Abstract

Suppose we observe a random vector $X$ from some distribution $P$ in a known family with unknown parameters. We ask the following question: when is it possible to split $X$ into two parts $f(X)$ and $g(X)$ such that neither part is sufficient to reconstruct $X$ by itself, but both together can recover $X$ fully, and the joint distribution of $(f(X), g(X))$ is tractable? As one example, if $X = (X_1, \ldots, X_n)$ and $P$ is a product distribution, then for any $m < n$, we can split the sample to define $f(X) = (X_1, \ldots, X_m)$ and $g(X) = (X_{m+1}, \ldots, X_n)$. Rasines and Young (2021) offers an alternative route of accomplishing this task through randomization of $X$ with additive Gaussian noise which enables post-selection inference in finite samples for Gaussian distributed data and asymptotically for non-Gaussian additive models. In this paper, we offer a more general methodology for achieving such a split in finite samples by borrowing ideas from Bayesian inference to yield a (frequentist) solution that can be viewed as a continuous analog of data splitting. We call our method data blurring, as an alternative to data splitting, data carving and p-value masking. We exemplify the method on a few prototypical applications, such as post-selection inference for trend filtering and other regression problems.

Contents

1 Introduction to data blurring
   1.1 Data blurring for post-selection inference ............................................. 4
   1.2 Related work on data splitting and carving ............................................. 6
   1.3 Outline of the paper .............................................................................. 6

2 Techniques to accomplish data blurring
   2.1 Decomposition using the “conjugate-prior machine” ................................. 7
   2.2 List of decompositions ......................................................................... 8
   2.3 Relationship between data splitting and data blurring ............................ 11

3 Application: selective confidence intervals after interactive multiple testing 12
   3.1 Review of some multiple testing methods and data blurring proposal ........ 13
   3.2 Empirical results ................................................................................. 14

4 Application: selective CIs in fixed-design linear regression 16
   4.1 Comparison with data splitting ............................................................. 18
   4.2 Empirical results ................................................................................. 19

5 Application to trend filtering and other nonparametric regression problems 22
   5.1 A recap of trend filtering ................................................................. 23
   5.2 Empirical results ................................................................................. 24
   5.3 Demonstration on spectroscopy datasets .............................................. 29

6 Concluding remarks .................................................................................. 30

7 Acknowledgements .................................................................................. 30
1 Introduction to data blurring

One of the most common practices in applied statistics is data splitting. Given a dataset $X = (X_1, \ldots, X_n)$ containing $n$ independent samples, an analyst wishes to divide the data into two smaller independent datasets in order to complete an analysis. The typical method for accomplishing this would be to choose some $1 \leq m < n$ and then form two new datasets: $f(X) = (X_1, \ldots, X_m)$ and $g(X) = (X_{m+1}, \ldots, X_n)$. However, there are alternative approaches for accomplishing this goal that may be preferable depending on one’s objectives.

As a simplified example, consider the setting where we only observe a single data point $X \sim N(0, 1)$, and we would like to “split” $X$ into two parts such that each part contains some information about $X$, $X$ can be reconstructed from both parts taken together, but neither part is sufficient by itself to reconstruct $X$, and yet the joint distribution of these two parts is known. The constraints mentioned in the previous sentence avoid trivial solutions like outputting $f(X) = X$ and $g(X) = 0$, or $f(X) = X/3$ and $g(X) = 2X/3$, and so on. In other words, we really must “partition the information” in $X$ into two complementary pieces.

Fortunately, this example has a simple solution involving the use of external randomization. Generate an independent $Z \sim N(0, 1)$, and set $f(X) = X + Z$ and $g(X) = X - Z$. Of course, with both parts, one can reconstruct $X$ by addition (and $Z$ by subtraction, but we care less about $Z$, which has no utility of its own). Importantly, just knowing one out of $f(X)$ or $g(X)$ does not allow one to reconstruct $X$, but both parts have nontrivial information about $X$. Formally, one may say that their mutual information with $X$ is nonzero or simply that neither part is independent of $X$. Most importantly, we know that $f(X)$ and $g(X)$ are actually independent, and their marginal distributions are also Gaussian, so their joint distribution is tractable and known. Note that this idea of adding and subtracting Gaussian noise has also been employed for orthogonal goals unrelated to selective inference, such as FDR control in regression problems using the concept of Gaussian mirrors as discussed in Xing et al. (2021) and to estimate risk in the normal means problem as discussed in Oliveira et al. (2021). Here we ask and answer different questions, with different goals in mind, and provide more methodological alternatives beyond the Gaussian case.

To that end, we can construct a family of pairs of functions $(f_\tau(X), g_\tau(X))_{\tau \in \mathcal{T}}$, for some totally ordered set $\mathcal{T}$ (typically a subset of the real line), such that we can smoothly trade off the amount of information that each part contains about $X$. When $\tau$ approaches $\tau^+ := \sup \{ \tau : \tau \in \mathcal{T} \}$, $f(X)$ will approach independence from $X$, while $g(X)$ will essentially equal $X$, but when $\tau$ approaches $\tau^- := \inf \{ \tau : \tau \in \mathcal{T} \}$, the opposite will happen. To see how to do this, simply choose $Z$ as before, and define $f(X) = X - \tau Z$ and $g(X) = X + \frac{\tau}{2}$ and let $\mathcal{T} := (0, \infty)$. We call this type of procedure “data blurring”, in the sense that each of $f(X)$ and $g(X)$ provide a blurred view of $X$; we may call $X$ as the original data and either $f(X)$ or $g(X)$ blurred data.

Data blurring is similar in spirit to data splitting. However, data blurring manages to achieve the same effect from just a single sample $X$ and not an $n$-dimensional vector. Nevertheless, the connection to data splitting in this case is more than a mere analogy, and it is possible to exactly quantify the relationship between $\tau$ and $m$, such that one can rigorously view data blurring as a continuous analog of data splitting in the Gaussian case; we do this in the next section.

Now, can the above ideas be generalized to other distributions? In other words, can we employ external randomization in order to “split” a single data point into two nontrivial parts when the distribution $P$ is not Gaussian? This is the topic of study for the rest of this paper. We provide a positive answer when $P$ is conjugate (in the standard Bayesian sense) to some other distribution $Q$, where the latter will be used (along with $X$) to determine the external randomization. In most cases, $f(X)$ and $g(X)$ will not simply be the sum/difference of $X$ with some $Z$; such a form was achieved only in the special case of Gaussians. Similarly, $f(X)$ and $g(X)$ will typically not be independent. Nevertheless, they will satisfy the conditions set out in the second paragraph of the paper and can be treated for inferential purposes as single-sample variants of data splitting, justifying the title of the paper.
1.1 Data blurring for post-selection inference

We primarily focus on demonstrating the applicability of these ideas in the context of (potentially high-dimensional) model selection and post-selection inference, which usually involves multiple samples. In a traditional data splitting setup, the analyst picks some fraction \( a \in \left\{ \frac{1}{n}, \ldots, \frac{n-1}{n}, 1 \right\} \) of the data to use for model selection and the remaining \( 1 - a \) fraction is used for inference as illustrated in Figure 1. Data blurring is similar in spirit to this idea but instead uses randomization so that part of the information contained in every data point is used for both selection and inference. The procedure broadly works in three stages.

1. Split \( X \) into \( f(X) \) and \( g(X) \) such that \( g(X) | f(X) \) is tractable to compute. The parameter \( \tau \) described above controls the proportion of the information to be used for model selection.

2. Use \( f(X) \) to select a model and/or set of hypotheses to test using any available method.

3. Use \( g(X) | f(X) \) to test the hypotheses and/or perform inference using the model chosen in step 2.

See Figure 2 for a graphical representation of the above steps. This approach can be thought of a generalization of the methodology discussed by Rasines and Young (2021). In our framework, their approach amounts to letting \( f(X) = X + Z \) and \( g(X) = X \) with \( Z \sim N(0, \sigma^2) \). In the case where \( X \sim N(\mu, \sigma^2) \) with known \( \sigma^2 \), the authors show that \( g(X) | f(X) \) has a tractable finite sample distribution. In cases where \( X \) is non-Gaussian, however, \( g(X) | f(X) \) can only be described asymptotically. In the next section, we will explore alternative ways of using this framework to construct \( g(X) \) that result in tractable finite sample distributions in the non-Gaussian case.

In some ways, these methodologies can be seen as a compromise between data splitting and the approach of data carving as introduced in Fithian et al. (2014). Data carving, as illustrated in Figure 3, improves on data splitting in cases where the conditional distribution of the data given a selection event is known by including the leftover portion of Fisher information that was not used to inform the model choice in the inference procedure. A key limitation of this approach, however, is that it confines the analyst to model selection techniques with tractable post-selective distributions, such as the lasso as described by Lee et al. (2016) or more general sequential regression procedures such as those discussed in Tibshirani et al. (2016). In many settings, ad hoc exploratory data analyses such as data plots or consideration of multiple competing selection techniques are ubiquitous and make the use of a data carving approach intractable.

Although data blurring conditions on \( g(X) \) rather than the selection event itself, it retains some similarities to data carving insofar as it uses a portion of every data point to inform both selection and inference. This has advantages relative to data splitting in at least two distinct ways. First, certain settings that involve sparse or rarely occurring features may result in a handful of data points having a disproportionate amount of influence — data blurring allows for the analyst to “hedge their bets” by including these points in both the selection and inference steps. Second, in settings where the selected model is defined relative to a set of fixed covariates, the theoretical justification for data splitting becomes less clear conceptually. (In a fixed-X setup, how can a model that has been selected based on its ability to estimate the conditional distribution \( Y | X^{\text{first half}} \) also be understood to model the distribution that conditions on the other half of the split dataset \( Y | X^{\text{second half}} \)?) This is perhaps most stark in the context of time series data, where splitting a sample may require the analyst to have a selection and inference dataset that span entirely different time periods.

On the other hand, similar to data splitting, data blurring affords the analyst complete flexibility in how they choose their model based on the information revealed in the selection stage. In particular, our procedure can accommodate a model selection process that relies on qualitative or heuristic methods such as visual examination of graphs and residual plots or the consultation of domain experts to inquire about the plausibility of certain uncovered relationships.

Although we anticipate that these ideas may have other downstream applications beyond selective inference such as data privacy, creation of fake datasets, and comparing machine learning algorithms, we do not fully explore these here.
Use some portion $a$ of the data for selection
Use the remaining portion $1 - a$ of the data for inference

Use any procedure to choose a model on the selection dataset and then use the inference dataset to test it

**Figure 1.** Illustration of typical **data splitting procedures** for post-selection inference. Splitting the data has the advantage of allowing the user to choose any selection strategy for model selection, but at the cost of decreased power during the inference stage.

Blur data using randomization scheme
Reveal full dataset at inference stage

Use any procedure to choose a model from $f(X)$

**Figure 2.** Illustration of our proposed **data blurring procedure**. Similar to data splitting, it allows for any selection procedure for choosing the model. However, it achieves this through randomization rather than a direct splitting of the data.

Choose some fixed selection procedure $S$

Conditioning on $S(X)$ needs to be tractable (e.g. LASSO, forward selection)

**Figure 3.** Illustration of **data carving procedure** as discussed in Fithian et al. (2014). Data carving has the advantage of using all unused information for inference, but requires the selection procedure to be fixed at the onset of investigation. Moreover, computing the conditional distribution needs to be tractable, either in closed form (e.g. LASSO as described in Lee et al. (2016)) or through numerical simulation. Thus data carving and blurring have complementary benefits and tradeoffs.
1.2 Related work on data splitting and carving

Our work is influenced heavily by the existing literature on procedures for selective inference after model selection. Although data splitting is perhaps the oldest and most commonly used method for ensuring valid coverage guarantees after model selection, rigorous examination of data splitting has only recently emerged in the literature. Rinaldo et al. (2019) is one such example — in this paper, the authors examine data splitting in an assumption-lean context with weak assumptions on the distribution of the data and no requirements that the model chosen during the selection stage is “correct”.

The idea of adding randomization for selective inference is extensively discussed in Tian and Taylor (2018) for Gaussian distributed data. They introduce a noise variable that is independent of the original data, and involve it into the selection procedure arbitrarily (for example, introducing perturbations to the gradient when performing logistic lasso to select features). Instead, we explore the perspective of constructing a random variable \( f(X) \) that depends on \( X \) but is not exactly equal to \( X \), and perform arbitrary selection procedures on \( f(X) \).

More recently, randomization was used in Li and Fithian (2021) to recast the knockoff procedure of Barber and Candès (2015) as a selective inference procedure for the linear Gaussian model that adds noise to the OLS estimates \( \hat{\beta} \) to create a “whitened” version of \( \hat{\beta} \) to use for hypothesis selection. The work of Sarkar and Tang (2021) explores similar ways of using knockoffs to split \( \hat{\beta} \) into independent pieces for hypothesis selection but uses a deterministic splitting procedure. Although conceptually quite similar to our approach, these methods focus on splitting the coefficient estimates for a fixed model into two independent pieces and then adaptively choosing the hypotheses to test during the inference stage. As such, it is most naturally applied in low dimensional settings where model selection is less necessary. Similar randomization schemes for Gaussian distributed data are also explored in Ignatiadis et al. (2021) but within the context of empirical Bayes estimation and not selective inference.

Of considerable interest to us is making these results applicable in assumption-lean settings where few requirements are placed on the true distribution of the data even when parametric models are ultimately used for inference. Buja et al. (2019) outlines a procedure for generic likelihood-based inference procedures but outside of a selective inference setting. Buja et al. (2019) focuses exclusively on assumption-lean linear regression but does not explore in detail its properties in the context of data splitting. Throughout our work, we try to place minimal assumptions on the conditional mean function \( \mu(x) = E[Y|X=x] \) of our data following a particular structural form but still require the distribution of the data to be correctly specified in order to ensure that our procedure produces an \( f(X) \) and \( g(X) \) that are truly independent.

1.3 Outline of the paper

Our proposed data blurring procedure is introduced in Section 2. We first illustrate how the procedure can be used in the context of Gaussian distributed data but then generalize this for the broader class of distributions where the data has a known conjugate prior. Examples are given across a variety of distributions commonly used for regression and data analysis such as Gaussian, Poisson, and Binomial.

The remainder of the paper explores the use of data blurring for Gaussian distributed data in three contexts: selective confidence intervals after interactive multiple testing in Section 3, linear regression in Section 4 and trend filtering in Section 5. A key limitation within all of these applications is that theoretical guarantees can only be given in the case of known variance and correct specification on the distribution of errors due to the need to ensure independence between \( f(X) \) and \( g(X) \). However, these guarantees are still assumption-lean in the sense of assuming an unknown form for the conditional mean function \( \mu(x) = E[Y|X=x] \). Situations where the variance is unknown and is estimated before data blurring are explored empirically, but we leave theoretical guarantees for these methods as an open avenue for future investigation. We provide some concluding remarks in Section 6.

In the Appendix A, we also explore additional applications of data blurring for constructing post-selective confidence intervals for generalized linear models (Poisson and logistic regression). We note significant gains in performance when compared with data splitting, but theoretical guarantees are again left open for future investigations.
2 Techniques to accomplish data blurring

With potential statistical applications in mind, we explore decompositions of \( X \) to \( f(X) \) and \( g(X) \) such that both parts can be used to infer the parameter \( \theta \) of our interest, and with either of the following two properties:

(P1). \( f(X) \) and \( g(X) \) are independent with known distributions (up to the same unknown parameter \( \theta \)); or

(P2). \( f(X) \) has a known marginal distribution and \( g(X) \) has a known conditional distribution given \( f(X) \) (up to knowledge of \( \theta \)).

The former property implies the latter but the former is more tractable for practical applications.

