

# Data Blurring Fission

## Splitting a Single Data Point

James Leiner

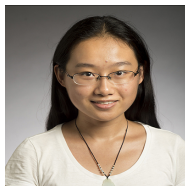
September 6, 2022

Joint work with Boyan Duan, Larry Wasserman, and Aaditya Ramdas

# Authors



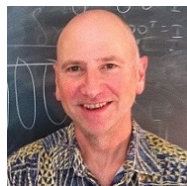
James Leiner



Boyan Duan



Aaditya Ramdas



Larry Wasserman

# Table of Contents

- 1 Introduction to data fission and related approaches
- 2 Application 1: Fixed-design linear regression
- 3 Application 2: Trend filtering
- 4 Application 3: Confidence intervals after interactive multiple testing
- 5 Conclusion

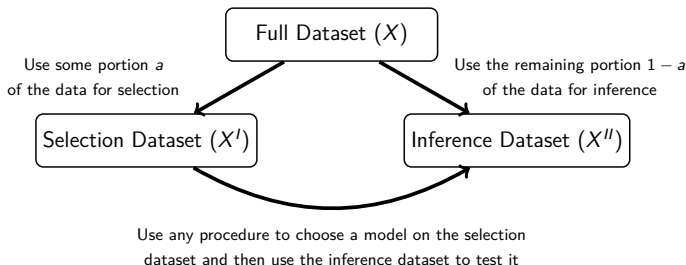
## Section 1

# Introduction to data fission and related approaches

- Many statistical procedures require the analyst to select a model or set of hypotheses to test before conducting inference.
- If the same set of data is used to select a model and conduct inference, statistical guarantees are not usually valid.
- Most common solution to this is sample splitting, but recent work has expanded the methodologies available (e.g. conditional selective inference, p-value masking).
- Data fission, a solution that involves external randomization, takes inspiration from all of the above approaches. It is uniquely suited for settings in which sample splitting does not make conceptual sense (e.g. time series data, graph data).

# Data splitting

- Choose some  $a \in [0, 1]$  to control the split of the data and randomly assign observation to selection and inference datasets.



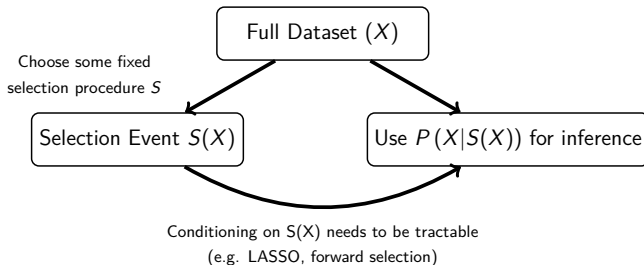
## Advantages

- Flexible to arbitrary choice of model selection procedure.
- No assumptions required on underlying data beyond iid.

## Disadvantages

- “Throws away” information that might be usable for inference.

# Data carving



## Advantages

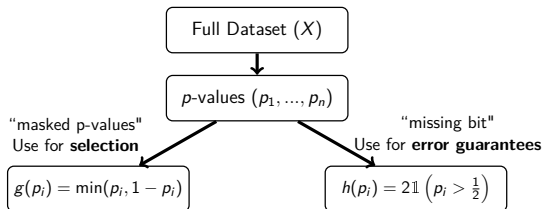
- More powerful than data splitting because all “leftover” information from selection is available for inference.

## Disadvantages

- Requires distributional assumptions on the underlying data.
- Requires pre-commitment to a specific selection rule.
- Post-selective distribution must be tractable in closed form or through numerical simulations.

# P-value masking

- AdaPT [Lei and Fithian (2018)] introduces the idea of masking a p-value to allow an analyst to interactively form a rejection set, while controlling FDR.
- Subsequent work generalizes to handle structural constraints [Lei, Ramdas, and Fithian (2020)], FWER control, global null testing and causal inference [Duan et al ('20, '21, '22)].



## Advantages

- Allows the analyst to interactively form a rejection set with the masked p-values in combination with generic side information.

