: variance estimation of the first principal component.

Randomisation as information averaging

In the linear Gaussian model this has a more transparent interpretation.

Assume that $Y \sim N(X\beta, \sigma^2 I_n)$, with $X^T X$ invertible and σ^2 known, and denote by $\hat{\beta}_{r^c}$ and $\hat{\beta}_V$ the maximum likelihood estimators of β based on Y^{r^c} and V, respectively.

When providing inference with a data split r^c, the estimation variance ought to be considered conditional on the split:

$$\mathrm{Var}(\hat{\beta}_{r^c} \mid R = r) = \mathcal{I}_{r^c}(\beta)^{-1},$$

rather than unconditionally, as R is an ancillary.

Taking $\varphi(A) = \eta^T A \eta$ for some $\eta \in \mathbb{R}^p \setminus \{0_n\}$, the previous result gives

$$\mathrm{Var}(\eta^T \hat{\beta}_V) < \mathbb{E}[\mathrm{Var}(\eta^T \hat{\beta}_{r^c} \mid R = r)].$$

Randomisation as information averaging

Therefore, randomisation produces, **on average over the data splits**, **smaller confidence intervals** for any linear combination $\eta^T \beta$ than the data splitting rule it is designed to improve upon.

By the law of total variance we also have the unconditional version of the result, where the variance is computed relative the data and the data splitting rule distributions:

$$\mathrm{Var}(\eta^T \hat{\beta}_V) < \mathrm{Var}(\eta^T \hat{\beta}_{R^c}).$$

This has a different interpretation: on repeated application of the method, estimates based on V will be, on average, more accurate than estimates based on Y^{R^c}.

Randomisation as information averaging

In summary:

- (U, V) decomposition (selection using U, inference using V) is better than data splitting simultaneously **for selection and inference**.
- Different choices of φ have direct interpretation in terms of inferential accuracy.
- Theoretical implications for selection are harder to pinpoint. They will be demonstrated empirically.

Data carving

The previous comparison concerns cases where all the information contained in the data used for selection is discarded (i.e. when inference is conditioned on $Y^{r^c} = y^{r^c}$ or $U = u$).

What about data carving?

We have two main results:

- Carved confidence intervals after randomisation have **bounded length uniformly over the observed data** y.
- Carved confidence intervals after data splitting have, in general, **infinite expected length**.

Data carving: randomisation

Let $Y \sim N(\mu, \sigma^2 I_n)$, $\mu \in \mathbb{R}^n$, and suppose that inference is sought for $\psi = \eta^T \mu$ for some $\eta \in \mathbb{R}^n$.

Let $U = Y + W$, $V = Y - \Sigma_W^{-1} W$, and for a model $M \subseteq \{1, \dots p\}$ write the selection event as $E = \{u \colon \hat{M}(u) = M\}$.

Let $P_\eta = \|\eta\|^{-2} \eta\eta^T$ be the projection matrix onto the line spanned by η, and $F_\psi(x) = P\{\hat{\psi} \leq x \mid U \in E, (I_n - P_\eta)Y = z\}$.

$\rightarrow$ A confidence interval of coverage $\alpha = q_2 - q_1$ is given by $[a(Y), b(Y)]$, where the endpoints solve $F_{a(Y)}(\hat{\psi}) = q_2$ and $F_{b(Y)}(\hat{\psi}) = q_1$.

Data carving: randomisation

If the interval was constructed from the marginal distribution of V alone, via the distribution of $\eta^T V$, it would have length

$$l(q_1, q_2) = \{\eta^T \Sigma_V \eta\}^{1/2} \{\Phi^{-1}(q_2) - \Phi^{-1}(q_1)\}, \tag{13}$$

where $\Sigma_V = \sigma^2\{I_n + \Sigma_W^{-1}\}$.

Since carved inference incorporates extra information coming from $U \mid \{U \in E\}$, the resulting intervals should intuitively not be larger than $l(q_1, q_2)$. This is in fact the case.

Proposition.

The confidence interval defined before has $b(Y) - a(Y) \leq l(q_1, q_2)$.

