Biometrika (2023), **110**, 3, *pp.* 597–614 https://doi.org/10.1093/biomet/asac070
Advance Access publication 21 December 2022

Splitting strategies for post-selection inference

By D. García Rasines and G. A. Young

Department of Mathematics, Imperial College London, London SW7 2AZ, U.K.
daniel.garcia-rasines16@imperial.ac.uk alastair.young@imperial.ac.uk

Summary

We consider the problem of providing valid inference for a selected parameter in a sparse regression setting. It is well known that classical regression tools can be unreliable in this context because of the bias generated in the selection step. Many approaches have been proposed in recent years to ensure inferential validity. In this article we consider a simple alternative to data splitting based on randomizing the response vector, which allows for higher selection and inferential power than the former, and is applicable with an arbitrary selection rule. We perform a theoretical and empirical comparison of the two methods and derive a central limit theorem for the randomization approach. Our investigations show that the gain in power can be substantial.

Some key words: Data splitting; Post-selection inference; Randomization; Regression; Variable selection.

1. Introduction

Suppose we have data $Y \sim N(\mu, \sigma^2 I_n)$, where $\mu \in \mathbb{R}^n$, $\sigma^2 > 0$ and I_n is the $n \times n$ identity matrix. We assume that the components of μ are modelled as a function of p covariates, $\mu_i = g(x_{i1}, \ldots, x_{ip})$ for some unknown $g\colon \mathbb{R}^p \to \mathbb{R}$, and denote by $X = (x_{ij}) \in \mathbb{R}^{n \times p}$ the known, fixed design matrix. In many situations, it is suspected that only a few covariates are truly active, and a preliminary variable-selection step is performed to identify these. Having screened a set of potentially relevant variables, we may want to provide inference for the regression coefficients of the best linear approximation of μ in the selected model or some other parameter depending on the output of the selection step.

If the same data that were used for selection are also used to provide inference for the selected parameter, standard inferential procedures are unreliable, typically leading to overoptimistic results; see, for example, Hong et al. (2018). Data-splitting techniques, whereby a portion of the data is reserved for uncertainty quantification, offer a simple yet effective way to circumvent this problem. Unfortunately, data splitting often leads to procedures with little power, both for identifying the active covariates and for providing inference for the selected parameters. In this paper we study an alternative to data splitting, motivated by the work of Tian & Taylor (2018), which provides a more efficient way of splitting the sample information, resulting in more powerful procedures, and which is easy to apply with a general variable-selection method. Our main objective is to provide a theoretical and empirical comparison of the two information-splitting strategies.

Data splitting is a popular tool in prediction problems, where the hold-out observations are used to assess the accuracy of a predictive model. When the goal is inference rather than

prediction, a frequent criticism of data splitting is that different splits can produce different selected models, and therefore two people analysing the same data may end up answering different questions. While this is a valid concern, we stress that selection based on the full data is not free from some level of arbitrariness, as the selection process always involves subjective decisions, including the choice of the selection rule itself and in many cases a random input independent of the data; see, for example, Wasserman & Roeder (2009), Meinshausen & Buhlmann (2010) and Candès et al. (2018). Nevertheless, it is important to keep the effect of the random components low, as failure to do so results in high uncertainty about the relevance of the selected variables.

The information-splitting technique considered in this paper can be viewed as a variant of data splitting, which produces datasets that are more similar to the full sample than those resulting from data splitting, and which is therefore potentially less affected by randomness. Furthermore, it entails no extra computational cost over data splitting. The method involves applying the variable-selection algorithm to a randomized version of the data, and then basing inference on the conditional distribution of the data given the randomized form, thereby avoiding any selection bias. The general idea of basing selection on an artificial perturbation of the data in this context was proposed by Tian & Taylor (2018), as a way of deriving uniformly consistent and powerful inferential procedures, and has become a popular device in the literature. In the original paper, the authors compare the resulting inferential power of randomization and data splitting in circumstances where it is possible to provide inference conditionally on the selection event, showing the superiority of the former. Here we compare the methods in circumstances where inference discards all the information of the selection split, and is therefore unaffected by the complexity of the selection rule. Concisely, the new procedure works by generating artificial noise W and transforming the data-noise pair (Y, W) into two independent components, both informative about the generative model, so that one is used for selection and the other for inference.

A large number of methods have been proposed in recent years to deal with selection bias. They can be broadly divided into two categories: those which assume that the selection algorithm is of a specific form, and those which provide guarantees for an arbitrary selection rule. An important class of methods in the first group comprises conditional procedures, as considered above, which are constructed by analysing the conditional distribution of the data given the specified selection event, when this is available. This line of work was started by Lockhart et al. (2014) and has subsequently been extended in multiple works such as Lee & Taylor (2014), Loftus & Taylor (2014), Lee et al. (2016), Fithian et al. (2017), Tibshirani et al. (2018) and Panigrahi et al. (2020). The second group of methods includes the post-selection inference approach of Berk et al. (2013) and extensions of it (Bachoc et al., 2017, 2020), which achieve uniformly valid inference by maximizing over all possible model-selection procedures and are very conservative as a result, as well as the data-splitting approach of Rinaldo et al. (2019), which provides model-free procedures with asymptotically valid guarantees in a random-design setting. Cox (1975) analysed data splitting in a simple inferential problem involving many normal means and found it to be competitive against a natural alternative. Methods based on data splitting have also been considered in more recent works such as Rubin et al. (2006), Wasserman & Roeder (2009), Ignatiadis et al. (2016) and DiCiccio et al. (2020). Fithian et al. (2017) observed that inference after data splitting based only on the hold-out observations is inadmissible, being always dominated by data-carving rules, which consider the sampling distribution of the full data conditional on the selection event. Such data-carving may, however, be very complicated to implement in many situations owing to the complexity of the conditional distribution.

2. Post-selection inference

Suppose that for a given data vector Y, a variable-selection algorithm selects a subset $s \subseteq \{1, \ldots, p\}$ of the covariates. In general, determining an appropriate inferential objective post-selection is not straightforward. Under a linearity assumption, $\mu = X\beta$ for some $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}} \in \mathbb{R}^p$, a natural possibility is to provide inference for the components of β associated with the selected variables, $\{\beta_i : i \in s\}$. However, this is a difficult problem when $p > n$ as the model is not identifiable. A popular alternative target of inference is the projection parameter, proposed by Berk et al. (2013). The projection parameter is the regression parameter of μ projected onto the subspace spanned by the selected columns of X:

$$\beta_s(X) = \underset{z \in \mathbb{R}^{|s|}}{\arg\min}\, E\{\| Y - X(s)z\|^2\} = \{X(s)^{\mathrm{T}}X(s)\}^{-1}X(s)^{\mathrm{T}}\mu,$$

where $X(s)$ is the submatrix of X that contains the selected columns and $|s|$ denotes the number of selected covariates; it is the best linear predictor of μ in the selected model. When the model is linear in the selected covariates, i.e., $\mu = X(s)\beta_s$ for some $\beta_s \in \mathbb{R}^{|s|}$, $\beta_s(X)$ is simply β_s. Otherwise, the interpretation of the projection parameter is less transparent: the jth component of $\beta_s(X)$ may be viewed as the average change of the response when the jth selected covariate increases by one unit, approximated in the selected model. An alternative interpretation is given in § 6.5; see also Berk et al. (2013, § 3.2). When the full model is not linear, one can also consider the projection parameter using the full design matrix, $\beta^{\mathrm{F}}(X) = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\mu$, assuming X has full column rank. Some authors refer to this parameter as the full target and to the previous one as the partial target. In this case, post-selection inference may be provided for the components of $\beta^{\mathrm{F}}(X)$ associated with the selected variables, $\{\beta^{\mathrm{F}}(X)_i : i \in s\}$. Another option, presumably more common in practice, is to proceed under the assumption that $\mu = X(s)\beta_s$ and carry out inference on β_s. In the random-design case, Rinaldo et al. (2019) developed inferential methods for other choices of the selected parameter that depend on the distribution of the covariates.

Let us denote a generic selected parameter, possibly depending on the design matrix as well as on the selected set, by $h_s(\mu; X)$. Adopting the conditional approach of Fithian et al. (2017), we deem an inferential statement about $h_s(\mu; X)$, valid if its error guarantees hold under the conditional distribution of the data given the event that $h_s(\mu; X)$ was selected. For example, a $1 - \alpha$ confidence set T for $h_s(\mu; X)$ is valid if it satisfies

$$\mathrm{pr}\{h_s(\mu; X) \in T \mid S = s\} \geqslant 1 - \alpha,$$

where S is the random set of selected covariates. For simplicity, we will assume that the choice of parameter of interest depends only on s, so that inference on a given $h_s(\mu; X)$ is required if and only if $S = s$.