2.1 Decomposition using the “conjugate-prior machine”

To construct \( f(X) \) and \( g(X) \) satisfying the second property: \( f(X) \) has a known marginal distribution, and \( g(X) \) has a known conditional distribution given \( f(X) \), we naturally can explore the class of conjugate priors. Suppose the distribution of \( X \) coincides with a conjugate prior distribution of the parameter in some likelihood. We can then construct a new random variable \( Z \) following that likelihood (with the parameter being \( \theta \)).

\[ f(X) = Z \text{ and } g(X) = X, \]

Then, because of the property of a conjugate prior, we know that the conditional distribution of \( g(X) \mid f(X) \) is Gamma(1+\( \sum_{i=1}^{B} f_i(X) \), \( \theta + n \)) (notice that Exponential distribution is a special case of the Gamma distribution), which can be used for inference on \( \theta \). On the other hand, we can also verify that \( f(X) \) has a known marginal distribution as desired, as \( \text{Geo}(\theta) \). In this construction, a larger \( B \) indicates more informative \( f(X) \). More examples of the decomposition inspired by this procedure, which we term the “conjugate-prior machine” are detailed in the next section.

Note that in the specific case of exponential family distributions, we can construct \( f(X) \) and \( g(X) \) in the following way.

**Theorem 1.** Suppose that for some \( A(\cdot, \cdot), \theta_1, \theta_2, H(\cdot, \cdot) \), the density of \( X \) is given by

\[
p(x \mid \theta_1, \theta_2) = H(\theta_1, \theta_2) \exp\{\theta_1^T x - \theta_2^T A(x)\}. \tag{1}
\]

Suppose also that there exists \( h(\cdot), T(\cdot), \theta_3 \) such that

\[
p(z \mid x, \theta_3) = h(z) \exp\{x^T T(z) - \theta_3^T A(x)\} \tag{2}
\]

is a well-defined distribution. First, draw \( Z \sim p(z \mid X) \), and let \( f(X) := Z \) and \( g(X) := X \). Then, \((f(X), g(X)) \) satisfy the data blurring property (P2). Specifically, note that \( f(X) \) has a known marginal distribution \( p(z \mid \theta_1, \theta_2, \theta_3) = h(z) \frac{H(\theta_1, \theta_2)}{H(\theta_1 + T(z), \theta_2 + \theta_3)}, \) while \( g(X) \) has a known conditional distribution given \( f(X) \), which is \( p(x \mid z, \theta_1, \theta_2, \theta_3) = p(x \mid \theta_1 + T(z), \theta_2 + \theta_3). \)

**Proof.** Note that because the density \( p(z \mid x) \) must integrate to 1, we can view the function \( H(\theta_1, \theta_2) \) as a normalization factor since

\[
H(\theta_1, \theta_2) = \frac{1}{\int_{-\infty}^{\infty} \exp\{\theta_1^T x - \theta_2^T A(x)\}dx}.
\]

Therefore, to compute the marginal density, we have

\[
p(z; \theta_1, \theta_2, \theta_3) = \int_{-\infty}^{\infty} h(z) H(\theta_1, \theta_2) \exp\{(T(z) + \theta_1)^T x - (\theta_2 + \theta_3)^T A(x)\}dx \\
= h(z) \frac{H(\theta_1, \theta_2)}{H(\theta_1 + T(z), \theta_2 + \theta_3)}. \]
Similarly, the computation of the conditional density is straightforward

\[
p(x|z, \theta_1, \theta_2, \theta_3) = \frac{h(z)H(\theta_1, \theta_2) \exp\{(T(z) + \theta_1)^T x - (\theta_2 + \theta_3)^T A(x)\}}{h(z)H(\theta_1, \theta_2)} = H(\theta_1 + T(z), \theta_2 + \theta_3) \exp\{x^T (\theta_1 + T(z)) - (\theta_2 + \theta_3)^T A(x)\}
\]

This completes the proof. \(\square\)

Note that typically \(\theta_3\) cannot be zero for \(p(z | x)\) to be a well-defined distribution for arbitrary \(x\). Also, the definition of \(p(x | \theta_1, \theta_2, \theta_3)\) may not be identifiable: \(x\) can be included in \(A(x)\), which gives us multiple options to construct \(p(z | x, \theta_1, \theta_2, \theta_3)\) according to the definition of \(A(x)\).

**Remark 1** (Trading off information in \(f\) and \(g\)). As an extension to the above result, we can draw \(B\) samples independently from (2) denoted as \(Z_i\) for \(i \in [B]\), in which case \(f(X) = Z \equiv \{Z_1, \ldots, Z_B\}\) has marginal distribution

\[
p(z|\theta_1, \theta_2, \theta_3) = \left[ \prod_{i=1}^{B} h(z_i) \right] \frac{H(\theta_1, \theta_2)}{H(\theta_1 + \sum_{i=1}^{B} T(z_i), \theta_2 + B\theta_3)},
\]

and \(g(X) = X\) has conditional distribution as

\[
p(x | z, \theta_1, \theta_2, \theta_3) = p(x | \theta_1 + \sum_{i=1}^{B} T(z_i), \theta_2 + B\theta_3).
\]

Intuitively, larger \(B\) indicates more information in \(f(X)\), and thus less left over in \(g(X)\).

### 2.2 List of decompositions

As an illustration of the above principles, we will describe decompositions for several commonly encountered distributions. When discussing each method, we will label them as follows so the reader will be able to relate them easily to the strategies discussed in the preceding sections:

- **(P1)** will indicate that the decomposition strategy is an instance of the first principle — \(f(X)\) and \(g(X)\) are independent with known distributions.

- **(P2)** will indicate that the decomposition strategy is an instance of the second principle — \(f(X)\) has a known marginal distribution, while \(g(X)\) has a known conditional distribution given \(f(X)\).

- **(P2 CP)** will indicate that the decomposition strategy is an instance of the second principle — but specifically using the “conjugate-prior machine” as described above.

Although we recognize readers may find a close reading of this section tedious, we feel it is important to include this information for two reasons. First, we wish to provide several concrete examples to demonstrate to the reader how the strategies for constructing \(f(X)\) and \(g(X)\) discussed in the prior sections can be practically instantiated. Second, we want to highlight the wide applicability of this approach to disparate kinds of problems within statistics — only providing a small number of examples would give the false impression that this methodology is only tractable within a few idealized settings.

- **Gaussian**

  1. **(P1)** Suppose \(X \sim N(\mu, \Sigma)\) is \(d\)-dimensional \((d \geq 1)\). Draw \(Z \sim N(0, \Sigma)\). Then \(f(X) = X + \tau Z\), where \(\tau \in (0, \infty)\) is a tuning parameter, has distribution \(N(\mu, (1 + \tau^2)\Sigma)\); and \(g(X) = X - \frac{1}{\tau} Z\) has distribution \(N(\mu, (1 + \tau^{-2})\Sigma)\); and \(f(X) \perp \perp g(X)\). Larger \(\tau\) indicates less informative \(f(X)\) (and more informative \(g(X) | f(X)\)).
2. **(P2 CP)** Alternatively, draw \( Z \sim N(X, \tau \Sigma) \), where \( \tau \in (0, \infty) \) is a tuning parameter. Then \( f(X) = Z \) has marginal distribution \( N(\mu, (1 + \tau)\Sigma) \); and \( g(X) = X \) has conditional distribution \( N\left(\frac{\mu}{\tau + 1} + f(X)/\tau, \frac{\Sigma}{\tau + 1}\right) \). Larger \( \tau \) indicates less informative \( f(X) \) (and more informative \( g(X) \mid f(X) \)).

3. **(P2)** More generally, we can add Gaussian noise with an arbitrary covariance matrix. Draw \( Z \sim N(0, \Sigma_0) \) and let \( f(X) = X - Z \) with \( g(X) = X + Z \) as before. For notational convenience, let \( \Sigma_1 = \Sigma + \Sigma_0 \) and \( \Sigma_2 = \Sigma - \Sigma_0 \) Then \( f(X) \sim N(\mu, \Sigma_1) \) and \( g(X) \mid f(X) \sim N\left(\mu + \Sigma_2 \Sigma_1^{-1} (f(X) - \mu), \Sigma_1 - \Sigma_2 \Sigma_1^{-1} \Sigma_2\right) \).

• **Gamma**

1. **Exponential (P2 CP)** Suppose \( X \sim \text{Exp}(\theta) \). Draw \( Z = (Z_1, \ldots, Z_B) \) where each element is i.i.d. \( Z_i \sim \text{Poi}(\theta) \) and \( B \in \{1, 2, \ldots\} \) is a tuning parameter. Then \( f(X) = Z \), where each element is independently distributed as \( \text{Geo}\left(\frac{\theta}{\theta + 1}\right) \). \( g(X) = X \) has conditional distribution \( \text{Gamma}(1 + \sum_{i=1}^B f_i(X), \theta + B) \). Larger \( B \) indicates more informative \( f(X) \) (and less informative \( g(X) \mid f(X) \)).

   **(P2 CP)** Alternatively, we can draw \( Z \sim \text{Poi}(\tau X) \), where \( \tau \in (0, \infty) \) is a tuning parameter. Then \( f(X) = Z \), marginally distributed as \( \text{Geo}\left(\frac{\tau}{\tau + 1}\right) \). \( g(X) = X \) has conditional distribution \( \text{Gamma}(1 + f(X), \theta + \tau) \). Here, \( f(X) \) is most informative when \( \tau \) is comparable with \( \theta \), and less informative when \( \tau \) approaches zero or infinity.

2. **(P2 CP)** Generally, suppose \( X \sim \text{Gamma}(\alpha, \beta) \). Draw \( Z = (Z_1, \ldots, Z_B) \) where each element is i.i.d. \( Z_i \sim \text{Poi}(\theta) \) and \( B \in \{1, 2, \ldots\} \) is a tuning parameter. Then \( f(X) = Z \), where each element is independently distributed as a negative binomial \( \text{NB}(\alpha, \frac{\theta}{\theta + 1}) \). \( g(X) = X \) has conditional distribution \( \text{Gamma}(\alpha + \sum_{i=1}^B f_i(X), \beta + B) \). Larger \( B \) indicates more informative \( f(X) \) (and less informative \( g(X) \mid f(X) \)).

3. **(P2 CP)** Alternatively, we can draw \( Z \sim \text{Poi}(\tau X) \), where \( \tau \in (0, \infty) \) is a tuning parameter. Then \( f(X) = Z \), marginally distributed as \( \text{NB}(\alpha, \frac{\tau}{\tau + 1}) \). \( g(X) = X \) has conditional distribution \( \text{Gamma}(\alpha + f(X), \beta + \tau) \). \( f(X) \) is most informative when \( \tau \) is comparable with \( \theta \), and less informative when \( \tau \) approaches zero or infinity.

4. Note: decomposition of the Gamma distribution implies decomposition of the Chi-square distribution \( \chi_k^2 \), as it is equivalent to Gamma\((k/2, 1/2)\).

• **Beta**

1. **(P2 CP)** Suppose \( X \sim \text{Beta}(\theta, 1) \). Draw \( Z \sim \text{Bin}(B, X) \), where \( B \in \{1, 2, \ldots\} \) is a tuning parameter. Then \( f(X) = Z \) has marginal distribution as a discrete uniform in \( \{0, 1, \ldots, B\} \) when \( \theta = 1 \), and stochastically larger (smaller) than a discrete uniform when \( \theta \) is larger (smaller) than one (the PMF of \( Z \) is \( p_\theta(z) = \frac{\theta^z (\theta + 1)^{B-1}}{\Gamma(\theta) \Gamma(B+1)} \)). \( g(X) = X \) has conditional distribution \( \text{Beta}(\theta + f(X), B - f(X) + 1) \). Larger \( B \) indicates more informative \( f(X) \) (and less informative \( g(X) \mid f(X) \)).

2. **(P2 CP)** Similarly, if \( X \sim \text{Beta}(1, \theta) \), we can draw \( Z \sim \text{Bin}(B, 1 - X) \), and the resulting \( f(X) \), \( g(X) \) have the same (conditional) distributions as above.

3. **Extension to multivariate case: Dirichlet (P2 CP)** Suppose \( X \sim \text{Dir}(\theta, 1, \ldots, 1) \), where \( (\theta, 1, \ldots, 1) \) is a \( d \)-dimensional vector with \( k \geq 2 \). Draw \( Z \sim \text{Multinom}(B, X) \), where \( B \in \{1, 2, \ldots\} \) is a tuning parameter. Then \( f(X) = Z \) has marginal distribution as a discrete uniform in its support \( \{z_i \in \{0, \ldots, B\} \) for \( i \in \{1, \ldots, d\} \) : \( \sum_{i=1}^d z_i = B \) \) when \( \theta = 1 \), and for other \( \theta \), the PMF is \( p_\theta(z) = \frac{\Gamma(\sum_{i=1}^d z_i + \theta) \Gamma(d-1+\theta)}{\Gamma(\theta) \Gamma(B+\theta) \Gamma(B+1)} \). \( g(X) = X \) has conditional distribution \( \text{Dir}(\theta + f_1(X), 1 + f_2(X), \ldots, 1 + f_k(X)) \). Larger \( B \) indicates more informative \( f(X) \) (and less informative \( g(X) \mid f(X) \)).
In the general case, where \( X \sim \text{Dir}(\theta_1, \theta_2, \ldots, \theta_d) \), we can use the same construction. Then \( f(X) = Z \) has marginal distribution

\[
p_\theta(z) = \frac{\Gamma \left( \sum_{i=1}^d \theta_i \right)}{\Gamma \left( n + \sum_{i=1}^d \theta_i \right)} \frac{\prod_{i=1}^d \Gamma(\theta_i) \prod_{i=1}^d z_i!}{\prod_{i=1}^d (\theta_i) z_i!}.
\]

\( g(X) = X \) has conditional distribution \( \text{Dir}(\theta_1 + f_1(X), \ldots, \theta_k + f_d(X)) \). Larger \( B \) indicates more informative \( f(X) \) (and less informative \( g(X) \mid f(X) \)).

- **Binomial (P2).** Suppose \( X \sim \text{Bin}(n, \theta) \). Draw \( Z \sim \text{Bin}(n, p) \) where \( p \in (0, 1) \) is a tuning parameter. Then \( f(X) = Z \) has marginal distribution \( \text{Bin}(n, p\theta) \); and \( g(X) = X - Z \) has conditional distribution as \( \text{Bin}(n - Z, \frac{1 - \theta}{1 - p\theta}) \). Larger \( p \) indicates more informative \( f(X) \). Note that the decomposition of Binomial is not trivially applicable to Bernoulli distribution since \( X = 1 \) with probability one if \( Z = 1 \).

- **Bernoulli (P2).** Suppose \( X \sim \text{Ber}(\theta) \). Draw \( Z \sim \text{Ber}(p) \) where \( p \in (0, 1) \) is a tuning parameter. Then \( f(X) = X(1 - Z) + (1 - X)Z \) has marginal distribution \( \text{Ber}(\theta + p - 2p\theta) \); and \( g(X) = X \) has conditional distribution as \( \text{Ber}(\frac{\theta}{\theta + (1 - \theta)p/(1 - p)} \theta) \) given \( f(X) \). Note that modeling \( \log \left( \frac{\theta + p - 2p\theta}{1 - \theta + 2p\theta} \right) \) as \( u \beta \) (for covariates \( u \)) does not imply \( \log \left( \frac{1}{\theta} \right) \) is \( u \beta' \).