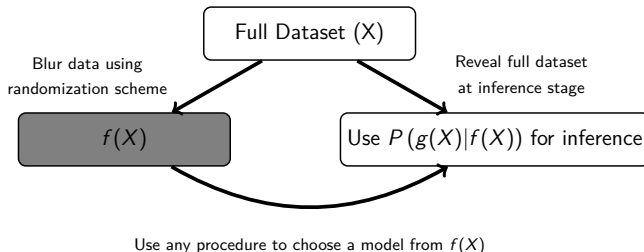
## Disadvantages

- Only usable for hypothesis rejection, not generic inferential tasks (e.g. forming confidence intervals).



# Data fission

Data fission can (loosely) be thought of as a compromise between all three of the preceding methods.



## Advantages

- Flexible to arbitrary choice of model selection procedure
- Allows for a piece of every data point to inform both selection and inference

## Disadvantages

- Requires distributional assumptions on the underlying data (but does not require the data to be distributed in any *particular way*).

# Long road to today's talk

- **Apr. 2018:** Tian and Taylor use external randomization for Gaussian conditional selective inference.
- **Nov. 2019:** Data fission research begins
- **Feb. 2021:** Rasines and Young apply Gaussian external randomization to non-Gaussian data with asymptotic error guarantees.
- **Dec. 2021:** Data fission preprint posted; initially unaware of [Rasines and Young (2021)].
- **Today:** Data fission still not submitted; being updated with new theory, methods, and applications.

# Procedure to accomplish data fission (stronger version)

- 1 Form two new random variable  $f(X)$  and  $g(X)$  from  $X$  such that:
  - 1  $f(X) \perp g(X)$
  - 2 There exists a function,  $h$ , such that  $h(f(X), g(X)) = X$ .
- 2 Use  $f(X)$  to inform model and/or hypothesis selection.
- 3 Use  $g(X)$  to conduct inference.

# Procedure to accomplish data blurring (weaker version)

Moving beyond independence, we can weaken the constraints on picking  $g(X)$  and  $f(X)$  to allow for more flexibility in randomization procedures.

- ① Form two new random variables  $f(X)$  and  $g(X)$  from  $X$  such that:
  - ①  $g(X)|f(X)$  is tractable to compute.
  - ② There exists a function,  $h$ , such that  $h(f(X), g(X)) = X$ .
- ② Use  $f(X)$  to inform model and/or hypothesis selection.
- ③ Use  $g(X)|f(X)$  to conduct inference.

# Trading off information between $f$ and $g$

- We will ideally want to construct  $f$  and  $g$  in such a way that we can smoothly trade off information between the two variables.
- Assuming that the distribution of  $X$  is known up to some unknown parameter of interest  $\theta$ , we can note how the Fisher information trades off between  $f$  and  $g$  through the following decomposition rules.

## Fisher information decomposition rule (strong version)

$$\mathcal{I}_X(\theta) = \mathcal{I}_{f(X)}(\theta) + \mathcal{I}_{g(X)}(\theta)$$

## Fisher information decomposition rule (weak version)

$$\mathcal{I}_X(\theta) = \mathcal{I}_{f(X)}(\theta) + E \left[ \mathcal{I}_{g(X)|f(X)}(\theta) \right]$$

# Example 1: Normally Distributed Data

(strong version)

- Similar construction as in [Tian and Taylor (2018)] and [Rasines and Young (2021)].

$$\begin{array}{ccc} & X \sim N(\theta, \Sigma) & \\ \swarrow & & \searrow \\ f(X) := X + \tau Z & \perp & g(X) := X - \frac{1}{\tau} Z \end{array}$$

where  $Z \sim N(0, \Sigma)$

$$f(X) \sim N(\theta, (1 + \tau^2)\Sigma) \text{ and } g(X) \sim N(\theta, (1 + \tau^{-2})\Sigma)$$

Choose  $\tau \in (0, \infty)$ :

- When  $\tau = 1$ ,  $\mathcal{I}_{f(X)}(\theta) = \mathcal{I}_{g(X)}(\theta)$ ,
- As  $\tau \rightarrow \infty$ ,  $\mathcal{I}_{g(X)}(\theta) \rightarrow \mathcal{I}_X(\theta)$ ,
- As  $\tau \rightarrow 0$ ,  $\mathcal{I}_{f(X)}(\theta) \rightarrow \mathcal{I}_X(\theta)$ .