For some popular variable-selection algorithms, such as the lasso or stepwise procedures with fixed tuning parameters, the conditioning event $\{S = s\}$ can be studied analytically. Often, it can be written as a union of affine sets; see, for example, Lee & Taylor (2014), Loftus & Taylor (2014) and Lee et al. (2016). In most cases, however, this event is too complicated to be explored analytically. Furthermore, even when they can be implemented, conditional methods tend to be very conservative; Kivaranovic & Leeb (2021a) showed, for instance, that in many cases confidence intervals constructed from the conditional distribution of $Y \mid \{S = s\}$ have infinite expected length. In such cases, data splitting offers an analytically simple and computationally light solution to the inference problem.

3. SPLITTING METHODS

The most common form of data splitting is simple data splitting. Here, a fraction $f = n_1/n$, where $1 \leqslant n_1 < n$, is specified, and a set of indices R is chosen uniformly at random from the subsets of $\{1, \ldots, n\}$ of size n_1. Then, for an outcome $R = r$, the sets of observations (Y^r, X^r) and (Y^{r^c}, X^{r^c}) are used for selection and for inference, respectively, where $Y^r = (Y_i)_{i \in r}$, $X^r = (x_{ij})_{i \in r}$ and r^c denotes the complement of r. Since Y^r and Y^{r^c} are independent by assumption, the conditional distribution of the inference set, given the output of the selection step is the same as the unconditional one, so classical procedures can be used to provide valid inference for a selected parameter. More elaborate data-splitting rules can be found in the prediction literature; see, e.g., Reitermanová (2010). These rules allocate the samples to the selection and inferential sets according to the observed values of the covariates, usually trying to divide them as evenly as possible, to ensure that the analyst has access to similar regions of the design space in both stages. Quite generally, then, a data-splitting rule can be formalized as a random variable R, possibly depending on X, taking values in the power set of $\{1, \ldots, n\}$.

Here we consider a different way of distributing the sample information between selection and inference via randomization. Suppose that W is a random quantity, possibly depending on X, and that in the selection step we allow ourselves to observe only the value of a function $U \equiv u(Y, W)$. Since selection depends on the data only through U, inference based on the conditional distribution of the data given the observed value, $Y \mid \{U = u\}$, is free of selection bias and does not require knowledge of the selection mechanism; by contrast, a conditional approach would base inference on $Y \mid \{S(U) = s\}$. In particular, we shall be concerned with cases where it is possible to define a quantity $V \equiv v(Y, W)$ which is independent of U and such that (U, V) is sufficient for Y, so that inference based on the conditional distribution of $Y \mid \{U = u\}$ is equivalent to inference based on the marginal distribution of V.

Tian & Taylor (2018) proposed randomization schemes of the form $U = Y + W$, where W is n-dimensional artificial noise whose variance controls the amount of information reserved for inference: small values assign most of the sample information for selection, while large values allocate most of it for inference. One clear advantage of this approach over data splitting is that it gives access to all the observed values of the covariates at both the selection and the inferential stages, whereas in data splitting we have access to only a subset of them at each stage. To ensure high inferential power, Tian & Taylor (2018) recommended that the distribution of W should have tails at least as heavy as the normal distribution. If the observation variance σ^2 is known, a common choice is $W \sim N(0_n, \sigma^2 \gamma I_n)$, where $\gamma > 0$ and 0_n denotes the n-dimensional vector of zeros. This allows for a remarkably simple analysis, as $U \sim N\{\mu, \sigma^2(1 + \gamma) I_n\}$ is of the same parametric form as the data, and basing inference on $Y \mid \{U = u\}$ amounts to basing it on the marginal distribution of $V = Y - \gamma^{-1} W \sim N\{\mu, \sigma^2(1 + \gamma^{-1}) I_n\}$. This follows because U and V are independent, as they are uncorrelated and normal, and jointly sufficient for μ.

The previous scheme can of course be generalized by considering an arbitrary normal noise vector $W \sim N(0_n, \sigma^2 \Sigma_W)$, with Σ_W positive definite. This produces the split $U = Y + W \sim N\{\mu, \sigma^2(I_n + \Sigma_W)\}$ and $V = Y - \Sigma_W^{-1} W \sim N\{\mu, \sigma^2(I_n + \Sigma_W^{-1})\}$. Henceforth, we refer to this randomization strategy as the (U, V) decomposition. Our goal is to show that this approach provides a better division of the available information than does data splitting.

If σ^2 is unknown, but can be estimated with reasonable precision, an approximate (U, V) decomposition can be achieved by substituting the variance estimate for the variance of

W. In § 5 we consider the asymptotic validity of this approach. In the linear case, if p is small relative to n, σ^2 can be estimated in the classical way. Otherwise, one has to resort to high-dimensional alternatives. In our simulation studies we used the estimator implemented in the `R` (R Development Core Team, 2023) package `selectiveInference` of Tibshirani et al. (2019), which estimates σ^2 using the residual sum of squares from a lasso fit with the penalty parameter tuned by cross-validation. The good performance of this estimator in sparse models was demonstrated by Reid et al. (2016). Other methods are available, such as those of Fan et al. (2012) and Bayati et al. (2013).

4. Theoretical analysis

4.1. *Randomization as information averaging*

An appealing feature of the (U, V) decomposition is that it provides a way of averaging information over multiple data splits using a single noise sample, as we show below. This supports the intuition that it provides a more balanced information split than does data splitting, and offers a possible way of selecting the randomization variance. Moreover, such representation points to a formal advantage in terms of inferential power over the data splits it averages over. In this section we extend the discussion to arbitrary regular parametric models for the sake of completeness, with the Gaussian case serving as an analytically workable example.

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 $\beta \in \mathbb{R}^p$; we will only require mild regularity conditions on the model, and in particular we do not assume that the components of Y are independent. Denote the Fisher information about β in Y by $\mathcal{I}_Y(\beta)$. A data split r distributes the total information between the selection and inferential tasks as

$$\mathcal{I}_Y(\beta) = \mathcal{I}_{Y^r}(\beta) + E_{Y^r}\{\mathcal{I}_{Y^{r^c}|\,Y^r=y^r}(\beta)\} \equiv \mathcal{I}_r(\beta) + \mathcal{I}_{r^c|r}(\beta),$$

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

$$\mathcal{I}_Y(\beta) = \mathcal{I}_U(\beta) + E_U\{\mathcal{I}_{Y|U=u}(\beta)\} \equiv \mathcal{I}_U(\beta) + \mathcal{I}_{Y|U}(\beta).$$

Consider a collection of data splits $\mathcal{R} = (r_1, \ldots, r_m)$ and a set of positive weights $\mathcal{P} = (p_1, \ldots, p_m)$ adding up to 1. We say that the randomization rule $U = u(Y, W)$ averages the information over the splits in $\mathcal{R}$ with respect to $\mathcal{P}$ if

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

This also implies that

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

For linear normal models with known covariance, the following result can be easily verified.

LEMMA 1. *Let $Y \sim N(X\beta, \Sigma)$, where Σ is invertible. For a given $(\mathcal{R}, \mathcal{P})$ such that $\bigcup_{i=1}^{m} r_i = \{1, \ldots, n\}$, a randomization scheme satisfying (1) is $U = Y + W$ and $V = Y - \Sigma\Sigma_W^{-1}\Sigma^{-1}W$, where $W \sim N(0_n, \Sigma\Sigma_W)$ with*

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

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

For the problem considered here, where $\Sigma = \sigma^2 I_n$, we get $\Sigma_W = \Gamma$, where Γ is diagonal with $\Gamma_{ii} = w_i^{-1} - 1$ and $w_i = \sum_{i\in r} p_r$. Furthermore, if $\mathcal{R}$ contains all subsets of $\{1, \ldots, n\}$ of size n_1 and all the weights are equal, then $w_i = n_1/n \equiv f$ so that $\Sigma_W = (1-f)f^{-1}I_n$.

In low-dimensional cases, the optimality of the Fisher information is commonly measured through summary statistics of its inverse. In such cases, an alternative interpretation of (1) becomes relevant. Suppose that the sets $\mathcal{R}$ and $\mathcal{P}$ represent a random data-splitting rule under which r_i is selected with probability p_i. Then any randomization rule that averages over $\mathcal{R}$, with respect to $\mathcal{P}$, provides a more efficient division of the information than the corresponding random data-splitting rule if the optimality of the inverse Fisher information is measured in a particular way. The idea is that for such U, $\mathcal{I}_U(\beta)$ is on average more optimal than $\mathcal{I}_r(\beta)$, and similarly $Y \mid U$ is on average more optimal than $Y^{r^{\mathrm{c}}} \mid Y^r$.