Data carving: data splitting

For data splitting, assume that selection has been carried out on a subset of the observations Y^r, $r \subset \{1, \ldots, n\}$, so that the selection event can be written as $Y^r \in E_r$ for some $E_r \subseteq \mathbb{R}^{|r|}$.

A carving approach would be based on the distribution $Y \mid \{Y^r \in E_r\}$, which involves $n - |r|$ observations unaffected by selection.

This is, however, not enough in general to avoid arbitrarily large confidence intervals.

Data carving: data splitting

Define $[a(Y), b(Y)]$ as before, but with

$$F_\psi(x) = P\{\hat{\psi} \leq x \mid Y^r \in E_r, (I_n - P_\eta)Y = z\}. \tag{14}$$

The following result extends the "infinite expected length result" for univariate Gaussian models.

Proposition.

Let $E_r \subseteq \mathbb{R}^{|r|}$ and the selected parameter be $\eta^T \mu = \eta_r^T \mu_r + \eta_{r^c}^T \mu_{r^c}$ for some $\eta \in \mathbb{R}^n$. For an observed y with $y^r \in E_r$, define $z = (I_n - P_\eta)y$. If $\inf\{w \in \mathbb{R} \colon z^r + w\eta^r \in E_r\} > -\infty$ or $\sup\{w \in \mathbb{R} \colon z^r + w\eta^r \in E_r\} < \infty$, then $E[b(Y) - a(Y)] = \infty$.

Interpretation: Each coordinate Y_i is only informative about its mean μ_i, so the only information about μ^r available in the conditional distribution $Y \mid Y^r \in E_r$ comes from a truncated Gaussian and the resulting intervals can be arbitrarily large as a consequence.

Model misspecification: CLT

In Gaussian models, efficient information splits are easily achievable via an additive perturbation of the data. However:

- If the errors are not Gaussian, the distributions of U and $Y \mid U$ are generally not available in closed form.
- If the observation variance has to be estimated or the normality assumption is mildly violated, basing inference on the marginal distribution of V is not formally justified.

Nonetheless, even when the model is not Gaussian, the (U, V) decomposition can still provide valid inferences asymptotically.

Model misspecification: CLT

Let $Y = \mu + \varepsilon$, where the components of $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$ are IID with mean zero, variance σ^2, and $\mathbb{E}(|\varepsilon_1|^3) < \infty$.

Set $U = Y + W$ and $V = Y - \gamma^{-1} W$, where $W = \hat{\sigma} Z$, $Z \sim N(0_n, \gamma I_n)$ independent of the data, and $\hat{\sigma}$ is an estimator of σ depending only on the first $[n/2]$ observations.

Assume that the selection event can be written as $\{M^T u \in \mathcal{E}\}$, where M is an $m \times n$ matrix and $\mathcal{E} \subseteq \mathbb{R}^m$ is convex.

Theorem

Under mild assumptions on M, if $\max(\eta)\|\eta\|^{-1} = O(n^{-1/2})$, $\mathbb{E}(|\hat{\sigma}^2 - \sigma^2|) = O(n^{-1/2})$ *and* $P(S = s)^{-1} = o(m^{-3/2} n^{1/2})$, *then*

$$(1 + \gamma^{-2})^{-1/2} \hat{\sigma}^{-1} \|\eta\|^{-1} (\eta^T V - \eta^T \mu) \mid \{\hat{M} = M\} \xrightarrow{d} N(0, 1). \quad (15)$$

Model misspecification: CLT

Some remarks on the assumptions of the result:

- The asymptotic condition on the selection probability ensures that $\hat{\sigma}$ is consistent for σ also conditionally on selection.
- Estimating σ using using only a subset of the observations limits the dependence between $\hat{\sigma}$ and $M^T Y$, ensuring that the distribution of the latter is asymptotically Gaussian.
- The asymptotic condition on η ensures that the asymptotic support of $\eta/\|\eta\|$ is unbounded. Projection parameters satisfy this condition.
- For some standard selection rules the selection event can be represented in the form described above with $M = X$. Furthermore, for these rules selection events can be written as convex polytopes after conditioning on the sign of the selected coefficients.

Model misspecification: CLT

What about more complicated selection rules?

Many complex variable-selection rules are built upon existing, more simple rules such as the lasso or marginal screening.