## Example 2: Poisson Distributed Data

(strong version)

- Alternatively, we can use *conditional* rather than *additive* randomization.

$$\begin{array}{ccc} & X \sim \text{Pois}(\theta) & \\ \swarrow & & \searrow \\ f(X) := Z & \perp & g(X) := X - Z \end{array}$$

where  $Z|X \sim \text{Bin}(X, \tau)$

$$f(X) \sim \text{Pois}(\tau\theta) \text{ and } g(X) \sim \text{Pois}((1 - \tau)\theta)$$

Choose  $\tau \in (0, 1)$ :

- When  $\tau = \frac{1}{2}$ ,  $\mathcal{I}_{f(X)}(\theta) = \mathcal{I}_{g(X)}(\theta)$ ,
- As  $\tau \rightarrow 0$ ,  $\mathcal{I}_{g(X)}(\theta) \rightarrow \mathcal{I}_X(\theta)$ ,
- As  $\tau \rightarrow 1$ ,  $\mathcal{I}_{f(X)}(\theta) \rightarrow \mathcal{I}_X(\theta)$ .

# Decomposition using “conjugate-prior” machine

## General cookbook

- Recall Bayesian inference:

$$\underbrace{p(\theta|X)}_{\text{Posterior}} \propto \underbrace{p(X|\theta)}_{\text{Data}} \underbrace{p(\theta)}_{\text{Prior}}.$$

- If we know that  $X$  is conjugate prior to some other density, we can “reverse” this:

$$\underbrace{p(X|Z)}_{\text{Dataset 2}} \propto \underbrace{p(Z|X)}_{\text{Dataset 1}} \underbrace{p(X)}_{\text{Original Data}}.$$

- Let  $Z = f(X)$  and  $g(X) = X$ . We then need to find

$$p(z) = \int p(z|x)p(x),$$

which has a closed form solution for exponential family distributions.

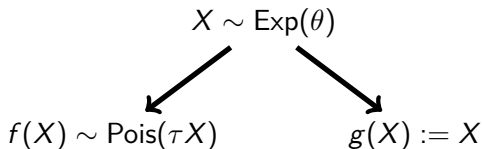
- Can continue to draw  $Z_i$  and form  $f(X) := (Z_1, \dots, Z_B)$  in order to increase amount of information available for selection.



## Example 3: Randomization with conjugacy

(weak version)

- The exponential distribution is conjugate prior to the Poisson distribution.



$$f(X) \sim \text{Geo}\left(\frac{\theta}{\theta + \tau}\right) \text{ and } g(X) | f(X) \sim \text{Gamma}(1 + f(X), \theta + \tau)$$

Choose  $\tau \in (0, \infty)$ :

- When  $\tau \approx \theta$ ,  $\mathcal{I}_{f(X)}(\theta)$  is largest,
- As  $\tau \rightarrow 0$  or  $\tau \rightarrow \infty$ ,  $E[\mathcal{I}_{g(X)|f(X)}(\theta)] \rightarrow \mathcal{I}_X(\theta)$ .

# Other decompositions

Very general methodology; see paper for larger list of decompositions (not exhaustive!)

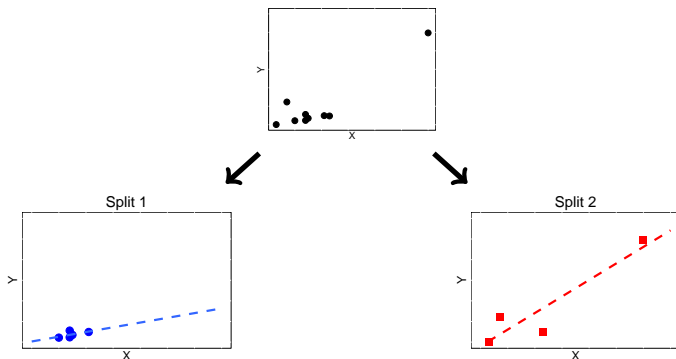
- Gaussian
- Gamma
- Exponential
- Beta
- Bernoulli
- Binomial
- Categorical
- Poisson
- Negative Binomial