- **Categorical (P2).** Suppose \( X \sim \text{Cat}(\theta_1, \ldots, \theta_d) \). Draw \( Z \sim \text{Ber}(p) \) where \( p \in (0, 1) \) is a tuning parameter. Also draw \( D \sim \text{Cat}(\frac{1}{2}, \ldots, \frac{1}{2}) \). Let \( f(X) = X(1 - Z) + DZ \) and \( g(X) = X \). Then \( f(X) \sim \text{Cat}(\phi_1, \ldots, \phi_d) \) with \( \phi_i = p\theta_i + (1 - p)\frac{1}{2} \). Furthermore \( g(X)|f(X) \) has distribution

\[
p_\theta(g(X) = s|f(X) = t) = \begin{cases} \frac{(1-p+\frac{x}{2})\theta_i}{(1-p+\frac{x}{2})\theta_i + \frac{x}{2}t} & \text{if } s = t \\ \frac{(1-p+\frac{x}{2})\theta_i}{(1-p+\frac{x}{2})\theta_i + \frac{x}{2}(t+1)} & \text{if } s \neq t \end{cases}
\]

Larger \( p \) indicates more informative \( f(X) \). Note that there is not anything special about choosing \( \frac{1}{2} \) above, the method generalizes to any known vector of probabilities for \( D \). Moreover, if we let \( d = 2 \) and substitute \( 1 - X \) for \( D \) in the above construction, we can recover the Bernoulli decomposition given above.

- **Poisson**

1. **(P1)** Suppose \( X \sim \text{Poi}(\mu) \). Draw \( Z \sim \text{Bin}(X, p) \) where \( p \in (0, 1) \) is a tuning parameter. Then \( f(X) = Z \) has marginal distribution \( \text{Poi}(p\mu) \); and \( g(X) = X - Z \) is independent of \( f(X) \) and distributes as \( \text{Poi}(1 - p)\mu \). Larger \( p \) indicates more informative \( f(X) \).

2. **(P2)** Alternatively, draw \( Z \sim \text{Poi}(p) \). We can then exploit the thinning property of the Poisson distribution (see Last and Penrose (2017)) to construct \( f(X) = X + Z \sim \text{Poi}(\mu + p) \).

Letting \( X = g(X) \) gives \( g(X)|f(X) \sim \text{Bin}(f(X), \frac{\mu}{\mu + p}) \). Larger \( p \) corresponds to a less informative \( f(X) \) (and more informative \( g(X)|f(X) \)).

- **Negative Binomial (P2).** Suppose \( X \sim \text{NB}(r, \theta) \). Draw \( Z \sim \text{Bin}(X, p) \) where \( p \in (0, 1) \) is a tuning parameter. Then \( f(X) = Z \) has marginal distribution \( \text{NB}(r, \frac{p\theta}{1 - (1 - p)\theta}) \); and \( g(X) = X - Z \) has conditional distribution as \( \text{NB}(r + Z, \theta(1 - p)) \). Larger \( p \) indicates more informative \( f(X) \).

A special case of Negative Binomial is geometric distribution where \( r = 1 \).

**Remark 2** (Relationship to infinite divisibility). **Note that we are able to decompose Poisson, Binomial, and Negative Binomial by drawing a Binomial distribution with size \( X \), which seems to be a generic formulation. Poisson and Negative Binomial belong to the class of discrete compound Poisson distribution (DCP) encoded by parameters \((\lambda\alpha_1, \lambda\alpha_2, \ldots,)\), which satisfies:**

\[
\mathbb{E}(t^X) \equiv \sum_{x=0}^{\infty} \mathbb{P}(X = x)t^x = \exp\left\{ \sum_{k=1}^{\infty} \alpha_k \lambda (t^k - 1) \right\},
\]

(3)

10
for $|t| \leq 1$. It is equivalent to saying that $X$ is infinitely divisible. We note that if $X$ follows a DCP with parameter $(\lambda_1, \lambda_2, \ldots)$ and we draw $Z \sim \text{Bin}(X, p)$, then $Z$ marginally follows a DCP with parameter $\left(\sum_{k=1}^{\infty} \alpha_k \lambda_k (1 - p)^{k-1} p, \ldots, \sum_{k=1}^{\infty} \alpha_k \lambda_k^k (1 - p)^{k-1} p, \ldots\right)$. However, it may not be true in general that $X \mid Z$ also follows a (shifted) DCP when $X$ is DCP. The above examples are simply three cases where the aforementioned conditional distribution is still tractable.

2.3 Relationship between data splitting and data blurring

A natural question that might arise is under which conditions data blurring yields equivalent estimators as data splitting. One possible way to think about this is to consider the conditions in which the proportion of Fisher information that is allocated to inference is the same for each approach.

Suppose we are given $n$ i.i.d. observations $X = (X_1, \ldots, X_n)$, where $X_i \sim p(\theta)$. The data splitting approach chooses $S$ as a random subset of $[n]$ of size $a$ where $a \in \{\frac{1}{n}, \ldots, \frac{n-1}{n}, 1\}$ is a tuning parameter. Letting $\mathcal{I}_X(\theta)$ denote the Fisher information for the complete sample and $\mathcal{I}_1(\theta)$ denote the Fisher information for a single observation, we then have

$$\mathcal{I}_X(\theta) = \frac{a}{n} \mathcal{I}_1(\theta) + (1 - a) \mathcal{I}_1(\theta).$$

In the case of data blurring, we have a slightly more complicated setup. First, note that the Hessian of the log likelihood can be split up as follows for smooth parametric models:

$$\nabla^2 \ell(\theta; X) = \nabla^2 \ell(\theta; f(X)) + \nabla^2 \ell(\theta; g(X) \mid f(X)).$$

(4)

Taking expectations above, and denoting $\mathcal{I}_{f(X)}, \mathcal{I}_{g(X) \mid f(X)}$ as the Fisher information associated with the selection and inference datasets respectively, yields

$$\mathcal{I}_X(\theta) = \mathcal{I}_{f(X)}(\theta) + E[\nabla^2 \ell(\theta; g(X) \mid f(X))]$$

$$= \mathcal{I}_{f(X)}(\theta) + E[\nabla^2 \ell(\theta; g(X) \mid f(X))]$$

$$= \mathcal{I}_{f(X)}(\theta) + E[\mathcal{I}_{g(X) \mid f(X)}(\theta)].$$

If we start with some parameter $a$ to construct an estimator using data splitting and then wish to think about what the equivalent way to partition the dataset using data blurring, one can choose $\tau$ such that

$$E[\mathcal{I}_{g(X) \mid f(X)}(\theta)] = (1 - a) \mathcal{I}_1(\theta).$$

(5)

Note that this selected $\tau$ will only guarantee that the inference datasets created by a data splitting approach and a data blurring approach contain the same information in expectation. For any particular realization of the blurring step, $\mathcal{I}_{g(X) \mid f(X)}$ may be higher or lower than $(1 - a) \mathcal{I}_1(\theta)$. In any situation where $g(X)$ and $f(X)$ are independent, however, you can modify (5) to hold exactly since $E[\mathcal{I}_{g(X) \mid f(X)}(\theta)] = \mathcal{I}_{g(X)}$.

Another, perhaps more consequential caveat to note, is that this technique is only practically applicable when (5) is not a function of any unknown parameters. This makes this methodology intractable for several of the decompositions listed in Section 2.2. However, in the Gaussian case with known variance, both of these caveats do not apply. This allows us to justify the assertion we made in Section 1 that data blurring in the Gaussian case can be viewed as a continuous analog of data splitting.

Finally, we note that the above discussion assumes the parameter of interest $\theta$ is fixed prior to blurring the data and is therefore the same for both the selection and inference datasets. In contexts such as selective inference where the selection dataset is used by the statistician to decide which parameters they wish to conduct inference on, there is no guarantee that the same set of parameters will be selected when using the data splitting and data blurring approaches. Therefore, this approach should be treated more as a heuristic guideline once we start to apply this set of ideas in more complex settings.
Remark 3 (Trading off information between $f$ and $g$, part II). In the list of decompositions and in Remark 1 we noted that we could tradeoff information between $f(X)$ and $g(X)$ by varying certain hyperparameters. The framework in this section allows us to state what this means more precisely. More formally, for a hyperparameter $p \in (a,b)$, to say that larger $p$ corresponds to more informative $f(X)$ and less informative $g(X) | f(X)$ means that

$$\lim_{p \to b} I_{f(X)}(\theta) = I_X(\theta)$$

and

$$\lim_{p \to b} E[I_{g(X) | f(X)}(\theta)] = 0.$$

Gaussian Datasets. Let $X_i \sim N(\mu, \sigma^2)$. Recall that the data splitting approach defines $f(X)$ and $g(X)$ as:

$$f^{\text{split}}(X) = \frac{1}{an} \sum_{i \in S} X_i, \quad g^{\text{split}}(X) = \frac{1}{(1-a)n} \sum_{i \notin S} X_i,$$

By design, $f(X)$ and $g(X)$ satisfy that

$$f^{\text{split}}(X) \sim N(\mu, \frac{1}{an}) \quad g^{\text{split}}(X) \sim N(\mu, \frac{1}{(1-a)n}) \quad f^{\text{split}}(X) \perp \perp g^{\text{split}}(X).$$

Larger $a$ leads to more information in $f(X)$ (because of smaller variance) and less information in $g(X) | f(X)$.

Alternatively, the data blurring approach decomposes each observation $X_i$ as $f(X_i)$ and $g(X_i)$. We first simulate $\{Z_i\}_{i=1}^n$ distributed as i.i.d. $N(0, \sigma^2)$, and define $f(X)$ and $g(X)$ as:

$$f^{\text{blur}}(X_i) := X_i + b Z_i, \quad g^{\text{blur}}(X_i) := X_i - \frac{1}{b} Z_i,$$

where $\tau \in (0, \infty)$ is a tuning parameter. Observe that $f^{\text{blur}}(X_i)$ and $g^{\text{blur}}(X_i)$ satisfy that

$$f^{\text{blur}}(X_i) \sim N(\mu_i, \sigma^2(1 + \tau^2)), \quad g^{\text{blur}}(X_i) \sim N(\mu_i, \sigma^2(1 + \frac{1}{\tau^2})), \quad f^{\text{blur}}(X_i) \perp \perp g^{\text{blur}}(X_i).$$

Averaging all of these data points together gives us

$$f^{\text{blur}}(X) = \frac{1}{n} \sum_{i=1}^n f^{\text{blur}}(X_i), \quad g^{\text{blur}}(X) = \frac{1}{n} \sum_{i=1}^n g^{\text{blur}}(X_i).$$

In the Gaussian setting, equating the Fisher information of $f^{\text{blur}}(X)$ and $f^{\text{split}}(X)$ is equivalent to equating the variance. Therefore, to compare the two approaches of generating data, we can fix $a$ for data splitting and then calculate the choice of $\tau$ that results in $f(X)$ and $g(X)$ having the same variance when generated via data blurring. Performing this calculation results in the relation

$$a = \frac{1}{1 + \tau^2}. \quad (6)$$

For example, $a = 0.5$ and $\tau = 1$.

3 Application: selective confidence intervals after interactive multiple testing

The rest of the paper will focus on the use of data blurring for selective inference. We explore this idea more extensively in Sections 4 and 5 but we begin with a simpler example to illustrate the concept.
3.1 Review of some multiple testing methods and data blurring proposal

Suppose we observe $y_i \sim N(\mu_i, \sigma^2), i \in [1,n]$ with known $\sigma^2$ alongside generic covariates $x_i \in \mathcal{X}$. After observing the data, the analyst has two goals. First, the analyst wishes to choose a subset of hypotheses $\mathcal{R}$ to reject from the set $\{H_{0,i} : \theta_i = 0\}$ while controlling the false discovery rate (FDR), which is defined as the expected value of the False discovery proportion (FDP): $\text{FDP} := \frac{|\{x_i \in \mathcal{R} : \mu_i = 0\}|}{\max(|\mathcal{R}|, 1)}$.

After selecting these hypotheses, the analyst then may wish to construct either:

- multiple confidence intervals with $1 - \alpha$ coverage of $\mu_i$ for each $i \in \mathcal{R}$; or
- a single confidence interval with $1 - \alpha$ coverage of $\bar{\mu} = \frac{1}{|\mathcal{R}|} \sum_{i \in \mathcal{R}} \mu_i$.

One method for rejecting hypotheses and constructing selective confidence intervals that achieve coverage for the individual $\mu_i$ would be to use the Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg (1995)) to form the rejection set and then construct confidence intervals with $1 - \alpha$ coverage as described by Benjamini and Yekutieli (2005)—which we refer to as BY-corrected confidence intervals in the remainder of this paper. In the context of the problem setup described above, these can be calculated as

$$y_i \pm \frac{z_{\alpha/2}}{\sqrt{1 + \frac{1}{\tau^2}}}$$  \hspace{1cm} (7)

A weakness of this approach, however, is that it affords the analyst no flexibility when forming the rejection set. The procedure forces the analyst to reject hypotheses according to a pre-defined rule instead of allowing them to interact with the data and form a rejection set in a data adaptive way. Moreover, even though we have a confidence interval with valid post-selective coverage for each individual $y_i$, we know of no methodology for aggregating the individual confidence intervals to form a single confidence interval that will cover $\bar{\mu}$.

An alternative general approach for selective inference is to compute a $p$-value for testing each $y_i$, and then decompose the $p$-value as proposed by the AdaPT (Lei and Fithian (2018)) and STAR (Lei et al. (2020)) procedures. These methodologies allow the data analyst to iteratively build a rejection set in a data adaptive way. Masked versions of $p$-values are given to the analyst for data exploration. The analyst then chooses hypotheses to reject one-by-one. After the analyst adds a hypothesis to the rejection set, the full $p$-value is revealed and can then be used to further fine-tune the selection rule for subsequent rejection decisions. A major drawback of these approaches, however, is that they only are designed to work with a $p$-value. Information about the individual $y_i$ is discarded when running these analyses, preventing the analyst from constructing confidence intervals from the data. Although constructing confidence intervals by following equation 7 appears to control the false coverage rate empirically, there is no known theoretical guarantee to ensure that confidence intervals constructed in this way have proper coverage. Moreover, to the best of our knowledge, there is no existing method to cover $\bar{\mu}$, even heuristically.

Data blurring offers one possible path forward. Our proposed methodology would be as follows:

1. Draw $z_i \sim N(0, \sigma^2)$ and let $f(y_i) = y_i + \tau z_i$ with $g(y_i) = y_i - \frac{1}{\tau} z_i$.
2. Use $f(y_i)$ to select a rejection set of suspected non-nulls using any desired error-control procedure. If the analyst believes that the set of rejections should form a convex region, they may wish to use the STAR algorithm. On the other hand, AdaPT or BH procedures may be chosen if no such belief is held.
3. After selecting hypotheses, we can choose to form confidence intervals to cover either the individual $\mu_i \in \mathcal{R}$ with $1 - \alpha$ coverage as

$$g(y_i) \pm z_{\alpha/2} \sqrt{1 + \frac{1}{\tau^2}}$$  \hspace{1cm} (8)
or we can form a $1 - \alpha$ confidence interval to cover $\mu_i$ as

$$
\frac{\sum_{i \in R} g(y_i)}{|R|} \pm z_{\alpha/2} \sqrt{\frac{1 + \frac{1}{n}}{|R|}}.
$$

### 3.2 Empirical results

Although the confidence intervals constructed by data blurring to cover the individual $\mu_i$ have proper coverage, they tend to be too wide to be useful in most contexts since their widths do not scale down with the size of the rejection set. If the analyst decides that they are interested in estimating the signal strength over the entire rejection region, however, data blurring offers good performance compared to existing methodologies.

To illustrate this, we adopt the setting of Lei et al. (2020). We let the covariates $x_i \in \mathbb{R}^2$ be arranged on a $25 \times 25$ grid in the area $[-100, 100] \times [-100, 100]$. We let the set of true non-nulls be arranged in a circle in the center of the grid as shown in Figure 4 and set $\mu_i = 2$ for each non-null and $\mu_i = 0$ for each true null. We then generate $y_i \sim N(\mu_i, 1)$.