PROPOSITION 1. *Let R be a random data-splitting rule induced by $(\mathcal{R}, \mathcal{P})$, and let 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 a randomization scheme that averages over $\mathcal{R}$ with respect to $\mathcal{P}$, and assume that $\mathcal{I}_r(\beta)$ and $\mathcal{I}_{r^{\mathrm{c}}|r}(\beta)$ 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*

$$\varphi\{\mathcal{I}_U(\beta)^{-1}\} < E\big[\varphi\{\mathcal{I}_R(\beta)^{-1}\}\big], \quad \varphi\{\mathcal{I}_{Y|U}(\beta)^{-1}\} < E\big[\varphi\{\mathcal{I}_{R^{\mathrm{c}}|R}(\beta)^{-1}\}\big].$$

In the linear Gaussian model, Proposition 1 has a direct interpretation in terms of inferential accuracy. Assume that $Y \sim N(X\beta, \sigma^2 I_n)$, with $X^{\mathrm{T}}X$ invertible and σ^2 known, and for a given $(\mathcal{R}, \mathcal{P})$ let U and V be defined as in Lemma 1. Denote by $\hat{\beta}_{r^{\mathrm{c}}}$ and $\hat{\beta}_V$ the maximum likelihood estimators of β based on $Y^{r^{\mathrm{c}}}$ and V, respectively. When providing inference with a data split r^{c}, the estimation variance ought to be considered conditional on the split, $\mathrm{var}(\hat{\beta}_{R^{\mathrm{c}}} \mid R = r) = \mathcal{I}_{r^{\mathrm{c}}}(\beta)^{-1}$, rather than unconditionally, as R is an ancillary. Taking $\varphi(A) = \eta^{\mathrm{T}}A\eta$ for some $\eta \in \mathbb{R}^p \setminus \{0_n\}$, we obtain

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

In particular, when σ^2 is known, randomization produces, on average over the data splits, smaller confidence intervals for any linear combination $\eta^{\mathrm{T}}\beta$ than the data splitting rule it is designed to improve upon. Equal-tailed confidence intervals for $\eta^{\mathrm{T}}\beta$ based on the data split $Y^{r^{\mathrm{c}}}$ are $[\eta^{\mathrm{T}}\hat{\beta}_{r^{\mathrm{c}}} \mp k\,\mathrm{var}(\eta^{\mathrm{T}}\hat{\beta}_{r^{\mathrm{c}}})^{1/2}]$ for some constant k, while intervals with the same coverage based on V are of the form $[\eta^{\mathrm{T}}\hat{\beta}_V \mp k\,\mathrm{var}(\eta^{\mathrm{T}}\hat{\beta}_V)^{1/2}]$. We can therefore state the result in terms of average confidence interval length.

COROLLARY 1. *In the current setting, let $L = \mathrm{var}(\eta^{\mathrm{T}}\hat{\beta}_V)$ and $L(r^{\mathrm{c}}) = \mathrm{var}(\eta^{\mathrm{T}}\hat{\beta}_{r^{\mathrm{c}}})$. Then $L < E\{L(R^{\mathrm{c}})\}$.*

This does not, however, say anything about any particular data split r^{c}, which can potentially produce smaller intervals.

Furthermore, since the maximum likelihood estimators of linear combinations are unbiased, by the law of total variance we also have the unconditional version of the result, where the variance is computed relative to the data and the data-splitting rule distributions:

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

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 $Y^{R^{\mathrm{c}}}$.

In other models, an analogous asymptotic interpretation may be given, provided that the maximum likelihood estimators of β based on U and $Y \mid \{U = u\}$ satisfy the central limit theorem. The precise implications of Proposition 1 at the selection stage are harder to pinpoint. In § 6 we conduct a simulation study to compare data splitting and randomization in terms of selection power.

4.2. *Data carving*

In § 4.1 we compared randomization and data splitting in situations where all the information contained in the data used for selection is discarded. In contexts where the selection event is known and tractable, it may also be possible to carry out inference in a fully conditional manner; this is done by basing inference on $Y \mid \{S(U) = s\}$ in the case of randomization and on $Y \mid \{S(Y^r) = s\}$ in the case of data splitting. In the former case the resulting procedures have been shown to avoid the low-power characteristic of nonrandomized approaches (Kivaranovic & Leeb, 2021b). When selection is applied to a subset of the observations, the approach is commonly known as data carving (Fithian et al., 2017), though for simplicity we will use the term carving to refer to the fully conditional approach based on either type of information split.

For randomization one proceeds as follows. Consider the unrestricted mean model $Y \sim N(\mu, \sigma^2 I_n)$ with $\mu \in \mathbb{R}^n$, and suppose that inference is sought for $\psi = \eta^{\mathrm{T}}\mu$ for some $\eta \in \mathbb{R}^n$. Consider a (U, V) decomposition $U = Y + W$ and $V = Y - \Sigma_W^{-1} W$, and for a selection set $s \subseteq \{1, \ldots p\}$ write the selection event as $E = \{u\colon S(u) = s\}$. Let $P_\eta = \|\eta\|^{-2}\eta\eta^{\mathrm{T}}$ be the projection matrix onto the line spanned by η. Carved confidence intervals for ψ can be obtained from the conditional distribution of $\hat{\psi} = \eta^{\mathrm{T}} Y$ given $\{U \in E, (I_n - P_\eta) Y = z\}$, which is free of nuisance parameters. Specifically, let $F_\psi(x) = \mathrm{pr}\{\hat{\psi} \leqslant x \mid U \in E, (I_n - P_\eta) Y = z\}$. Then a confidence interval of coverage $\alpha = q_2 - q_1$ is $[a(Y), b(Y)]$, where the endpoints solve $F_{a(Y)}(\hat{\psi}) = q_1$ and $F_{b(Y)}(\hat{\psi}) = q_2$. Our construction differs slightly from that of Kivaranovic & Leeb (2021b, p. 13), which conditions instead on the observed value of $(I_n - P_\eta) U$. A plausible criticism of the latter approach is that the interval is not a function of Y alone and is therefore in violation of the sufficiency principle.

If the interval for ψ were constructed from the marginal distribution of V alone, via the distribution of $\eta^{\mathrm{T}} V$, it would have length $l(q_1, q_2) = (\eta^{\mathrm{T}}\Sigma_V \eta)^{1/2}\{\Phi^{-1}(q_2) - \Phi^{-1}(q_1)\}$, where $\Sigma_V = \sigma^2(I_n + \Sigma_W^{-1})$. Since carved inference incorporates extra information coming from $U \mid \{U \in E\}$, intuitively the resulting intervals should be no larger than $l(q_1, q_2)$. The following result, an extension of Theorem 1 in Kivaranovic & Leeb (2021b), confirms this intuition.

PROPOSITION 2. *The confidence interval defined above has* $b(Y) - a(Y) \leqslant l(q_1, q_2)$.

For data splitting the matter is more delicate. Assume that selection has been carried out on a subset of the observations, Y^r for $r \subset \{1, \ldots, n\}$, so that the selection event can be written as $Y^r \in E_r$ for some $E_r \subseteq \mathbb{R}^{|r|}$. Then a fully conditional 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. Indeed, define $[a(Y), b(Y)]$ as before with $F_\psi(x) = \mathrm{pr}\{\hat{\psi} \leqslant x \mid Y^r \in E_r, (I_n - P_\eta)Y = z\}$. The following proposition follows trivially from Proposition 1 of Kivaranovic & Leeb (2021a).

PROPOSITION 3. *Let $E_r \subseteq \mathbb{R}^{|r|}$, and let the selected parameter be $\eta^{\mathrm{T}}\mu = \eta_r^{\mathrm{T}}\mu_r + \eta_{r^c}^{\mathrm{T}}\mu_{r^c}$ 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 $\inf\{w \in \mathbb{R}: z^r + w\eta^r \in E_r\} > -\infty$ or $\sup\{w \in \mathbb{R}: z^r + w\eta^r \in E_r\} < \infty$, then $E\{b(Y) - a(Y)\} = \infty$.*

A point of confusion could arise from the fact that carved intervals in the random-sample setting, where n independent copies of $N(\mu, \sigma^2 I)$ are available and a subset of them is used for selection, do in fact have finite expected length by the analysis of Kivaranovic & Leeb (2021b). In the regression setting, however, each coordinate Y_i is informative only 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.

5. ASYMPTOTIC VALIDITY OF THE (U, V) DECOMPOSITION

In Gaussian models, efficient information splits are easily achievable via an additive perturbation of the data. However, when the observation error and the randomization noise are not normal, the distributions of U and $Y \mid U$ are generally not available in closed form, complicating the selection and inferential analyses. Furthermore, when the observation variance has to be estimated or the normality assumption is mildly violated, basing inference on the marginal distribution of V, as described above, is not formally justified. In this section we provide a set of conditions under which the (U, V) decomposition is asymptotically valid in more general settings.

