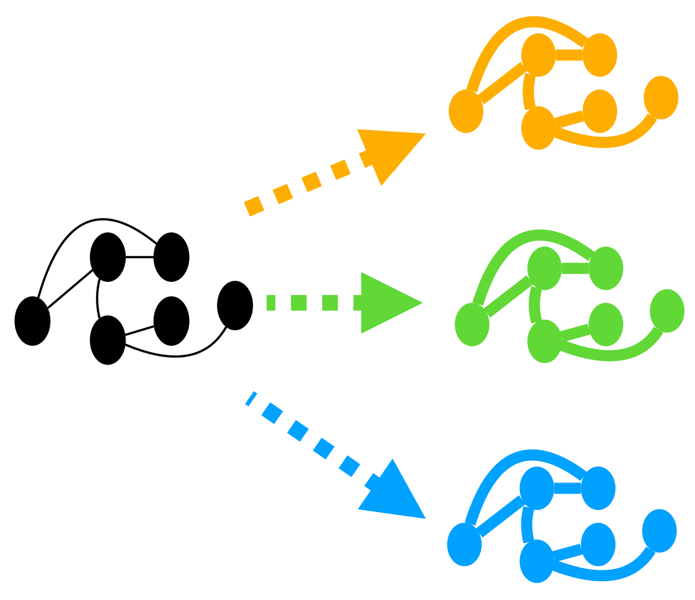
Graph fission and cross-validation

James Leiner[1]
Aaditya Ramdas[1,2]

Department of Statistics[1] and Machine Learning[2]
Carnegie Mellon University

Motivation

- We observe a graph $\mathcal{G} = (V, E, Y)$ with a known vertex (V) and edge (E) set, alongside observations (Y). Let $y_i = \mu_i + \epsilon_i$, where $\mu_i = E[y_i]$ and ϵ_i is a mean 0 random variable.
- If an analyst needs to select a model or tune hyper parameters over the graph, it may be useful to divide the data into multiple independent copies. However, because the data is not iid, **sample splitting** is not available.
- We use external randomization to create m independent copies of the graph $\mathcal{G}_1, \ldots, \mathcal{G}_m$ with corresponding observations $Y^{\mathcal{G}_1}, \ldots, Y^{\mathcal{G}_m}$, such that:
  1. $\mathcal{G}_i$ has the same vertex and edge set as $\mathcal{G}$.
  2. Taken together, the individual datasets recover the original data Y in the sense that there exists a known deterministic function h such that $\mathcal{G} = h\left(\mathcal{G}_1, \ldots, \mathcal{G}_m\right)$.
  3. The information contained in Y is divided across $\mathcal{G}_1, \ldots, \mathcal{G}_m$ in any proportion desired.



Graph Fission in P1 Regime

- We leverage techniques called Data Fission (Leiner et al., 2023), and Data Thinning (Neufeld et al., 2022) to decompose the graph into multiple copies.

Desiderata:
Desiderata:

- $\mathbb{E}[Y^{\mathcal{G}_j}] = f(\mu)$ for some known function f
- $Y^{\mathcal{G}_1}, \ldots, Y^{\mathcal{G}_m}$ are all mutually independent

Convolution Closed Definition (Joe, 1996)

- Let F_θ be a distribution indexed by a parameter θ
- Drawing $X' \sim F_{\theta_1}$ and $X'' \sim F_{\theta_2}$ independently , then F is **convolution-closed**, if $X' + X'' \sim F_{\theta_1 + \theta_2}$.

Generic Formulation (Neufeld et al., 2022)

Choose $\tau_1, \ldots \tau_m$ such that $\sum_{i=1}^{m} \tau_i = 1$

Let $G_{\theta_1, \ldots, \theta_m}$ be the joint distribution of $(Y^{\mathcal{G}_1}, \ldots, Y^{\mathcal{G}_1}) \mid \sum_{j=1}^{m} Y_{\mathcal{G}_i} = Y$,

Example: Gaussian Data

- Assume $y_i \sim N(\mu_i, \sigma^2)$
- Draw $y_i^{\mathcal{G}_1}, \ldots, y_i^{\mathcal{G}_m}$ from the distribution

$$N\left(\begin{bmatrix} y_i \\ \vdots \\ y_i \end{bmatrix}, \sigma^2 \begin{bmatrix} (m-1) & -1 & \ldots & -1 \\ -1 & (m-1) & \ldots & \vdots \\ \vdots & & \ddots & \\ -1 & -1 & \ldots & (m-1) \end{bmatrix}\right)$$

- Marginally, $y_i^{\mathcal{G}_j} \sim N(\mu_i, m\sigma^2)$, all mutually independent.