We then form a rejection set using the BH, AdaPT, and STAR procedures, both with and without data blurring. For the non-blurred versions of these procedures using the full dataset, the (one-sided) p-values that are used are calculated as $p_i^{\text{null}} = 1 - \Phi(y_i)$. Similarly, for the blurred versions of these procedures, the (one-sided) p-values are calculated as $p_i^{\text{blur}} = 1 - \Phi(f(y_i))$. Note that although AdaPT and STAR control the false discovery rate regardless of how the analyst chooses to reject their hypotheses, for the purposes of this simulation, we rely on the algorithmic procedures described in the original papers to form the rejection set. Please see Section 4.2 of Lei et al. (2020) and Section 4 of Lei and Fithian (2018) for full details.

An example of a rejection set for each of the six methodologies for a single trial run can be seen in Figure 4. We note that the increased variance introduced through blurring will necessarily decrease the power of these procedures in the selection phase, but the level of degradation is relatively minor for small values of $\tau$. After forming the rejection sets for the non-blurred dataset, we form confidence intervals to cover the individual $\mu_i \in R$ using equation (7). After forming the rejection sets for the blurred versions of the datasets, we form confidence intervals to cover $\mu$ using equation (9). We again reiterate that the BY-corrected confidence intervals designed to cover $\mu_i$ only have known theoretical coverage guarantees in the case of the BH procedure. Nonetheless, we compute them for the AdaPT and STAR procedures as well in order to have some baseline of comparison for the blurred versions of these procedures.

We repeat this simulation over 500 trial runs for a variety of different $\tau$ to observe how information trades off between the selection and inference steps. To measure the effectiveness of our procedure, we track the following metrics for the selection stage:

- **Power** := $\frac{|x_i \in R : \mu_i \neq 0|}{|\{i : \mu_i \neq 0\}|}$, and
- **false discovery proportion (FDP)** := $\frac{|x_i \in R : \mu_i = 0|}{\max\{|R|, 1\}}$.

For the inference stage, we denote the confidence interval created through the data blurring procedures that covers $\mu$ as $\text{CI}$ with lower and upper bounds $\text{CI}(1), \text{CI}(2)$ and compute the following metrics for each trial run

- **Miscoverage rate** := $\text{Ind}(\mu \not\in \text{CI})$, and
- **CI Length** := $|\text{CI}(2) - \text{CI}(1)|$.

We denote the confidence intervals created through the non-blurred procedures that cover the individual $\mu_i$ as $\text{CI}_i$ with lower and upper bounds $\text{CI}_i(1)$ and $\text{CI}_i(2)$ and compute the following additional metrics

- **False coverage rate (FCR)** := $\frac{|x_i \in R : \mu_i \not\in \text{CI}_i|}{\max\{|R|, 1\}}$, and
- **CI Length** := $\frac{1}{R} \sum_{i \in R} |\text{CI}_i(2) - \text{CI}_i(1)|$.

The results of these experiments are shown in Figures 6 and 7 with target CI coverage at level $1 - \alpha = 0.8$ and target FDR control at 0.1. As $\tau$ increases, we see a smooth tradeoff between the
Figure 4. True underlying region of non-nulls for hypothesis detection simulation. Each point is a signal, with black representing the non-nulls with $\mu_i = 2$. Grid size for simulations is $25 \times 25$. The region for detection is convex, making this problem most naturally suited for the STAR algorithm. However, we also compare performance for BH and AdaPT procedures as a point of comparison.

Figure 5. Example rejection region for a single trial run. STAR, AdaPT, and BH procedures are used to try and recover the true signal pattern shown in Figure 4 with a target FDR of 0.1. Blurring the data (with $\tau = 0.1$) to allow for the estimation of signal strength only mildly impairs the ability of these procedures to correctly identify the rejection region.
Figure 6. Numerical results for the selection stage averaged over 500 trials for a 25 × 25 grid of hypotheses with target FDR level chosen at 0.1 and τ varying over \{0.1, 0.3, 0.5, 0.7, 0.9\}. Solid lines denote metrics for the rejection sets formed using the full dataset and dotted lines denote metrics calculated using the rejection sets formed through data blurring. All methods control the false discovery rate at the desired coverage level but we can see that increasing τ decreases the power of the blurred procedures as some of the information gets reserved for inference.

power in the selection stage and the confidence interval length in the inference stage. Across most procedures and choice of τ, covering the entire selected region with data blurring allows for a more precise estimate of the signal strength, as judged by CI length, compared with trying to cover each individual data point using the Benjamini-Yekutieli (BY) correction.

4 Application: selective CIs in fixed-design linear regression

We now turn to applying data blurring to fixed-design Gaussian linear regression. This is largely similar to the discussion in Rasines and Young (2021) but we reinforce the findings here as we build on several of the results in this section in our treatment of trend filtering in Section 5. We assume that \( y_i \) is the dependent variable and \( x_i \in \mathbb{R}^p \) is a vector of \( p \) features for \( i = 1, \ldots, n \) samples. For simplicity, we also denote \( X = (x_1, \ldots, x_n)^T \) as the model design matrix and \( Y = (y_1, \ldots, y_n)^T \) with the following structural equation:

\[
Y = \mu + \epsilon \quad \text{with} \quad \epsilon = (\epsilon_1, \ldots, \epsilon_n)^T \sim N(0, \Sigma),
\]

where \( \mu = E[Y | X] \in \mathbb{R}^n \) is a fixed unknown quantity and \( \epsilon \in \mathbb{R}^{n \times n} \) is a random quantity with a known covariance matrix \( \Sigma \).

During the blurring phase, we introduce the independent quantities \( f(Y) \) and \( g(Y) \) created by adding Gaussian noise \( Z \sim N(0, \Sigma) \) as described in Section 2.3:

\[
f(Y) = \mu + \epsilon + \tau Z, \quad \text{and} \quad g(Y) = \mu + \epsilon - \frac{1}{\tau} Z.
\]

We use \( f(Y) \) to select a model \( M \subseteq [p] \) that, in turn, defines a model design matrix \( X_M \) which is itself a submatrix of \( X \). After selecting these covariates, we then use the independent dataset \( g(Y) \) for inference.

We conduct inference by fitting a linear regression on \( g(Y) \) against the selected covariates \( X_M \). However, in an assumption-lean setting where \( \mu = E[Y | X] \) is not guaranteed to be a linear combination of the chosen covariates, it is not entirely clear what the fitted coefficients and corresponding confidence intervals represent. For our purposes, we will use the same problem setup of Buja et al. (2019), but
Figure 7. Numerical results for the inference stage averaged over 500 trials for a 25 × 25 grid of hypotheses. Target FDR level chosen at 0.1 and target CI coverage chosen at 1 − α = 0.8 with and τ varying over {0.1, 0.3, 0.5, 0.7, 0.9}. The solid lines represent the false coverage rate (FCR) of µ for the confidence intervals computed using the BY correction on the full dataset. The dotted lines represent the average miscoverage rate and confidence interval length for the confidence interval created through data blurring to cover µ. All confidence intervals have the desired coverage but we see that the CI length decreases monotonically as τ increases and more of the dataset gets reserved for inference.

with one significant modification. The above paper focuses on the random-X setting where (x_i, y_i) are sampled as random pairs from a common joint density. In the context of data blurring, it is only possible to make such an assumption during the blurring stage. When conducting inference, X has already been observed once during model selection so we necessarily must condition on the realized values of X. This restricts us to a fixed-X setting during the inference stage.

Setting aside these concerns for a moment, we will introduce notation. We will let \( \hat{\beta}(M) \) be defined in the usual way as
\[
\hat{\beta}(M) = \arg\min_{\tilde{\beta}} \| g(Y) - X_M \tilde{\beta} \|_2 = (X_M^T X_M)^{-1} X_M^T g(Y).
\]
(10)

Our target parameter is then denoted as
\[
\beta_*(M) = \arg\min_{\tilde{\beta}} E \left[ \| Y - X_M \tilde{\beta} \|_2^2 \right] = (X_M^T X_M)^{-1} X_M^T \mu.
\]
(11)

We are then able to form confidence intervals that guarantee 1 − α coverage of \( \beta_*(M) \) as follows.

**Theorem 2.** Let \( \hat{\beta}(M) \) be defined as in (10) and \( \beta_*(M) \) be defined as in (11). Then the following holds:

\[
\hat{\beta}(M) \sim N(\beta_*(M), (1 + \tau^{-2})(X_M^T X_M)^{-1} X_M^T \Sigma X_M (X_M^T X_M)^{-1}).
\]

Furthermore, we can form a 1 − α confidence interval for the kth element of \( \beta_*(M) \) as
\[
\hat{\beta}(M)_k \pm z_{\alpha/2} \sqrt{(1 + \tau^{-2})(X_M^T X_M)^{-1} X_M^T \Sigma X_M (X_M^T X_M)^{-1}}.
\]

**Proof.** This follows as a standard application of OLS properties. First, write
\[
\hat{\beta}(M) = (X_M^T X_M)^{-1} X_M^T g(Y) = (X_M^T X_M)^{-1} X_M^T [\mu + \epsilon - \frac{1}{\tau} Z]
\]
\[
= \beta_*(M) + (X_M^T X_M)^{-1} X_M^T [\epsilon - \frac{1}{\tau} Z].
\]
From Section 2.3, we know that $(\epsilon - \frac{1}{\tau}Z) \sim N(0, (1 + \tau^{-2})\Sigma)$. Therefore
\[
\hat{\beta}(M) \sim N(\beta_*(M), (1 + \tau^{-2})(X_M^T X_M)^{-1}X_M^T \Sigma X_M (X_M^T X_M)^{-1}).
\]
The coverage statement then follows straightforwardly from the definition of a confidence interval and the properties of the multivariate Gaussian distribution.

A key assumption of this procedure is that the variance is known in order to do the initial split between $f(Y)$ and $g(Y)$ during the blurring phase. In the case of unknown variance, one can form an estimate $\hat{\sigma}$ in order to create a split, but this forces the analyst to then condition on both $f(Y)$ and $\hat{\sigma}$, which may not have a tractable distribution in high dimensional settings. In the case where $p$ is fixed and $n \to \infty$, we explore an extension of this methodology to account for unknown variance in B.1. Extending this approach to finite samples and high-dimensional regimes remains an open question for future lines of research.

### 4.1 Comparison with data splitting

We can give an analogous statement about how this data blurring procedure relates to data splitting as the one made in Section 2.3. First, let us recap what the distributions are in each case.

**Data Splitting** Here we pick $a \in \{\frac{1}{n}, \ldots, \frac{n-1}{n}, 1\}$ and set aside the first $an$ observations for selection and the remaining $(1-a)n$ observations for inference. We then select some model $M \subseteq [p]$ during the selection stage. This restricts us to a smaller model design matrix $X_{M,\text{inf}} \in \mathbb{R}^{(1-a)n \times |M|}$ during the inference stage.

Also denote $\Sigma^{\text{inf}}$ the covariance matrix for the error term of these $(1-a)n$ observations. We then have that
\[
\hat{\beta}_{\text{split}} \sim N(\beta_*(M), (X_{M,\text{inf}}^T X_{M,\text{inf}})^{-1}X_{M,\text{inf}}^T \Sigma_{\text{inf}} X_{M,\text{inf}} (X_{M,\text{inf}}^T X_{M,\text{inf}})^{-1}).
\]

**Data Blurring** Denote the model chosen during the selection stage as $M^* \subseteq [p]$ and the corresponding model design matrix $X_{M^*} \in \mathbb{R}^{n \times |M^*|}$. In this setting we have:
\[
\hat{\beta}_{\text{blur}} \sim N(\beta_*(M), (1 + \tau^{-2})(X_{M^*}^T X_{M^*})^{-1}X_{M^*}^T \Sigma X_{M^*} (X_{M^*}^T X_{M^*})^{-1}).
\]

Both quantities are unbiased, but determining whether $\hat{\beta}_{\text{blur}}$ or $\hat{\beta}_{\text{split}}$ has lower variance depends on too many unknown quantities to say definitively without making additional assumptions. In particular:

- There is no guarantee that $M^*$ and $M$ will be the same — i.e. different models may be selected during the first stage for each of these two procedures.
- Nothing is known about the model design matrix $X$ so it is hard to say much about how much smaller the quantity $(X^T X)^{-1}$ becomes if restricted to only a subset of rows/columns (e.g. $X_{M,\text{inf}}$).

Intuitively, the estimator increases variance when compared with data splitting via randomization through the $1 + \tau^{-2}$ term but decreases variance by increasing the number of rows within the model design matrix. Although not very realistic, we can explore a more restrictive example to illustrate the application of the above equations.

**Example 1.** We assume that we have:

1. **Orthogonal Covariates:** $X_i X_j = 0$ for all $i \neq j$.
2. **Homoscedasticity:** $\sigma^2 = \sigma_i^2$ for all $i$.
3. We select the same model during the selection stage for each procedure, so $M = M^*$.
4. As in the above, \( \Sigma \) is known.

These assumptions simplify things because we now have that
\[
\hat{\beta}_{\text{blur}} \sim N(\beta_*(M),(1 + \tau^{-2})\sigma^2(X_MX_M^T)^{-1}),
\]
\[
\hat{\beta}_{\text{split}} \sim N(\beta_*(M),\sigma^2(X_{\text{inf}}X_{\text{inf}}^T)^{-1}).
\]
Moreover the matrices are easily invertible since the orthogonality assumptions yields \( X^TX \) diagonal. Therefore, we have:
\[
\text{Var}(\beta_{\text{blur}}) = (1 + \tau^{-2})\sigma^2 \sum_{i=1}^n X_{ij}^2,
\]
\[
\text{Var}(\beta_{\text{split}}) = \sigma^2 \sum_{i=an+1}^n X_{ij}^2.
\]

For a given choice of \( a \), the parameter \( \tau \) that equates these two variances is
\[
\tau = \left( \frac{\sum_{i=1}^n X_{ij}^2}{\sum_{i=an+1}^n X_{ij}^2} \right)^{1/2}.
\]

Although idealized, this example draws attention to a key potential weakness that data splitting has in fixed-design linear regression. Data blurring is always able to smoothly tradeoff the information between the selection and inference datasets through the parameter \( \tau \) but data splitting is more limited in the ways that information can be divided since it requires that the analyst allocate points discretely. In cases where the sample size is large and the distribution of covariates is well behaved, data splitting may be able to tradeoff information relatively smoothly by changing the proportion of data points in each sample (\( a \)). However, data blurring will tend to outperform data splitting in settings with small sample size where the distribution of covariates is skewed such that there are a handful of points with high leverage. Data splitting has an inherent disadvantage in this setting because the analyst is forced to choose to allocate each of these leverage point to either the selection or inference dataset. In contrast, data blurring enables the analyst to “hedge their bets” so that a piece of the information contained in each leverage point is allocated to both the selection and inference datasets. Figure 8 offers an illustration of this tendency on an example dataset.

4.2 Empirical results

We now demonstrate the advantages that data blurring has over data splitting through a large-scale empirical study. We conduct inference on some vector \( Y \) given a set of covariates \( X \) and a known covariance matrix \( \Sigma = \sigma^2 I_n \) as follows.

1. Decompose \( Y_i \) into \( f(Y_i) = Y_i - Z_i \) and \( g(Y_i) = Y_i + Z_i \) where \( Z_i \sim N(0,\sigma^2) \).
2. Fit \( f(Y_i) \) by linear regression with lasso regularization to select features, denoted as set \( M \subseteq [p] \) (in our experiments, we use \texttt{cv.glmnet} in R package \texttt{glmnet} and choosing the tuning parameter \( \lambda \) by the 1 standard deviation rule, which can be found in the value of \texttt{lambda.1se})

3. Fit \( g(Y_i) \) by another linear regression without regularization with only the selected features (for example, using \texttt{lm} in R package \texttt{stats}).