## Section 2

### Application 1: Fixed-design linear regression

# Data splitting may not perform well in cases where a few data points have very high leverage

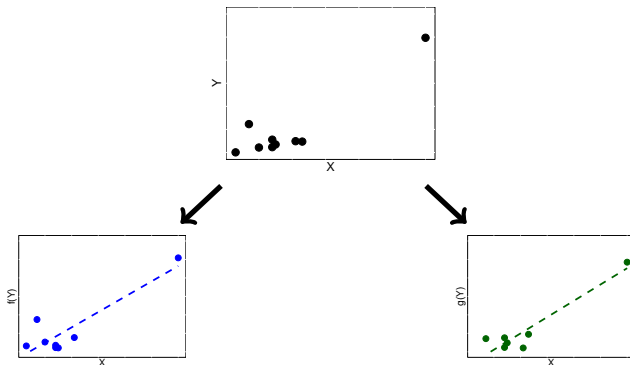
- Consider using a model  $Y = \beta_0 + \beta_1 X + \epsilon$  with  $\epsilon \sim N(0, \sigma^2)$  on the below dataset.



- Total Fisher information available is  $\mathcal{I}(\beta_1) = \frac{\sum (x_i - \bar{x})^2}{\sigma^2}$ . We cannot smoothly split this quantity if the sum is weighted heavily towards a few data points.

# Data fission allows for a smooth trade-off between selection and inference

- We now randomize with  $f(Y) = Y + \tau Z$  and  $g(Y) = Y - \frac{1}{\tau}Z$  for  $Z \sim N(0, \sigma^2)$ .



- Information available in inference stage controlled by  $\tau$ :  $\mathcal{I}(\beta_1) = \frac{\sum (x_i - \bar{x})^2}{(1 + \tau^{-2})\sigma^2}$ .

# Problem setup and notation

- We assume that  $y_i$  is the dependent variable and  $x_i \in \mathbb{R}^p$  is a vector of  $p$  features for  $i = 1, \dots, n$  samples.
- We assume that  $Y$  follows the equation

$$Y = \mu + \epsilon \text{ with } \epsilon = (\epsilon_1, \dots, \epsilon_n)^T \sim N(0, \sigma^2 I_n),$$

where  $\mu = E[Y|X] \in \mathbb{R}^n$  is a fixed unknown quantity and  $\epsilon \in \mathbb{R}^n$  is a random quantity with known variance  $\sigma^2$ .

- We fission the data into  $f(Y)$  and  $g(Y)$ :
  - From  $f(Y)$ , we select a subset of covariates  $M \subseteq [p]$  using any arbitrary selection procedure.
  - Using  $g(Y)$ , we perform inference using the restricted model design matrix  $X_M$ .

# Forming confidence intervals

- The estimator  $\hat{\beta}$  is defined as

$$\hat{\beta}(M) = \operatorname{argmin}_{\tilde{\beta}} \|g(Y) - X_M \tilde{\beta}\|^2 = (X_M^T X_M)^{-1} X_M^T g(Y).$$

- Like conditional selective inference, the target parameter is

$$\beta_*(M) = \operatorname{argmin}_{\tilde{\beta}} E \left[ \|Y - X_M \tilde{\beta}\|^2 \right] = (X_M^T X_M)^{-1} X_M^T \mu.$$

## Distribution of $\hat{\beta}(M) | f(Y)$

$$\begin{aligned} \hat{\beta}(M) &= (X_M^T X_M)^{-1} X_M^T g(Y) \\ &= (X_M^T X_M)^{-1} X_M^T [\mu + \epsilon - \tfrac{1}{\tau} Z] \\ &= \beta_*(M) + (X_M^T X_M)^{-1} X_M^T [\epsilon - \tfrac{1}{\tau} Z] \\ &\sim N(\beta_*(M), (1 + \tau^{-2}) \sigma^2 (X_M^T X_M)^{-1}) \end{aligned}$$

- From the above, we can form  $1 - \alpha$  confidence intervals as

$$\hat{\beta}^k(M) \pm z_{\alpha/2} \sqrt{(1 + \tau^{-2}) \sigma^2 (X_M^T X_M)^{-1}_{kk}}.$$