In those cases, if the base rule admits the required linear representation, the same usually holds for the complex rule.

For example, starting from the fixed-penalty lasso, one can derive more interesting selection rules by:

- **Cross-validation**, which optimises for prediction accuracy.
- **Fixed-X knockoffs**, which controls the false discovery rate.
- **Stability selection**, which controls the expected number of false discoveries.

Simulation

Empirical comparison between data splitting and the (U, V) decomposition.

- The data was generated according to the model $Y = X\beta + N(0_n, \sigma^2 I_n)$, where $\sigma^2 = 1$ but is treated as unknown.
- We used a lasso-based estimator of σ^2.
- Covariates were generated as $X_i \sim N(0, \Gamma)$, where $\Gamma_{ij} = \rho^{|i-j|}$.
- For data splitting we used the DUPLEX algorithm.
- We compare data splitting with splitting fraction f with randomised procedure with $\Sigma_W = (1 - f)f^{-1} I_n$.

Simulation

For selection, we considered two algorithms:

- Fixed-X knockoff.

  FDR control set at 0.3.

- Stability selection with lasso.

  Expected number of false discoveries ≤ 3.

Simulation: selection power

- True values of β generated by sampling 10 non-zero positions uniformly at random and filling them with independent random variables distributed uniformly in the set $\{-1, -0.9, -0.8, \ldots, -0.1, 0.1, \ldots, 0.9, 1\}$.

- We compare the selection ability compared according to:

  - True positive rate: average number of correct discoveries divided by the total number of active covariates.

  - Power: average number of times a coefficient with absolute value $|\beta_i|$ is selected.

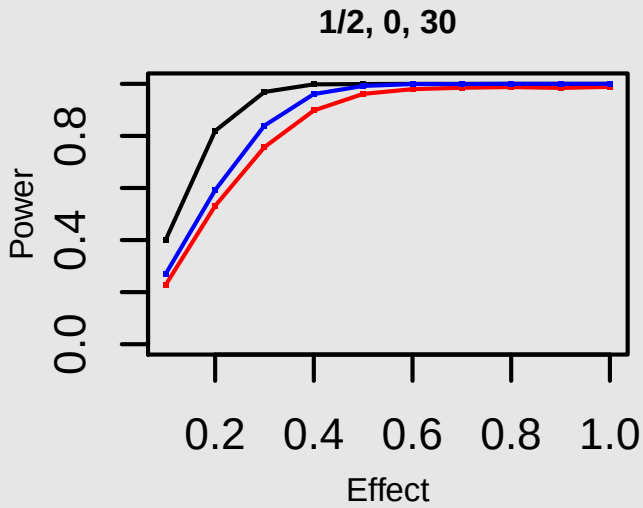
Simulation: selection power

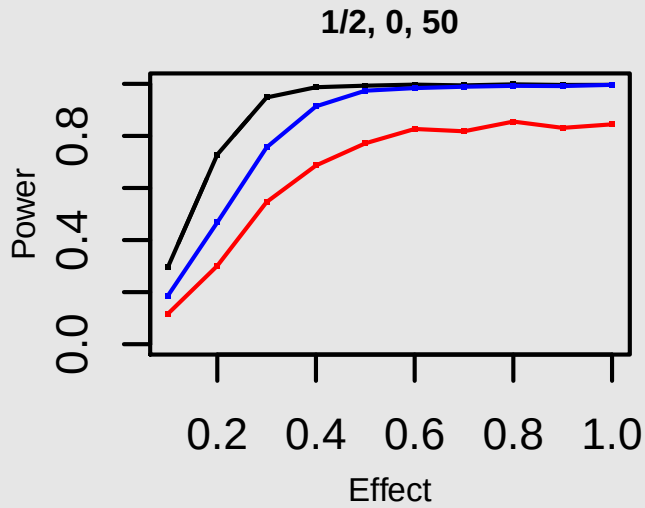
Table: Normalised **true positive rate** of the selection algorithms applied after data splitting (DS) and randomisation (R); $n = 200$.

