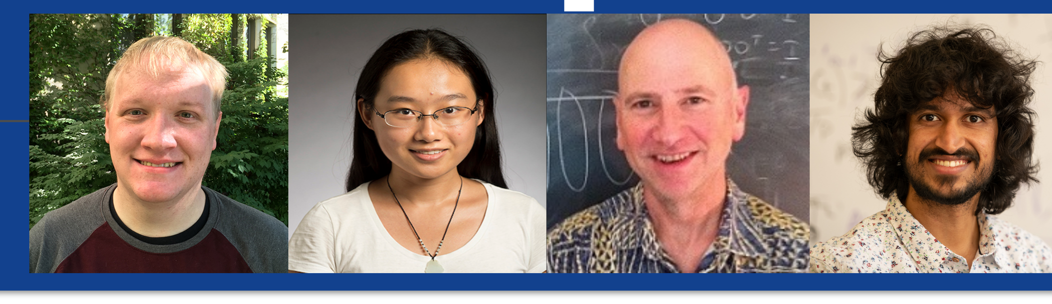
Data fission: splitting a single data point

James Leiner[1], Boyan Duan[2], Larry Wasserman[1], Aaditya Ramdas[1]

[1]Department of Statistics and Data Science, Carnegie Mellon University

[2]Google



Introduction

Suppose we observe data with a known distribution, up to an unknown variable of interest θ: $X \sim P_X(\theta)$.

We explore decompositions of X into $f(X)$ and $g(X)$ such that:

1. $f(X)$ is not sufficient to reconstruct X by itself
2. There exists a function h such that $h(f(X), g(X)) = X$
3. One of the following two properties holds:
   - $f(X) \perp g(X)$ with known marginal distributions ("strong version")
   - $f(X)$ and $g(X) \mid f(X)$ have known and tractable distributions ("weak version")

Additive randomization

Consider observing $X \sim N(\theta, \sigma^2)$. We can then draw $Z \sim N(0, \sigma^2)$ and randomize by:

$X \sim N(\theta, \sigma^2)$

$f(X) := X + \tau Z \sim N(\theta, (1+\tau^2)\sigma^2)$ $\perp$ $g(X) := X - \frac{1}{\tau} Z \sim N(\theta, (1+\tau^{-2})\sigma^2)$

As τ increases, more information gets allocated to $f(X)$ and less information gets allocated to $g(X)$.

Conditional randomization

Alternatively, we can use *conditional* rather than *additive* randomization and then choosing a a $\tau \in (0,1)$:

$X \sim \text{Pois}(\theta)$

$f(X) \sim \text{Bin}(X, \tau) \sim \text{Pois}(\tau\theta)$ $\perp$ $g(X) := X - f(X) \sim \text{Pois}((1-\tau)\theta)$

When $\tau = \frac{1}{2}$, $f(X)$ and $g(X)$ are roughly comparable. As $\tau \to 0$, $g(X) \xrightarrow{d} X$. As $\tau \to 1$, $f(X) \xrightarrow{d} X$.

Conjugate Prior "Reversal"

Bayesian Inference: $\underbrace{p(\theta \mid X)}_{\text{Posterior}} \propto \underbrace{p(X \mid \theta)}_{\text{Data}} \underbrace{p(\theta)}_{\text{Prior}}$

Is $X \sim P_X(\theta)$ conjugate to some other distribution P_Z? Then draw $Z \sim P_Z(X)$

"Reverse" the prior to accomplish data fission $\underbrace{p(X \mid Z)}_{\text{Dataset 2}} \propto \underbrace{p(Z \mid X)}_{\text{Dataset 1}} \underbrace{p(X)}_{\text{Original Data}}$

$f(X) := Z$ Known marginals for exponential family distributions

$g(X) := X$ Distribution of $g(X) \mid f(X)$ known from Bayesian inference

Need more information for selection? Draw $Z_1, \ldots, Z_B$ and let $f(X) := (Z_1, \ldots, Z_B)$.

Randomization with conjugacy

The exponential distribution is conjugate prior to the Poisson distribution.

$X \sim \text{Exp}(\theta)$

$f(X) \sim \text{Pois}(\tau X)$ $g(X) := X$

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

When $\tau \approx \theta$, the most information is contained in $f(X)$. As $\tau \to 0$ or $\tau \to \infty$, $g(X) \mid f(X) \xrightarrow{d} X$.

Selective Inference

We primarily focus on the use of data fission for (potentially high dimensional) model selection and inference.

Full Dataset (X)

Randomize dataset | Reveal full data at inference step

Use any procedure to select a model from $f(X)$ → Use $P\left(g(X) \mid f(X)\right)$ for inference

Information splitting

Assuming that $P_X(\theta)$ is known up to unknown parameter of interest θ, data fissions smoothly trades off Fisher information between selection and inference by varying τ.