# Unknown variance?

- Assume we have access to a weakly consistent estimator of  $\sigma$ .
  - If  $p$  is fixed and  $n \rightarrow \infty$ , use the standard error of the residuals obtained from a model using the full suite of covariates [Tian and Taylor (2017)].
  - In higher dimensional settings, it is typically easier to get an overestimate of the variance (to be explored in future work).
- Draw  $\hat{Z} \sim N(0, \hat{\sigma}^2)$  and randomize as

$$\hat{f}(Y) = Y + \tau \hat{Z} \text{ and } \hat{g}(Y) = Y - \frac{1}{\tau} \hat{Z}.$$

- Continuous mapping theorem combined with Slutsky's theorem can be used to show that

$$(\hat{f}(Y), \hat{g}(Y)) \xrightarrow{d} (f(Y), g(Y)).$$

- Repeat the arguments in prior slide, but now the coverage is asymptotic rather than exact.



# 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 covariates  $X_i \in \mathbb{R}^{15}$  generated from independent standard Gaussians.
- For the last data point, we set

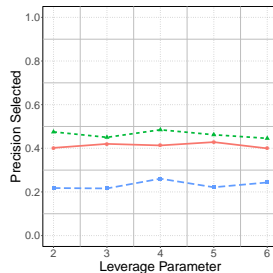
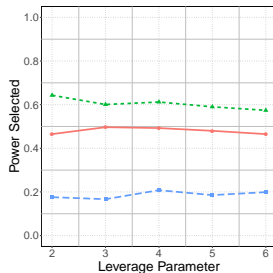
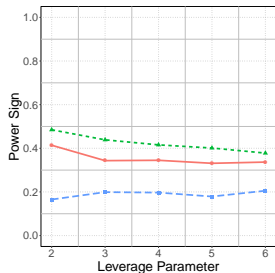
$$X_{16} = \alpha (|X_1|_\infty, \dots, |X_p|_\infty),$$

where  $\alpha$  is a parameter controlling how much higher leverage the last data point has compared to the first 15.

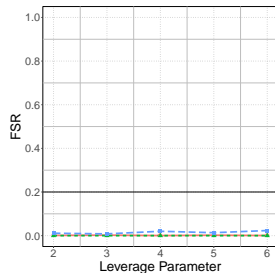
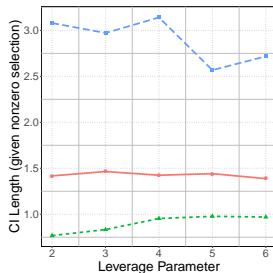
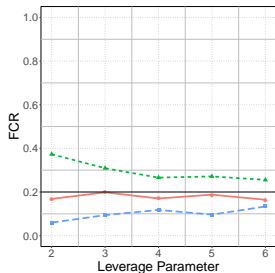
- LASSO used to select features during selection stage.
- Performance of data splitting and data fission tracked alongside the (invalid) procedure which both selects variables and trains the model on the full dataset.

- At the *selection* stage:
  - Power: percentage of variables with non-zero coefficients that are selected.
  - Precision: percentage of selected variables that have non-zero coefficients.
- At the *inference* stage:
  - False coverage rate (FCR): percentage of confidence intervals that do not cover the true parameter.
  - Confidence interval length.
  - Power (Sign): for non-zero parameters, the percentage of confidence intervals that cover the sign of the parameter.
  - False sign rate (FSR): percentage of confidence intervals that do not contain parameters with the correct sign.

In this setting, data fission has more power than data splitting at the selection stage...



# ...along with tighter confidence intervals at the inference step



## Section 3

### Application 2: Trend filtering

# Problem setup and notation

- We assume the data is distributed as

$$y_t = f_0(t) + \epsilon_t,$$

where  $\epsilon_t \sim N(0, \sigma^2)$ .  $f_0$  is not assumed to belong to any model class.