4. Construct confidence intervals for the coefficients trained in the third step, each at level \( \alpha \) and the standard error is estimated by the sandwich estimator (using \texttt{conf.int} in R package \texttt{clubSandwich} with \texttt{vcov = "CR2"})

**Simulation setup**  We choose \( \sigma^2 = 1 \) and generate \( n = 16 \) data points with \( p = 20 \) covariates. For the first 15 data points, we have an associated vector of covariate \( X_i \in \mathbb{R}^{15} \) generated from independent standard Gaussians. The last data point, which we denote \( X_{\text{lev}} \), is generated in such a way as to ensure it is likely to be more influential than the remaining observations due to having much larger leverage. We define \( X_{\text{lev}} = \alpha (|X_1|_{\infty}, \ldots, |X_p|_{\infty}) \) where \( \alpha \) is a parameter that we will vary within these simulations that reflects the degree to which the last data point has higher leverage than the first set of data points. \( Y_i \) follows a Gaussian distribution with expected value \( \gamma^T X_i \). The parameter \( \gamma \) is nonzero for 4 features: \((\gamma(1), \gamma(16), \gamma(17), \gamma(18)) = (1, 1, -1, 1)\).

We repeat over 500 trials and then summarize the performance using the following metrics. For the selection stage, we compute the power and precision of selecting features with a nonzero parameter:

\[
\text{power}_{\text{selected}} := \frac{|j \in [p] : \gamma_j \neq 0 \text{ is selected and nonzero}|}{|j \in [p] : \gamma_j \neq 0|}, \quad (12)
\]

\[
\text{precision}_{\text{selected}} := \frac{|j \in [p] : \gamma_j \neq 0 \text{ is selected and nonzero}|}{|j \in [p] : \gamma_j \neq 0|}. \quad (13)
\]

To track performance during the inference stage, we compute the false coverage rate and confidence interval length within the selected model, defined as

\[
\text{FCR} := \frac{|k \in M : \gamma_k \notin CI_k|}{\max\{|M|, 1\}} \quad \text{and} \quad \text{CI length} := \frac{1}{|M|} \sum_{k \in M} |CI_k(2) - CI_k(1)|, \quad (14)
\]

where \( CI_k \) is the CI for \( \gamma_k \) and \( CI_k(1), CI_k(2) \) are the lower and upper bound of \( CI_k \). Several other metrics we explore include the averaged proportion of falsely reported confidence intervals among those indicating a non-zero parameter (the false sign rate):

\[
\text{FSR} := \frac{|k \in M : (\gamma_k < 0 \text{ and } CI_k(1) > 0) \text{ OR } (\gamma_k > 0 \text{ and } CI_k(2) < 0)|}{\max\{|k \in M : \gamma_k \neq 0\}, 1\}}, \quad (15)
\]

and the power of correctly reporting confidence intervals that indicates a nonzero parameter with the correct sign:

\[
\text{power}_{\text{sign}} := \frac{|k \in M : (\gamma_k > 0 \text{ and } CI_k(1) > 0) \text{ OR } (\gamma_k < 0 \text{ and } CI_k(2) < 0)|}{\max\{|j \in [p] : \gamma_j \neq 0\}, 1\}}, \quad (16)
\]

\footnote{This sandwich estimator slightly modifies the original one proposed by Huber (1967) and White (1982) to include a small sample correction.}
Figure 9. An instance of the selected feature (blue crosses) and the constructed confidence intervals using blurred data (left), full data twice (middle), and split data (right) with \( S_\delta = 0.2 \) and target FCR set at 0.2. The selected features are marked by blue crosses, which include all of the nonzero coefficients (corresponding to almost 100% power for selection) and also a few zero coefficients (corresponding to around 70% precision for selection). Confidence intervals which do not cover the parameters correctly are marked in red.

As an illustration, Figure 9 shows an instance of the selected features and corresponding confidence intervals for an example trial run. As a point of comparison, we compare the confidence intervals constructed using data blurring with those constructed using data splitting (when 50% of the dataset is used for selection and the remaining for inference). We also compare these results to the (invalid) algorithm where the original dependent variable is used twice to both select features and construct intervals (replacing \( f(Y_i) \) and \( g(Y_i) \) both by \( Y_i \) in the above algorithm). We note that the third methodology will not have adequate coverage guarantees but it is still a useful point of comparison for evaluating the performance of the other two (valid) methodologies.

Figure 10. FCR, length of the confidence intervals, FSR, power for the sign of parameters, and power and precision for the selected features, when varying the leverage parameter \( \alpha \) in \( \{2, 3, 4, 5, 6\} \). The results are averaged over 500 trial runs. Both data splitting and data blurring still control the FCR as before, but data blurring now has higher power and precision, as well as tighter confidence intervals than data splitting.

Figure 10 shows results average over 500 trial runs. We see that data splitting and data blurring both control the FCR, but data blurring dominates data splitting across every other metric—significantly tighter confidence interval and higher power and precision. We note that in other simulation settings with larger sample sizes and less skewed covariates, data blurring and data splitting have
roughly comparable performance—please see Appendix B for further details. We also note relatively large gains in power over data splitting for non-Gaussian fixed design regression, even when sample sizes are quite large and covariate distributions are not skewed. These findings are discussed in greater length in Appendices A.1 and A.2.

5 Application to trend filtering and other nonparametric regression problems

We consider a standard nonparametric setup where we observe \( y_i \) and a corresponding set of covariates \( x_i \) following the equation

\[
y_i = f_0(x_i) + \epsilon_i
\]

where \( f_0 \) is the underlying function to be estimated and \( \epsilon_i \) is random noise. As before we denote \( Y = (y_1, ..., y_n)^T \) and \( \epsilon = (\epsilon_1, ..., \epsilon_n)^T \). For now, we assume that \( \epsilon \sim N(0, \Sigma) \) for some known \( \Sigma \). We can apply the methodologies of sections 2 and 3 to this problem as follows.

1. Decompose \( Y \) into \( f(Y) = Y + \tau Z \) and \( g(Y) = Y - \frac{1}{\tau} Z \) where \( Z \sim N(0, \Sigma) \).
2. Use \( f(X) \) to choose a set of basis functions \( a_1, ..., a_p \) for the series expansion of \( x_i \) and denote

\[
a(x_i) = (a_1(x_i), ..., a_p(x_i))^T.
\]

Let \( A \) denote the matrix of basis vectors for all \( n \) data points.
3. Use \( g(X)|f(X) \) to construct pointwise or uniform confidence intervals as described below.

Pointwise confidence interval We first note that the fitted line

\[
\hat{\beta}(A) = \arg\min_{\beta} ||g(Y) - A\beta||^2 = (A^T A)^{-1} A^T g(Y)
\]

(18)
corresponds to \( A\hat{\beta}(A) \) which is just the projection of \( g(Y) \) onto the basis chosen during the model selection stage. Meanwhile, we define the projected regression function as

\[
\hat{\beta}^*(A) = \arg\min_{\beta} E \left[ ||Y - A\beta||^2 \right] = (A^T A)^{-1} A^T f_0(X).
\]

(19)

We are then interested in using the fitted estimates \( \hat{\mu}_{x_i}(A) = a(x_i)^T \hat{\beta}(A) \) to construct intervals that guarantee coverage for the underlying function \( f_0(x_i) \) projected onto the chosen basis which we denote as \( \mu^*_x(A) = a(x_i)^T \beta^*(A) \) and refer to as the projected mean.

**Corollary 1.** Recall the definitions of \( y_i, \epsilon_i, \hat{\beta}(A), \beta^*_x(A) \) from (17), (18) and (19). When \( \epsilon \) has known variance \( \Sigma \), then the following holds:

\[
\hat{\mu}_{x_i}(A) \sim N \left( \mu^*_x(A), (1 + \tau^{-2})a(x_i)^T(A^T A)^{-1}A^T \Sigma A(A^T A)^{-1}a(x_i) \right).
\]

Furthermore, we can form a \( 1 - \alpha \) confidence interval for \( y^*_{x_i}(A) \) as

\[
\hat{\mu}_{x_i}(A) \pm z_{\alpha/2} \sqrt{(1 + \tau^{-2})a(x_i)^T(A^T A)^{-1}A^T \Sigma A(A^T A)^{-1}a(x_i)}.
\]

Proof. It follows directly from Theorem 2 that

\[
\hat{\beta}(A) \sim N \left( \beta^*_x(A), (1 + \tau^{-2})(A^T A)^{-1}A^T \Sigma A(A^T A)^{-1} \right).
\]

To conclude, simply multiply by \( a(x_i)^T \) and apply standard properties of the multivariate Gaussian distribution. \( \square \)
Uniformly valid confidence intervals  Alternatively, we can construct uniformly valid confidence interval to cover the projected mean. In this case, the goal is to construct a set of confidence intervals \( \{CI_i\}_{i=1}^{n} \) that control the simultaneous type I error rate defined as

\[
P(\exists i \in [n] : y_{x_i}^*(A) \notin CI_i).
\]

We use equation (5.5) in Koenker (2011) for constructing the uniform confidence interval for additive models. Note that in this case, we further assume homoscedasticity in the error terms so \( \Sigma = \sigma^2 I_n \).

Construct \( CI_i := (\hat{y}_{x_i}(A) - w, \hat{y}_{x_i}(A) + w) \), where the width \( w := c(\alpha) \cdot \sigma \sqrt{(1 + \tau^{-2})a(x_i)^T(A^T A)^{-1}a(x_i)} \), where \( c(\alpha) \) is a multiplier. By equation (5.5) in Koenker (2011), \( c(\alpha) \) is the solution of

\[
\frac{|\gamma|}{2\pi} e^{-\gamma^2/2} + 1 - \Phi(c) = \alpha/2,
\]

where \( |\gamma| \) is the length of the curve connected by \( \tilde{a}(x_i) := \frac{(A^T A)^{-1/2}a(x_i)}{||(A^T A)^{-1/2}a(x_i)||} \). In other words, \( |\gamma| = \sum_{i=1}^{n-1} ||\tilde{a}(x_{i+1}) - \tilde{a}(x_i)||_2 \).

**Fact 1.** The above constructed CIs will cover the projected mean uniformly with simultaneous type I error controlled at level \( \alpha \):

\[
P(\exists i \in [n] : y_{x_i}^*(A) \notin CI_i) \leq \alpha,
\]

where \( \{y_{x_i}^*(A)\}_{i=1}^{n} \) denotes the projected mean.

*Proof.* This is proven in Koenker (2011), using techniques discussed in Knowles (1987).

### 5.1 A recap of trend filtering

We investigate the use of data blurring in the context of nonparametric regression empirically through the lens of trend filtering as proposed by Kim et al. (2009) and Steidl et al. (2006).

We consider the problem of estimating the underlying smooth trend of a time series \( y_t \in \mathbb{R} \) with \( t = 1, \ldots, n \). We have a similar equation as in the standard nonparametric setup, but now let \( x_t = t \) for all \( i \):

\[
y_t = f_0(t) + \epsilon_t.
\]

(22)

Our goal is to estimate the underlying trend \( (f_0(1), \ldots, f_0(n)) \). The approach of trend filtering is to fit a piecewise polynomial of degree \( k \) to the data with adaptively chosen breakpoints or knots. Formally, the \( k \)-th order trend filtering estimator is defined to be \( \hat{x} = (\hat{x}_1, \ldots, \hat{x}_n) \), which is the solution to the minimization problem

\[
\hat{x} = \arg\min_{x \in \mathbb{R}^n} \frac{1}{2} \|Y - x\|_2^2 + \lambda \left\| D^{(k+1)} x \right\|_1, \tag{23}
\]

where \( \lambda \geq 0 \) is a tuning parameter and \( k \) is the order of the piecewise polynomial that is being chosen to fit the data. \( D^{(k+1)} \in \mathbb{R}^{(n-k) \times n} \) is the \( k \)-th order difference matrix defined recursively by defining

\[
D^{(1)} = \begin{bmatrix}
-1 & 1 & 0 & \ldots & 0 & 0 \\
0 & -1 & 1 & \ldots & 0 & 0 \\
\vdots & \ddots & \ddots & \ddots & \ddots & \ddots \\
0 & 0 & 0 & \ldots & -1 & 1
\end{bmatrix} \in \mathbb{R}^{(n-k-1)},
\]

and \( D^{(k+1)} \in \mathbb{R}^{(n-k) \times n} \) as

\[
D^{(k+1)} = D^{(1)} D^{(k)}.
\]

From this definition, we can see that setting \( k = 0 \) yields \( D^{(k+1)} = D^{(1)} \) which corresponds to fitting a piecewise constant function. This is also known as the 1-dimensional fused lasso problem of Tibshirani et al. (2005). Choosing \( k = 1 \) corresponds to fitting a piecewise linear function across \( t \), choosing \( k = 2 \) corresponds to fitting a piecewise quadratic function across \( t \) and so on.
Although trend filtering is a relatively recently developed tool in nonparametric statistics, it has gained substantial popularity over the last several years due to it converging to the true underlying function at the minimax rate (see Tibshirani (2014)) in addition to being computationally efficient. The specialized alternation direction method of multipliers (ADMM) algorithm of Ramdas and Tibshirani (2014) converges at the $O(n^{1.5})$ rate in the worst case but in practice tends to scale extremely well for large datasets. Most other nonparametric estimators that can be computed efficiently such as smoothing splines and uniform-knot regression splines are not locally adaptive and therefore do not converge at the minimax rate. Other methods that are locally adaptive such as the locally adaptive regression spline of Mammen and van de Geer (1997) can be shown to converge at the minimax rate but are not computationally tractable. For a more complete overview of these competing methods and the various ways in which they may tradeoff between their theoretical properties, the interested reader may wish to consult Politsch et al. (2020a).

A key limitation of this approach, however, is that parametric uncertainty quantification is not generally tractable even when strong distributional assumptions are made. Even when it is assumed that the data follows a Gaussian distribution with known variance, the fact that knots are chosen adaptively turns the post-selective distribution into something non-Gaussian. In this section, we will demonstrate how data blurring can be used to create valid confidence intervals when there is an underlying assumption of Gaussian distributed data with known variance $\Sigma$.

In some specific settings, data carving approaches may also offer a viable path forward. When the (fused or generalized) lasso is used to select the choice of knots, methods such as those described in Chen et al. (2021), Duy and Takeuchi (2021), and Hyun et al. (2018) can be used to construct test and confidence intervals with valid post-selective distributions. Although tests constructed through these methodologies will likely have higher power than a data blurring approach due to conditioning on a smaller amount of information, the drawback of these approaches is that they offer the analyst very little flexibility during the selection stage. If the analyst wishes to change the degree of the fitted polynomial or decrease the number of chosen knots after seeing preliminary results, a data blurring approach allows them this flexibility. Although such actions are common in applied data analysis, data carving approaches do not easily handle these sorts of ad-hoc changes to the prespecified selection methodology.

As discussed in Section 4, an assumption of known variance is required when using data blurring in order to ensure the selection and inference datasets are truly independent. With that being said, we do discuss possible approaches for extending this methodology in cases when the variance is unknown in Appendix C.2. Although we do not have theoretical guarantees in this case, a heuristic approach for using data blurring in this case demonstrates promise based on the strength of early empirical results.

5.2 Empirical results

Using the blurring for Gaussian distributed data in Section 2.2, we propose a procedure to get pointwise and uniform confidence as follows.

1. Decompose $Y$ into $f(Y) = Y + \tau Z$ and $g(Y) = Y - \frac{1}{\tau} Z$ where $Z \sim N(0, \Sigma)$ (assuming $\Sigma$ is known).