Suppose that the model is $Y = \mu + \varepsilon$ with $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathrm{T}}$, where the errors are a random sample from an unknown distribution. We assume that the parameter of interest can be written as $\eta^{\mathrm{T}}\mu$ for some vector $\eta \in \mathbb{R}^n$. If the errors are $N(0, \sigma^2)$ and we know σ^2, exact post-selection inference for $\eta^{\mathrm{T}}\mu$ after selection based on $U = Y + W$, with $W \sim N(0_n, \sigma^2\gamma I_n)$, is provided via $\eta^{\mathrm{T}}V \sim N\{\eta^{\mathrm{T}}\mu, \sigma^2(1+\gamma^{-1})\|\eta\|^2\}$, where $\|\cdot\|$ denotes the Euclidean norm. For simplicity we assume that the randomization variance is of the form $\Sigma_W = \gamma I_n$. Theorem 1 gives conditions under which this approach is asymptotically valid when the errors are not normal and the observation variance is unknown, but can be estimated with enough precision. Unless otherwise stated, all the elements involved in the analysis depend on the sample size n. Also, for a vector or matrix A, $\max(A)$ denotes the maximum absolute entry of A; and for a square matrix A, $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the smallest and largest eigenvalues of A.

THEOREM 1. *Let $Y = \mu + \varepsilon$, where $\mu \in \mathbb{R}^n$ and the components of $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathrm{T}}$ are independent and identically distributed with mean zero and variance σ^2 and $E(|\varepsilon_1|^3) < \infty$. Define $U = Y + W$ and $V = Y - \gamma^{-1}W$, where $W = \hat{\sigma}Z$ with $Z \sim N(0_n, \gamma I_n)$ 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 $\{u: S(u) = s\}$ can be written as $\{M^{\mathrm{T}}u \in \mathcal{E}\}$, where M is an $m \times n$ matrix and $\mathcal{E} \subseteq \mathbb{R}^m$ is convex. Write $M^{\mathrm{T}} = [M_1^{\mathrm{T}} M_2^{\mathrm{T}}]$, where M_1 contains the first $[n/2]$ rows*

of M, and assume that for $A \in \{M_1, M_2\}$, $A^{\mathrm{T}}A$ is invertible for all n, $\lambda_{\max}(A^{\mathrm{T}}A) = O(n)$, $\lambda_{\max}\{(A^{\mathrm{T}}A)^{-1}\} = O(n^{-1})$ and $\max(A) = O(1)$; also assume that $\max(\eta)\|\eta\|^{-1} = O(n^{-1/2})$. If, as $n \to \infty$, $E(|\hat{\sigma}^2 - \sigma^2|) = O(n^{-1/2})$ and $\mathrm{pr}(S = s)^{-1} = o(m^{-3/2}n^{1/2})$, then

$$(1 + \gamma^{-2})^{-1/2}\hat{\sigma}^{-1}\|\eta\|^{-1}(\eta^{\mathrm{T}}V - \eta^{\mathrm{T}}\mu) \mid \{S = s\} \to N(0, 1)$$

in distribution as $n \to \infty$. When $\varepsilon_i \sim N(0, \sigma^2)$, the convexity requirement is not needed and the asymptotic requirement on the selection probability can be relaxed to $\mathrm{pr}(S = s)^{-1} = o(m^{-1/2}n^{1/2})$.

Importantly, the requirements on $\hat{\sigma}$ are independent of the selection events: the assumed lower-bound asymptotic condition on the selection probability ensures that $\hat{\sigma}$ is consistent for σ also conditionally on selection. The reason for estimating σ using only a subset of the observations is to limit the dependence between $\hat{\sigma}$ and $M^{\mathrm{T}}Y$, ensuring that the distribution of the latter is asymptotically Gaussian. For some standard selection rules such as the lasso, lars and stepwise regression with fixed hyperparameters, the selection event can be represented in the form described above with $M = X$, in which case the conditions on M are standard. Furthermore, for these rules selection events can be written as convex polytopes after conditioning on the sign of the selected coefficients. The asymptotic condition on η is very natural and ensures that the asymptotic support of $\eta/\|\eta\|$ is unbounded. When inference is sought for a projection parameter, this condition is satisfied under very mild conditions; see the Supplementary Material.

Recently there have been two important proposals for controlling Type I error in variable selection: stability selection (Meinshausen & Buhlmann, 2010; Shah & Samworth, 2013) and fixed-X knockoffs (Barber & Candès, 2015). These methods are not themselves variable-selection algorithms, but rather modes of implementation of existing ones that ensure error control. Thus far the resulting selection algorithms have been deemed beyond the reach of analytical conditional methods owing to the untraceability of their selection events. Simulation approaches have been proposed to bypass this problem, but they can be very computationally expensive (Markovic et al., 2019). In the following proposition we show that, when applied in conjunction with the lasso, the selection events of the respective algorithms can be written in the form required by Theorem 1 after some appropriate conditioning. We refer the reader to the respective articles for details about the algorithms.

PROPOSITION 4. *Consider the selection functions $S(y; X)$ for the stability selection and knockoff algorithms paired with the lasso. For all $s \subseteq \{1, \ldots, n\}$ the following hold.*

(i) *Stability selection: suppose that the lasso is applied with a fixed penalty $\lambda > 0$, and fix the set of splits $I_1, \ldots, I_B \subseteq \{1, \ldots, n\}$ to which the algorithm is applied. The events $\{y\colon S(y; X) = s\} \cap \{M^b(y) = m^b\colon b = 1, \ldots, B\}$ are convex for all $\{m^b\colon b = 1, \ldots, B\}$, where $M^b(y) = \mathrm{sign}\{\hat{\beta}^b(\lambda)\}$ and $\hat{\beta}^b(\lambda)$ is the lasso solution for the data (y_{I_b}, X_{I_b}). Furthermore, $S(y; X)$ is a function of My, with $M = (A_1X \ldots A_BX)$ where A_i is a diagonal matrix such that $(A_i)_{jj} = 1$ if $j \in I_i$ and $(A_i)_{jj} = 0$ otherwise, and if $\max(X) = O(1)$ and*

$$\min_{i \neq j} \lambda_{\min}(X^{\mathrm{T}}_{I_i \cap I_j^{\mathrm{c}}} X_{I_i \cap I_j^{\mathrm{c}}}) \geqslant cn$$

for some constant c, then $\max(M) = O(1)$ and $\lambda_{\max}\{(M^{\mathrm{T}}M)^{-1}\} = O(n^{-1})$. In this case, $m = Bp$.

(ii) *Knockoffs: the events* $\{S(y;X)=s\}\cap\{A(y)=A, s_A(y)=s_A\}$ *are convex for all* (A,s_A)*, where* $A(y)$ *and* $s_A(y)$ *are the image of the active set of the lasso solution and the image of the set of active signs as* λ *goes from* ∞ *to* 0*. Furthermore,* $S(y;X)$ *is a function of* My*, with* $M=(X\tilde{X})$ *where* $\tilde{X}$ *is the knockoff copy of* X*, and if* $\max(X)=O(1)$*,* $\lambda_{\max}(X^{\mathrm{T}}X)=O(n)$ *and* $s_i\leqslant 2\lambda_{\max}(X^{\mathrm{T}}X)-cn$ *for all* i *and some universal constant* $c>0$*, where* s_i *is the* i*th diagonal entry of* $X^{\mathrm{T}}X-X^{\mathrm{T}}\tilde{X}$*, then* $\max(M)=O(1)$ *and* $\lambda_{\max}\{(M^{\mathrm{T}}M)^{-1}\}=O(n^{-1})$*. In this case,* $m=2p$*.*

6. Simulation study

6.1. *Setting*

We compared data splitting with the (U,V) decomposition in the context of the normal linear model. The data-generating process was taken to be $Y=X\beta+\epsilon$, with ϵ distributed as $N(0_n,I_n)$ and $\beta\in\mathbb{R}^p$ an unknown sparse vector of coefficients. At each replication of the simulations, the rows of the design matrix were generated as independent samples from the distribution $N(0_p,\Gamma)$, where $\Gamma\in\mathbb{R}^{p\times p}$ is a Toeplitz matrix with (i,j) entry $\rho^{|i-j|}$ for some $\rho\geqslant 0$. The observation variance, fixed at 1, was assumed to be unknown, and was estimated in the classical way when $p<n/4$ and by using the high-dimensional alternative otherwise, as per § 3. We compared the two information-splitting strategies in terms of selection power, selection stability and inferential power in low- and high-dimensional settings.

The data-splitting rule considered was the duplex (Snee, 1977). In the even case, $|r|=n/2$, duplex finds the two covariate observations that are farthest apart and assigns them to the selection set. Then it finds the next two observations that are farthest apart and assigns them to the inference set. Finally, it allocates the remaining samples one at a time, rotating between the selection and inference sets, by selecting the remaining sample that is farthest apart from the rest of the observations in the given set, starting with the selection set. If the selection and inferential sets have unequal sizes, the algorithm is applied until the smaller set is filled, and all the remaining observations are assigned to the other one. For a given splitting fraction $f=|r|/n$, we compared the performance of duplex with the performance of the (U,V) decomposition with $\Sigma_W=(1-f)f^{-1}I_n$. Recall from § 4.1 that such a randomization strategy can be interpreted as a form of averaging information over data splits of the same size $|r|$ as the data split.