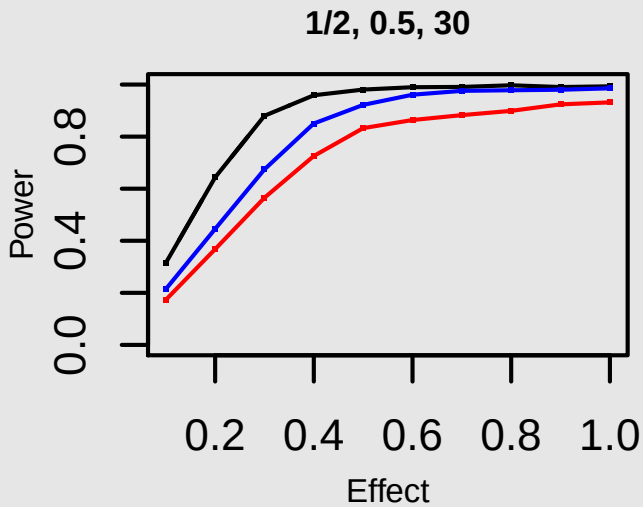
Knockoff			Split		**Stability**			Split	
f	ρ	p	DS	R	f	ρ	p	DS	R
1/2	0	30	0.90	0.94	1/2	0	200	0.69	0.87
1/2	0	50	0.74	0.92	1/2	0	1000	0.49	0.82
1/2	0.5	30	0.82	0.91	1/2	0.5	200	0.69	0.86
1/2	0.5	50	0.65	0.89	1/2	0.5	1000	0.48	0.82
3/4	0	30	0.97	0.98	3/4	0	200	0.90	0.95
3/4	0	50	0.94	0.97	3/4	0	1000	0.83	0.93
3/4	0.5	30	0.94	0.96	3/4	0.5	200	0.89	0.95
3/4	0.5	50	0.89	0.97	3/4	0.5	1000	0.83	0.93

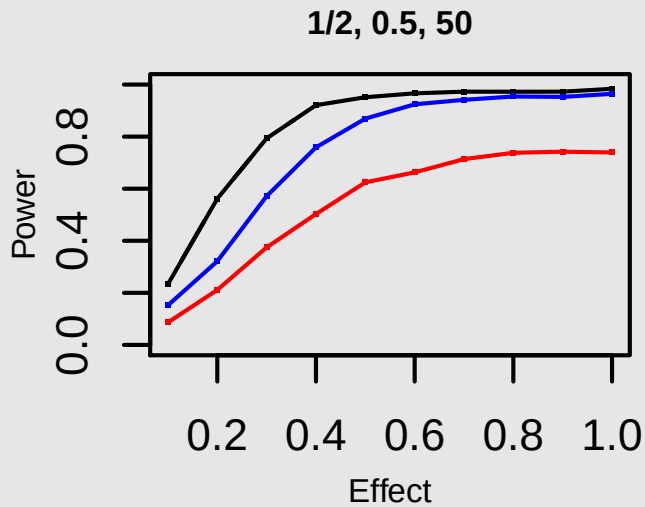
Simulation: selection power

Knockoff: randomisation, data splitting, full dataset.