For "weak version"

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

For "strong version"

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

Linear Regression

We assume that y_i is the dependent variable and $x_i \in \mathbb{R}^p$ is a vector of features with corresponding design matrix $X \in \mathbb{R}^{n \times p}$.

$Y = \mu + \epsilon$ with $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T \sim N(0, \sigma^2 I_n)$ where σ^2 is known and $\mu = E[Y \mid X] \in \mathbb{R}^n$ is unknown

Draw $Z \sim N(0, \sigma^2)$

Use $f(Y) := Y + \tau Z$ to select a model $M \subseteq [p]$ and corresponding design matrix X_M

Use $g(Y)$ to conduct inference on $\hat{\beta}(M) = \text{argmin}_{\tilde{\beta}} \left\| g(Y) - X_M \tilde{\beta} \right\|^2 = (X_M^T X_M)^{-1} X_M^T g(Y)$

Target Parameter:

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

Forming confidence intervals

$$\hat{\beta}(M) \sim N\left(\beta_\star(M), (1+\tau^{-2})\sigma^2 (X_M^T X_M)^{-1}\right)$$

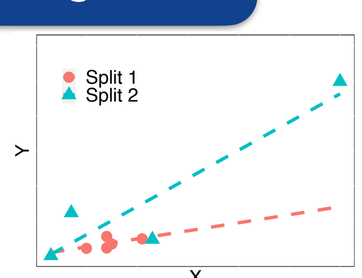
Form $1-\alpha$ confidence intervals as: $\hat{\beta}^k(M) \pm z_{\alpha/2} \sqrt{(1+\tau^{-2})\sigma^2 (X_M^T X_M)^{-1}_{kk}}$

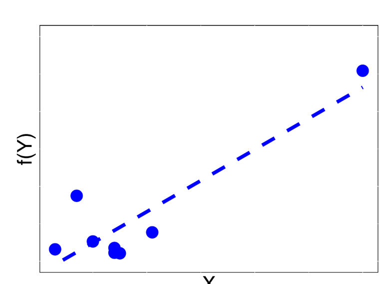
Unknown σ? If there exists $\hat{\sigma} \xrightarrow{p} \sigma$, calculate this before fissioning the data and use throughout—all guarantees above still hold asymptotically.

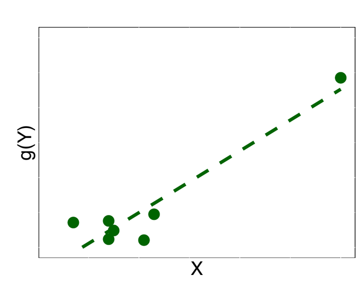
Why not data splitting?

Consider a dataset with low n and a handful of high leverage points.

Data splitting cannot allocate this information to both datasets, so information tradeoff is not "smooth".



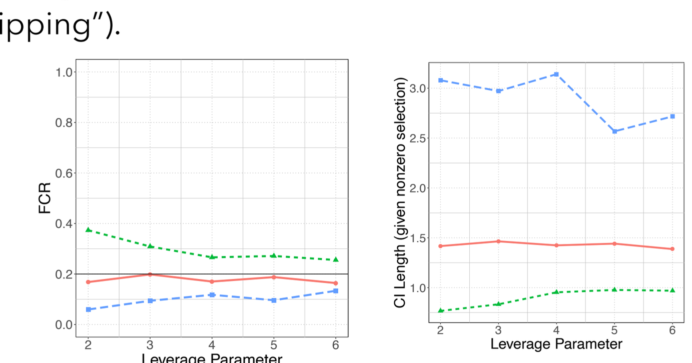




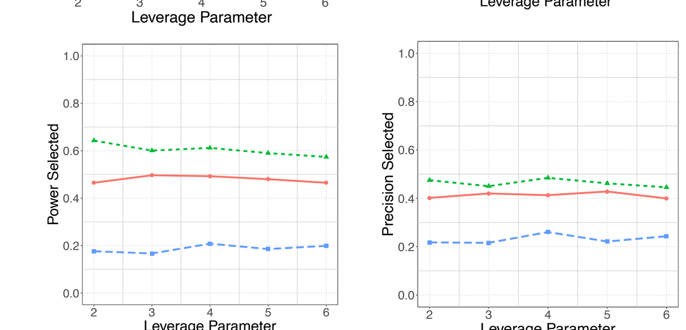
Data fission allows the information in that single point to be "split" evenly across $f(X)$ and $g(X)$.

We simulate a small dataset ($n = 20$, $p = 16$) with a single high leverage point and repeat the data fission procedure for 500 trials, selecting a model with LASSO. We compare data fission along with data splitting, and reusing the full dataset for both selection and inference ("double dipping").

Splitting and fission both control FCR but fission allows for tighter confidence intervals. Double dipping violates FCR control.