2. Train a $k$-th order trendfilter using $f(Y)$ (using cv.trendfilter in R package genlasso and select the tuning parameter $\lambda$ by selecting the set of knots with the minimum cross-validation error. Denote the chosen knots $\tau_k$ for $k \in 1, \ldots, m$. Note that although we restrict ourselves to a fixed selection rule for simulation purposes, data blurring allows for the use of any arbitrary method of choosing the set of knots. For a discussion of alternative methods of choosing the tuning parameter $\lambda$ such as the 1 standard deviation rule, please see Appendix C.3.2

3. Construct a $k$th degree falling factorial basis, as described in Tibshirani (2014), with knots at $\{\tau_k\}_{k=1}^m$ and use this to regress against $g(Y)$. Note that when $k = 0$ or $k = 1$, this reduces to generating the more familiar truncated power basis.
4. Get \( \text{pointwise} \) confidence intervals \( \text{CI}_t \) for each data point \( t = 1, \ldots, n \) at level \( \alpha \) as described in Corollary 1. Alternatively, in cases where \( \Sigma = \sigma^2 I_n \), we can construct uniform confidence intervals as described in described in Fact 1.

Note that the above constructed CIs will cover the projected mean, which in this case is the projection of the function \( f_0(t) \) onto the linear spline with knots at \( \{\tau_k\}_{k=1}^m \). The pointwise confidence intervals constructed through this procedure for \( t = 1, \ldots, n \) should have FCR control at level \( \alpha \), while the uniform confidence intervals will control the probability of not covering the projected mean at any time point at level \( \alpha \).

For our simulations, we focus on first order trend filtering where \( k = 1 \), although the results in the preceding section are fully general and can be applied to polynomial splines of any arbitrary degree. In this setting, (23) simplifies to finding

\[
\hat{x} = \arg\min_{x_t \in \mathbb{R}} 1/2 \sum_{t=1}^{n} (Y_t - x_t)^2 + \lambda \sum_{t=2}^{n-1} |x_{t-1} - 2x_t + x_{t+1}|. 
\]

**Simulation setup.** We generate simulated data as \( Y_t = f(t) + Z_t \) for \( t \in [n] \), where \( Z_t \) is noise following \( N(0, \sigma^2) \), and \( f_0(t) \) is the true underlying trend that is generated as \( f_0(t+1) = f_0(t) + v_t \). The trend slopes \( v_t \) are chosen from a simple Markov process (independent of \( \{z_t\}_{t=1}^n \)). With probability \( 1 - p = 0.95 \), we have \( v_{t+1} = v_t \), i.e., no slope change in the underlying trend. With probability \( p \), we choose \( v_{t+1} \) from a uniform distribution on \( [-0.5, 0.5] \). We choose the initial slope \( v_1 \) from a uniform distribution on \( [-0.5, 0.5] \) and set \( f_0(0) = 0 \).

An example of how this methodology performs when compared to the (invalid) approach of using the full dataset twice for both selection and inference is shown in Figure 11 for pointwise confidence intervals. As there is not an obvious way to apply a data splitting approach in the context of trend filtering, this view is excluded from our comparison. Although reusing the full dataset twice is always invalid, we see that it performs worse for datasets that are more volatile, either in terms of the underlying structural trend (i.e. more knots and/or larger slopes) or in terms of the underlying noise level. For a similar demonstration of how uniform confidence intervals perform for a single example run, please see Appendix C.1.
Figure 11. Two instances of the observed points (in yellow) and the pointwise confidence intervals (in blue if correctly cover the trend, in red if not; the time points with false coverage is also amplified in the bar at the bottom) using two types of methods: full data twice (left), and data blurring (right). The underlying projected mean is marked in cyan, which mostly overlaps with the true underlying trend. The true knots are marked by vertical lines. In all cases, we attempt to set FCR with $\alpha = 0.2$. The first instance (top) has smaller noise ($\sigma = 0.05$) and low probability of new knots ($p = 0.05$). The FCP when reusing the full dataset twice is 0.285, which appears to violate FCR control at 0.2 and is higher than the FCP when using the blurred data (with FCP being 0.1). The second instance (bottom) has larger noise ($\sigma = 0.5$) and higher probability of new knots ($p = 0.3$). The FCP using the full data is 0.365, which appears to violated the FCR control at 0.2 and is higher than that using the blurred data (with FCP being 0.225). In general, as the underlying noise and trend become more volatile, reusing the full dataset results in increasingly worse performance at the inference stage.

We repeat this experiment over 500 trial runs, using the same set of metrics as discussed in Section 5.2 to track performance. Figure 12 tracks the performance of pointwise confidence intervals in terms of FCR, while Figure 13 tracks the performance based on the length of the constructed confidence intervals. The increase in confidence interval length for the blurred procedure enables us to have valid FCR control. In cases where both the number of new knots and the amount of noise is relatively small, the confidence interval length when using data blurring is not much larger than when using the full data twice. However, when either of these values increases by even a small amount, the confidence intervals constructed by reusing the full dataset no longer have valid FCR control.

Figure 12. FCR for the pointwise confidence intervals when varying the probability of having knots $p$ in $\{0.01, 0.55, 0.1, 0.145, 0.19\}$ and the noise SD in $\{0.05, 0.1, 0.15, 0.2\}$ (the blue circled cell represents the setting for the first shown instance in Figure 11). The confidence intervals generated using full data twice do not have valid FCR control, especially when $p$ is large (more knots) and the noise SD is small; whereas data blurring always ensures valid FCR error control.
(a) Difference in the averaged CI length when reusing the full dataset, as compared against data blurring.

(b) CI lengths when noise SD is 0.05.

(c) CI lengths when the probability of new knots is 0.2.

Figure 13. The width of pointwise confidence intervals when varying the probability of having knots $p$ in $\{0.01, 0.55, 0.1, 0.145, 0.19\}$ and the noise SD in $\{0.05, 0.1, 0.15, 0.2\}$. The widths for CIs constructed using data blurring range from 0.03 to 0.43, and are larger than the CIs constructed when the full dataset is reused for inference. Both methods increase the confidence interval length as the noise SD increases, but when reusing the full dataset, the length is still not adequate to ensure FCR control. Interestingly, only the data blurring approach increases the length of the confidence interval as the number of true knots increases.

We report a similar set of metrics to track the performance of uniform confidence intervals. However, in addition to the metrics described above, we also track an indicator function for each trial that denotes whether the confidence intervals actually achieve simultaneous type I error control in that instance. The major findings that emerged for pointwise confidence intervals remain when extended to the uniform setting. Results tracking the FCR and simultaneous type I error rates for uniform confidence intervals can be found in Figure 14. Blurring the data guarantees error control for both the FCR and simultaneous type I error rate at confidence level $1 - \alpha$. Using the full data twice tends not to violate the coverage guarantee when noise is relatively high and the number of knots is small but this method quickly deteriorates when the noise is relatively small and the probability of new knots is high. The confidence interval lengths for both methods are also tracked in Figure 15.
Figure 14. Simultaneous type I error and FCR for uniform confidence intervals when varying the probability of having knots $q$ in $\{0.01, 0.05, 0.1, 0.145, 0.19\}$ and the noise SD in $\{0.05, 0.1, 0.15, 0.2\}$. The uniform confidence intervals generated using the full data twice do not have valid type I error control especially when $p$ is large (more knots) and the noise SD is small; whereas confidence intervals constructed using data blurring always have valid type I error control.
5.3 Demonstration on spectroscopy datasets

We conclude this section by demonstrating the utility of these methodologies for conducting inference on a real-world data example in observational astronomy. In Politsch et al. (2020b), the authors introduce trend filtering as a useful tool for astronomical data analysis because of multiple settings within astronomy where a key step in the analysis pipeline reduces to one dimensional data compression. One extended numerical example discussed in the aforementioned paper is the use of trend filtering to aid in spectral classification. Here, astronomers observe a spectrum consisting of wavelengths ($\lambda$) and measurements of the coadded flux ($f(\lambda)$) at each wavelength for an object of interest. Trend filtering is then used to create a “spectral template”—that is, a smoothed line of best fit for the observed spectrum. This template can then be combined with emission-line parameter estimates to classify the astronomical object.

In the original paper, the authors demonstrate that trend filtering empirically performs quite well compared to the existing state-of-the-art for creating spectral templates which revolve around low-dimensional principal component analysis. We use this analysis as a jumping off point to further demonstrate how pointwise and uniform confidence intervals constructed using data blurring appear visually when used on a real data example.

The dataset used for this analysis is the twelfth data release of the Baryon Oscillation Spectroscopic Survey (Alam et al. (2015)). We focus on estimating a smoothed line of best fit using trend filtering, along with pointwise and uniform confidence bands designed to cover the conditional mean $\mu = E[f(\lambda)|\lambda]$ of the observed spectra of the same three objects discussed in Politsch et al. (2020b):

1. A quasar. DR12, Plate = 7130, MJD = 5659, Fiber = 58. Located at (RA, Dec, z) $\approx (349.737^\circ, 33.414^\circ, 2.399)$.

2. A galaxy. DR12, Plate= 7140, MJD = 56569, Fiber=68. Located at (RA, Dec, z) $\approx (349.374^\circ, 33.617^\circ, 0.138)$.

3. A star. DR12, Plate= 4055, MJD = 55359, Fiber=84. Located at (RA, Dec, z) $\approx (236.834^\circ, 0.680^\circ, 0.000)$.

We use same procedure as discussed in Section 5 to construct a trend filter estimate for these objects. In this setting, the flux measurement variances are known a priori as they are constructed from the BOSS spectroscopy pipeline (Bolton et al. (2012)) and account for the statistical uncertainty introduced by photon noise, CCD read noise, and sky-subtraction error. One point of departure from the procedure used in Section 5 is that we also consider quadratic trend filtering in addition to linear trend filtering to construct a fitted curve.
The results are shown in Figure 16. Since the confidence bands displayed are covering the conditional mean and not the observed data, there is no objective “ground truth” to compare the outputs of the model to in order to assess goodness of fit. Therefore, the results need to be judged more holistically. In particular, we note an interesting tradeoff between the degree of the polynomial that is used and the smoothness of the estimated function. Larger degree polynomial result in fewer knots being chosen during the selection step which leads to a smoother-looking confidence band, but at the expense of not capturing some of the more volatile pieces of the data. Smaller degree polynomials results in more knots being chosen during the selection stage which leads to a less smooth band but also a function that tracks the overall volatility of the data more closely.

6 Concluding remarks

We have proposed a novel method for selective inference that can be viewed as a continuous analog to data splitting through randomization. We note numerous potential applications of this procedure but focus on two key applications in this paper: Gaussian linear regression and trend filtering. In the case of Gaussian linear regression, we note slightly tighter confidence intervals compared to traditional data splitting. For trend filtering, data blurring enables uncertainty quantification while affording the user of these procedures flexibility in choosing the number of knots and degree of the polynomial based on heuristic criteria.

Numerous avenues for follow-up work exists. Although we note promising empirical results that suggest this procedure can be generalized to situations where the variance is unknown, additional work needs to be done to establish theoretical guarantees in this circumstance. More broadly, this procedure lends itself to non-Gaussian distributed data such as GLMs as well. Unfortunately, coverage guarantees can only be given in this setting for the relatively uninteresting case where the correct model has been chosen during the selection stage. In the case of an incorrectly chosen model, the existing literature tends to provide guarantees only in a random-X setting. In a data blurring setup, we necessarily require a fixed-X setup since the covariates have already been examined once during model selection. Unfortunately, the literature investigating model misspecification in fixed-X settings is significantly more sparse. Nonetheless, we note promising empirical results for logistic and poisson regressions in Appendices A.1 and A.2. In both these cases, the gains in power are significantly greater than in the Gaussian applications. We hope to explore this finding in greater detail in future lines of work.

Finally, we note that it is possible to repeat the data blurring procedure multiple times in parallel, but aggregating the varied results is nontrivial, so we leave this direction for future investigation.

7 Acknowledgements

We would like to thank Collin Politsch for his extensive commentary on previous drafts and help in the collection of data for Section 5.3.
Figure 16. Fitted values as well as uniform and pointwise confidence intervals for three different spectra—identifiable as a quasar, galaxy, and star. The quasar spectrum is fit using linear trend filtering and the galaxy and stellar spectra are fit using quadratic trend filtering. We note that the higher degree polynomial result in fewer knots being chosen (and thus smoother-looking functions) but tend not to capture as many peaks as the lower-order polynomial methods which select more knots. The top view shows the trend filter over the entire spectra, but the bottom view offers a “zoomed in” view on a smaller subset of the data to aid in visual identification.
References

Alam, S., F. D. Albareti, C. A. Prieto, F. Anders, S. F. Anderson, T. Anderton, B. H. Andrews, E. Armengaud, E. Aubourg, S. Bailey, and et al. (2015). The eleventh and twelfth data releases of the sloan digital sky survey: final data from SDSS-III. The Astrophysical Journal Supplement Series 219(1).

Barber, R. F. and E. J. Candès (2015). Controlling the false discovery rate via knockoffs. The Annals of Statistics 43(5), 2055 – 2085.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological) 57(1), 289–300.

Benjamini, Y. and D. Yekutieli (2005). False discovery rate–adjusted multiple confidence intervals for selected parameters. Journal of the American Statistical Association 100(469), 71–81.

Bolton, A., D. Schlegel, E. Aubourg, S. Bailey, V. Bhardwaj, J. Brownstein, S. Burles, Y.-M. Chen, K. Dawson, D. Eisenstein, J. Gunn, G. Knapp, C. Loomis, R. Lupton, C. Maraston, D. Muna, A. Myers, M. Olmstead, N. Padmanabhan, and W. Wood-Vasey (2012, 07). Spectral classification and redshift measurement for the sdss-iii baryon oscillation spectroscopic survey. The Astronomical Journal 144.

Buja, A., L. Brown, R. Berk, E. George, E. Pitkin, M. Traskin, K. Zhang, and L. Zhao (2019). Models as Approximations I: Consequences Illustrated with Linear Regression. Statistical Science 34(4), 523 – 544.

Buja, A., L. Brown, A. K. Kuchibhotla, R. Berk, E. George, and L. Zhao (2019). Models as approximations II: A model-free theory of parametric regression. Statistical Science 34(4), 545–565.

Chen, Y., S. Jewell, and D. Witten (2021). More powerful selective inference for the graph fused lasso. arXiv preprint arXiv:2109.10451.

Duy, V. N. L. and I. Takeuchi (2021). Parametric programming approach for more powerful and general lasso selective inference. arXiv preprint arXiv:2004.09749.

Fithian, W., D. Sun, and J. Taylor (2014). Optimal inference after model selection. arXiv preprint arXiv:1410.2597.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Volume 1, pp. 221–233. University of California Press.

Hyun, S., M. G’Sell, and R. J. Tibshirani (2018). Exact post-selection inference for the generalized lasso path. Electronic Journal of Statistics 12(1), 1053 – 1097.

Ignatiadis, N., S. Saha, D. L. Sun, and O. Muralidharan (2021). Empirical bayes mean estimation with nonparametric errors via order statistic regression on replicated data. Journal of the American Statistical Association (eprint), 1–13.

Kim, S.-J., K. Koh, S. Boyd, and D. Gorinevsky (2009). $\ell_1$ trend filtering. SIAM Review 51(2), 339–360.

Knowles, M. (1987). Simultaneous confidence bands for random functions. Ph. D. thesis, Stanford University.

Koenker, R. (2011). Additive models for quantile regression: Model selection and confidence bandaids. Brazilian Journal of Probability and Statistics 25(3), 239–262.
Last, G. and M. Penrose (2017). *Lectures on the Poisson Process*, Volume 7. Cambridge University Press.

Lee, J. D., D. L. Sun, Y. Sun, and J. E. Taylor (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics* 44(3), 907 – 927.

Lei, L. and W. Fithian (2018). AdaPT: an interactive procedure for multiple testing with side information. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80(4), 649–679.

Lei, L., A. Ramdas, and W. Fithian (2020). A general interactive framework for false discovery rate control under structural constraints. *Biometrika* 108(2), 253–267.