Example: Poisson Data

- Assume $y_i \sim \text{Pois}(\mu_i)$
- Draw $y_i^{\mathcal{G}_1}, \ldots, y_i^{\mathcal{G}_m}$ from the distribution

$$\text{Multinomial}\left(y_i, \left(\frac{1}{m}, \ldots, \frac{1}{m}\right)\right)$$

- Marginally, $y_i^{\mathcal{G}_j} \sim \text{Pois}\left(\frac{\mu_i}{m}\right)$, all mutually independent.

Graph Fission in P2 Regime

- The decomposition rules in the **P1 Regime** are clean, but sometimes require knowledge of a nuisance parameter (e.g. σ^2 in the Gaussian case) which may be inconvenient.

We create two synthetic graphs such that $Y^{\mathcal{G}_1}, Y^{\mathcal{G}_2}$

- The law of $Y^{\mathcal{G}_2} \mid Y^{\mathcal{G}_1}$ is known and tractable.
- There exists a function h such that $\mathcal{G} = h(\mathcal{G}_1, \mathcal{G}_2)$.

Example: Gaussian Data

Assume $y_i \sim N(\mu, \sigma^2 I_n)$ and draw $Z \sim N(0, \sigma_0^2 I_n)$

$$y_i \sim N(\mu_i, \sigma^2)$$

$$Y^{\mathcal{G}_1} = Y + Z \sim N(\mu_i, \sigma^2 + \sigma_0^2)$$

$$Y^{\mathcal{G}_2} = Y \qquad Y \mid Y^{\mathcal{G}^1} \sim N\left(\mu(1-\tau) + \tau Y^{\mathcal{G}^1}, \sigma^2(1-\tau) I_n\right)$$

$$\tau := \frac{\sigma^2}{\sigma^2 + \sigma_0^2}$$

Background: Structural Trend Estimation

- We consider estimating a structural trend as a running example to consider across two applications: **cross-validation** and **post-selection inference**.

$$\hat{\mu} := \text{argmin}_{\beta \in \mathbb{R}^n} \underbrace{\ell(Y, \beta)}_{\text{Loss}} + \underbrace{D(\beta)}_{\text{Penalty}}$$

Forming a penalty:

- We consider penalties of the form $D(\beta) := \lambda \left\| \Delta^{(k+1)} \beta \right\|_1$ or $D(\beta) := \lambda \left\| \Delta^{(k+1)} \beta \right\|_2$.
- $\Delta^{(1)} \in \{-1, 0, 1\}^{n \times p}$ and consists of one row per edge

$$\Delta_v^{(1)} = (0, \ldots \underset{i}{\underset{\uparrow}{-1}}, \ldots \underset{j}{\underset{\uparrow}{1}}, \ldots 0)$$

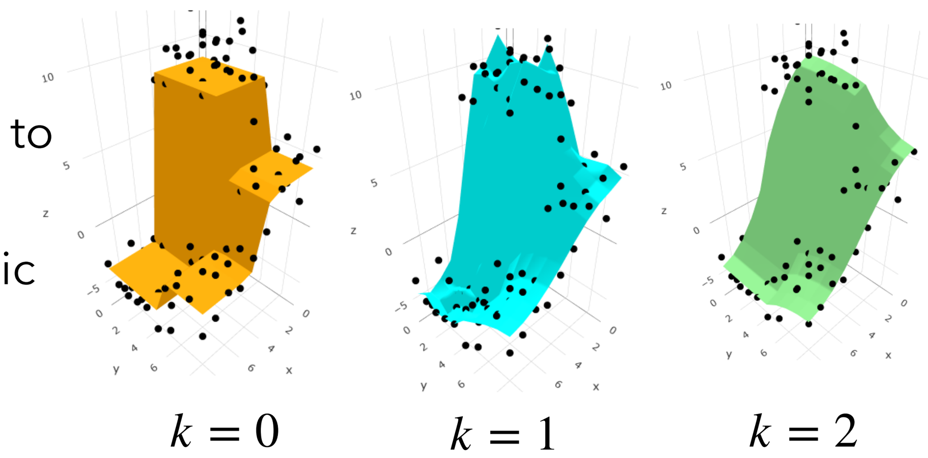
Row corresponding to edge (i, j)