Data fission also enables improved power/precision at the selection stage when using LASSO for variable selection.



Interactive Hypothesis Testing

Full Dataset (X) → $f(X)$, $g(X)$

$f(X)$ → p-values $(p_1, \ldots, p_n)$ → $g(p_i) = \min(p_i, 1 - p_i)$, $h(p_i) = 2\mathbb{I}\left(p_i > \frac{1}{2}\right)$

$g(X)$: Use for CI construction on rejection set

Form rejection set adaptively using BH, AdaPT (Lei and Fithian '18), or STAR (Lei et al. '20) procedures.

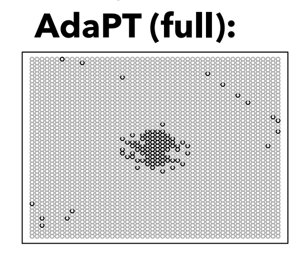
Simulation results

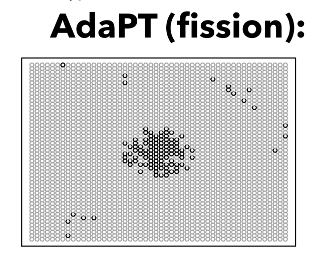
We assume $x_i \sim N(\mu_i, 1)$ with $\mu_i = 0$ for nulls and $\mu_i = 2$ for nulls (arranged in a circle). After forming rejection set ($\mathcal{R}$) from fissioned data (using same Gaussian decomposition as above), we:

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

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

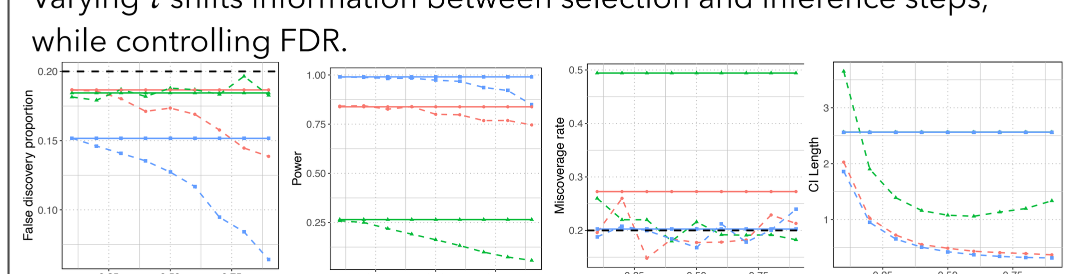
AdaPT (full): **AdaPT (fission):**





Example rejection regions, with and without data fission for $\tau = 0.1$

Varying τ shifts information between selection and inference steps, while controlling FDR.



Misspecified GLMs

We assume that y_i follows some distribution in the exponential dispersion family and attempt to model $\mu_i := E[y_i \mid x_i]$ through covariates $x_i \in \mathbb{R}^p$ under the assumption that $m(\mu_i) = \beta^T x_i$ for some known link function m.

Problem! Even if the distribution of y_i is known, it is unlikely that μ_i is actually a linear combination of the selected covariates for realistic selection rules.

Solution

Assumption: The analyst fissions the data such that $g(y_i) \perp g(y_k) \mid X, f(Y)$ for all $i \neq k$.

Use $f(Y)$ to select a model $M \subseteq [p]$ which induces a quasi-likelihood function on g(Y), for some working model p:

$$L_n := \sum_{i=1}^{n} \log p(g(y_i) \mid \beta, f(Y), X_M),$$

Denote $\hat{\beta}_n(M)$ to be the empirical maximizer of L_n

Target parameter now is the KL minimizer between the working model and true distribution q:

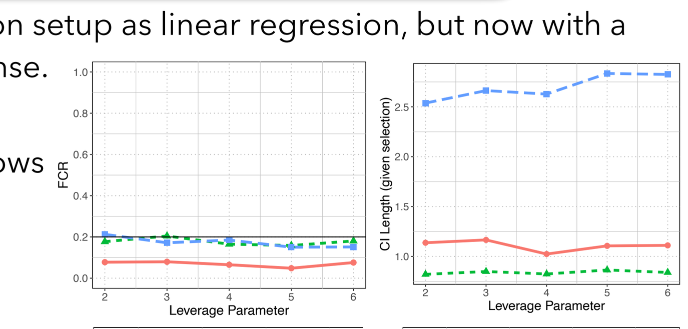
$$\beta_n^\star(M) = \text{argmin}_\beta D_{KL}\left(\prod_{i=1}^{n} q(g(y_i) \mid X, f(Y)) \,\Big|\Big|\, \prod_{i=1}^{n} p(g(y_i) \mid \beta, f(Y), X_M)\right)$$

Under mild regularity conditions (Fahrmeir 1990), $\hat{\beta}_n(M) \xrightarrow{p} \beta_n^\star(M)$ and asymptotically **conservative** CIs can be constructed (see arxiv preprint for details).