- Trend filtering estimates  $f_0$  by constructing a piecewise polynomial  $\hat{x}$  of degree  $k$  through the minimization problem

$$\hat{x} = \operatorname{argmin}_{x \in \mathbb{R}^n} \frac{1}{2} \|Y - x\|_2^2 + \lambda \|D^{(k+1)}x\|_1, \quad (3)$$

where  $\lambda \geq 0$  is a tuning parameter.  $D^{(k+1)} \in \mathbb{R}^{(n-k) \times n}$  is the  $k$ -th order difference matrix defined recursively as

$$D^{(1)} = \begin{bmatrix} -1 & 1 & 0 & \dots & 0 & 0 \\ 0 & -1 & 1 & \dots & 0 & 0 \\ \vdots & & & & & \\ 0 & 0 & 0 & \dots & -1 & 1 \end{bmatrix} \in \mathbb{R}^{(n-k-1)},$$

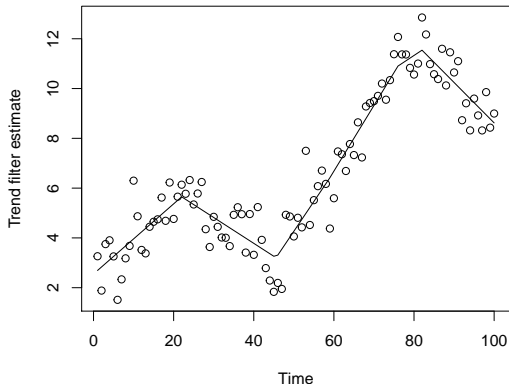
and  $D^{(k+1)} = D^{(1)}D^{(k)}$ .

# Trend filtering example (for $k=1$ )

In the case where  $k = 1$ , the above optimization problem reduces to

$$\hat{x} = \operatorname{argmin}_{x \in \mathbb{R}^n} \frac{1}{2} \|Y - x\|_2^2 + \lambda \|(x_{t+1} - x_t) - (x_t - x_{t-1})\|_1.$$

**Trend filtering example**



# Trend filtering in two stages

The solution set  $\hat{x} = (\hat{x}_1, \dots, \hat{x}_n)$  can also be conceptualized as occurring in two stages: selection of knots and then least squares minimization.

## Stage 1: Knot Selection

- The kink points at which  $\hat{x}$  switches direction are called *knots*.
- A specific set of knots  $t_1, \dots, t_r$  implicitly defines a **falling factorial basis**, which is a set of functions whose *discrete* derivatives are constant for adjacent design points up to order  $k - 1$ .

## Stage 2: Minimization

- Denote  $A$  to be a matrix with entries corresponding to the selected falling factorial basis evaluates at the design points  $x$ .
- We then have that

$$\hat{x} = A(A^T A)^{-1} A^T Y$$



# Target of estimation

- As before, we fission into two variables:  $f(Y)$  in order to select knots and  $g(Y)$  in order to perform inference.
- We wish to cover the *projected mean*

$$\mu^*(A) = A(A^T A)^{-1} A^T \mu,$$

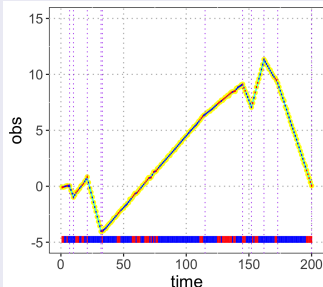
where  $\mu = (f_0(1), \dots, f_0(n))^T$ .

- Two types of confidence intervals available:
  - **Pointwise confidence intervals:**  $\mathbb{P}(\mu^*(A)_i \notin \text{CI}(\mu^*)_i) \leq \alpha$  for all  $i$ .  
Construction follows exactly as described in prior section.
  - **Uniform confidence intervals:**  $\mathbb{P}(\exists i \in [n] : \mu^*(A)_i \notin \text{CI}(\mu^*)_i) \leq \alpha$ .  
See [Koenker (2011)] for construction methodology.

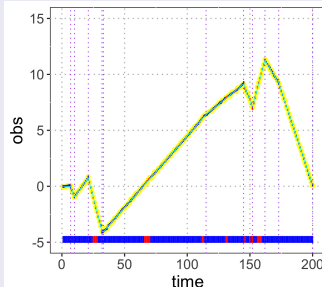
# Simulation setup

- We generate simulated data as  $Y_t = f_0(t) + Z_t$  where  $Z_t \sim N(0, 1)$ .
- $f_0(t + 1) = f_0(t) + v_t$ .
  - With probability 0.95,  $v_{t+1} = v_t$ .
  - With probability 0.05,  $v_{t+1} = u$  with  $u \sim \text{Unif}(-0.5, 0.5)$ .