(orientation of -1 and 1 is arbitrary)

$$\Delta^{(k+1)} = \begin{cases} \left(\Delta^{(1)}\right)^\top \Delta^{(k)} = L^{\frac{k+1}{2}} & \text{for odd } k \\ \Delta^{(1)} \Delta^{(k)} = \Delta^{(1)} L^{\frac{k}{2}} & \text{for even } k \end{cases}$$

Iterative formula for constructing $\Delta^{(k+1)}$

$k = 0$ corresponds to a piecewise constant trend, $k = 1$ corresponds to piecewise linear trend, and $k = 2$ corresponds to piecewise quadratic trends. See left examples when square loss is used



Application: Cross Validation

- Consider choosing λ in the above structural trend estimation problem.

Graph Cross Validation Approach

- Assume $Y \sim F_\theta$ is convolution-closed and we construct $Y^{\mathcal{G}_1}, \ldots, Y^{\mathcal{G}_m}$ under **P1**

Use to train model

$$Y^{\mathcal{G}_{-j}} := \sum_{j \neq i} Y^{\mathcal{G}_j} \sim F_{\theta \frac{m-1}{m}}$$

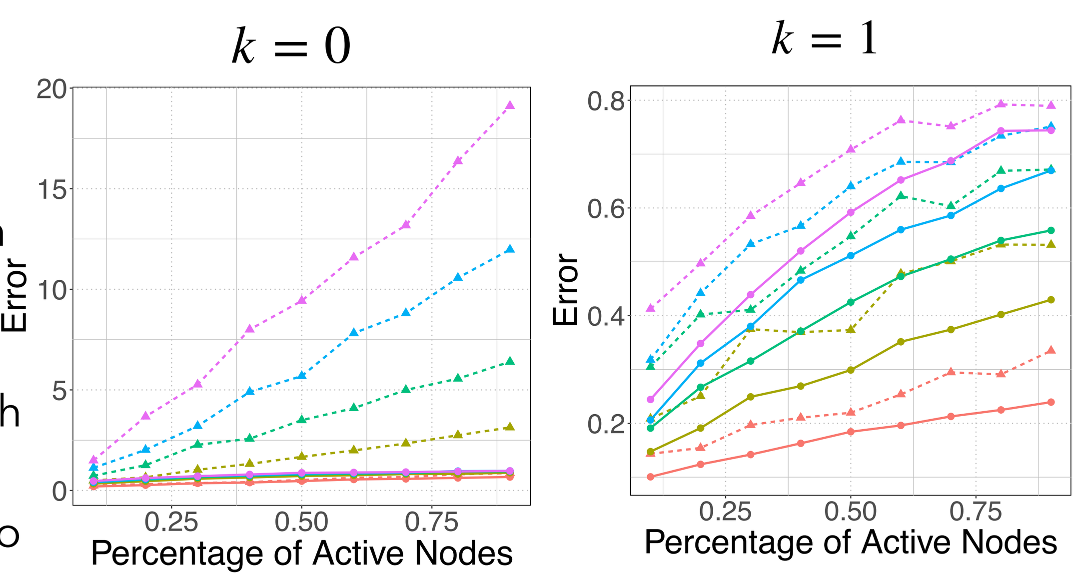
Use for evaluation

$$Y^{\mathcal{G}_j} \sim F_{\theta \frac{1}{m}}$$

Ordinary Cross Validation Approach

- Select a subset of nodes $I \subseteq V$.
- Train $\hat{\beta}_{-I}$ by excluding these nodes and running STE
- Denote $\hat{\beta}_I$ as the average of fitted values across adjacent nodes for each $i \in I$.
- Evaluate $\hat{\beta}_I$ performance using held out

- We vary the size of jumps at breakpoints along with the percentage of active nodes (i.e. number of breakpoints) in the graph, and compare graph cross-validation against ordinary cross-validation.
- The relative performance of graph cross-validation (dotted) compared to ordinary cross-validation (solid) increases with both the size of jumps and number of breakpoints, indicating that less smooth trends benefit the most from using graph fission to tune λ.



Link to paper
(arXiv: 2401.15063)

Application: Inference after Structural Trend Estimation

- We use graph fission to construct **confidence intervals** around a fitted trend $\hat{\mu}$ when a square loss function is used, and $D(\beta) := \lambda \left\| \Delta^{(k+1)} \beta \right\|_1$.