Li, X. and W. Fithian (2021). Whiteout: when do fixed-X knockoffs fail? *arXiv preprint arXiv:2107.06388*.

Mammen, E. and S. van de Geer (1997). Locally adaptive regression splines. *The Annals of Statistics* 25(1), 387 – 413.

Oliveira, N. L., J. Lei, and R. J. Tibshirani (2021). Unbiased risk estimation in the normal means problem via coupled bootstrap techniques. *arXiv preprint arXiv:2109.10451*.

Politsch, C. A., J. Cisewski-Kehe, R. A. C. Croft, and L. Wasserman (2020a). Trend filtering – I. A modern statistical tool for time-domain astronomy and astronomical spectroscopy. *Monthly Notices of the Royal Astronomical Society* 492(3), 4005–4018.

Politsch, C. A., J. Cisewski-Kehe, R. A. C. Croft, and L. Wasserman (2020b). Trend filtering – II. Denoising astronomical signals with varying degrees of smoothness. *Monthly Notices of the Royal Astronomical Society* 492(3), 4019–4032.

Ramdas, A. and R. J. Tibshirani (2014). Fast and flexible ADMM algorithms for trend filtering. *Journal of Computational and Graphical Statistics* 23, 839 – 858.

Rasines, D. G. and G. A. Young (2021). Splitting strategies for post-selection inference. *arXiv preprint arXiv:2102.02159*.

Rinaldo, A., L. Wasserman, and M. G’Sell (2019). Bootstrapping and sample splitting for high-dimensional, assumption-lean inference. *The Annals of Statistics* 47(6), 3438 – 3469.

Sarkar, S. K. and C. Y. Tang (2021). Adjusting the Benjamini-Hochberg method for controlling the false discovery rate in knockoff assisted variable selection. *arXiv preprint arXiv:2102.09080*.

Steidl, G., S. Didas, and J. Neumann (2006). Splines in higher order TV regularization. *Journal of Computational and Graphical Statistics* 15, 241 – 255.

Tian, X. and J. Taylor (2017). Asymptotics of selective inference. *Scandinavian Journal of Statistics* 44(2), 480–499.

Tian, X. and J. Taylor (2018). Selective inference with a randomized response. *The Annals of Statistics* 46(2), 679–710.

Tibshirani, R., M. Saunders, S. Rosset, J. Zhu, and K. Knight (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67(1), 91–108.

Tibshirani, R. J. (2014). Adaptive piecewise polynomial estimation via trend filtering. *The Annals of Statistics* 42(1), 285 – 323.

Tibshirani, R. J., J. Taylor, R. Lockhart, and R. Tibshirani (2016). Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association* 111(514), 600–620.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* 50(1), 1–25.

Xing, X., Z. Zhao, and J. S. Liu (2021). Controlling false discovery rate using gaussian mirrors. *Journal of the American Statistical Association* 0(0), 1–20.
A Applications to selective confidence intervals in generalized linear models

We also investigate the applicability of data blurring to statistical problems outside of Gaussian distributed data here. Although the empirical results in this section seem to indicate larger gains in performance from data blurring than we observe in the Gaussian examples, we are not yet able to provide theoretical guarantees for coverage of fitted parameters outside of the relatively uninteresting case where the correct model has been chosen during the selection stage. In the case of model misspecification or omitted covariates, the majority of the literature, such as Buja et al. (2019) and Huber (1967) focuses on the random-\(X\) setting which is inapplicable in this case. Nonetheless, we include these results here so the reader may understand heuristically how data blurring procedures behave in a non-Gaussian setting.

A.1 Poisson regression

Setup. Let \(Y_i\) be the dependent variable and \(X_i \in \mathbb{R}^p\) be a vector of \(p\) features. Suppose we have \(n = 1000\) samples. We have collected \(p = 100\) features \(X_i \in \{0, 1\}^2 \times \mathbb{R}^{98}\), where the first two follow \(\text{Ber}(1/2)\) and the rest follow independent Gaussians. Suppose \(Y_i\) follow a Poisson distribution with the expected value \(\exp\{\gamma^T X_i\}\), where the parameter \(\gamma\) is nonzero for 29 features: \((\gamma(1), \gamma(3), \ldots, \gamma(22), \gamma(93), \ldots, \gamma(100)) = S_\Delta \cdot (1, \ldots, 1_2, \ldots, 2_8)\) and \(S_\Delta\) encodes the signal strength.

Proposed procedure. Following Section 2.2, we can use data blurring for constructing selective confidence intervals as follows.

1. Decompose each \(Y_i\) as \(f(Y_i) \sim \text{Bin}(Y_i, 0.5)\), and \(g(Y_i) = Y_i - f(Y_i)\).

2. Fit \(f(Y_i)\) by the GLM with log-link and Lasso regularization to select features, denoted as \(M \in [p]\). (in our examples, we use \text{cv.glmnet} in the \text{R} package \text{glmnet} and choose the tuning parameter \(\lambda\) by the 1 standard deviation rule, which can be found in the value of \text{lambda.1se}).

3. Fit \(g(Y_i)\) by another GLM without regularization with log-link using only the selected features (we use \text{glm} in the \text{R} package \text{stats}).

4. Construct confidence intervals for the coefficients trained in the third step, each at level \(\alpha\) and with the standard errors estimated using the sandwich estimator (using \text{conf.int} in \text{R} package \text{clubSandwich} with \text{vcov = "CR2"}).

Note that with the above decomposition, \(f(Y_i)\) and \(g(Y_i)\) are independent Poisson with mean \(\mu_i/2\) where \(\mu_i = \mathbb{E}(Y_i \mid X_i)\).

Simulation results with mutually independent covariates. The resulting confidence intervals should have FCR control at level \(\alpha\). As before, we compare the confidence intervals constructed through data blurring with the confidence intervals constructed via data splitting (where 50\% of the observations are allocated to selection and 50\% for inference) as well as the (invalid) approach where the original dependent variable is used twice to both select features and construct intervals (replacing \(f(Y_i)\) and \(g(Y_i)\) both by \(Y_i\) in the above algorithm).

As an illustration of the procedure, Figure 17 shows an instance of selected features and the constructed confidence intervals for a single trial run. The selected features are marked by blue crosses, which include all of the nonzero coefficients (corresponding to almost 100\% power for selection) and also many zero coefficients (corresponding to around 50\% precision for selection). The constructed confidence intervals that do not contain the true value are marked in red.

\footnote{This sandwich estimator slightly modifies the original one proposed by Huber (1967) and White (1982) to include a small sample correction.}
Figure 18 shows results average over 500 trial runs. Compared with the results using data splitting, the confidence interval using data blurring is significantly tighter. This is expected since data splitting uses $n/2$ samples following $\text{Poi}(\mu(X_i))$ whereas data blurring uses $n$ samples following $\text{Poi}(\mu(X_i)/2)$, which has smaller variance. Although the precision during the selection step is not high for both methods (around 50% of selected features do not have true signals), we are able to identify the true signals by constructing confidence intervals in the second step with FCR control. Note that the confidence interval width decreases with the signal strength because the variance of the coefficient estimates also decreases their expected value in Poisson regression.

Figure 17. An instance of the selected feature (blue crosses) and the constructed confidence intervals using blurred data (left), full data twice (middle), and split data (right). Confidence intervals which do not cover the parameters correctly are marked in red.

Figure 18. FCR, length of the confidence intervals, FPR and power for the sign of parameters, and power and precision for the selected features, when varying the signal strength $S_\Delta$ in $\{0, 0.125, 0.25, 0.375, 0.5\}$. The results are averaged over 500 repetitions. The confidence intervals by using the original data twice do not have FCR control guarantee due to selection bias. In contrast, our proposed procedure using blurring has valid FCR, without inflating the length of confidence intervals much or reducing the power of selecting non-zero features.

Simulation results with dependent covariates. The above procedure will still be meaningful even if the expected value of $Y$ given the selected features is not a linear combination of the selected
features. Denote the vector of selected features for sample \( i \) as \( X_i(M) \). The constructed confidence intervals should cover the “projected” parameters \( \beta_* \) that minimize the KL divergence between the true and the modeled distribution:

\[
\beta_* := \arg\min_{\beta} D_{KL} \left( \prod_{i=1}^{n} f(\exp\{\gamma^T X_i\}) \bigg\| \prod_{i=1}^{n} f(\exp\{\beta^T X_i(M)\}) \right), \tag{24}
\]

where \( f(\mu_i) \) is the distribution of Poisson distribution with expected value as \( \mu_i \). The projected \( \beta_* \) is equivalent to the solution of maximizing the expected log-likelihood:

\[
\beta_* \equiv \arg\max_{\beta} \sum_{i=1}^{n} \mathbb{E} \left[ \beta^T X_i(M)Y_i - \log(Y_i!) - \exp\{\beta^T X_i(M)\} \right], \tag{25}
\]

and the solution satisfies:

\[
\sum_{i=1}^{n} X_{ik}(M) \exp\{\gamma^T X_i\} = \sum_{i=1}^{n} X_{ik}(M) \exp\{\beta_*^T X_i(M)\}, \tag{26}
\]

for all selected features \( k \in M \).

Our simulation considers dependent covariates \( X \), generated from a multivariate Gaussian with zero mean. The covariance matrix is a five-block diagonal matrix, each block a \( 20 \times 20 \) toeplitz matrix:

\[
\begin{bmatrix}
1 & \rho & \cdots & \rho^{d-2} & \rho^{d-1} \\
\rho & 1 & \rho & \cdots & \rho^{d-2} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\rho^{d-2} & \cdots & \rho & 1 & \rho \\
\rho^{d-1} & \rho^{d-2} & \cdots & \rho & 1 \\
\end{bmatrix}, \tag{27}
\]

where \( d = 20 \).

Results averaged over 500 trial runs are shown in Figure 19 for several different possible values of \( \rho \). We note that the performance of the three methodologies are relatively consistent regardless of the level of correlation between the features in the dataset.
Figure 19. FCR, length of the confidence intervals, FSR/power for the sign of parameters, and power and precision for the selected features, when varying the correlation parameter $\rho$ in \{-0.5, -0.25, 0, 0.25, 0.5\} (with $\rho = 0$ corresponding to mutually independent covariates). The performance of the three methods is relatively similar under different degrees of dependence.

A.2 Logistic regression

Setup. We consider a similar application in logistic regression. Let $Y_i$ be the dependent variable and $X_i \in \mathbb{R}^p$ be a vector of $p$ features. Suppose we have $n = 1000$ samples. We have collected $p = 100$ features $X_i \in \{0,1\}^2 \times \mathbb{R}^{98}$, where the first two follow Ber(1/2) and the rest follow independent Gaussians. Suppose the conditional distribution of $Y_i$ given $X_i$ is a Bernoulli distribution with expected value $(1 + \exp{-\gamma^T X_i})^{-1}$, where the parameter $\beta$ is nonzero for 30 features: $(\gamma(1), \gamma(3), \ldots, \gamma(22), \gamma(92), \ldots, \gamma(100)) = S_\Delta \cdot (1, \ldots, 1, 2, \ldots, 2)$ and $S_\Delta$ encodes the signal strength.

Proposed procedure. Following Section 2.2, we can use data blurring for constructing selective confidence intervals as follows.

1. Draw $Z_i \sim \text{Ber}(p)$ where the “flip probability” $p \in (0, 1)$ is a tuning parameter; and let $f(Y_i) = Y_i(1 - Z_i) + (1 - Y_i)Z_i$, and $g(Y_i) = Y_i$.

2. Fit $f(Y_i)$ with a GLM with a logit link function and lasso regularization to select features, denoted as $M \in [p]$ (in our examples, we use cv.glmnet in the R package glmnet and choose the tuning parameter $\lambda$ by the 1 standard deviation rule, which can be found in the value of lambda.1se).

3. Fit $g(Y_i)$ with another GLM with a logit link function and no regularization using only the selected features (we use glm in the R package stats).

4. Construct confidence intervals for the coefficients trained in the third step, each at level $\alpha$ and with the standard error estimated by the sandwich estimator (using conf.int in R package clubSandwich with vcov = "CR2").
With the above decomposition, \( f(Y_i) \) has distribution \( \text{Ber}(\mu_i + p - 2p\mu_i) \) where \( \mu_i = \mathbb{E}(Y_i \mid X_i) \); and \( g(Y_i) \) is distributed as \( \text{Ber}\left(\frac{\mu_i}{\mu_i + (1 - \mu_i)[p/(1-p)]^2 \tau(R_i, t)}\right) \).

Note that fitting a logistic model for \( g(Y_i) \) is not the same as fitting a logistic model for \( Y_i \); however, we can still interpret the fitted parameters as projection onto the working model. In the fixed \( X \) setting, the optimal solution \( \beta_* \) satisfies:

\[
\sum_{i=1}^{n} X_{ik}(M)\mathbb{E}[g(Y_i) \mid X_i, f(Y_i)] = \sum_{i=1}^{n} X_{ik}(M) \left(1 + \exp\left\{\beta_*^T X_i(M)\right\}\right)^{-1},
\]

for all selected features \( k \in M \). Although this is a similar argument at the one discussed in Section A.1, we note that we must additionally condition on \( f(Y_i) \) in this instance. This procedure is illustrated for a single trial run in Figure 20. As usual, we compare the confidence intervals constructed using data blurring with data splitting and the (invalid) approach of reusing the full dataset for both selection and inference. Note that in this case, the confidence intervals are now are covering the projected parameters \( \beta_* \) as opposed to the true underlying parameters \( \beta \).

![Figure 20](image)

**Figure 20.** An instance of confidence intervals for an example set of selected features using data blurring, data splitting, and the (invalid) approach of using using the full dataset twice for both selection and inference. The upper left-hand graph shows the true coefficients, but the confidence interval cover the projected coefficients as described above. However, the projected coefficients are quite close to the original coefficient values in this instance (as well as most instances encountered in simulation). We note that both data splitting and data blurring have the valid FCR control (set at target level = 0.2), but the confidence intervals constructed using data blurring are much tighter.
Simulations varying signal strength. We repeat this experiment over 500 trial runs while varying signal strength from 0 to 0.5. We note that across all observed signal strengths, data blurring offers tighter confidence intervals when compared to both data splitting and the (invalid) approach that reuses the full dataset for both selection and inference. This is a remarkable feat since it is typically the case (e.g. in the Gaussian and Poisson applications of data blurring) that larger confidence intervals are the price that is required in order to ensure coverage guarantees for data blurring.

![Graphs showing FCR, CI width, FSR, power for the sign of parameters, and power and precision for the selected features](image)

Figure 21. FCR, CI width, FSR, power for the sign of parameters, and power and precision for the selected features, when varying the signal strength $S_{\Delta}$ in {0, 0.1, 0.2, 0.3, 0.4, 0.5}. The hyperparameter $p$ is chosen as 0.2 and target FCR for confidence intervals is set at 0.2. We note that data blurring offers smaller confidence interval widths when compared to both data splitting and the (invalid) approach that reuses the full dataset for both selection and inference.

Simulations varying blurring hyperparameter $p$. We also examine the performance of these experiments as we vary the blurring hyperparameter $p$. We note that values of $p$ close to either 0 or 1 correspond to having more information reserved for the selection step, while setting $p = 0.5$ maximizes the amount of information reserved for inference. In Figure 22, we can see this behavior manifested in the “dip” in the power and precision of the selection event as $p$ ranges from 0.4 to 0.6.