## Example: Using Full Data

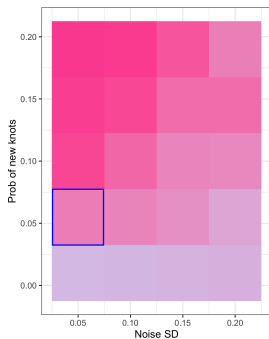


## Example: Using Data Fission

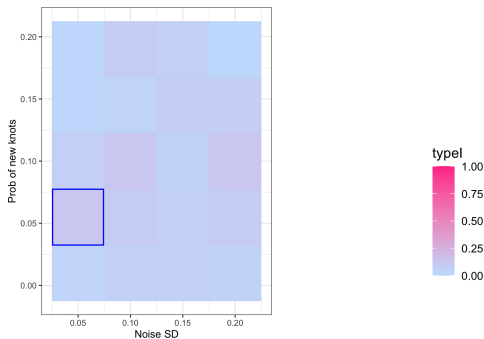


# Type I error control is guaranteed using data fission

- The below graphs compare type I error control using data fission with the (invalid) approach of using the full dataset for both knot selection and inference.
- Red entries indicate that type I error control is violated.



(a) Simultaneous type I error using full data twice.

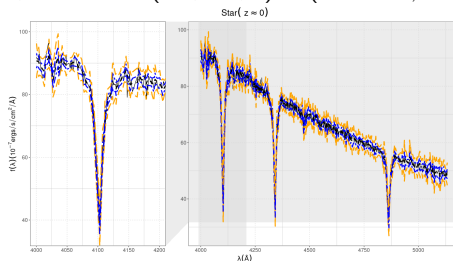


(b) Simultaneous type I error using data fission.

# Real Data Example: Spectroscopy

- Astronomers observe a spectrum consisting of wavelengths ( $\lambda$ ) and measurements of the coadded flux  $f(\lambda)$  for an object of interest.
- Data fission used to construct confidence interval after fitting a smooth trend.

Star DR12, Located at (RA,Dec, z)  $\approx$  (236.834 $^\circ$ , 0.680 $^\circ$ , 0.000).



— Pointwise CI    — Uniform CI    — Actual    - - Fitted

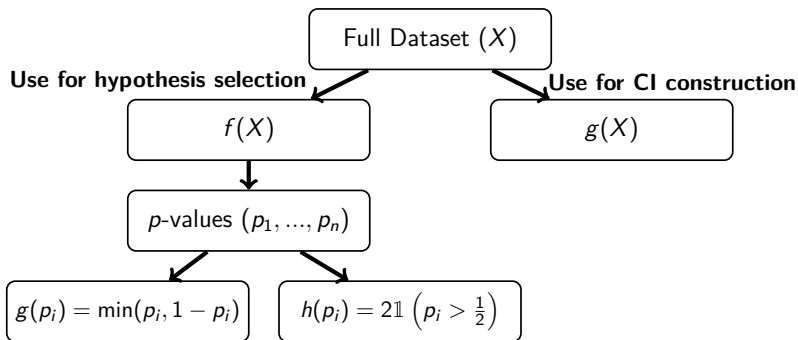
Sourced from the twelfth data release of the Baryon Oscillation Spectroscopic Survey [Alam et al. (2015)]. Thanks to Collin Politsch for assistance.

## Section 4

### Application 3: Confidence intervals after interactive multiple testing

# Problem Setup

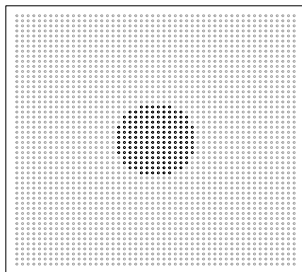
- AdaPT and STAR allow for FDR control after interactive multiple testing, but do not offer a way of estimating effect sizes.
- Combining these methods with data fission allows us to construct confidence intervals after hypotheses selection.