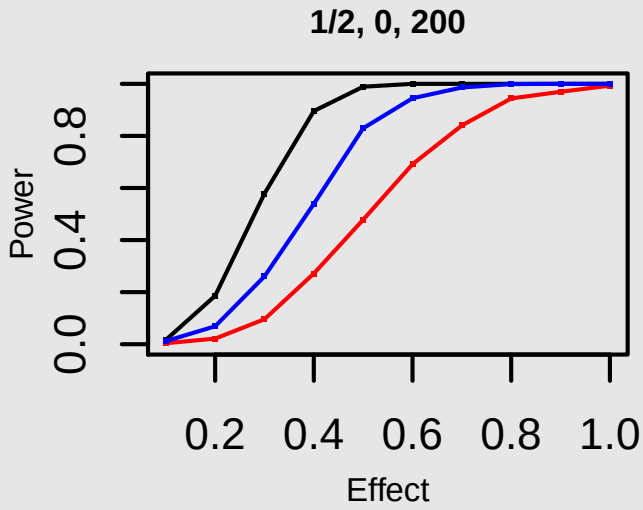


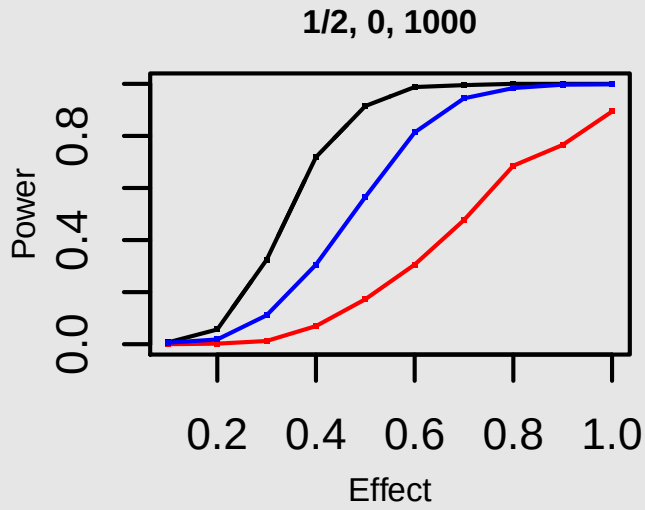


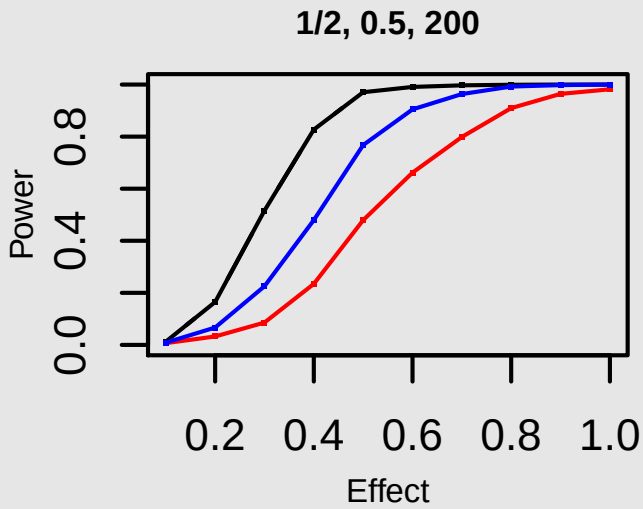


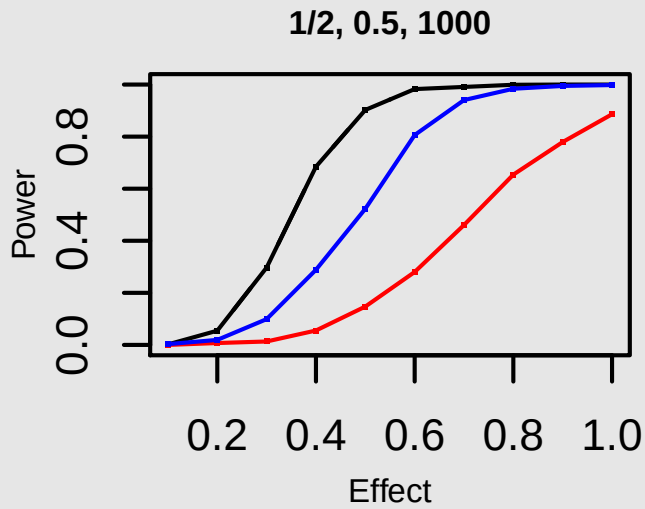
Simulation: selection power

Stability: randomisation, data splitting, full dataset.









Simulation: selection stability

- The selection output should not depend strongly on the random split, so we compare the robustness of the two splitting methods with respect to this criterion.