Thus, for a given f and dataset (Y,X), we considered two procedures. The first bases selection on (Y^r,X^r) and inference on (Y^{r^c},X^{r^c}), and the second bases selection on $(U,X)=(Y+W,X)$ and inference on $(V,X)=(Y-\gamma^{-1}W,X)$, with $W=\hat{\sigma}Z$ where $Z\sim N(0_n,I_n)$ is artificially generated noise independent of the data and $\hat{\sigma}^2$ is the estimate of the model variance.

Regarding selection, we considered the fixed-X knockoff algorithm of Barber & Candès (2015) and the stability selector paired with the lasso proposed by Meinshausen & Buhlmann (2010), and improved by Shah & Samworth (2013), as implemented in the `R` packages `knockoff` (Barber et al., 2020) and `stabs` (Hofner & Hothorn, 2017). As mentioned earlier, these are selection rules for which conditional inference is analytically intractable and computational approaches are very demanding, so they constitute an example where information-splitting approaches would likely be preferred in practice. These algorithms aim to identify the set of active coefficients while keeping the number of false discoveries under control. The knockoff provides a guarantee on the false discovery rate, while stability selection controls the expected number of false discoveries. In all simulations we set the

Table 1. *True positive rate of the selection algorithms applied after data splitting and randomization, normalized by the true positive rate of selection applied to the full dataset*

Knockoff			Split		Stability			Split	
f	ρ	p	DS	R	f	ρ	p	DS	R
1/2	0	30	0.903	0.942	1/2	0	200	0.694	0.867
1/2	0	50	0.738	0.923	1/2	0	1000	0.486	0.821
1/2	0.5	30	0.820	0.914	1/2	0.5	200	0.691	0.858
1/2	0.5	50	0.648	0.890	1/2	0.5	1000	0.478	0.820
3/4	0	30	0.972	0.978	3/4	0	200	0.895	0.948
3/4	0	50	0.935	0.971	3/4	0	1000	0.826	0.933
3/4	0.5	30	0.940	0.964	3/4	0.5	200	0.886	0.948
3/4	0.5	50	0.890	0.970	3/4	0.5	1000	0.826	0.934

DS, data splitting; R, randomization.

false discovery rate of the knockoff algorithm to 0.3 and set the expected number of false discoveries and cut-off threshold of the stability algorithm to 3 and 0.7, respectively. Since the knockoff is applicable only in settings with fewer covariates than observations, it was considered only in the lower-dimensional cases.

6.2. *Selection power*

In this simulation we generated 10^3 triplets (β, Y, X) independently for each combination of the following parameters: $n = 200$; $\rho = 0, 0.5$; $f = 1/2, 3/4$ and correspondingly $\gamma = 1, 3^{-1/2}$; $p = 30, 50$ for the knockoff algorithm and $p = 200, 1000$ for the stability selection algorithm. For each combination of the parameters and each repetition, the true β was generated by sampling 10 nonzero positions uniformly at random, and filling them with independent random variables distributed uniformly in the set $\{-1, -0.9, -0.8, \ldots, -0.1, 0.1, \ldots, 0.9, 1\}$. Then, for each β a pair (Y, X) was generated, and the corresponding selection algorithm was applied to (Y^r, X^r) in the case of data splitting and to (U, X) in the case of the randomized procedure.

The selection abilities of the methods were compared according to two criteria: true positive rate and power. The true positive rate is the average number of correct discoveries divided by the total number of active covariates, and the power is defined, for each possible value of $|\beta_i| = 0.1, \ldots, 1$, as the average number of times a coefficient with absolute value $|\beta_i|$ is selected, averaged over all the generated coefficients of all the β_i. We also computed, for comparison, the results from applying the selection algorithms to the full dataset, (Y, X).

Table 1 shows the observed true positive rates of data-splitting and randomization divided by the observed true positive rates of full-data selection, and Figs. 1 and 2 show the empirical power functions of the three methods. Only plots for $\rho = 0.5$ are shown; the plots for $\rho = 0$ are almost identical. The results clearly favour randomization over data splitting, particularly in the cases with larger values of p. Quite remarkably, despite the fact that the choices of f and γ were balanced, selection based on a randomized split yielded performance which was in some cases closer to full-data selection than to data-splitting selection.

6.3. *Selection stability*

In addition to having high power, it is important that the selection method not depend strongly on ancillary components of the analysis: in the case of data splitting, on the choice

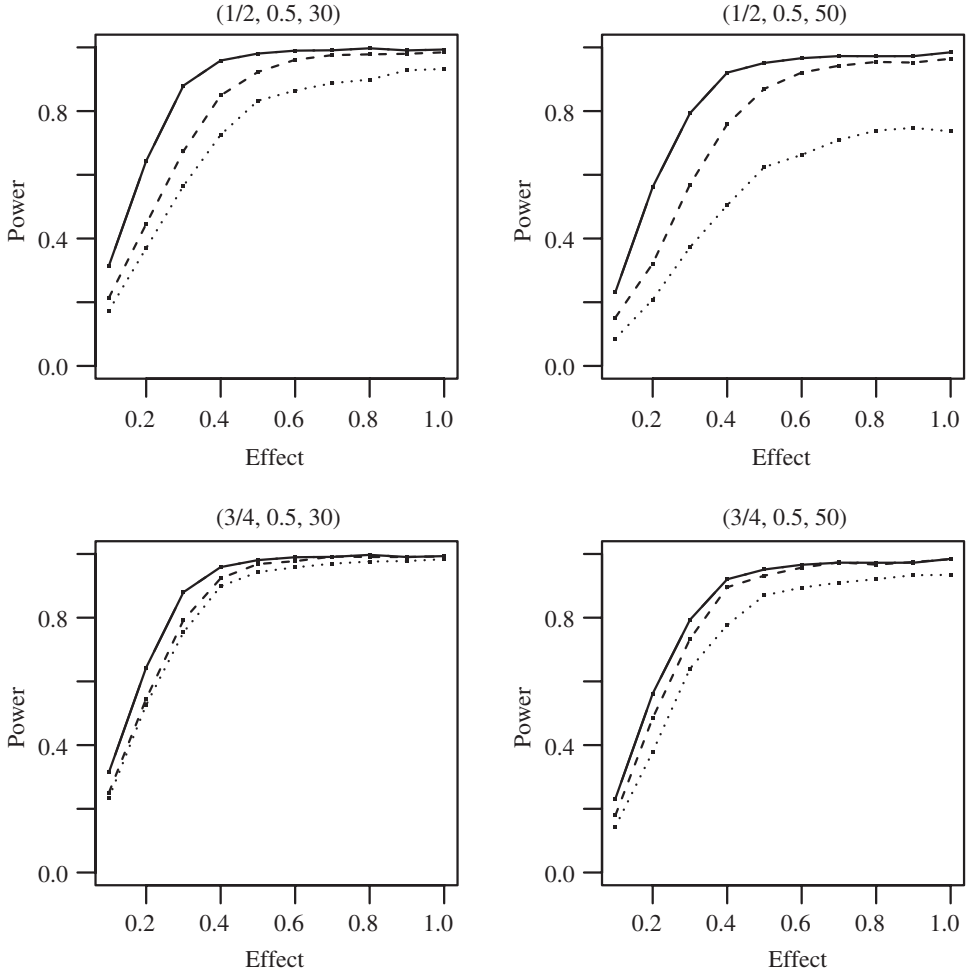

Fig. 1. Power of the knockoff applied to the complete dataset (solid), a randomized version of the data (dashes) and a data split (small dashes). The title above each panel indicates the values of (f, ρ, p).

of the selection and inference sets; and in the case of randomization, on the observed value of the artificial noise W. To measure the stability of the selection strategies with respect to these elements, a simulation was conducted in which the selection algorithms were applied to multiple splits of the same dataset. Since DUPLEX is a deterministic allocation rule, in this section we considered a simple data-splitting scheme instead.

We set $\beta = (1, 0.9, \ldots, 0.1, 0, \ldots, 0)^{\mathrm{T}}$, $\rho = 0.5$, $f = 1/2, 3/4$ and $p = 50$ for the knockoff, and set $p = 400$ for stability selection. For each (f, p) we generated 100 pairs (Y, X), and for each pair we sampled 50 selection sets uniformly at random and 50 realizations of the randomization noise W, and then applied the selection algorithms to each data split and perturbed instance of the data. For each (Y, X) we recorded the average number of times each active covariate was selected across the different splits, an estimate of $\mathrm{pr}(i \in S \mid Y, X)$ $(i = 1, \ldots, 10)$. The empirical means and standard deviations of the 100 estimated averages are reported in Table 2.