$\mathcal{G}$ → $\mathcal{G}_{\text{sel}}$: Use to select a basis **B** and choose inferential target $\eta^T \mu := e_j^T B(B^T B)^{-1} B^T \mu$

$\mathcal{G}$ → $\mathcal{G}_{\text{inf}}$: Use for inference

Step 1: Basis Selection

Fit $\hat{\mu}$ on $\mathcal{G}_{\text{sel}}$ for some choice of k

When k is even:

1. $C \leftarrow L^{\frac{k}{2}} \hat{\beta}$
2. Identify unique values of C: $c_1, \ldots, c_\ell$
3. $B \leftarrow \left(L^\dagger\right)^{\frac{k}{2}} \left[c_1^T \quad \ldots \quad c_\ell^T\right]$
4. $B \leftarrow [1 \quad B]$

When k is odd:

1. $C \leftarrow L^{\frac{k+1}{2}} \hat{\beta}$
2. Identify $A \subseteq \{1, \ldots n\}$ corresponding to the non-zero rows of C.
3. Let B be $\left(L^\dagger\right)^{\frac{k+1}{2}}$ with only the columns corresponding to A included
4. $B \leftarrow [1 \quad B]$

Step 2: Inference

- In the **P1** regime, standard inferential procedures can be used (e.g. least squares), because the selection and inference graphs are independent
- The **P2** regime may be necessary when $G_{\theta_1, \ldots, \theta_m}$ is a function of unknown nuisance parameters. Consider the case where $Y \sim N(\mu, \sigma^2 I_n)$ with σ^2 unknown and $Z \sim N(0, \sigma_0^2 I_n)$, with $Y^{\mathcal{G}}\text{sel} = Y + Z$.
- In many cases, consistent estimates of σ^2 are not available, introducing further complication. In these cases, Theorem 1 can be used for inference.

Theorem 1

Assume we have access to $\hat{\sigma}_{\text{high}}$ and $\hat{\sigma}_{\text{low}}$ such that $\lim_{n \to \infty} \mathbb{P}\left(\sigma^2 \in [\hat{\sigma}^2_{\text{low}}, \hat{\sigma}^2_{\text{high}}] \mid Y^{\mathcal{G}\text{sel}}\right) = 1$.

Also define: $\hat{\tau}_{\text{low}} = \frac{\hat{\sigma}^2_{\text{low}}}{\hat{\sigma}^2_{\text{low}} + \sigma_0^2}$, $\hat{\tau}_{\text{high}} = \frac{\hat{\sigma}^2_{\text{high}}}{\hat{\sigma}^2_{\text{high}} + \sigma_0^2}$

$$A_1 = \min\left\{\frac{\eta^T Y - \hat{\tau}_{\text{low}} \eta^T Y^{\mathcal{G}\text{sel}}}{1 - \hat{\tau}_{\text{low}}}, \frac{\eta^T Y - \hat{\tau}_{\text{high}} \eta^T Y^{\mathcal{G}\text{sel}}}{1 - \hat{\tau}_{\text{high}}}\right\} \quad A_2 = \max\left\{\frac{\eta^T Y - \hat{\tau}_{\text{low}} \eta^T Y^{\mathcal{G}\text{sel}}}{1 - \hat{\tau}_{\text{low}}}, \frac{\eta^T Y - \hat{\tau}_{\text{high}} \eta^T Y^{\mathcal{G}\text{sel}}}{1 - \hat{\tau}_{\text{high}}}\right\}.$$

Then, a conservative asymptotic $1 - \alpha$ CI for $\eta^T \mu$l is given by:

$$C_{1-\alpha} := \left[A_1 - z_{\alpha/2} \frac{\|\eta\|_2 \hat{\sigma}_{\text{high}}}{\sqrt{1 - \hat{\tau}_{\text{high}}}}, A_2 + z_{\alpha/2} \frac{\|\eta\|_2 \hat{\sigma}_{\text{high}}}{\sqrt{1 - \hat{\tau}_{\text{high}}}}\right]$$

Experimental Results

- We compare confidence intervals constructed by Theorem 1 compared to the naive approach that assumes consistent estimates for σ^2.
- Confidence intervals using naive estimates for σ^2 undercover, but Theorem 1 CIs are conservative.