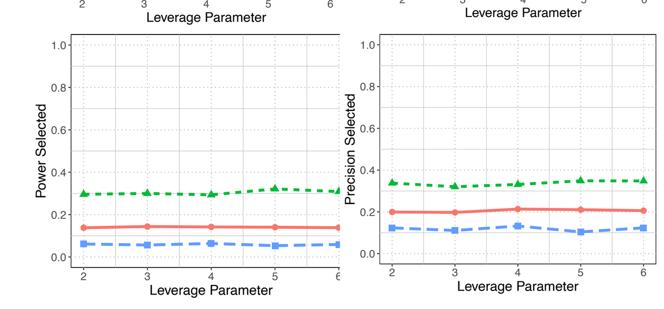
Fission for Poisson regression

We use the same simulation setup as linear regression, but now with a Poisson-distributed response.

Splitting and fission both control FCR but fission allows for tighter confidence intervals.

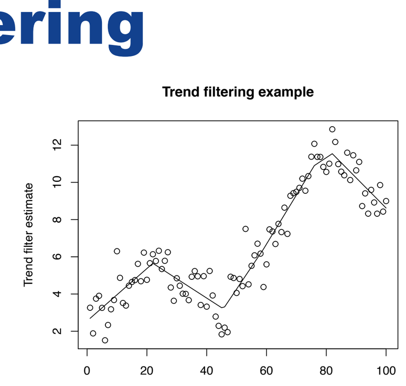


... and also enables improved power/precision at the selection stage when using LASSO for variable



Selective Inference for Trend Filtering

We observe a time series $y_t = f_0(t) + \epsilon_t$ where $\epsilon_t \sim N(0, \sigma^2)$ and f_0 is not assumed to belong to any model class. Trend filtering estimate $\hat{f}_0$ by constructing a piecewise linear function as:



$\hat{f}_0 = \text{argmin}_{x \in \mathbb{R}^n} \frac{1}{2} \| Y - x \|_2^2 + \lambda \| (x_{t+1} - x_t) - (x_t - x_{t-1}) \|_1$ where λ is a tuning parameter

Equivalently, we can conceptualize trend filtering in two stages:

Stage 1: Knot Selection

The kink points at which $\hat{f}_0$ switches direction are called knots

A specific set of knots $t_1, \ldots, t_r$ implicitly defines a falling factorial basis which is a set of functions whose discrete derivatives are constant for adjacent design points up to order $k-1$

Use $f(Y)$ to select basis A

Stage 2: Minimization

Denote A to be a matrix with entries corresponding to the selected falling factorial basis

We then have that:

$$\hat{f}_0 = A(A^T A)^{-1} A^T Y$$

Use $g(Y)$ to conduct inference

Forming confidence intervals

Target parameter is the **projected mean:** $\mu^\star(A) = A(A^T A)^{-1} A^T \mu$ where $\mu = (f_0(1), \ldots, f_0(n))^T$

Pointwise CIs:

$\mathbb{P}(\mu^\star(A)_i \notin \text{CI}(\mu^\star)_i) \leq \alpha$ for all i.

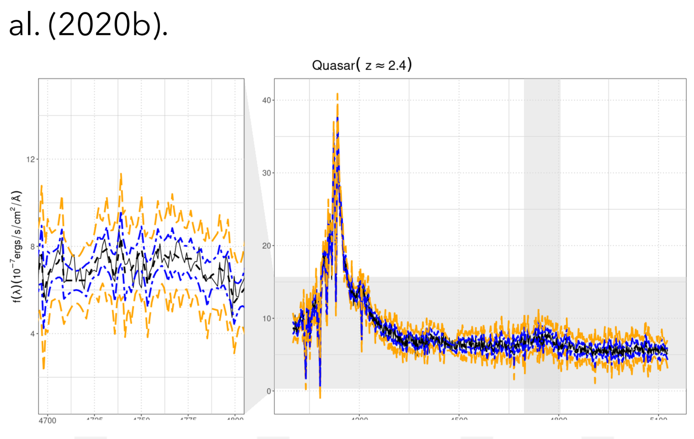
Can be constructed exactly the same as in the linear regression example

Uniform CIs:

$\mathbb{P}(\exists i \in [n] : \mu^\star(A)_i \notin \text{CI}(\mu^\star)_i) \leq \alpha$

Using Koenker (2011) for construction methodology (see arxiv preprint for details).

The above construction will control the FCR (for pointwise CIs) or simultaneous type I error rate (for uniform CIs). To test this procedure, we run it on a real data example. For an astronomical object of interest, we model the coated flux $f(\lambda)$ as a function of wavelengths λ (Politsch et al. (2020b).



Data fission appears to model the underlying trend well, while still allowing for enough information to construct tight confidence intervals