Randomization was more stable than data splitting, because the corresponding selection probabilities of the active covariates were more concentrated around higher values. For example, for the knockoff with $f = 1/2$, for most values of (Y, X) we had a more than 90%

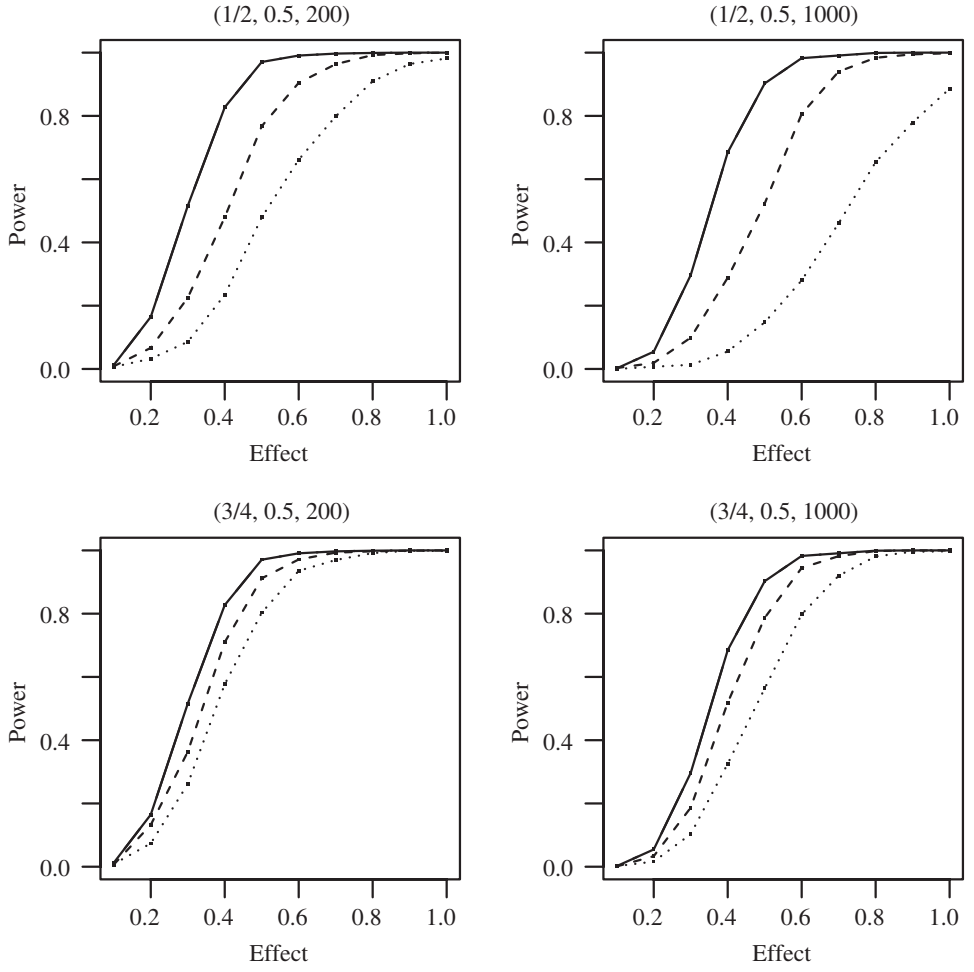

Fig. 2. Power of stability selection applied to the complete dataset (solid), a randomized version of the data (dashes) and a data split (small dashes). The title above each panel indicates the values of (f, ρ, p).

probability of selecting β_3 under repeated sampling of W, as opposed to the approximately 77% for data splitting.

6.4. *Inference for selected coefficients*

In this simulation we consider the problem of constructing confidence intervals for the selected coefficients $\{\beta_i : i \in s\}$. Firstly, we show that a face-value approach, which reports the standard confidence intervals ignoring selection, undercovers coefficients with small effect size $|\beta_i|$, while the information-splitting techniques yield valid intervals. This gives an illustration of the need to take selection into account at the inferential stage. Secondly, we compare the lengths of the intervals derived from the randomized procedure with those obtained by data splitting. To avoid complications related to confidence interval construction in nonidentifiable settings, we consider only low-dimensional cases here.

Assume that X has full rank and let $\hat{\beta}_i = e_i^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y$ be the ordinary least squares estimator of β_i. In a classical, nonselective setting, where the inferential goals are determined prior to the data collection, marginal inference for β_i is based on the pivot $T_i = (\hat{\beta}_i - \beta_i)/\hat{\sigma}$, which follows a scaled t distribution with $n - p$ degrees of freedom, where $\hat{\sigma}^2$ is the classical

Table 2. *Means and standard deviations of the estimated selection probabilities for fixed values of* (Y, X)

		β_i									
f	Split	1	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
		Knockoff									
1/2	DS	0.81 (0.08)	0.76 (0.09)	0.77 (0.09)	0.74 (0.09)	0.71 (0.09)	0.67 (0.11)	0.62 (0.11)	0.53 (0.16)	0.38 (0.16)	0.15 (0.13)
	R	0.98 (0.06)	0.92 (0.18)	0.98 (0.06)	0.95 (0.10)	0.95 (0.11)	0.92 (0.13)	0.89 (0.13)	0.77 (0.23)	0.57 (0.26)	0.25 (0.22)
3/4	DS	0.98 (0.14)	0.95 (0.22)	0.97 (0.17)	0.97 (0.17)	0.96 (0.20)	0.88 (0.33)	0.89 (0.31)	0.80 (0.40)	0.62 (0.49)	0.25 (0.44)
	R	0.99 (0.04)	0.93 (0.19)	0.99 (0.04)	0.97 (0.10)	0.98 (0.08)	0.96 (0.10)	0.96 (0.11)	0.86 (0.22)	0.69 (0.29)	0.34 (0.32)
		Stability									
1/2	DS	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.01)	0.99 (0.02)	0.95 (0.10)	0.87 (0.17)	0.58 (0.28)	0.15 (0.18)	0.01 (0.02)
	R	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.01)	0.98 (0.05)	0.95 (0.10)	0.76 (0.25)	0.30 (0.27)	0.04 (0.07)
3/4	DS	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.98 (0.14)	0.88 (0.33)	0.31 (0.47)	0.08 (0.27)
	R	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.01)	1.00 (0.02)	0.90 (0.20)	0.44 (0.37)	0.06 (0.14)

DS, data splitting; R, randomization.

variance estimator. If, however, inference on a given β_i is provided only for some data samples, T_i is no longer pivotal and the resulting confidence intervals can be miscalibrated. To exemplify this, we ran the following simulation. We fixed $n = 200$, $p = 30$, the true parameter $\beta = (1, -1, 0.5, -0.5, 0.2, -0.2, 0, 0, \ldots, 0)^{\mathrm{T}}$, $\rho = 0, 0.5$ and $f = 1/2, 3/4$, and for each combination of (ρ, f) we generated 5×10^3 pairs (Y, X). We then applied the selection algorithm to each dataset, as in the previous subsections, and constructed classical equal-tailed confidence intervals based on T_i for each selected coefficient, with a nominal coverage of 90%. Table 3 shows the observed coverages of these intervals in the rows indicated by face value, averaged across coefficients with equal absolute effect. Clearly, the face-value intervals are unreliable when $|\beta_i| = 0$ or 0.2, with actual coverages that are significantly lower than the nominal one in most cases.

The rows labelled HD contain the coverages of the intervals constructed using only the hold-out split of the data. In the case of data splitting, the intervals were derived from the t-pivots of the hold-out observations, $\hat{\beta}_i^{\mathrm{DS}} = e_i^{\mathrm{T}}(X^{r^{\mathrm{c}}T} X^{r^{\mathrm{c}}})^{-1} X^{r^{\mathrm{c}}T} Y^{r^{\mathrm{c}}}$. With the current simulation parameters, the number of remaining observations is always greater than the number of covariates, so $X^{r^{\mathrm{c}}}$ has full rank almost surely. For the randomized procedure, we used the approximation $\hat{\beta}_i^{\mathrm{R}} = e_i^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}V \dot{\sim} N\{\beta_i, (1+\gamma^{-2})\sigma^2 e_i^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}e_i\}$. The intervals were thus constructed by studentization of $\hat{\beta}_i^{\mathrm{R}}$. Despite the distributional approximation, we see that the coverages of the resulting confidence intervals were very close to the nominal one.

For the hold-out methods we also recorded the average lengths of the intervals, which are presented in Table 4. The intervals provided by the randomized procedure were always shorter, on average, than those provided by the data-splitting procedure. In the cases where the amount of information reserved for inference was small, i.e., $f = 3/4$, the matrices $X^{r^{\mathrm{c}}}$ were of dimension 50×30 and the resulting intervals were very wide as a consequence.