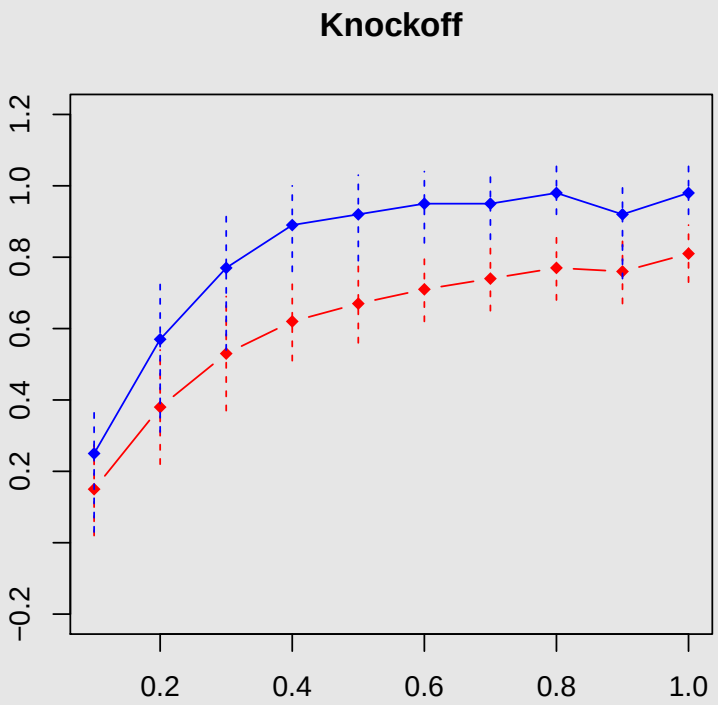
- Set $\beta = (1, 0.9, \ldots, 0.1, 0, \ldots, 0)^T$.

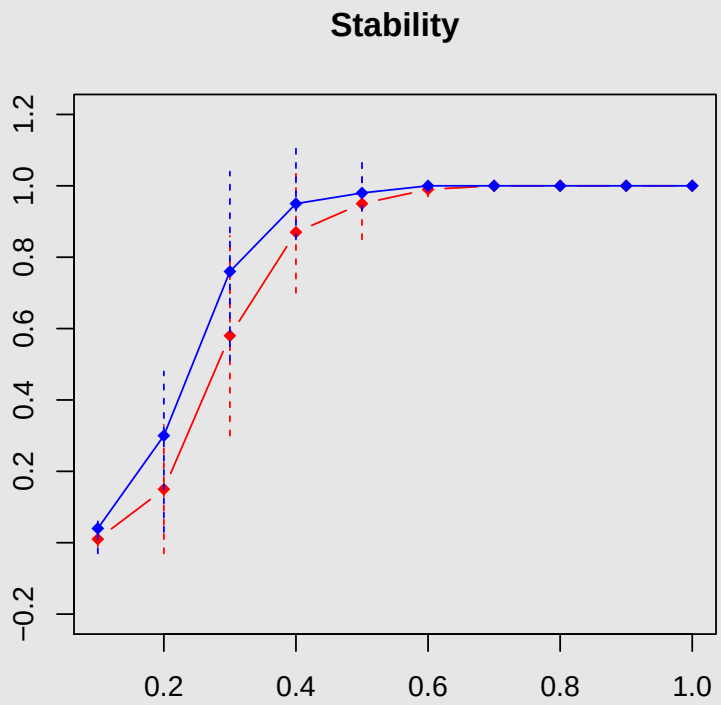
- For each (f, p), we generated 100 pairs (Y, X) and, for each pair, we sampled 50 selection sets uniformly at random and 50 realisations of W.

- For each (Y, X) we estimate $\mathbb{P}(i \in M | Y, X)$.

Simulation: selection stability

Figure: Mean $\pm$ standard dev. of $\mathbb{P}(i \in M | Y, X)$; randomisation, data splitting; $n = 200$, $p = 50,400$, $f = 1/2$.





Simulation: inferential power

Finally, we compare the two methods in terms of the power available at the inferential stage.

- For a given M we construct CI's for β_i with $i \in M$ in the model $Y = X\beta + \varepsilon$.
- We compare the coverage and average length of equal-tailed 90% intervals obtained by both methods.

Simulation: inferential power

Table: **Coverage** of CIs for selected coefficients: knockoff, $n = 200$, $p = 30$.

			$\lvert\beta_i\rvert$			
f	ρ	Split	0	0.2	0.5	1
1/2	0	DS	89.8	91.3	90.0	90.0
		R	89.8	89.2	90.0	90.4
1/2	0.5	DS	90.0	90.8	89.9	90.2
		R	89.6	89.9	89.9	90.3
3/4	0	DS	90.3	89.7	89.7	89.6
		R	89.8	90.2	90.0	89.6
3/4	0.5	DS	90.6	90.1	90.4	89.8
		R	89.2	90.4	89.7	89.6

Simulation: inferential power

Table: **Average length** of CIs for selected coefficients: knockoff, $n = 200$, $p = 30$.