![Graphs showing CI width, power and precision for flip probability](image)

39
Furthermore, we can form an (approximate) distribution. The coverage statement for the confidence interval then follows from standard properties of a Gaussian

\[ \text{Theorem 3. Let } \hat{\beta}(M) \text{ be defined as in (10) and } \beta_s(M) \text{ be defined as in (11). If } f(Y) \text{ and } g(Y) \text{ are defined as in (29) and } \Sigma = \sigma^2 I_n, \text{ then the following holds:} \]

\[ \hat{\beta}(M) \xrightarrow{p} N(\beta_s(M), (1 + \tau^{-2})\sigma^2 (X_M^T X_M)^{-1}). \]

Furthermore, we can form an (approximate) $1 - \alpha$ confidence interval for the kth element of $\beta_s(M)$ as

\[ \hat{\beta}_k(M) \pm \hat{\sigma} z_{\alpha/2} \left[ \sqrt{(1 + \tau^{-2})(X_M^T X_M)^{-1}} \right]_{kk}. \]

**Proof.** By assumption, we know $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$. By Slutsky’s theorem, $\dfrac{Y}{\hat{\tau}} \xrightarrow{d} N(\frac{\tau}{\sigma}, I_n)$. Draw an independent $Z \sim N(0, I_n)$ and define $f_0(Y) = \frac{Y}{\hat{\tau}} + \tau Z$ and $g_0(Y) = \frac{Y}{\hat{\tau}} - \frac{1}{\hat{\tau}} Z$. Applying the continuous mapping theorem to the joint density of $(\frac{Y}{\hat{\tau}}, Z)$ implies that

\[ (f_0(Y), g_0(Y))^T \xrightarrow{d} N \left( \mu, \begin{pmatrix} \frac{Y}{\hat{\tau}} & 0 \\ 0 & \frac{1}{\hat{\tau}} \end{pmatrix} \right). \]

Applying Slutsky’s theorem once more and noting that $f(Y) = \hat{\sigma} f_0(Y)$ and $g(Y) = \hat{\sigma} g_0(Y)$ gives us that

\[ (f(Y), g(Y))^T \xrightarrow{d} N \left( \mu, \begin{pmatrix} \sigma^2 & 0 \\
0 & \sigma^2 \end{pmatrix} \right). \]

Applying the continuous mapping theorem once more to the definition of $\hat{\beta}(M)$ and repeating the arguments in Section 2 gives us that

\[ \hat{\beta}(M) \xrightarrow{d} N(\beta_s(M), (1 + \tau^{-2})\sigma^2 (X_M^T X_M)^{-1}). \]

The coverage statement for the confidence interval then follows from standard properties of a Gaussian distribution.

---

**B Supplemental materials for linear regression**

**B.1 Extension to unknown variance**

In the case where $p$ is fixed, $n \to \infty$, and $\Sigma = \sigma^2 I_n$ we can extend the argument in Section 4 to accommodate unknown variance. In this setting, if we take $\hat{\sigma}^2$ to be the variance estimated from the residuals by fitting the full model with all $p$ covariates, it is a consistent estimate of $\sigma^2$. See section 5.3 of Tian and Taylor (2017) for further details.

We can then modify the definition of $f(Y)$ and $g(Y)$ as follows:

\[ f(Y) = Y + \hat{\sigma} \tau Z, \quad \text{and} \quad g(Y) = Y - \frac{\hat{\sigma}}{\tau} Z. \]  

where $Z \sim N(0, I_n)$. We can then modify 2 in a natural way to get an asymptotic coverage guarantee:

**Theorem 3.** Let $\hat{\beta}(M)$ be defined as in (10) and $\beta_s(M)$ be defined as in (11). If $f(Y)$ and $g(Y)$ are defined as in (29) and $\Sigma = \sigma^2 I_n$, then the following holds:

\[ \hat{\beta}(M) \xrightarrow{d} N(\beta_s(M), (1 + \tau^{-2})\sigma^2 (X_M^T X_M)^{-1}). \]

Furthermore, we can form an (approximate) $1 - \alpha$ confidence interval for the kth element of $\beta_s(M)$ as

\[ \hat{\beta}_k(M) \pm \hat{\sigma} z_{\alpha/2} \left[ \sqrt{(1 + \tau^{-2})(X_M^T X_M)^{-1}} \right]_{kk}. \]

**Proof.** By assumption, we know $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$. By Slutsky’s theorem, $\dfrac{Y}{\hat{\tau}} \xrightarrow{d} N(\frac{\tau}{\sigma}, I_n)$. Draw an independent $Z \sim N(0, I_n)$ and define $f_0(Y) = \frac{Y}{\hat{\tau}} + \tau Z$ and $g_0(Y) = \frac{Y}{\hat{\tau}} - \frac{1}{\hat{\tau}} Z$. Applying the continuous mapping theorem to the joint density of $(\frac{Y}{\hat{\tau}}, Z)$ implies that

\[ (f_0(Y), g_0(Y))^T \xrightarrow{d} N \left( \mu, \begin{pmatrix} \frac{Y}{\hat{\tau}} & 0 \\
0 & \frac{1}{\hat{\tau}} \end{pmatrix} \right). \]

Applying Slutsky’s theorem once more and noting that $f(Y) = \hat{\sigma} f_0(Y)$ and $g(Y) = \hat{\sigma} g_0(Y)$ gives us that

\[ (f(Y), g(Y))^T \xrightarrow{d} N \left( \mu, \begin{pmatrix} \sigma^2 & 0 \\
0 & \sigma^2 \end{pmatrix} \right). \]

Applying the continuous mapping theorem once more to the definition of $\hat{\beta}(M)$ and repeating the arguments in Section 2 gives us that

\[ \hat{\beta}(M) \xrightarrow{d} N(\beta_s(M), (1 + \tau^{-2})\sigma^2 (X_M^T X_M)^{-1}). \]

The coverage statement for the confidence interval then follows from standard properties of a Gaussian distribution. \(\square\)
B.2 Additional simulation studies

The empirical results discussed in 4.2 demonstrate the advantage of data blurring in fixed-design settings with small sample sizes and a handful of points with high leverage. In cases where the sample size is larger and the distribution of covariates are less irregular, data splitting and data blurring have roughly similar empirical properties. The following additional simulations help to elucidate this.

Simulation with independent covariates
We repeat the simulation discussed in 4.2 but now have $p = 100$ features $X_i \in \mathbb{R}^{100}$. The vector of covariates $X_i$ follows independent standard Gaussians; and $Y_i$ follow a Gaussian distribution with the expected value as $\gamma^T X_i$, where the parameter $\gamma$ is nonzero for 30 features: $(\gamma(1), \gamma(3), \ldots, \gamma(22), \gamma(92), \ldots, \gamma(100)) = S_\Delta \cdot (1, \ldots, 1, -1, \ldots, -1)$ and $S_\Delta$ encodes the signal strength. Figure 23 shows the result of these simulations averaged over 500 repetitions. Data splitting and data blurring appear to have roughly comparable performance in this case, as the covariates are now all drawn from the same distribution.

Simulation with dependent covariates
We repeat the simulation discussed in 4.2 but now have dependent covariates $X_i$. We let $X_i$ be generated from a multivariate Gaussian with zero mean. The covariance matrix is a five-block diagonal matrix, each block a $20 \times 20$ Toeplitz matrix:

$$
\begin{bmatrix}
1 & \rho & \cdots & \rho^{d-2} & \rho^{d-1} \\
\rho & 1 & \rho & \cdots & \rho^{d-2} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\rho^{d-2} & \cdots & \rho & 1 & \rho \\
\rho^{d-1} & \rho^{d-2} & \cdots & \rho & 1
\end{bmatrix},
$$

(30)

Figure 23. FCR, length of the confidence intervals, FPR and power for the sign of parameters, and power and precision for the selected features, when varying the signal strength $S_\Delta$ in \{0, 0.05, 0.1, 0.15, 0.2\}. The results are averaged over 500 repetitions. The confidence intervals by using the original data twice do not have FCR control guarantee due to selection bias. In contrast, our proposed procedure using the masking idea has valid FCR, without inflating the length of confidence intervals much or reducing the power of selecting non-zero features.

Simulation with dependent covariates
We repeat the experiment described above but with dependent covariates $X$. We let $X$ be generated from a multivariate Gaussian with zero mean. The covariance matrix is a five-block diagonal matrix, each block a $20 \times 20$ Toeplitz matrix:
where $d = 20$. Results are shown in Figure 24. We note that in this setting the relative ordering of the three methods is unchanged, but the power of all three methods decreases sharply when there is negative dependence among the covariates.

Figure 24. FCR, length of the confidence intervals, FPR and power for the sign of parameters, and power and precision for the selected features, when varying the correlation parameter $\rho$ in $\{-0.5, -0.25, 0, 0.25, 0.5\}$ (with $\rho = 0$ being independent covariates). It appears that negative correlation leads to lower power for all methods.

C Supplemental materials for trend filtering

C.1 Additional results for uniform confidence intervals

Understanding the relationship of $c(\alpha)$ to changing confidence interval lengths One obstacle in understanding how the widths of uniform confidence bands change as noise increases and the underlying structural trend changes is that the length is controlled by the multiplier $c(\alpha)$ defined in Fact 1 which is somewhat opaque. To aid in forming an intuition behind the empirical trends noted in Figure 15, we plot some intermediate statistics in Figure 25.

Figure 25. The CI width using data blurring method increases with noise variance and the probability of new knots (first), because SE increases (second) compared to the change in the multiplier $c(\alpha, m)$ (third). Both changes can be traced back to change in the number of knots (fourth): SE decreases and $c(\alpha, m)$ increases with the number of knots.
The above plots are medians over repetitions since the distribution of CI width is skewed to large values; thus, the mean may not summarize the pattern clearly. The mean of CI width does not have a consistent trend because there are few extremely large CI widths, due to few extremely large multipliers $c(\alpha, m)$. Upon closer look, such large value exist when the number of knots is large. In Figure 26, we show the value of left-hand side of (20) for solving $c(\alpha, m)$ when the number of knots is 197 (vs 200 time points). The solution for $c(\alpha, m)$ is over 400 (at the intersection of the black path and the red level of $\alpha$).

(a) Trajectory of the left-hand side of (20) when there are 197 knots.  
(b) Log of the multiplier when the number of knots increases from 190 to 197.

Figure 26: The multiplier $c(\alpha)$ increases dramatically when the number of knots is larger than 190.

Example trial runs for uniform confidence bands  
Similar to the examples shown in Figure 11, we visualize the coverage of uniform confidence bands as well over four examples in Figure 27. The confidence interval using the full data fails to cover the projected mean at many time points because of double dipping, whereas data blurring leads to a uniformly-valid confidence interval with type I error control.
Figure 27. Four instances of the observed points (in yellow) and the uniform confidence interval (in blue if correctly cover the trend, in red if not) using two types of methods: full data twice (left), and data blurring (right). The underlying projected mean is marked in cyan, which also mostly overlaps with the original true trend because both methods can find turning points easily under small noise.

C.2 Estimate variance before data blurring

The procedure discussed in Section 5 uses the knowledge of variance $\sigma^2$, which is usually not known in practice. Alternatively, we can estimate the variance as $\hat{\sigma}^2 = \frac{1}{2(n-1)} \sum_{i=1}^{n-1} (Y_{i+1} - Y_i)^2$ before step 1,
and simulate $Z_t \sim N(0, \sigma^2)$. We must also modify equation (20) in step 4 slightly by replacing it with equation (5.6) in Koenker (2011) and instead choose $c(\alpha)$ to be the solution of

$$\frac{|\gamma|}{2\pi}(1 + c^2/v)^{-v/2} + P(t_v > c) = \alpha/2$$

(31)

where $t_v$ denotes a $t$-distribution with $v = n - m - 1$ degrees of freedom with $m$ being the number of knots.

Notice that because we use the observed data to estimate variance, it could break the independence between $Y_t + Z_t$ and $Y_t - Z_t$. However, we notice in simulation that the error control still seems to hold in most cases. First, we evaluate how well the variance can be estimated across different settings in Figure 28. The proposed methodology tends to overestimate the noise SD in general, with this overestimation increasing as the probability of new knots increases or the slope increases, because the formula to compute $\hat{\sigma}^2$ given above treats all the change in adjacent time points as noise.

Figure 28. The noise SD is often over estimated with a range in (0.05, 0.3) and the true one varies in (0.05, 0.2). The over-estimation fades when the noise SD increases and the slope decreases (and slightly fades as the probability of new knots decreases).

Luckily, a tendency to overestimate the variance leads to confidence intervals with conservatively chosen widths, meaning errors are still controlled at the appropriate level in most cases. We see in Figure 29 that data blurring still offers simultaneous type I error control empirically when using uniform confidence bands. This is achieved at the expense of having overly conservative confidence intervals in cases where the underlying structural trend is variable — either because of many knots points or because of the size of the slope. We investigate how the level of conservatism, as measured by the average widths of the constructed confidence bands, increases as the underlying trend varies more in Figure 30).
(a) data blurring method when varying prob of new knots. (b) data blurring method when varying slope range. (c) Full data twice when varying prob of new knots. (d) Full data twice when varying slope range.

Figure 29. Simultaneous type I error for the uniform CI constructed using data blurring method, and full data twice, when varying the probability of new knots, the slope range, and the true noise variance. Data blurring method seems to have lower simultaneous type I error than the target level (0.2) in most cases, except cases circled in red (with a max simultaneous type I error of 0.26).

Figure 30. The CI width for uniform confidence bands using either data blurring (right) or full data twice (left) have similar trends with respect to noise SD and probability of new knots. The difference in confidence interval length between the two methods decreases when the noise SD increases (such that the effect of double dipping is smaller) or the probability of new knots gets smaller (such that the CI from data blurring is tighter).

C.3 Alternative method for selecting knots

In Section 5.2, we selected the knots by choosing the regularization parameter with the smallest cross-validation error, which often leads to selecting more knots than the underlying truth when the full dataset is used for both selection and inference. Although data blurring allows for an analyst to guarantee error control under arbitrary selection rules, some selection rules tend to be more robust to the analyst reusing the data in terms of their empirical performance even if statistical guarantees are not available.

C.3.1 Stein’s Unbiased Risk Estimate (SURE)

An alternative methodology for selecting knots would be to minimize the SURE formula — which provides an unbiased estimate of mean-squared percentage error in a fixed-design setting. In the context of trend filtering, this can be computed as

$$\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + 2\sigma^2 \frac{df(\hat{y})}{n}$$
where \( \text{df}(\hat{y}) \) is the number of knots chosen to fit \( \hat{y} \). Interestingly, using this formula seems to result in type I error being controlled even when reusing the full data for both selection and inference. Please see Politsch et al. (2020a) for further background on how SURE can be used to aid in knot selection.

**Figure 31.** Simultaneous type I error for uniform confidence intervals as we vary the probability of having new knots \( q \in \{0.01, 0.55, 0.1, 0.145, 0.19\} \) and the noise SD in \( \{0.05, 0.1, 0.15, 0.2\} \). The error control violation when using the full data twice is no longer as stark in the simulation results when using SURE as the selection rule, with the highest simultaneous type I error at 0.2 given target level \( \alpha = 0.2 \).

### C.3.2 1-SD Rule

To mitigate the possibility of over selection of knots, an alternative approach is to choose the regularization parameter to be the the one with smallest error plus one standard deviation. As we observe in Figure 32, adding a standard deviation to the selection rule does not change the error much. However, this rule appears to be more robust when the full data is reused for inference—in Figure 33, we can see that simultaneous type I error control is not as seriously violated when using this selection rule.

**Figure 32.** In the path of regularization parameter, adding an extra standard deviation does not change the error much.
Simultaneous type I error using full data twice.

Simultaneous type I error using data blurring.

Figure 33. Simultaneous type I error for uniform confidence intervals as we vary the probability of having new knots $q \in \{0.01, 0.55, 0.1, 0.145, 0.19\}$ and the noise SD in $\{0.05, 0.1, 0.15, 0.2\}$. The error control violation when using the full data twice is no longer as stark in the simulation results when using this new selection rule, with the highest simultaneous type I error at 0.3 given target level $\alpha = 0.2$. 

48