Table 3. *Coverages of confidence intervals for the selected coefficients*

				$\lvert\beta_i\rvert$							
				Knockoff				Stability			
f	ρ	Split	Method	0	0.2	0.5	1	0	0.2	0.5	1
1/2	0	DS	FV	68.7	90.1	90.8	89.8	39.1	80.2	91.1	89.8
		R	FV	67.3	89.9	90.6	89.8	34.9	82.6	91.1	89.8
		DS	HD	89.8	91.3	90.0	90.0	90.3	90.6	90.0	90.0
		R	HD	89.6	89.9	89.9	90.3	90.1	89.4	89.8	89.7
1/2	0.5	DS	FV	73.9	85.2	91.0	89.9	61.7	81.6	89.5	90.0
		R	FV	72.0	84.0	91.2	89.7	55.8	80.1	89.5	89.9
		DS	HD	90.0	90.8	89.9	90.2	90.6	92.5	89.9	90.1
		R	HD	89.8	89.2	90.0	90.4	90.8	89.6	90.5	89.9
3/4	0	DS	FV	59.2	91.3	90.5	89.8	21.9	83.7	90.7	89.8
		R	FV	57.4	91.8	90.4	89.8	14.2	85.0	90.6	89.8
		DS	HD	90.3	89.7	89.7	89.6	90.8	90.2	89.6	89.6
		R	HD	89.8	90.2	90.0	89.6	88.5	90.4	89.9	90.3
3/4	0.5	DS	FV	66.4	85.8	91.5	90.0	46.9	79.9	90.3	89.9
		R	FV	65.0	85.2	91.4	89.9	40.0	80.8	90.6	89.9
		DS	HD	90.6	90.1	90.4	89.8	88.9	89.7	90.3	89.8
		R	HD	89.2	90.4	89.7	89.6	90.8	91.0	89.6	90.3

DS, data splitting; R, randomization; FV, face value; HD, coverage constructed using the hold-out split of the data.

Table 4. *Average lengths of confidence intervals for the selected coefficients*

				$\lvert\beta_i\rvert$							
				Knockoff				Stability			
f	ρ	Split	Method	0	0.2	0.5	1	0	0.2	0.5	1
1/2	0	DS	HD	0.391	0.390	0.391	0.391	0.389	0.391	0.391	0.392
		R	HD	0.356	0.355	0.358	0.358	0.360	0.354	0.358	0.358
1/2	0.5	DS	HD	0.508	0.507	0.510	0.482	0.506	0.506	0.509	0.483
		R	HD	0.457	0.455	0.459	0.436	0.459	0.458	0.457	0.438
3/4	0	DS	HD	0.653	0.651	0.650	0.651	0.657	0.648	0.651	0.651
		R	HD	0.503	0.502	0.506	0.506	0.507	0.500	0.507	0.507
3/4	0.5	DS	HD	0.899	0.904	0.903	0.846	0.890	0.904	0.902	0.847
		R	HD	0.644	0.645	0.650	0.617	0.648	0.648	0.646	0.620

DS, data splitting; R, randomization; HD, coverage constructed using the hold-out split of the data.

The randomized procedure performed significantly better in these cases, giving intervals that were approximately 75% shorter than those of data splitting. The maximum observed standard deviations of the values shown in the tables were, respectively, 1.9 and 0.007.

6.5. *Inference for projection parameters*

We now consider settings with the number of covariates exceeding the sample size. In these situations the full model is not identifiable, and the methods used in the previous section are not applicable. Instead, we consider the problem of constructing confidence intervals for the coefficients of a projection parameter $\beta_s(X) = \{X(s)^{\mathrm{T}}X(s)\}^{-1}X(s)^{\mathrm{T}}X\beta$.

If s had been fixed in advance and we knew the value of σ^2, inference on the coefficients of $\beta_s(X)$ would be based on the components of $\hat{\beta}_s(X) = \{X(s)^{\mathrm{T}}X(s)\}^{-1}X(s)^{\mathrm{T}}Y \sim$

Table 5. *Coverages of confidence intervals for the coefficients of the projection parameters: stability selection*

			$\lvert\beta_i\rvert$							
			$f = 1/2$				$f = 3/4$			
ρ	Split	Method	0	0.2	0.5	1	0	0.2	0.5	1
0	DS	FV	38.6	70.9	91.2	89.4	11.7	71.2	90.9	89.1
	R	FV	25.0	71.6	91.1	89.2	5.5	72.8	90.3	89.1
	DS	HD	89.7	86.1	89.4	90.0	88.2	89.8	90.0	90.0
	R	HD	87.9	89.5	89.8	89.9	89.4	90.3	90.0	89.9
0.5	DS	FV	41.7	75.0	83.9	90.6	18.3	71.3	88.5	90.5
	R	FV	32.5	75.3	87.5	90.8	12.9	70.1	89.9	90.1
	DS	HD	91.2	88.0	88.0	90.4	90.0	92.3	91.4	90.5
	R	HD	88.4	88.9	89.2	90.2	89.0	91.3	90.4	90.3

DS, data splitting; R, randomization; FV, face value; HD, coverage constructed using the hold-out split of the data.

Table 6. *Average lengths of confidence intervals for the coefficients of the projection parameters: stability selection*

				$\lvert s\rvert$							
f	ρ	Split	Method	1	2	3	4	5	6	7	8
1/2	0	DS	HD	–	0.34	0.34	0.34	0.33	0.33	0.34	–
		R	HD	–	0.35	0.34	0.33	0.33	0.33	0.33	0.30
1/2	0.5	DS	HD	0.34	0.38	0.37	0.37	0.36	–	–	–
		R	HD	0.35	0.39	0.37	0.36	0.36	0.35	0.34	–
3/4	0	DS	HD	–	0.48	0.48	0.48	0.48	0.47	0.47	0.48
		R	HD	–	0.52	0.49	0.47	0.47	0.46	0.46	0.45
3/4	0.5	DS	HD	0.48	0.55	0.54	0.52	0.52	0.53	–	–
		R	HD	0.49	0.56	0.54	0.52	0.51	0.48	0.49	–

DS, data splitting; R, randomization; HD, coverage constructed using the hold-out split of the data.

$N[\beta_s(X), \sigma^2\{X(s)^{\mathrm{T}}X(s)\}^{-1}]$. So, for a nominal coverage of $1-\alpha$, the confidence intervals would be given by $[\hat{\beta}_s(X)_i \mp q_{1-\alpha/2}\sigma[e_i^{\mathrm{T}}\{X(s)^{\mathrm{T}}X(s)\}^{-1}e_i]^{1/2}]$, where $q_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of a standard normal distribution. When σ^2 is unknown, we can plug in the high-dimensional estimate used for the (U, V) decomposition, $\hat{\sigma}^2_{\mathrm{HD}}$, say, and report $[\hat{\beta}_s(X)_i \mp q_{1-\alpha/2}\hat{\sigma}_{\mathrm{HD}}[e_i^{\mathrm{T}}\{X(s)^{\mathrm{T}}X(s)\}^{-1}e_i]^{1/2}]$.

When s is data-dependent, hold-out inference can be provided similarly, with the difference that for data splitting, the inferential target is not the full-projection parameter $\beta_s(X)$, but rather the projection parameter based on the hold-out observations. Inference is thus provided for $\beta_s(X^{r^{\mathrm{c}}}) = \{X^{r^{\mathrm{c}}}(s)^{\mathrm{T}}X^{r^{\mathrm{c}}}(s)\}^{-1}X^{r^{\mathrm{c}}}(s)^{\mathrm{T}}\mu_{r^{\mathrm{c}}}$, where $\mu_{r^{\mathrm{c}}} = E(Y^{r^{\mathrm{c}}})$, and is based on $\hat{\beta}_s^{\mathrm{DS}}(X^{r^{\mathrm{c}}}) = \{X^{r^{\mathrm{c}}}(s)^{\mathrm{T}}X^{r^{\mathrm{c}}}(s)\}^{-1}X^{r^{\mathrm{c}}}(s)^{\mathrm{T}}Y^{r^{\mathrm{c}}} \sim N[\beta_s(X^{r^{\mathrm{c}}}), \sigma^2\{X^{r^{\mathrm{c}}}(s)^{\mathrm{T}}X^{r^{\mathrm{c}}}(s)\}^{-1}]$. In the case of randomization, the target parameter is still $\beta_s(X)$, and inference is based on the normal approximation to the distribution of $\hat{\beta}_s^{\mathrm{R}}(X) = \{X(s)^{\mathrm{T}}X(s)\}^{-1}X(s)^{\mathrm{T}}V \dot{\sim} N[\beta_s(X), (1+\gamma^{-2})\sigma^2\{X(s)^{\mathrm{T}}X(s)\}^{-1}]$. In both cases, σ^2 was approximated by $\hat{\sigma}^2_{\mathrm{HD}}$.

