**Data fission: splitting a single data point**

James Leiner¹, Boyan Duan², Larry Wasserman¹, Aaditya Ramdas¹

¹Department of Statistics and Data Science, Carnegie Mellon University
²Google

**Introduction**

Suppose we observe data with a known distribution, up to an unknown variable of interest \( \theta \sim P(\theta) \). 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. It is possible to define \( k(X) \) such that \( k(X) \sim X \).
3. One of the following two properties holds:
   - \( f(X) \sim \tau(X) \) with known marginal distributions ("strong version")
   - \( f(X) \) and \( g(X) \) have known and tractable distributions ("weak version")

**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 \( X \in \mathbb{R}^{n \times p} \).

\[
Y = \beta^T X + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \quad \text{where } \sigma^2 \text{ is known and } \beta \sim \mathcal{N}(0, \Sigma) \quad \text{is unknown}
\]

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

**Misspecified GLMs**

We assume that \( y_i \) follows some distribution in the exponential dispersion family and attempt to model \( y_i = f(x_i) \) through covariates \( x_i \in \mathbb{R}^P \) under the assumption that \( \mu(x) = x^T \beta \) for some known link function \( f \).

**Problem**

Even if the distribution of \( y_i \) is known, it is unlikely that \( \beta \) is actually a linear combination of the selected covariates for realistic selection rules.

**Solution**

The analyst fissions the data such that \( (y_1, x_1; \tau), (y_2, x_2; \tau), \ldots, (y_n, x_n; \tau) \) for all \( \tau \).

**Fission for Poisson regression**

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

**Selective Inference for Trend Filtering**

We observe a vector \( y = (y_0, y_1, \ldots, y_n) \) with \( y_i \sim \text{Exp}(\theta) \) and \( \theta \) is not assumed to belong to any model class. Trend filtering estimates \( \theta \) by constructing a piecewise linear function as:

\[
\hat{\theta}_i = \arg\min_{\theta} \sum_{i=1}^n \left( y_i - f_i(\theta) \right)^2
\]

Equivalently, we can conceptualize trend filtering in two stages:

**Stage 1: Knot Selection**

The knots points at which \( f \) switches direction are called knots. A specific set of knots \( \mathcal{K} = \{k_1, k_2, \ldots, k_m\} \) implicitly defines a finite set of piecewise functions whose discrete derivatives are constant for adjacent design points up to order 1–4.

**Stage 2: Minimization**

Denote \( \Sigma \) to be the matrix of entries corresponding to the selected knotting factor.

**Uniform CIs**

\[
\text{CI}_L(\mu) = \hat{\mu} - 1.96 \sqrt{\text{Var}(\hat{\mu})} \quad \text{and} \quad \text{CI}_U(\mu) = \hat{\mu} + 1.96 \sqrt{\text{Var}(\hat{\mu})}
\]

The above construction will control the FDR (or simultaneous CI) or a single ratio error term (for uniform CIs). To test this procedure, we run it on real data examples. For an astronomical object of interest, we modeled the coated flux \( n_{\text{cov}} \) as a function of wavelength \( \lambda \) (Pichon et al. 2020b).

**Target pointwise CIs**

**Uniform CIs**

\[
P(\text{CI}_L(\mu) \leq CI_{\alpha} \leq \text{CI}_U(\mu)) \leq \frac{1}{2} \alpha
\]

The above construction will control the FDR (or simultaneous CI) or a single ratio error term (for uniform CIs). To test this procedure, we run it on real data examples. For an astronomical object of interest, we modeled the coated flux \( n_{\text{cov}} \) as a function of wavelength \( \lambda \) (Pichon et al. 2020b).