# Simulation Setup

- Each point is a signal, with non-null region arranged in a circle within the center of a grid.
- Entry points  $x_i \sim N(\mu_i, 1)$
- Signal strength  $\mu_i = 0$  for nulls and  $\mu_i = 2$  for non-nulls.

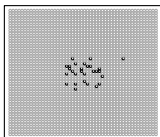
## Example of non-null region for detection



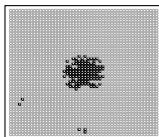
# For small values of $\tau$ , hypothesis selection using fissioned data is almost unchanged

Example rejection region for a single trial run shown for BH and STAR procedures using  $\tau = 0.1$ .

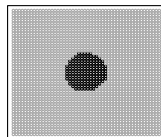
**BH (Full)**



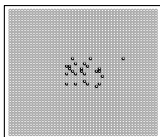
**AdaPT (Full)**



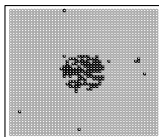
**STAR (Full)**



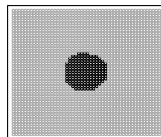
**BH (Fission)**



**AdaPT (Fission)**



**STAR (Fission)**





# Inference Procedure

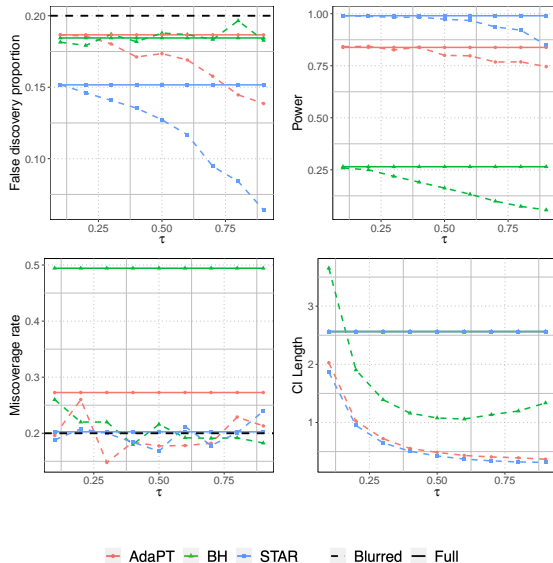
- We fission the data using Gaussian randomization (with parameter  $\tau$ ) and select hypotheses using  $f(X)$ .
- We then run inference on the average signal in the rejection set  $\mathcal{R}$ :

$$\bar{\mu} = \frac{1}{|\mathcal{R}|} \sum_{i \in \mathcal{R}} \mu_i.$$

Using the confidence interval construction:

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

# Results



## Section 5

### Conclusion

# Conclusion

- An alternative to data-splitting and data-carving (and p-value masking)
- Makes sense if you want flexible model/hypothesis selection rules (cannot condition on selection) and splitting makes less conceptual sense.
- Ideas from Bayesian conjugacy allow us to fission many types of data (Gaussian, Poisson, Negative Binomial, Gamma, etc).
- Works well in fixed design linear regression problems, like trend filtering (and extensions to graph trend filtering).
- Still some further study to be done on poisson/binomial regression problems to make the idea work well in practice.

# References I



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).*



Duan, B., A. Ramdas, and L. Wasserman (2020).

Familywise Error Rate Control by Interactive Unmasking

*Proceedings of the 37th International Conference on Machine Learning 255.*



Duan, B., A. Ramdas, S. Balakrishnan, and L. Wasserman (2020).

Interactive martingale tests for the global null

*Electronic Journal of Statistics 14(2), 4489–4551.*



Duan, B., A. Ramdas, and L. Wasserman (2021).

Interactive identification of individuals with positive treatment effect while controlling false discoveries

*arXiv preprint arXiv:2102.10778.*

# References II



Koenker, R. (2011).

Additive models for quantile regression: Model selection and confidence bands.  
*Brazilian Journal of Probability and Statistics* 25(3), 239–262.



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.




Rasines, D. G. and G. A. Young (2021).

Splitting strategies for post-selection inference.  
*arXiv preprint arXiv:2102.02159*.

# References III

 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.