The simulation parameters were set as in § 6.4 except for the number of covariates, which was taken to be $p = 400$. Here the knockoff was not considered, as it is not applicable when $p > n$. The results are given in Tables 5 and 6. In Table 5 the columns indicate the absolute

value of the full-model coefficients β_i, not the coefficients of the projection parameters, which in general depend on s. The coverage results were similar to those for the lower-dimensional case, with the difference that the effect of selection was more pronounced here. In one of the cases considered, the coefficients of the projection parameters associated with the null covariates in the full model were almost guaranteed to be missed by the face-value intervals. The hold-out intervals, on the other hand, remained well-calibrated across the different coefficients. Regarding interval lengths, in this case we computed the average of all the intervals produced for a given selection set s, and averaged the results over selection sets of equal size in order to establish a more equitable comparison. We see that the average lengths were very similar in this case. Recall, however, that the selection power of the latter method is substantially higher in high-dimensional settings, so in conjunction randomization dominates data splitting. The maximum standard deviation of the coverage figures were 1.4, 4.5, 1.2 and 0.4 for $|\beta_i| = 0, 0.2, 0.5$ and 1, respectively. The maximum standard deviation of the length values was 0.04. A dash in Table 6 indicates that no selection set of the corresponding size was selected in the simulation.

7. Discussion

Randomization offers an alternative to data splitting which divides the sample information more efficiently. We have compared randomization and data splitting with respect to their selection stability, as well as their selection and inferential power, showing that randomization can be substantially superior in settings with a limited amount of information. Our overall conclusion is that inference, as we have analysed here, based on the marginal distribution of V in the (U, V) decomposition of information offered by randomization provides a pragmatic and effective approach in the current context. One limitation of randomization, however, is that implementation requires an effective estimate of the model variance.

In this article we have considered only post-selection inference with conditional requirements. It would be interesting to explore how this idea fits within the framework of Berk et al. (2013), where inferential guarantees are unconditional and valid for arbitrary model-selection procedures. In this framework the authors constructed a constant K such that

$$\mathrm{pr}\big(e_j^{\mathrm{T}}\beta_S \in \big[e_j^{\mathrm{T}}\hat{\beta}_S \mp K\{(X_S^{\mathrm{T}}X_S)^{-1}\}_{jj}^{1/2}\hat{\sigma}\big]\ \forall j = 1, \ldots, |S|\big) \geqslant 1 - \alpha$$

for all variable-selection functions S, where $\hat{\sigma}$ is an estimator of σ satisfying certain conditions. In the context discussed here, where S depends on Y only through U, it is to be expected that there exists a valid constant $K(\gamma)$ with analogous guarantees which is decreasing in γ and smaller than K for all $\gamma > 0$. A similar idea has been fruitfully considered by Zrnic & Jordan (2022) with other types of randomization noise, and further work in this direction would be desirable.

Supplementary material

The Supplementary Material contains proofs of the theoretical results.

References

Bachoc, F., Leeb, H. & Pötscher, B. M. (2017). Valid confidence intervals for post-model-selection predictors. *Ann. Statist*. **47**, 1475–504.

Bachoc, F., Preinerstorfer, D. & Steinberger, L. (2020). Uniformly valid confidence intervals post-model-selection. *Ann. Statist.* **48**, 440–63.

Barber, R. F. & Candès, E. (2015). Controlling the false discovery rate via knockoffs. *Ann. Statist.* **43**, 2055–85.
Barber, R. F., Candès, E., Janson, L., Patterson, E. & Sesia, M. (2020). *knockoff: The Knockoff Filter for Controlled Variable Selection*. R package version 0.3.3.
Bayati, M., Erdogdu, M. A. & Montanari, A. (2013). Estimating LASSO risk and noise level. In *Proc. 26th Int. Conf. Neural Information Processing Systems (NIPS'13)*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani & K. Q. Weinberger, eds. Red Hook, New York: Curran Associates, pp. 944–52.
Berk, R., Brown, L., Buja, A., Zhang, K. & Zhao, L. (2013). Valid post-selection inference. *Ann. Statist.* **41**, 802–37.
Candès, E., Fan, Y., Janson, L. & Lv, J. (2018). Panning for gold: 'Model-X' knockoffs for high dimensional controlled variable selection. *J. R. Statist. Soc.* B **80**, 551–77.
Cox, D. R. (1975). A note on data-splitting for the evaluation of significance levels. *Biometrika* **62**, 441–4.
DiCiccio, C. J., DiCiccio, T. J. & Romano, J. P. (2020). Exact tests via multiple data splitting. *Statist. Prob. Lett.* **166**, article no. 108865.
Fan, J., Guo, S. & Hao, N. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *J. R. Statist. Soc.* B **74**, 37–65.
Fithian, W., Sun, D. L. & Taylor, J. E. (2017). Optimal inference after model selection. *arXiv:* 1410.2597v4.
Hofner, B. & Hothorn, T. (2017). *stabs: Stability Selection with Error Control*. R package version 0.6-3.
Hong, L., Kuffner, T. A. & Martin, R. (2018). On overfitting and post-selection uncertainty assessments. *Biometrika* **105**, 221–4.
Ignatiadis, N., Klaus, B., Zaugg, J. & Huber, W. (2016). Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature Meth.* **13**, 577–80.
Kivaranovic, D. & Leeb, H. (2021a). On the length of post-model-selection confidence intervals conditional on polyhedral constraints. *J. Am. Statist. Assoc.* **116**, 845–57.
Kivaranovic, D. & Leeb, H. (2021b). A (tight) upper bound for the length of confidence intervals with conditional coverage. *arXiv:* 2007.12448v2.
Lee, J. D., Sun, D. L., Sun, Y. & Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *Ann. Statist.* **44**, 907–27.
Lee, J. D. & Taylor, J. E. (2014). Exact post model selection inference for marginal screening. In *Proc. 27th Int. Conf. Neural Information Processing Systems (NIPS'14)*, vol. 1. Cambridge, Massachusetts: MIT Press, pp. 136–44.
Lockhart, R., Taylor, J. E., Tibshirani, R. & Tibshirani, R. (2014). A significance test for the lasso. *Ann. Statist.* **42**, 413–68.
Loftus, J. R. & Taylor, J. E. (2014). A significance test for forward stepwise model selection. *arXiv:* 1405.3920v1.
Markovic, J., Taylor, J. & Taylor, J. (2019). Inference after black box selection. *arXiv:* 1901.09973v1.
Meinshausen, N. & Buhlmann, P. (2010). Stability selection. *J. R. Statist. Soc.* B **72**, 417–73.
Panigrahi, S., Taylor, J. & Weinstein, A. (2020). Integrative methods for post-selection inference under convex constraints. *Ann. Statist*. **49**, 2803–24.
R Development Core Team (2023). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, http://www.R-project.org.
Reid, S., Tibshirani, R. & Friedman, J. (2016). A study of error variance estimation in lasso regression. *Statist. Sinica* **26**, 35–67.
Reitermanová, Z. (2010). Data splitting. In *WDS'10 Proceedings of Contributed Papers*, vol. I. Prague: Matfyzpress, pp. 31–6.
Rinaldo, A., Wasserman, L. & G'Sell, M. (2019). Bootstrapping and sample splitting for high-dimensional, assumption-lean inference. *Ann. Statist.* **47**, 3438–69.
Rubin, D., Dudoit, S. & van der Laan, M. (2006). A method to increase the power of multiple testing procedures though sample splitting. *Statist. Appl. Genet. Molec. Biol.* **5**, article no. 19.
Shah, R. & Samworth, R. (2013). Variable selection with error control: Another look at stability selection. *J. R. Statist. Soc.* B **75**, 55–80.
Snee, R. D. (1977). Validation of regression models: Methods and examples. *Technometrics* **19**, 415–28.
Tian, X. & Taylor, J. E. (2018). Selective inference with a randomized response. *Ann. Statist.* **46**, 679–710.
Tibshirani, R., Rinaldo, A., Tibshirani, R. & Wasserman, L. (2018). Uniform asymptotic inference and the bootstrap after model selection. *Ann. Statist.* **46**, 1255–87.
Tibshirani, R., Tibshirani, R., Taylor, J., Loftus, J., Reid, S. & Markovic, J. (2019). *Selective Inference: Tools for Post-selection Inference*. R package version 1.2.5.
Wasserman, L. & Roeder, K. (2009). High-dimensional variable selection. *Ann. Statist.* **37**, 2178–201.
Zrnic, T. & Jordan, M. I. (2022). Post-selection inference via algorithmic stability. *arXiv:* 2011.09462v2.

[*Received on* 3 *February* 2021. *Editorial decision on* 29 *November* 2022]