			$\lvert\beta_i\rvert$			
f	ρ	Split	0	0.2	0.5	1
1/2	0	DS	0.39	0.39	0.39	0.39
		R	0.36	0.36	0.36	0.36
1/2	0.5	DS	0.51	0.51	0.51	0.48
		R	0.46	0.46	0.46	0.44
3/4	0	DS	0.65	0.65	0.65	0.65
		R	0.50	0.50	0.51	0.51
3/4	0.5	DS	0.90	0.90	0.90	0.85
		R	0.64	0.65	0.65	0.62

Randomisation: extension

Instead of relying on an asymptotic result for quantitative data, we might want to devise a randomisation strategy that adapts better to different types of observations.

In general, given data Y with distribution $F(y;\theta)$, known up to the parameter θ, we want to find a $U = u(Y, W)$ such that, for all θ:

- The marginal distribution of U is tractable.
- The conditional distribution of $Y \mid U$ is tractable.

Randomisation: extension

Leiner et al. (2023) propose a "conjugate prior reversal" idea that works with any exponential family.

Suppose Y follows a distribution that is a conjugate prior distribution of the parameter in some likelihood.

The idea is to generate U from that likelihood, treating Y as a parameter.

Then, by construction, the conditional distribution $Y \mid U$ will be of the same form as Y (with a different parameter depending on the value of U).

Randomisation: extension

Suppose Y has density/mass function of the form

$$f(y \mid \theta_1, \theta_2) = H(\theta_1, \theta_2) \exp\{\theta_1^T y - \theta_2^T A(y)\}, \tag{16}$$

and suppose we can find $h(\cdot)$, $T(\cdot)$ and θ_3 such that

$$f(u \mid y, \theta_3) = h(u) \exp\{y^T T(u) - \theta_3^T A(y)\} \tag{17}$$

is a well-defined distribution

Then, if $U \sim f(u \mid y, \theta_3)$, we have

$$f(u \mid \theta_1, \theta_2, \theta_3) = h(u) \frac{H(\theta_1, \theta_2)}{H(\theta_1 + T(u), \theta_2 + \theta_3)}; \tag{18}$$

$$f(y \mid u, \theta_1, \theta_2, \theta_3) = f(y \mid \theta_1 + T(u), \theta_2 + \theta_3). \tag{19}$$

As a trivial instance of this result we can recover the Gaussian decomposition, but other examples are more interesting.

Randomisation: extension

– **Binary data**: If $Y \sim \text{Bernoulli}(\theta)$, let $W \sim \text{Bernoulli}(p)$, where $p \in (0,1)$ is a tuning parameter, and set $U = Y(1-W) + (1-Y)W$.

Then, $U \sim \text{Bernoulli}(\theta + p - 2p\theta)$, and

$$Y \mid U \sim \text{Bernoulli}\left(\frac{\theta}{\theta(1-\theta)[p/(1-p)]^{2U-1}}\right).$$

Note: Small values of p allocate more information to U.

– **Count data**: If $Y \sim \text{Poisson}(\theta)$, let $U \sim \text{Binomial}(Y, p)$, where $p \in (0,1)$ is a tuning parameter.

Then, $U \sim \text{Poisson}(p\theta)$, and $V = Y - U \sim \text{Poisson}((1-p)\theta)$ is independent of U.

Note: Large values of p allocate more information to U.

Randomisation: extension

Some remarks:

- This method is adaptive: if we deem the information in U insufficient for selection, we can simply generate more samples from the "posterior distribution" until we have enough information.

- The reverse is not true: if we have randomised too little, we cannot increase the information available for inference post-hoc.

- A drawback of this method with respect to data splitting and the Gaussian (U, V) decomposition is that, for a prespecified proportion $f \in (0, 1)$, we cannot manually allocate $f100\%$ of the information to selection, as the information depends on the unknown parameter.

  Note that with data splitting we can simply use fn samples for selection and the remaining $(1 - f)n$ for inference.